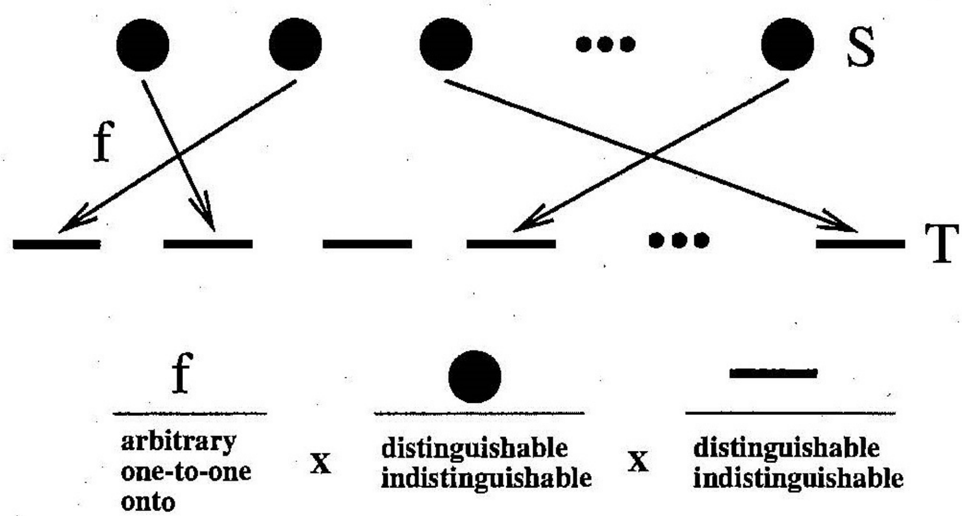
Probability



John N. Guidi

Lecture Notes - Spring 1998

MIT Course 18.313

Professor Gian-Carlo Rota

Volume 1 of 2

(Lectures 1–21)

This publication of *Probability: Volumes 1 & 2* (combined) is dedicated to John N. Guidi (1954-2012) whose almost "verbatim" notes in Prof. Gian-Carlo Rota's courses in 1998 at MIT faithfully reproduce both the content and erudition of Rota's famous lectures just before Rota's premature death in 1999.

John N. Guidi
Visiting Scholar in Applied Mathematics
Department of Mathematics
Room 2-392
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139
Email: guidi@mit.edu

These *Lecture Notes* originated from the lectures presented by Gian-Carlo Rota, Professor of Applied Mathematics, for course 18.313 - Probability, which he taught at MIT, during the Spring 1998 semester. The *Lecture Notes* were produced from notes I made during class, audio recordings I made of lectures, as well as clarifications and expansions I made of the material presented, after the fact. Although no attempt was made to provide a literal transcription of the lectures, an effort was made to present the material here in the same spirit and I hope to have introduced only a limited number of errors. The *Lecture Notes* are comprised of a two volume set. Volume 1 contains lectures 1–22, volume 2 contains lectures 22–35 and super class lectures 1–4.

Sara Billey, Assistant Professor of Applied Mathematics, who was the course director and whose recitations I attended, presented lectures 20, 21, and 27. The other recitation instructors were Alex Perlin and Ioana Popescu. Their efforts clearly contributed to the success of this course.

I was absent from lecture 25 and my thanks go to Jeff Lieberman, who taped the lecture for me and provided his notes, as well as to Ira Gerhardt and Rick Monte, who also provided their notes of this class.

During the course, I found that preparation of these *Lecture Notes* was a particularly useful way to really understand the material (or, as Professor Rota is fond of saying, "*to really rub it in*"). My gratitude goes to him for imparting his enthusiasm and sharing his profound knowledge of mathematics, as well as for presenting these superb lectures on probability.

John N. Guidi
October 15, 1998

Within the text, pagination is of the form *lecture_date.page*. Note that a different convention is used in both the table of contents and the index, where pagination is of the form *lecture_number.page*. The table of contents provides the mapping between *lecture_number* and *lecture_date*. For example, page 1.1 in the table of contents and index corresponds to page 2/4/98.1 in the text, where the topic anomaly detection is discussed.

Contents

Lecture 3 - [2/9/98]

Lecture 4 - [2/11/98]

Lecture 5 - [2/13/98]

Lecture 6 - [2/17/98]

Lecture 7 - [2/18/98]

Lecture 8 - [2/20/98]

Lecture 9 - [2/23/98]

Lecture 10 - [2/25/98]

Lecture 11 - [2/27/98]

Lecture 12 - [3/2/98]

Lecture 13 - [3/6/98]

Lecture 14 - [3/9/98]

Lecture 15 - [3/11/98]

Lecture 16 - [3/13/98]

After Notes Lecture AN16 - [3/16/98]

Lecture 17 - [3/16/98]

Lecture 18 - [3/18/98]

Lecture 19 - [3/20/98]

Volume 2

Lecture 35 - [5/11/98]

Super Class Lecture SC1 - [2/27/98]

Super Class Lecture SC2 - [3/13/98]

Super Class Lecture SC3 - [4/10/98]

Within the text, pagination is of the form *lecture_date.page*. Note that a different convention is used in both the table of contents and the index, where pagination is of the form *lecture_number.page*. The table of contents provides the mapping between *lecture_number* and *lecture_date*. For example, page 1.1 in the table of contents and index corresponds to page 2/4/98.1 in the text, where the topic anomaly detection is discussed.

Research Problems

The idea of probability is one of the great ideas developed in this century + you can find it in all the sciences.

Unsolved Problems of Probability

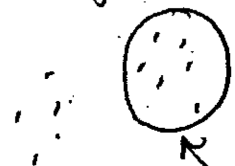
- Cluster Analysis
- Anomaly Detection

} There are literally millions, perhaps billions, of dollars to be made if someone were to make an important scientific advance in any of these problems.

These problems are not just hard to solve - they are hard to state
We know roughly what we want, but we don't know precisely what we are going to look at.

Cluster Analysis

You are given sets of multidimensional data in the plane.



Some how, these data cluster. You want a taxonomical criteria for separating data into clusters.

How do they do it today?
Like this.
They take a pencil and go like this.
That's all we know to this day.
There is no categorization theory.

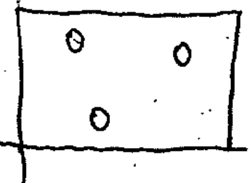
Anomaly Detection

A very closely related problem.
You have data (for example, income tax data) and you want some sort of criteria for spotting anomalous data - someone who is cheating.
The IRS has allocated $8 Billion

These are problems that are very much talked about in our time.
Some older problems of probability that come closer to the exact sciences:

- Hard Spheres - the name has been given to this problem by physicists.

The hard spheres problem is best stated for the plane.
You have a supply of identical pennies and you drop them on the carpet.
What is the probability that no two will overlap?



2-D

$$\underbrace{0 \quad 0 \quad \cdots \quad 0}_{n \text{ pennies}}$$

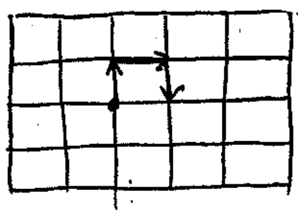
Q: How big is the carpet?
A: A big carpet

So the probability that the two of them overlap is unknown can be very difficult
It has been solved for <u>n=3</u> + <u>n=4</u> pennies.

This is a typical example of an <u>extremely simple problem</u> in probability for which very little is known.
In 3-dimensions, this would be <u>spheres</u>. People working in statistical mechanics would love to have the solution to the analogous problem.

- <u>Self avoiding random walk</u>

Another famous problem that arises in chemistry.



W, N, E, S
die

You have a grid of points (in the plane, or in space) w/ integer coordinates.
You start at the origin.
You have a <u>fair</u> die w/ 4 sides.
You flip it and, according to which side it shows, you take a step in the corresponding direction. Etc.

This is called a <u>random walk</u>.
It is one of the most studied stochastic processes

<u>Self avoiding random walk</u> problem is, again, an extremely simple problem to state.
It asks for the probability that <u>after n steps</u>, the random walk never visits a site that it has already occupied.

The basic definition of probability is the notion of a sample space.

Sample Space

plus

- Ω = a set } which, by abuse of speech, is sometimes called a sample space
- a family of $\boxed{\text{subsets}}$ of Ω, called $\boxed{\text{events}}$, say $\mathcal{E}$

Events satisfy the following properties: ..

1) ϕ is an event
   empty set

2) if A is an event, then so is $A^c = \Omega - A$

3) if A and B are events (i.e., if $A, B \in \mathcal{E}$)
   then $A \cap B$ is also an event
   $A + B$ are members of the family of events.

4) if $A_1, A_2, A_3, \ldots$ is a finite or infinite sequence of disjoint events (i.e., $A_i \cap A_j = \phi$, $i \neq j$) then
   $A_1 \cup A_2 \cup A_3 \cup \ldots$ is also an event

   This is called the Countable Additivity Axiom

A sample space is a set together w/ a family of events

To describe a sample space means not just to describe a set Ω, but a set Ω together w/ a family of subsets (called events), which satisfy these properties.
I'm not giving you an example, because this whole course will be examples of this.

A Probability P on a sample space Ω (with events $\mathcal{E}$) is a function that assigns to every event a number between 0 and 1 (i.e., $0 \leq P(A) \leq 1$ for $A \in \mathcal{E}$) subject to:

axioms

(1) $P(\Omega) = 1$ — whole sample space

(2) $P(A^c) = 1 - P(A)$

(3) if $A_1, A_2, A_3, \ldots$ is a finite or infinite sequence of disjoint events, then:

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots$$

This completes the definition of the notion of probability.

A sample space with a probability is sometimes called a probability space.

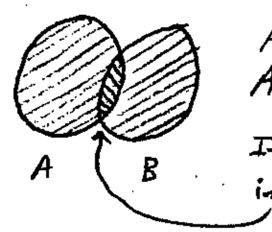
Q: $P(\phi)$?
A: By combining axioms (1) + (2) we have:

$$\left.\begin{array}{l} A = \Omega \\ A^c = \phi \end{array}\right\} \Rightarrow \begin{array}{l} P(A^c) = 1 - P(A) \\ P(\phi) = 1 - 1 \\ \quad\quad = 0 \end{array}$$

Consequences of the axioms of probability

Example $\underbrace{P(A \cup B)}_{\text{it is understood that } A+B \text{ are events}} = P(A) + P(B) - P(A \cap B)$

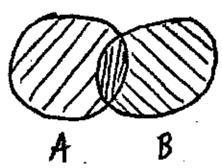
This is intuitively clear.
If you visualize events as sets and
if you visualize probability as some sort of measure, as in area,
then you have:



$A \cup B$ ///
$A \cap B$ \\\\

If you add $P(A) + P(B)$, you have overcounted the intersection.

Proof



$$A \cup B = \underbrace{(A \cap B^c) \cup (B \cap A^c) \cup (A \cap B)}_{\text{these 3 events are all disjoint}}$$

Thus, By axiom 3:
$P(A \cup B) = P(A \cap B^c) + P(B \cap A^c) + P(A \cap B)$

$$\left.\begin{array}{l} A = (A \cap B^c) \cup (A \cap B) \\ B = (B \cap A^c) \cup (A \cap B) \end{array}\right\} \begin{array}{l} P(A) = P(A \cap B^c) + P(A \cap B) \\ P(B) = P(B \cap A^c) + P(A \cap B) \end{array}$$

again, by axiom 3, as sets are disjoint

Therefore:

$$P(A)+P(B) = \underbrace{P(A\cap B^c) + P(B\cap A^c) + \overbrace{P(A\cap B)}^{\text{term appears twice}}}_{P(A\cup B)} + P(A\cap B)$$

$$\boxed{P(A\cup B) = P(A) + P(B) - P(A\cap B)}$$

This is a typical manipulation of events.

The next point is the most important concept of our book.
So far, everything can be read as sets.
But we come to the notion that set theory alone is not adequate.
We come to the notion of independent events.

Basic notion of probability

Independence

Events $A + B$ (in a $\overset{\text{understood}}{\text{sample space}}$) are said to be independent when

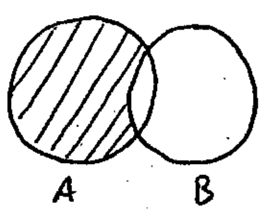
$$P(A\cap B) = P(A)P(B)$$

We will see many interpretations of this notion.

Observation:

- If $A + B$ are independent events, then $A + B^c$ are also independent events

Proof



$$A\cap B^c = A - (A\cap B)$$

the subtraction of 2 sets is well defined only when the second set is a subset of the first.

$$P(A\cap B^c) = P(A - A\cap B)$$ we'll show this next

$$\cancel{P(A) - P(A)P(B)}$$

$$= P(A) - P(A\cap B)$$

$$= P(A) - P(A)P(B)$$ (A+B are independent events, thus: $P(A\cap B) = P(A)$

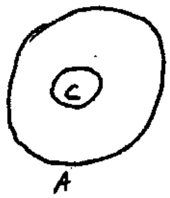
$$= P(A)[1 - P(B)]$$

$$P(A\cap B^c) = P(A)P(B^c) \quad \text{(by axiom 2)}$$

$A + B^c$ are independent events

$\overbrace{\text{C is contained in A}}$

- If A + C are events and if $C \subseteq A$
  then $P(A-C) = P(A) - P(C)$



$$A - C = A \cap C^c$$

$$A = \underbrace{(A \cap C^c) \cup C}_{\text{union of } \underline{\text{disjoint}} \text{ events}}$$

Q/A: if you wish, you could condense this to:

$A = (A-C) \cup C$

$$P(A) = P(A \cap C^c) + P(C) \quad \text{(by axiom 2)}$$

$$= P(A - C) + P(C) \quad \text{(from above)}$$

$$\therefore P(A-C) = P(A) - P(C)$$

That's it my friends.
From now on we see examples of this.
That's it for the <u>grammar</u>

{we'll discuss later the use of one word versus the other.}

2/4/98.7

Example 1 Coin Tossing $\left(= \text{Bernoulli} \begin{cases} \text{sample space} \\ \text{process} \end{cases} \right)$

Biased coin that gives Heads w/ probability p $(0 \le p \le 1)$
Tails " " $q = 1-p$

If we toss the coin infinitely many times, what will happen?

Now you say – we can't toss the coin infinitely many times. By imagining tossing the coin infinitely many times, you include all possible finite tosses of the coin. Therefore, all questions about tossing the coin 1 thousand versus 1 million times are included in our study of tossing the coin infinitely many times.

Here we come to the crucial step. We set up a sample space that reflects all properties of what can happen when we toss a coin infinitely many times. We can work this to answer any question about tossing the coin. These questions can be extremely complicated.

Let me give you an example of one question we will answer shortly. What is the probability that a run of 3 Heads will occur before a run of 7 Tails?

A run is a continuous, uninterrupted sequence of Heads or of Tails.

The event that a run of 3 Heads will occur before a run of 7 Tails has a definite probability and our mathematical setup should enable us to compute this probability easily.

What is our mathematical setup?
Here we come to a beginning of understanding of those mysterious properties we have been writing so far.

- Sample space Ω ← consists of all possible outcomes of infinite sequences of tosses of the coin.

$\omega \in \Omega$
↑ a point in the sample space is called a sample point

The sample space for coin tossing for the Bernoulli process consists of all possible sequential tosses of the coin at integer times.

We represent this by ω with an infinite sequence that tells us whether you get Heads or Tails. We could use H/T, but it has become much more customary to use 1/0.

1 = Head
0 = Tail

So a sample point is just an infinite sequence of 0 + 1.
All possible sequences of 0 + 1 are sample points.

We will sometimes write a sample point as:

$$\omega = (\omega_1, \omega_2, \omega_3, \ldots)$$

$$\text{where } \omega_n = \begin{cases} 0 & \text{tail} \\ 1 & \text{head} \end{cases}$$

Observe that a sample point does not look like a point
The points of probability do <u>not</u> look like a point of geometry.
In the case of coin tossing, the sample point is an <u>infinite</u> sequence of 0 + 1. This is a long way from being a point.

Q/A : If you consider 0/1 as <u>binary expansion</u>, something very interesting happens. You get the <u>Cantor Set</u>.

Have we defined a sample space? No!
We have to specify <u>which subsets</u> of the sample space are the <u>events</u>.

Intuitively, that corresponds to specifying all possible events whose probability we may ever wish to compute.
How shall we do this?
We first single out the <u>most obvious event</u>

- Event

H_n = event that n^{th} trial results in Heads

technical language

Translated into <u>set theoretic language</u>:

H_n = the set of all sample points $\omega \in \Omega$ with property that $\omega_n = 1$

The idea of probability is <u>not</u> to view this as a set.
It is to view it as something that either <u>happens</u> or does <u>not happen</u>.
In probability, you have a <u>novel way</u> of <u>visualizing sets</u>.
You visualize this as the event that the n^{th} toss resulted in Heads is psychologically <u>very different</u> from thinking the set of all sequences of 0+1 with the $n^{th} = 1$. Quite different.

Learning to rethink all of set theory is <u>exactly</u> what you are supposed to do in this course. It's a new way of thinking.
That's why this course is hard.

Sample Spaces – The Bernoulli Process (cont'd)

A sample space is the basic structure that models probabilistic phenomena. To set up sample space for various probabilistic phenomena has the property, because it is, in fact, sample spaces, that you will be able to compute the answers to obvious, real life questions about these phenomena

sample space : Ω = a set plus a family $\mathcal{E}$ of subsets of Ω, called events

You specify, inside the sample space, a family of subsets called events. These are the subsets whose probabilities might be computed.

Why doesn't it take every subset to be an event? We'll discuss this later.

Events satisfy the following properties:

(1) $\phi \in \mathcal{E}$

(2) if A is an event, then $A^c \in \mathcal{E}$

(3) if $A, B \in \mathcal{E}$, then $A \cap B \in \mathcal{E}$

(4) if $A_1, A_2, \ldots$ is a finite or infinite sequence of pairwise disjoint events (i.e., if $A_n \in \mathcal{E}$ for all n) then $\bigcup_n A_i \in \mathcal{E}$

> Events specifying a sample space means :
> 1) specifying the set Ω
> and 2) somehow specifying which subsets are the events
> This specification must be made in such a way that the events satisfy the above properties.
> It will take us all term to see how this is done.
> This is an entirely new concept. Don't expect to see the entire meaning of this concept when you see one or two examples. You have to see 60.

Q: For finite sets, is there any reason why you wouldn't take all of the subsets for your events?

A: If Ω is finite set, then it is almost always the case that the events are also.

Probability : $P : \mathcal{E} \to [0,1]$

A function P that associates to every event a number btw 0 and 1 with the properties:

(1) $P(\phi) = 0$ ← Probability that nothing happens is 0

(2) $P(A^c) = 1 - P(A)$

Probability that something does not happen is $\left(1 - \text{probability that it does happen}\right)$

Last time we stated things in terms of sets.
Now we rephrase things in probabilistic language.
↳ another way of interpreting set theory.

{Countable Additivity Axiom}

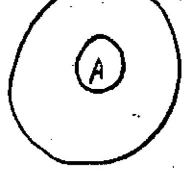
(3) if $A_1, A_2, \ldots$ is a finite or infinite sequence of pairwise disjoint events, then $P\left(\bigcup_n A_n\right) = \sum_n P(A_n)$

pairwise disjoint = any 2 events are disjoint

These axioms are intuitive. What is not so quickly clear, at first, is why is it that we must admit the possibility of an infinite sequence here. That we have to learn from examples, which are coming.

We have seen last time that from these axioms, one derives all the intuitive and obvious properties of probability.

Ex: $\underbrace{\text{if } B \supseteq A}_{\text{if } A \text{ happens, then } B \text{ happens}}$ then $P(B) \geq P(A)$



B

$$B = (B \cap A^c) \cup A$$

$$P(B) = \underbrace{P(B \cap A^c)}_{\geq 0 \text{, since } P: \mathcal{E} \to [0,1]} + P(A) \qquad \text{(by axiom 3)}$$

$$\therefore P(B) \geq P(A)$$

This fact is a consequence of the axioms.

Another fact that can be obtained as a consequence of the axioms, which may be on the quiz:

Ex: If we have a sequence of events such that:

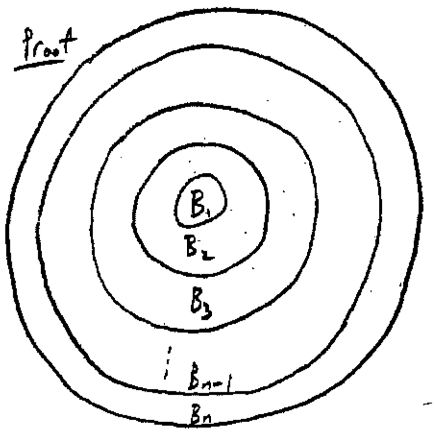
$$B_1 \subseteq B_2 \subseteq B_3 \subseteq \ldots$$

then $P(\bigcup_n B_n) = \lim_{n\to\infty} P(B_n)$

This follows from countable additivity

This is sometimes called the <u>Continuity Property of Probability</u>

<u>Proof</u>



Written as the union of <u>pairwise disjoint events</u>:

$$\bigcup_n B_n = (B_n - B_{n-1}) \cup (B_{n-1} - B_{n-2}) \cup \ldots \cup (B_3 - B_2) \cup (B_2 - B_1) \cup B_1$$

The <u>countable additivity axiom</u> gives:

$$P(\bigcup_n B_n) = P(B_n - B_{n-1}) + P(B_{n-1} - B_{n-2}) + \cdots + P(B_3 - B_2) + P(B_2 - B_1) + P(B_1)$$

On [2/4/98.6] we proved that if $A + C$ are events and if $C \subseteq A$, then $P(A-C) = P(A) - P(C)$

Thus we have:

$$P(\bigcup_n B_n) = \underbrace{P(B_n) - P(B_{n-1}) + P(B_{n-1}) - P(B_{n-2})} + \ldots + \underbrace{P(B_3) - P(B_2) + P(B_2) - P(B_1)} + P(B_1)$$

$$P(\bigcup_n B_n) = P(B_n)$$

Thus, as $n \to \infty$,

$$P(\bigcup_n B_n) = \lim_{n\to\infty} P(B_n)$$

//

It will be interesting later to see examples where these axioms are not satisfied. Perfectly innocent probabilities that do not satisfy these axioms.

This is the abstract grammatical scheme of probability, on which all of probability theory is based.
To this, we add:
The Fundamental Notion of Probability – Notion of Independence

Events $A + B$ are said to be independent when { I pass over, in silence, in a given sample space and with a given probability }

$$P(A \cap B) = P(A)P(B)$$

probability of their joint happening

This is a great discovery of mathematics. That the intuitive idea of 2 events being independent is rendered mathematically with this extremely simple equation. It's an incredible discovery.
That this corresponds, as we will convince ourselves gradually in the next few weeks, exactly to our intuitive notion of 2 things happening independent of each other. It's a great thing. We tend to dismiss things because we are jaded. But every once in a while you should reflect on how extraordinary these discoveries are. You should recover your sense of wonder from time to time.

One remark that I omitted last time:
This is the definition for independence of two events. What about 3 events?

More generally, events A, B, C are independent when any two of them are independent:

$$\begin{cases} P(A \cap B) = P(A)P(B), \quad P(A \cap C) = P(A)P(C), \quad P(B \cap C) = P(B)P(C) \\ \underline{\text{AND}} \\ P(A \cap B \cap C) = P(A)P(B)P(C) \end{cases}$$

(probability of all 3 events happening is the product of the probabilities)

For 3 events to be independent, it's not enough for any 2 to be independent. You have to add this last condition.

Q: Couldn't we just use the last condition: $P(A \cap B \cap C) = P(A)P(B)P(C)$?
A: No. The last condition does not imply the others.
You can construct artificial sample space and artificial probabilities to show that all these conditions are needed. Can show that only some, not all, conditions can be satisfied by artifically assigning probabilities to sample sizes.

2/6/98.5

For more than 3 events, similarily, for <u>all possible intersections of subsets</u>, when you take the probability, it is the product.

Similarily – for any <u>finite</u> family of events [It's pedantic to write this down]

• ∈ all possible $\cap$ $\Longrightarrow$ $P(\cdot)$ = product

Q/A: For 4 events to be independent:

$$\left.\begin{array}{l}\text{all pairs}\\ \text{all triples}\\ \text{all quadruples}\end{array}\right\} P(\cdot) = \text{product}$$

Otherwise, no independence.

Open question:

It is an open question as to which of these conditions imply which. In other words, to build an implication table, if I take a certain family of all possible intersections, does it imply the others? This has not been settled to this day.

That's it.
The intuitive meaning of independence will gradually be found.

- Bernoulli Process – mathematical model for all questions pertaining to coin tossing.
  Any question pertaining to coin tossing is answered by computation with the Bernoulli Process. Some of them can be extremely difficult.

Sample points $\omega = (\omega_1, \omega_2, \dots)$ ← infinite sequences of 0 and 1

$$\omega_n = \begin{cases} 1 \\ 0 \end{cases}$$

(all possible histories of tossing a coin
List all possible histories of an infinite sequence of tosses.)

Sample space Ω = set of all sample points

plus

Secondly, we have to define events

Events

Define events in 2 steps.

1) Pick out events whose probability is obvious

H_n = event that n^{th} toss is Head (1)

The set of all sample points whose n^{th} entry is 1, but we don't say it that way.

As a set, it's an infinite set. ← {there are ∞ many sample points whose n^{th} entry is 1}

2) Take these elementary events + take all events obtained by taking $\cup$, $\cap$, ^{c}, and unions of infinite sequences of disjoint events. In all possible ways.

Any subset of Ω obtained from $H_1, H_2, \dots$ by taking $\cup$, $\cap$, ^{c}, and unions of infinite sequences of disjoint events is an event. } ⇐ By definition

You may ask – Doesn't this include all subsets of a sample space?
This is an interesting question.
To prove that this does not include all subsets, you need the axiom of choice.

The subsets in (2) above are the only subsets whose probabilities we can hope to compute – because they are obtained from elementary events by operations where the probability can be computed.

Probability for the Bernoulli Process

1) $P(H_n) = p \qquad \Longrightarrow P(T_n) = 1 - p = q$

the bias of the coin

2) Every finite set of H_n is independent
e.g., $P(H_2 \cap H_3) = P(H_2)P(H_3)$
$P(H_1 \cap H_2^c \cap \ldots \cap H_{10}) = P(H_1)P(H_2^c) \ldots P(H_{10})$ } by definition of independence

{extension of [2/4/98.5]. If A + B are independent events, then A + B^c are independent events.}

Intuitively, we tell this system that sucessive tosses of the coin are independent events.

We have the following result, which we are not going to prove:
Theorem (1) and (2) uniquely specify P on all events

Once you assume (1) + (2) and the axioms of probability then the one and only one way of computing the probability of each event is set.
We will not prove this. It's hopelessly dull. Proof as research problem.

In this way, we have defined the Bernoulli Process.

What do we do with it?

Example: Let's compute the probability of some event related to coin tossing, using this model.

✗ A = $\underbrace{\text{event that a run of } \overset{HH}{\text{Two Heads}} \text{ occurs before a run of } \overset{OO}{\text{Two Tails}}}_{\text{this is not well stated}}$

instead:

$\Downarrow$

✓ A = event that the first run of 2 is a run of $\overset{HH}{\text{2 Heads}}$

What is $P(A)$?

We have to express, some how, the event A in terms of the events H_n and their intersections. From [2/6/98.7], all we know is:

(1) $P(H_n) = p$

(2) Every finite set of H_n is independent
e.g., $P(H_1 \cap H_2) = P(H_1)P(H_2)$

So we need to reduce A to H_n and its intersections.

$$A = (H_1 \cap H_2) \cup (H_1 \cap T_2 \cap H_3 \cap H_4) \cup (H_1 \cap T_2 \cap H_3 \cap T_4 \cap H_5 \cap H_6) \cup \dots \leftarrow$$ exhaustively enumerate all that start w/ Head

$$\cup (T_1 \cap H_2 \cap H_3) \cup (T_1 \cap H_2 \cap T_3 \cap H_4 \cap H_5) \cup (T_1 \cap H_2 \cap T_3 \cap H_4 \cap T_5 \cap H_6 \cap H_7) \cup \dots \leftarrow$$ w/ Tai

This accounts for all possibilities.
We have expressed event A as a union of intersections of elementary events H_n.
Any of these intersections, we know the probability of (from (2) above) since they are independent.
Furthermore, if you take any two of the union terms, they are disjoint.
Why are they disjoint? They differ by the number of H_n, so they can't happen together.

Therefore the Axiom of Countable Additivity allows us to compute the probability by adding these infinite events.
Now you see where the axiom of countable additivity comes in.
If we didn't have this axiom, we couldn't even compute the probability of this event A.

$$P(A) = p^2 + (pq)p^2 + (pq)^2p^2 + \dots + (pq)^n p^2 + \dots$$
$$+ qp^2 + q(pq)p^2 + q(pq)^2p^2 + \dots + q(pq)^n p^2 + \dots$$
$$= p^2(1 + pq + (pq)^2 + \dots + (pq)^n + \dots)$$
$$+ qp^2(1 + pq + (pq)^2 + \dots + (pq)^n + \dots)$$

geometric series

$$= p^2\left(\frac{1}{1-pq}\right) + qp^2\left(\frac{1}{1-pq}\right) \qquad 1 + pq + (pq)^2 + \dots = \frac{1 - (pq)^{\infty} \to 0}{1-pq}$$

$$P(A) = \frac{p^2 + qp^2}{1-pq}$$

Probability that the first run of length 2 is a run of 2 Heads.

How do you double check this?

Suppose the coin is fair. Then $P(A)$ better be $\frac{1}{2}$.
If you don't see this, I can not explain it.
The Principle of Ignorance

$$p = q = \frac{1}{2} \Rightarrow \quad P(A) = \frac{p^2 + p^3}{1-p^2}$$
$$= p^2 \frac{\cancel{(1+p)}}{\cancel{(1+p)}(1-p)}$$
$$= \frac{1}{2} \checkmark$$

Even in computing one of the simplest questions about coin tossing, we have used the full power of the axioms.

Let's do another one.
Jack it up.

Example: B = event that a run of $\overbrace{9 \text{ Heads}}^{\text{IIIIIIIII}}$ ever occurs

$$P(B) = 1 - P(B^c)$$

B^c = event that 9 successive Heads never occurs.

If the event B never occurs, then it doesn't occur in the first 9 tosses
It doesn't occur in tosses 1 – 9
10 – 18
19 – 27
etc.

B_1 = event that $\omega_1 = \omega_2 = \dots = \omega_9 = 1$

B_2 = event that $\omega_{10} = \omega_{11} = \dots = \omega_{18} = 1$

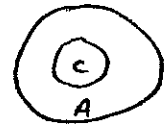
B_3 = event that $\omega_{19} = \omega_{20} = \dots = \omega_{27} = 1$

etc.

I spaced them disjointly. It looks a little weird at first, because it can happen that by straddling you can get 9 successive Heads

If B doesn't occur, then none of $B_1, B_2, \dots$ can occur.

$$B \supseteq B_1 \cup B_2 \cup B_3 \cup \dots$$
$$B^c \subseteq B_1^c \cap B_2^c \cap B_3^c \cap \dots$$



$A \supseteq C \implies A^c \subseteq C^c$

Recall [2/6/98.2]
If $C \supseteq A$ then $P(C) \geq P(A)$
Applied to above gives:

$$P(B^c) \leq P(\underbrace{B_1^c \cap B_2^c \cap B_3^c \cap \dots}_{\substack{\text{all disjoint} \\ \text{So they are independent}}})$$

$$= P(B_1^c)\,P(B_2^c)\,P(B_3^c)\dots$$

Note that:

$$P(B_n) = p^9$$
$$P(B_n^c) = 1 - P(B_n)$$
$$= 1 - p^9$$

$$P(B^c) \leq (1-p^9)(1-p^9)(1-p^9)\dots$$

$p \in [0,1]$ If $p \neq 0$, $(1-p^9) < 1$

$$\lim_{n\to\infty} P(B^c) = 0$$

Therefore, with probability $P(B) = 1$ (in the limit) a run of 9 Heads occurs. Thus we conclude:

$$P(B) = 1$$

- We can even jazz it up.
  I say that this is not a strong enough statement.

Suppose 9 Heads occurs only once.
Then you start the next one. It doesn't know what happened before it.
It occurs <u>again</u>.
It can not happen just once. Because then you repeat the argument from this point on. So it occurs twice. But then you can start again. Etc.

So it occurs <u>infinitely often.</u>

C = event that a run of $\overbrace{\text{9 Heads}}^{111111111}$ occurs <u>infinitely often</u>

Claim $P(C) = 1$

PBC

<u>Proof</u>

C^c = event that a run of $\overbrace{\text{9 Heads}}^{111111111}$ occurs <u>finitely</u> many times

If the run of 9 Heads occurs only finitely many times, then it <u>doesn't</u> occur from a certain toss onward.

(i.e., <u>no run</u> 111111111 occurs <u>after n^{th} toss</u>, for some n)

But if we start tossing the coin after the n^{th} toss, its as if we started from scratch.
And we've already proved (previous page) that with probability 1 there <u>has</u> to be one occurrence of the run of 111111111.

<u>Contradicts</u> previous argument unless $P(C^c) = 0$

Therefore $P(C) = 1$

The event (that a run of 9 Heads occurs <u>infinitely often</u>) occurs with certainty.

- Note - instead of 9 Heads, we could have chosen any <u>finite pattern</u> whatsoever. Exactly the same argument works.
  The same pattern will work for any pattern whatsoever.

Take any finite pattern (e.g., 00100111).
The probability that this pattern <u>ever</u> occurs is 1
<u>Whats more</u>, this pattern will <u>occur infinitely often</u>

This is the famous assertion that if a monkey starts typing randomly on a computer, sooner or later the monkey will type the bible. And, it will type it infinitely often.

John Guidi
guidi@mit.edu

Q/A: Usually $\cap$ takes precedence over $\cup$.
↑ not always the case

The Bernoulli Process (cont'd)

{After this we begin the theory of distribution of occupancy.}

The Bernoulli Process, w/ the sample space Ω, consists of all sample points which are infinite

Set

- $\Omega = \{ \omega = (\omega_1, \omega_2, \omega_3, \ldots), \ \omega_n = \{ {1 \atop 0} \} \}$

1 = Heads
0 = Tails

↑ sample point = infinite sequence of tosses

Philosophically - sample points are all the possibilities
events are all the properties

We try to use sample points as little as possible.
We try to reason about everything in terms of properties ← i.e., in terms of events

Events

- $\mathcal{E}$: take all subsets of Ω (sample space) "generated" by H_n, $n = 1, 2, \ldots$

By generated, we mean we can take $\cup$, $\cap$, ^{c}, and infinite unions of sequences that are disjoint.
Anything that can be obtained that way is an event.
There is no other description of an event - unless you want to go into excruciating detail for mathematical specialists.

Probability

- a) $P(H_n) = p$
furthermore
b) any finite set of H_n's is independent

[Research Problem]
These data determine a $\underbrace{\text{unique probability}}_{\text{which satisfies the 3 axioms of probability.}}$ on all the elements.

This is the model of anything you want whatsoever from coin tossing.
That's the Bernoulli Process.

We then proceeded to establish the following fact:

The probability that a finite run $\underbrace{111\ldots11}_{k}$ occurs infinitely often equals one.

This we established by a very simple argument, which we will shortly summarize again.
A highly non-trivial result, as we remarked at the end of our last meeting.
It means that if a monkey starts typing on the computer, the monkey will type the bible infinitely often. Or any message whatsoever.

We also remarked that exactly the same argument that we use to establish that the probability that a run occurs infinitely often equals one can be used to show that: the probability that any pattern whatsoever of 0's + 1's, provided it is finite, occurs infinitely often is one
Exactly the same reasoning.

By the same reasoning, we establish the Borel-Cantelli Lemma.
The probability that any finite pattern, say 010011...01, occurs infinitely often equals one.

The proof is exactly like we did it for a run of 1's [2/6/98.9-11]
It is intuitively obvious that it is the same proof.

Let's restate this in terms of sets.
You realize now that if you switch from probabilistic language to set theoretic language, this is harder to state.
This is the probability of what event?
The event that a given pattern occurs infinitely often.
→ the set of all sample points that contain infinitely many occurrences of that pattern

Now you begin to see that constructing the probability from set theoretic language is more complicated than for probabilistic language.
Even though the two languages are completely similar.

You say:
"The event that the pattern 01110 occurs infinitely often"
You have an intuitive feeling for that.

But if you say:
"The set of all sample points that contain infinitely many occurrences of this pattern"
It's not as intuitive. That's the way it is.
So we are gradually shifting our language.

Let's do a little grammar.
This is pure mathematics at it's best (or worst) depending on what your view is.
The pure mathematician at this point will say:

Just a minute. Is this really the most general way of stating the Borel - Cantelli theorem? Can we jazz it up and make it comprehensible? Even though it's the same statement?
Sure.

General Formulation of Borel-Cantelli Lemma

This is where you find it in some probability books that never explain to you what is going on. You have to figure it out.

Given Ω and P {as usual, we pass over in silence the description of events. Of course there are events.}

Sample space with probability P

Given a sequence $A_1, A_2, A_3, \ldots$ of independent events

Assume that $\sum_n P(A_n) = \infty$

Then, with probability one, infinitely many events A_n occur.

Where did we get this? The secret crib.
The secret crib goes like this.
Imagine the Bernoulli process. Imagine a pattern of length k.

$$\underbrace{\overbrace{\cdot\ \cdot\ \cdot\ \cdot\ \cdot}^{k}}_{A_1}\ \underbrace{\cdot\ \cdot\ \cdot\ \cdot\ \cdot}_{A_2}\ \underbrace{\cdot\ \cdot\ \cdot\ \cdot\ \cdot}_{A_3}$$

A_1 is event that pattern occurs in first k tosses
A_2 " " " " " second k tosses
etc.

} non overlapping

This is uneconomical because you do not consider the possibility of the pattern occurring in between.
But, as we saw last time, it is enough to establish the result we want.

The events $A_1, A_2, A_3, \ldots$ are secretly copied from the list of ___?___ of events of non overlapping successive paterns.

In this case, of course, since the pattern of 1's is probability p^k then the sum of the probabilities is infinite.

$$P(A_n) = p^k$$

$$\sum_n P(A_n) = \infty$$

This is the general formulation of the Borel-Cantelli Lemma.

Any independent event whatsoever, which has any pattern, and the sum of the probabilities is divergent, then infinitely many of them occur.

Since we are on a pure math kick, let's really rub it in.
Now you say "Excuse me. What does it mean for infinitely many of the A_n to occur?" We know intuitively. How do you this mathematically? This is how you say it:

First, we give it a name

Say A_{io} is the event that infinitely many of the A_n occur
(io = infinitely often)

(How do we know A_{io} is an event? We have to express it in terms of elementary events)

Then

$$A_{io} = \bigcap_{k=1}^{\infty} \left(\bigcup_{n \geq k} A_n \right)$$

We will convince ourselves in a minute that that expression is exactly the event that infinitely many of the events A_n occur.
There are many ways of convincing ourselves of what I just said.

One of them is the extremely pure way. You just take this - you don't care what it means.

Lets take the complement and see what this means.

De Morgan Laws ← ["There is nothing wrong with not knowing something. What's wrong is not admitting that you don't know. That is wrong"] *

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

You convince yourself of DeMorgan Laws by drawing little pictures

The complement of A_{io} gives:

$$A_{io}^c = \left[\bigcap_{k=1}^{\infty} \left(\bigcup_{n \geq k} A_n \right) \right]^c$$

$$= \bigcup_{k=1}^{\infty} \left(\bigcap_{n \geq k} A_n^c \right)$$

Let's write this out: 2/9/98.5

$$= \left(\bigcap_{n\geq 1} A_n^c\right) \cup \left(\bigcap_{n\geq 2} A_n^c\right) \cup \left(\bigcap_{n\geq 3} A_n^c\right) \cup \ldots$$

Let's read the first 3 terms:

$$\underbrace{\left(\bigcap_{n\geq 1} A_n^c\right)}_{\text{none of the events } A_n \text{ occur}} \quad \underbrace{\left(\bigcap_{n\geq 2} A_n^c\right)}_{\text{none of the events occur after the first } k \text{ tosses}} \quad \underbrace{\left(\bigcap_{n\geq 3} A_n^c\right)}_{\text{none of the events occur after the second } k \text{ tosses}}$$

How can we read the infinite union?
It means that none of the events occurs after some point or other.
Which means that only finitely many of the events occur.

Therefore, the complement of our set is the event that only finitely many of the A_n occur.
Therefore A_{io} is the event that infinitely many occur.
There you are.

$$P(A_{io}^c) = P\Big(\underbrace{\left(\bigcap_{n\geq 1} A_n^c\right) \cup \left(\bigcap_{n\geq 2} A_n^c\right) \cup \left(\bigcap_{n\geq 3} A_n^c\right) \cup \ldots}\Big)$$

Well, we don't know whether these will be disjoint.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$P(A \cup B) \leq P(A) + P(B)$$

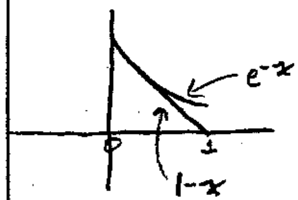
$$P(A_{io}^c) \leq P\left(\bigcap_{n\geq 1} A_n^c\right) + P\left(\bigcap_{n\geq 2} A_n^c\right) + P\left(\bigcap_{n\geq 3} A_n^c\right) + \ldots$$

At this point, the mathematician asks: we haven't used the assumption that these A_n are independent.
It has to come in somewhere.

$$P\left(\bigcap_{n\geq 1} A_n^c\right) = \prod_{n\geq 1} P(A_n^c)$$
$$= \prod_{n\geq 1} (1 - P(A_n))$$

$$= \prod_{n\geq 1}(1-P(A_n)) + \prod_{n\geq 2}(1-P(A_n)) + \prod_{n\geq 3}(1-P(A_n)) + \ldots$$

For $0 \le x \le 1$,
$1 - x \le e^{-x}$



Maclaurin series for e^{-x}

$$1 - x \le 1 - x + \underbrace{\underbrace{\frac{x^2}{2!} - \frac{x^3}{3!}}_{\ge 0} + \underbrace{\frac{x^4}{4!} - \frac{x^5}{5!}}_{\ge 0} + \cdots}_{\ge 0}$$

every pair ≥ 0

$$\therefore\ 1 - x \le e^{-x}$$

$$P(A_{io}^c) \le \prod_{n\ge1}(1-P(A_n)) + \prod_{n\ge2}(1-P(A_n)) + \prod_{n\ge3}(1-P(A_n)) + \cdots$$

$$\le \prod_{n\ge1} e^{-P(A_n)} + \prod_{n\ge2} e^{-P(A_n)} + \prod_{n\ge3} e^{-P(A_n)} + \cdots$$

$$= e^{-\sum_{n\ge1} P(A_n)} + e^{-\sum_{n\ge2} P(A_n)} + e^{-\sum_{n\ge3} P(A_n)} + \cdots$$

But the sum of the probabilities of A_n is assumed to be infinite:

$$\sum_n P(A_n) = \infty$$

$$= 0 + 0 + 0 + \cdots$$

a countable number of times

Therefore,

$$P(A_{io}^c) = 0$$

Therefore,

$$P(A_{io}) = 1, \text{ as desired}$$

That is what mathematicians make out of the Borel-Cantelli Lemma. If you think about it concretely, you must always think about the patterns. But with the lemma couched this way, you can apply it to a number of situations.

Q: What are the A_n above?
A: A_n are events with the properties:
1) A_n are <u>independent</u>
2) $\sum_n P(A_n) = \infty$ sum of the probabilities is infinite.

Recap:
- First, we showed Borel-Cantelli with a crib (concrete example)
- Then, we proved Borel-Cantelli <u>abstractly</u>. The advantage of doing it abstractly is that you can find <u>unexpected examples</u>. Like the follow

[Research Problem] this is pretty cute

Example : Let A_n = event that at $\overbrace{2n}^{\text{even}}$ th toss, the number of 0's equals the number of 1's. (Bernoulli Process)

Here is a sequence of events.
What is the probability that infinitely many of the A_n occur?

Note that the A_n are not independent.
Q/A : Why?
Deferred. If they are independent, so much the better. But it's pointless to spend time showing something is not true.

If they were independent - better.
But we don't want to worry about it.

$$A_{io} = \bigcap_{k=1}^{\infty} \left(\bigcup_{n \geq k} A_n \right)$$

This is the event that infinitely many times, you have equalization of Heads + Tails,

Now this is a sensible and important question to ask.

Q: Will this happen infinitely often or not?

A: The answer is very interesting

$$P(A_{io}) = \begin{cases} 0 & \text{if } p \neq \frac{1}{2} \\ 1 & \text{if } p = 1/2 \end{cases}$$

for an absolutely fair coin, you get infinitely often,
If the coin is just a little bit unfair - then no.

Before we leave this preliminary study of the Bernoulli Process, let's examine one of the most extraordinary properties of coin tossing.
A property that goes <u>against</u> <u>intuition</u>:

Kolmogorov Zero-one law

Let's philosophize about A_{io}. ← infinitely often
If we <u>leave out</u> any <u>finite number</u> of the A_n, then the event A_{io} has the same solution.

Also, this event A_{io} doesn't depend on what happened at those coin tosses

→ The probability is <u>not</u> affected.

An event B in the Bernoulli Process is said to be a <u>tail event</u> when B is independent of $H_1, H_2, \ldots, H_n$ for every n

What are examples of <u>tail events?</u>
The occurrence of a pattern <u>ever</u>.

The event that you <u>eventually</u> get a run of 15 Heads is a <u>tail event</u>.
It's probability doesn't change if the probability is dependent on the first coin toss.

Anything to do w/ <u>eventually</u> is likely to be a <u>tail event</u>

[Research Problem]

- <u>Kolmogorov Zero-One Law</u>

  Every Tail Event has probability 0 or 1 } not particularly hard, but hopelessly dull

Example

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots = \infty$$

Harmonic Series is divergent

You all know this. If you didn't, you do now.

Keep the terms of the Harmonic Series, but instead of making them all +, I make them + or −, at random.

The probability of this series will converge

Let's make this more precise:

$$1 \pm \frac{1}{2} \pm \frac{1}{3} \pm \frac{1}{4} \pm \dots$$

$\omega_1 = \begin{cases} 1 \text{ if } + \\ 0 \text{ if } - \end{cases}$ $\omega_2 = \begin{cases} 1 \\ 0 \end{cases}$ $\omega_3 = \begin{cases} 1 \\ 0 \end{cases}$

You look at each entry of the sample point and put + or − according to whether entry is Heads or Tails.

Sample point:

$$\omega = (\omega_1, \omega_2, \omega_3, \dots)$$

Rule for assigning the signs.
Toss a fair coin.
So $P(+) = P(-) = \frac{1}{2}$
Assign the n^{th} sign depending on what the n^{th} toss is.
You get a random series

C = event that series is convergent

What is $P(C)$?

We apply the Kolmogorov Zero-One Law.
The convergence of the series is not affected if you cross out the first n terms, for any n.
A divergent series - if you cross out first 15,000 terms is still divergent.

Therefore, C is a Tail Event ← in the sense of the Kolmogorov Zero-One Law

$$\therefore\ P(C) = \begin{cases} 1 \\ \text{-or-} \\ 0 \end{cases} \Leftarrow$$ I leave it to you to figure out which one it is.

Actually it is:

$$P(C) = 1$$ Rademacher's Theorem

Now you say - But that's impossible. What if I make up some weirder series. Ex:

$$\sum \frac{(\sin n\pi / \log n)}{n^{1/10}}$$

If you assign +/- at random, the series converges by Kolmogorov Zero-One L with either probability 0 or 1.
No matter how weird the series.

If you establish that the probability is stated at .00001 then you are through, probability is 1.

Q/A: "Anything that makes sense is an event"

Lastly, before we completely leave the Bernoulli Process:

Observe that we have, so far, passed over in silence, the phenomena of events of probability 0

To say that $P(A)=0$ does not mean that the event is impossible. For example

$$\left.\begin{array}{l} \text{event } \omega = (1111\ldots) \text{ has} \\ \underbrace{P(\omega)=0}_{\text{probability zero}} \quad \text{if } 0<p<1 \end{array}\right\} \text{But it may occur}$$

Similarily, if we assert that $P(\text{event})=1$, this does not mean the event is certain.

The point we are making now is what is the philosophical interpretation of probability?

You must be aware that there are events with:

$$\left.\begin{array}{l} P(\text{event}) = 0 \ \text{— not impossible} \\ P(\text{event}) = 1 \ \text{— not certain} \end{array}\right\} \text{but pretty close}$$

This concludes our preliminary study of the Bernoulli Process.

Theory of Distribution and Occupancy

This is a big chapter of probability.
And it's very bread and buttery - this stuff has tremendous applications.
What's it all about?

Probabilistic analysis of the following situation:

We have a set of k balls (set of balls = B)

n boxes (set of boxes = U (urns))

then we place the balls into the boxes, subject to various restrictions.

And we ask for the probability of various events happening.
You think this is easy?
You'll see in a minute, this can be awfully tough.
Some fundamental results of physics depend crucially on estimating the probability of certain events relating to placements of balls into boxes.

In fact, many people whisper behind my back that all I did in my life is place balls into boxes.

There are really 3 cases:

<u>statistics</u> here is another word for <u>sample space</u>

- Maxwell - Boltzman statistics
  k ordinary balls, n boxes and you place balls into boxes.
- Bose - Einstein statistics
  Strikingly different from Maxwell-Boltzman, as you will see.
  The balls are <u>indistinguishable</u>. Identical balls. You will see by example that this does happen in real life. <u>Identical</u> balls.
- Fermi - Dirac statistics
  <u>Identical</u> balls, <u>but</u>
  no more than one ball per box

All these occur in real life.
They also occur in the wonderful theory land of Quantum Mechanics.
Hence the names. They were first identified in Q.M., but there are examples from concrete, everyday life, as you will see.

For the next couple of lectures, we will study these things to death.
And you have to learn it cold.

I will guarantee you one thing.
Every word in this course will be used later in your life.
You are not wasting any time.
On the other hand, you can not afford to fall behind. If you fall one week behind, you will be unable to follow the lecture.
You should take careful notes. Part of your training is to take careful notes of this course. The material in the lectures is strongly at variance with the material in the so-called text, which I am trying to rewrite.
A lot of this material is not covered in any textbook. It is in research papers.
This is not the thing you are going to find in any old book.
If you miss these classes, fine. You won't find this material anywhere else.
For the kind of tuition you pay, that's what you deserve.
Otherwise, you take any old probability book + go to Oshkosh College. It will cost you half as much.

Take down everything I say. And treat it for later references. You may get a job on one remark in this course. I'm not joking.

The theory of distribution and occupancy (beg'g)

This is a central chapter of probability and one that is very easy to visualize. Because this has to do w/ placing balls into boxes.
The setup is:

k B = set of k balls

n U = set of n boxes

You place the balls into boxes in an arbitrary way.

Every ball goes into one box.
The same box may receive several balls.

Ω_{MB} = Maxwell-Boltzmann sample space

This is a finite sample space

sample space: a sample point is a function ω from the set B to the set U.

{ Placing balls into boxes is taking a function from the set of balls to the set of boxes. }

How many such functions are there?

k balls $\left\{ \begin{array}{l} \text{The first ball can be placed in any of } n \text{ boxes.} \\ \text{The second ball can be placed in any of } n \text{ boxes,} \\ \qquad\qquad \text{etc} \end{array} \right.$

Therefore, there are n^k sample points

In the absense of any further information, we will assume all sample points have the same probability.

$$P(\underbrace{\omega}) = \frac{1}{n^k}$$

event consisting of <u>one</u> sample point

<u>Events</u>: all subsets of Ω_{MB}

as happens for all finite sample spaces, where the events are all subsets

If A is an event then $P(A) = \sum_{\omega \in A} P(\omega)$ ← {computed in the almost obvious customary fashion}

$$= \frac{|A|}{n^k}$$ ← number of elements of A

All this is completely obvious.

In order to perform some interesting computations with the Maxwell-Boltzmann sample space, also know as Maxwell-Boltzman statistics, as we said last time, we need to review some classical quantities.
We assume, as known, the following facts:

$n!$ = number of permutations of a set with n elements

Read: "binomial n k" $\binom{n}{k}$ = binomial coefficient
number of $\underbrace{k \text{ subsets}}_{\text{subset w/ } k \text{ elements}}$ of an n set

There are formulas for the binomial coefficient.
Let's introduce one formula for the binomial coefficient:

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

Note that:

$$0! = 1$$

The number of subsets of the empty set:

$$\binom{0}{0} = 1$$

the empty set ϕ has one subset. Itself.

We also use the following notation:
Lower factorial

$$\underbrace{(n)_k}_{} = \overbrace{n\,(n-1)(n-2)\cdots(n-k+1)}^{k \text{ terms}}$$

Read:
"n lower factorial k"

We have the binomial coefficient:

$$\binom{n}{k} = \frac{n!}{(n-k)!}\left(\frac{1}{k!}\right)$$

$$= \frac{(n)_k}{k!}$$

Observe that the derivative $D = \frac{d}{dn}$ is:

$$D(n^k) = k\,n^{k-1}$$

Similarly, we can define the difference operator $\Delta f(n) = f(n+1) - f(n)$
There is a very close analogy to the derivative with the difference operator.

$$\Delta (n)_k = (n+1)_k - (n)_k$$

$$= (n+1)\underbrace{n \cdots (n-k+2)}_{(n)_{k-1}} - \underbrace{n\,(n-1)\cdots(n-k+2)}_{(n)_{k-1}}(n-k+1)$$

$$= \left((n+1) - (n-k+1)\right)(n)_{k-1}$$

$$\Delta (n)_k = k\,(n)_{k-1}$$

Rising factorial

$$\langle n \rangle_k = \overbrace{n\,(n+1)(n+2)\cdots(n+k-1)}^{k \text{ terms}}$$

Read: "Bra n ket k"

Backwards Difference Operator

$$\nabla f(n) = f(n) - f(n-1)$$

$$\nabla \langle n \rangle_k = \langle n \rangle_k - \langle n-1 \rangle_k$$

$$= \underbrace{n(n+1)\cdots(n+k)}_{\langle n \rangle_{k-1}}(n+k-1) - (n-1)\underbrace{n \cdots (n+k)}_{\langle n \rangle_{k-1}}$$

$$= \left[(n+k-1) - (n-1)\right] \langle n \rangle_{k-1}$$

$$\nabla \langle n \rangle_k = k \langle n \rangle_{k-1}$$

Multiset Coefficients – We define the multiset coefficients. Soon to be interpreted.

$$\left\langle {n \atop k} \right\rangle = \frac{\langle n \rangle_k}{k!}$$

"bra n over k ket"

These are quantities that will be interpreted extensively over the next couple of lectures. I thought it better for me to write them out so you know what is coming.

Multinomial coefficient

$$\binom{k}{i_1, i_2, \ldots, i_n} = \begin{cases} 0 & \text{if } i_1 + i_2 + \cdots + i_n \neq k \\ \dfrac{k!}{i_1!\, i_2! \cdots i_n!} & \text{otherwise} \end{cases}$$

Binomial Theorem

$$(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}$$

In the course of these lectures, we will see 17 different proofs of the binomial theorem.

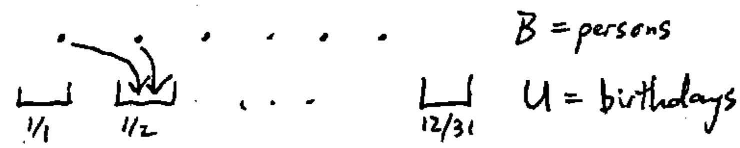
Let's go back to Maxwell-Boltzmann statistics and do some computation.

<u>Example with Ω_{MB}</u>

A = event that no two persons in room 4-370 have the same $\overbrace{\text{birthday}}^{\text{month, day}}$

$P(A)$?

The answer is obtained by setting up the Maxwell-Boltzmann statistics, where the <u>balls are the persons</u> and the <u>boxes are the birthdays.</u>



Two people could have the same birthday.
Each person has only one birthday.

The way it is set up is a <u>function</u> from persons to birthdays. The event that we want is the event that a sample point will be a function that has the property that no two balls will be in the same box.

How many sample points are there in Ω_{MB} with no two balls in the same box?

Let's count.
You take the 1st ball and it can go into any one of the n boxes.
There are n ways of doing things.

Now we take the second ball. It can go into any box <u>except</u> the box occupied by the 1st ball. Which means there are <u>$(n-1)$</u> ways of doing things.

$$\underbrace{n(n-1)(n-2)\cdots(n-k+1)}_{k \text{ terms}} = (n)_k$$

Therefore, the probability of our event is:

$$P(A) = \frac{(n)_k}{n^k}$$

Let's go back to the title of this lecture.
The title of this lecture is <u>The Theory of Distribution and Occupancy</u>

Why <u>distribution</u> and <u>occupancy</u>?

<u>Distribution</u> is the interpretation of a function from the set B to the set U

$$B \longrightarrow U$$

which consists of placing the balls of B into the boxes of U.

There is, however, another interpretation of this same sample space.

<u>Occupancy interpretation</u>

label the balls $1, 2, \ldots, k$

label the boxes (by letters of the alphabet) $\underbrace{\sqcup_a \ \sqcup_b \ \cdots \ \sqcup_c}_{n \text{ boxes}}$ $\longleftarrow$ (The Scottish way of labelling.)

Now, the function from the balls to the boxes means that you take the letter b and place it in position 1
Then letter a in position 2
etc.

baca bbd

You visualize a function as a <u>word</u> make of an alphabet.
This is mathematically the same, but a completely diferent visualization

That's the occupancy interpretation of Maxwell-Boltzmann statistics.

We can ask: what's the probability that a word starts w/ ab?

B = event that a word starts with ab

$$\left.\begin{array}{l} 1 \longrightarrow a \\ 2 \longrightarrow b \\ \text{remaining } k-2 \text{ letters can go anywhere} \end{array}\right\}$$ Therefore, there are n^{k-2} ways of making such a word.

And each one has probability $\frac{1}{n^k}$.
Thus:

$$P(B) = \frac{n^{k-2}}{n^k} = \frac{1}{n^2}$$

Now let's look at a further problem.
For every placement of balls into boxes, we have occupation numbers.
Namely:

Occupation Numbers

$\sqcup \quad \sqcup \quad \dots \quad \sqcup$
$1 \quad 2 \quad \quad \quad n$

we label the boxes $1, 2, \dots n$.
(We could label these $a, b, \dots c$, if you wish.)

θ_1 = number of balls in box 1

for every sample point (for every way of placing the balls into the boxes) there are a certain number of balls in box 1.
That's the occupation number, by definition, of box 1.

θ_2 = number of balls in box 2
etc.

what is $P(\theta_1 = i)$?
Placing the balls into the boxes at random, what's the probability that i balls fall into box 1.

First, out of the set of balls, we choose i balls.
This is done in $\binom{k}{i}$ ways. ← Read: "binomial k i"

Then the remaining balls must go into boxes labelled 2 to n.
Therefore, you have:
$(n-1)^{k-i}$ ways of placing the remaining balls into the remaining boxes.

Since each of these has probability $\frac{1}{n^k}$, we have:

$$P(\theta_1 = i) = \frac{\binom{k}{i}(n-1)^{k-i}}{n^k}$$

Let's fool around with this a bit.
We notice that the events:

$$\underbrace{(\theta_1=0) \cup (\theta_1=1) \cup (\theta_1=2) \cup \dots \cup (\theta_1=k)} = \Omega_{mB}$$

The union of these events is the whole sample space.

These events are disjoint.
If you have 3 balls in the 1st box, you can't have 2 balls in the 1st box.

Therefore: the probability of the union of these events is the <u>sum</u> of the probabilities.

$$P(\theta_1=0)+P(\theta_1=1)+P(\theta_1=2)+\dots+P(\theta_1=k) = P(\Omega_{mB})$$

Q/A: Why do we not have $\dots P(\theta_1=k+1)+\dots+P(\theta_1=n)$?
Because there are only <u>k balls</u>. n is the number of boxes.
We all fall into this trap. We exchange the balls for the boxes. Sometimes it's hard to set up the problem.

We already have the answer to $P(\theta_1=i)$.
So this is:

$$\sum_{i=0}^{k} \frac{\binom{k}{i}(n-1)^{k-i}}{n^k} = 1$$

Let's interpret this.

$$\sum_{i=0}^{k} \binom{k}{i}(n-1)^{k-i} = n^k$$

We just proved the <u>binomial theorem.</u>
The 1st of 17 proofs.

Why is this the binomial theorem?
Because $n^k = (\underbrace{(n-1)}_{\text{this is the } a} + \underbrace{1}_{\text{this is the } b})^k$

$$\boxed{\sum_{i=0}^{k} \binom{k}{i}(n-1)^{k-i} = ((n-1)+1)^k}$$

Q: Isn't this the special case of the binomial theorem?
A: It's special, but you easily see that it implies the general form.
It's only seemingly special.
You split the set of boxes into 2.

Next, let's do a extension of this problem.
What's the probability of having the 1^{st} box have i balls and the 2^{nd} box have j balls?

From the set of k balls, we take i balls and place them in box 1.
You are left w/ $k-i$ balls; from these, you take j balls and place them in box 2.
You are left w/ $k-i-j$ balls, which have to go into boxes labelled 3 to n.
That means that you have only $n-2$ boxes into which $k-i-j$ balls can be placed.

$$P\big((\theta_1=i)\cap(\theta_2=j)\big) = \frac{\binom{k}{i}\binom{k-i}{j}(n-2)^{k-i-j}}{n^k}$$

Next - let's now compute the following probability.
Let us find __all__ the occupation numbers.

probability of each sample point

$$P\big((\theta_1=i_1)\cap(\theta_2=i_2)\cap\dots\cap(\theta_n=i_n)\big)$$
$$= \binom{k}{i_1}\binom{k-i_1}{i_2}\binom{k-i_1-i_2}{i_3}\cdots\binom{k-i_1-i_2-\dots-i_{n-1}}{i_n}\frac{1}{n^k}$$

First, from the set of k balls, take i_1 balls and place them in the 1^{st} box.
We are left w/ $k-i_1$ balls. From these, we pick a subset of i_2 and place them in the 2^{nd} box.
We are left w/ $k-i_1-i_2$ balls. From these, we pick a subset of i_3 and place them in the 3^{rd} box.

if $i_1+i_2+\dots+i_n \neq k$
then you get a contradiction.
The probability of the event is 0 because no ball can be thrown away.

We can simplify the above expression in an extraordinary manner.
Let's write the expressions for the binomial coefficients:

$$= \frac{k!}{i_1!\,\cancel{(k-i_1)!}}\cdot\frac{\cancel{(k-i_1)!}}{i_2!\,\cancel{(k-i_1-i_2)!}}\cdot\frac{\cancel{(k-i_1-i_2)!}}{i_3!\,\cancel{(k-i_1-i_2-i_3)!}}\cdots\frac{\cancel{(k-i_1-i_2-\dots-i_{n-1})!}}{i_n!\,\underbrace{(k-i_1-i_2-\dots-i_{n-1}-i_n)!}_{0!=1}}\cdot\frac{1}{n^k}$$

Now, miracles of cancellation occur.

$$= \frac{k!}{i_1!\,i_2!\cdots i_n!}\cdot\frac{1}{n^k}$$

(this is the multinomial coefficient [2/11/98, 4])

$$= \binom{k}{i_1,i_2,\dots,i_n}\frac{1}{n^k}$$

This is the probability of us finding all of the occupation numbers.

This is really the interpretation of the multinomial coefficient.
The number of ways of placing balls into boxes when the occupation number of each box is assigned.

$$\binom{k}{i_1, i_2, \ldots, i_n} = \text{number of ways of placing } k \text{ balls into } n \text{ boxes with } i_1 \text{ balls in box 1, } i_2 \text{ balls in box 2, etc.}$$

This is a very important result.

In real life, you get Maxwell-Boltzmann statistics thrown at you under various disguises. Being an applied mathematician means being apply to recognize them.
Someone comes in w/ some weirdo statement - but it's nothing but balls into boxes. Nobody can teach you that. You have to learn it by yourself the hard way.

For example, a physicist comes in and says: "Oh, we have particles and states. We want to know how many ways these particles can be in these states"
The particles are the balls.
The states are the boxes.

The only mistakes that I make and that you will make, probably, is to sometimes confuse the balls w/ the boxes.

(*) The way you test yourself (Always remember this)
A ball can only go into one box.
A box may receive several balls.

This is the test for particles and states:
You can have several particles in the same state.
But, to each particle there is only one state.
Which means:
particles are the balls
states are the boxes

Or else, someone comes in w/ questions of linguistics.
You have an alphabet.
What's the probability that the letter "a" does not occur in a word of length n?
It's just balls into boxes.
It is the occupancy interpretation.
In this case, the letters are the boxes
the positions are the balls

pos 1 2 .... n

A B ... Z

We can go on and on w/ all sorts of interpretations.
Chemistry, management, jobs and people, etc.

2/11/98.11

We have time to go to our next sample space in our theory of distribution and occupancy. This is the sample space of dispositions.
This is another finite sample space.

Dispositions Ω_D

We have, again, k balls and n boxes. {Before you are through w/ this course, you will remember this case.}

$\circ \quad \circ \quad \cdots \quad \circ \qquad k$

$\sqcup \quad \sqcup \quad \cdots \quad \sqcup \quad n$

A disposition is a funny thing.
It is the placement of the balls into the boxes - but, after placing the balls into the boxes, then you linearly order the balls in each box.

So a disposition is a placement of the balls into the boxes plus a linear order of the balls in the boxes.

Placement of B to U, but the set of balls in each box is linearly ordered.

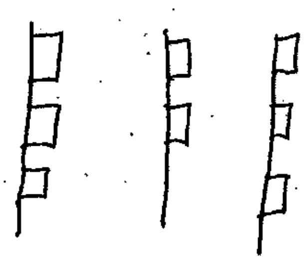
What would be a model of disposition?
My favorite model of disposition is:

balls $\rightarrow$ flags
boxes $\rightarrow$ poles.

You have to raise the flags on different poles,
And the flags on a pole are linearly ordered - some are higher than others.

This is a disposition.
The number of ways of putting k flags on n poles is the number of dispositions of k flags. Because you have to arrange linearly the flags on each pole.

You see the idea. It's not just balls into boxes. It's placing balls into boxes with some extra structure.



k flags
n poles

A Mickey Mouse example with 2 balls and 3 boxes.

Example: $k = 2$, $n = 3$

Let's divide the blackboard into 2 and compare Maxwell-Boltzman with Dispositions.

MB $\overset{a}{\bullet}$ $\overset{b}{\bullet}$

1	2	3
a	b	.
a	.	b
b	a	.
b	.	a
.	a	b
.	b	a
ab	.	.
.	ab	.
.	.	ab

Maxwell-Boltzman has $n^k = 3^2 = 9$ sample points

D $\overset{a}{\bullet}$ $\overset{b}{\bullet}$

1	2	3
a	b	.
a	.	b
b	a	.
b	.	a
.	a	b
.	b	a
ab	.	.
ba	.	.
.	ab	.
.	ba	.
.	.	ab
.	.	ba

So far, it looks just like ordinary placement of balls into boxes.

But now we have this. ab and ba are different now.

These are different dispositions because of the poles.

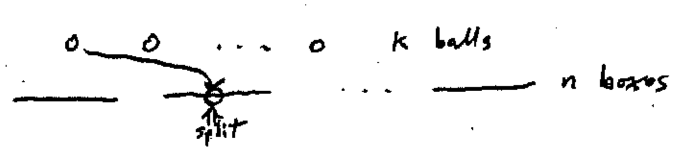
Each member of this list is a sample point of the space of dispositions.
As I told you already, sample points of probability do not look like points.

How many have we got? What will be the formula?
Let's conclude today's meeting with the formula for the number of dispositions.

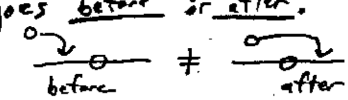
Read: "n, rising factorial k"

There are $\langle n \rangle_k$ dispositions of k balls into n boxes

Proof This is pretty cute. Watch.
I draw the boxes this way ——

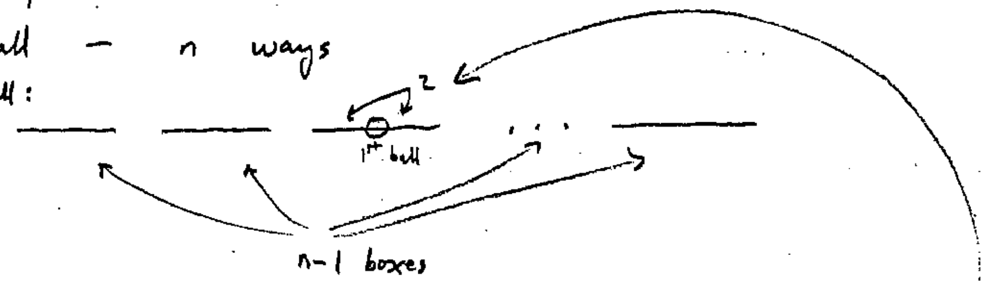


I take the 1st ball and put it into a box.
Now, this act of placing the first ball into the box is the same as if the box were to be split into 2 boxes.
Because the successive ball will give a different disposition, according to whether it goes before or after.



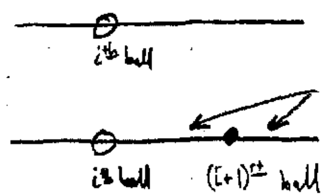
Therefore, it's as if we have one more box for the next ball, to get the disposition.

1^{st} ball — n ways
2^{nd} ball:



The second ball can go any of the $n-1$ boxes plus any of these two, which is $n+1$.

Wherever the 2^{nd} ball goes, it can be assumed to split the portion of the box where it goes into two. Because the remaining balls have a choice to go between.
So even if it goes in the same box as the 1^{st} ball, it cuts the box into two:



So every ball cuts a box into two.

$$\# \text{ of dispositions of } k \text{ balls into } n \text{ boxes} = \overbrace{n(n+1)(n+2)\cdots(n+k-1)}^{k \text{ terms}}$$

$$= \langle n \rangle_k$$

So there are $\langle n \rangle_k$ ways of disposing k balls into n boxes.
Therefore, we define in Ω_D (the sample space of disposition) a probability by setting:

$$\underbrace{P_D(\omega)}_{\substack{\text{probability of} \\ \text{disposition of a} \\ \text{single sample point}}} = \frac{1}{\langle n \rangle_k}$$

And, we have again that the probability of any event A is:

$$P(A) = \sum_{\omega \in A} P_D(\omega) = \frac{|A|}{\langle n \rangle_k}$$

Next time we will see other models for dispositions.
I would appreciate if you think of some interesting other models for disposition.
Because I only have two models in my book.
I'd love to have 4 or 5.

John Guidi
guidi@mit.edu

You have a new problem set today.
It's extremely difficult. 5 problems were taken from research papers.
Don't wait until the last night, the night before the problem set is due.
Work on it. And try to do them jointly as much as possible.
If you can get a sucker to work out these problems for you, you deserve an A. There are very few people on this campus willing to spend his or her time working out these problems.
Remember, the problems in this course are completely different from the quizzes.
The quizzes test strictly the material from the lectures.
The problems are, in general, expansions of material in the lectures.
When you study for the quiz, when the time comes, you will have to study your notes. The quiz will cover strictly the material we cover in the lectures.

As I say, this is an intensive course. With the kind of tuition you pay, I feel morally obliged to give you the hardest problems I can find. What else can I do? Otherwise it's not worth your tuition. So I spent 4 hours yesterday going through research papers preparing your problems. I'm not kidding.

I've given out a number of research problems. The research problems you've received so far are not hard and well worth working on. You can consult any published material for the research problems. You can crib them, if you can find them written up somewhere.

Q/A: I'd love to have a really elegant proof of the Kolmogorov Zero-One Law. Something cute. So I can put it in my book.

The Theory of Distribution and Occupancy (cont'd)

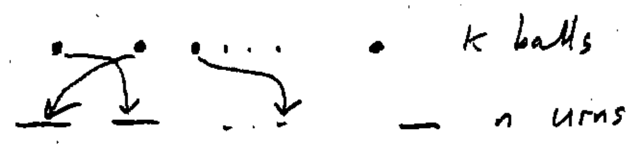
We've seen so far 2 finite sample spaces that pertain to the theory of distribution and occupancy.

1) Maxwell-Boltzman statistics (sample space) Ω_{MB}

In this sample space, you have n^k sample points.

$$P_{MB}(\underset{\substack{\uparrow \\ \text{sample point}}}{\omega}) = \frac{1}{n^k}$$

The sample points correspond to the functions from a set of k balls to a set of n urns:



A sample point is a function from the set of balls to the set of urns.

Such functions can be given a <u>dual interpretation</u>.
The <u>distribution</u> interpretation and the
<u>occupancy</u> interpretation.

The distribution interpretation is the usual interpretation as particles going to states.

The <u>occupancy</u> interpretation is the following:

the balls are numbered $1, \ldots, k$
the boxes are viewed as the letters of an alphabet

Last time we considered the sample points corresponding to the situation where:

$$|B| = 2$$
$$|U| = 3$$

Last time, we considered this from a <u>distribution</u> point of view.

Let's now do the same from an <u>occupancy</u> point of view

$$\underset{a}{\overset{1}{\bullet}} \quad \underset{b}{\overset{2}{\bullet}} \quad \underset{c}{}$$

list of sample points from an <u>occupancy</u> point of view:

$$\left.\begin{matrix} aa \\ ab \\ ac \\ ba \\ bb \\ bc \\ ca \\ cb \\ cc \end{matrix}\right\} \text{words of the sample space}$$

Mathematically the <u>same</u>, but $\underbrace{\text{psychologically \underline{different}}}_{\text{eventually translates into a mathematical statistic}}$

Let's write down some of the results we obtained for Maxwell-Boltzmann statistics

Occupation number of the 1^{st} box:

$$P(\theta_1 = i) = \frac{\binom{k}{i}(n-1)^{k-i}}{n^k}$$

← numerator is the number of sample points.

← $\frac{1}{n^k}$ is probability of a single sample point.

recall that probability of event A is:

$$P(A) = \frac{|A|}{n^k} \longleftarrow \text{number of elements in } A$$

$P(\Theta_1=i)$

This is the occupation number distribution, which is an important datum about Maxwell-Boltzmann, as well as balls into boxes.

We also computed:

$$P\left((\Theta_1=i_1)\cap(\Theta_2=i_2)\cap\dots\cap(\Theta_n=i_n)\right) = \frac{k!}{i_1!\,i_2!\cdots i_n!}\cdot\frac{1}{n^k}$$

<u>if</u> $i_1+i_2+\dots+i_n=k$

This term is also know as the <u>multinomial coefficient</u>:

$$\binom{k}{i_1,i_2,\dots,i_n}$$

As a matter of fact, the computation is the combinatorial definition of the multinomial coefficient.

2) Now let's do the same w/ the sample space of <u>dispositions</u>

Ω_D = Dispositions k balls into n boxes

We have seen that the number of elements in the sample space is:

$$|\Omega_D| = \langle n\rangle_k = n(n+1)(n+2)\dots(n+k-1)$$

This gives a combinatorial meaning to the rising factorial.

Therefore, we define the disposition probability of a sample point as:

$$P_D(\omega) = \frac{1}{\langle n\rangle^k}$$

Let's now perform computations of occupation numbers for dispositions. Occupation numbers are Very Important Data. Because that's what you usually observe.

Confusing balls w/ boxes is a very common mistake. There are situations where you don't know which are the balls + which are the boxes. That is called applied mathematics. Sometimes this is extremely difficult. Sometimes the balls become boxes.

$$P_D(\theta_1 = i) = \frac{\binom{k}{i} \,\square\, \langle n-1\rangle_{k-i}}{\langle n\rangle_k}$$

binomial k i: we place i balls into box 1

we are left w/ $k-i$ balls, which can be <u>disposed</u> in any of the other $n-1$ boxes, $\langle n-1\rangle_{k-i}$

I haven't ordered the balls in box 1 yet. Remember, a disposition doesn't just mean put the balls into the boxes. In each box, you order the balls. If you order them differently, it is different. Therefore, I have to count the number of ways of ordering the balls in box 1. This is just $i!$

$$P_D(\theta_1 = i) = \frac{\binom{k}{i}\, i!\, \langle n-1\rangle_{k-i}}{\langle n\rangle_k}$$

It's an interesting exercise, which I did and I leave for you to work out that:

$$P(\theta_1 = 0) + P(\theta_1 = 1) + \dots + P(\theta_1 = k) = 1$$

and you get a fancy combinatorial identity proved cheapo. You don't have any computations.

$$\sum_{i=0}^{k} \frac{\binom{k}{i}\, i!\, \langle n-1\rangle_{k-i}}{\langle n\rangle_k} = 1$$

$$\boxed{\sum_{i=0}^{k} \binom{k}{i}\, i!\, \langle n-1\rangle_{k-i} = \langle n\rangle_k}$$

a combinatorial alternative for the rising factorial $\langle n\rangle_k = n(n+1)\cdots(n+k-1)$

2/13/98.5

Now, we could do the occupation numbers of box 1 + box 2 simultaneously. As we have done it already for Maxwell-Boltzmann statistics. But that's kind of dull + you have the picture anyway.

Rather, let's jump ahead to the <u>most striking</u> and <u>unexpected phenomenon</u> pertaining to dispositions.

Namely, let's consider this event - that each of the boxes is given an occupation number.

$$P\Big((\theta_1=i_1)\cap(\theta_2=i_2)\cap\ldots\cap(\theta_n=i_n)\Big) = ?$$

$$\text{where } i_1+i_2+\cdots+i_n = k$$

$$(\text{otherwise } P()=0)$$

Let's compute it brutally and then pretend we have a simple proof.
Once we get a result, then we immediately see there ought to have been a simpler way of getting it.

Take i_1 balls and dispose in box 1.
I'm left w/ $k-i_1$ balls and dispose i_2 in box 2.
I'm left w/ $k-i_1-i_2$ balls and dispose i_3 in box 3.
etc

$$= \binom{k}{i_1} i_1!\binom{k-i_1}{i_2} i_2!\binom{k-i_1-i_2}{i_3} i_3!\cdots\binom{k-i_1-i_2-\cdots-i_{n-1}}{i_n} i_n!\cdot\frac{1}{\langle n\rangle_k}$$

Let's simplify this. A wonder of wonders happens.

$$= \frac{k!}{\cancel{i_1!}\cancel{(k-i_1)!}}\cancel{i_1!}\,\frac{\cancel{(k-i_1)!}}{\cancel{i_2!}\cancel{(k-i_1-i_2)!}}\cancel{i_2!}\,\frac{\cancel{(k-i_1-i_2)!}}{\cancel{i_3!}\cancel{(k-i_1-i_2-i_3)!}}\cancel{i_3!}\cdots\frac{\cancel{(k-i_1-\cdots-i_{n-1})!}}{\cancel{i_n!}\underbrace{(k-i_1-\cdots-i_n)!}_{0!=1}}\cancel{i_n!}\cdot\frac{1}{\langle n\rangle_k}$$

$$= \frac{k!}{\langle n\rangle_k}$$

Q/A: Note that $\dfrac{k!}{\langle n\rangle_k} = \dfrac{1}{\left\langle {n \atop k} \right\rangle}$ ← multiset coefficient

You're spoiling my act.

What is remarkable about this result is that it is <u>independent</u> of the <u>occupation numbers</u>.

Balls into boxes with <u>any prescription of occupation numbers</u>, the probability is <u>not</u> dependent on $i_1, i_2, \ldots, i_n$.

This is a very fundamental fact

2/13/98.6

The probability

$$\boxed{P\big((\theta_1=i_1)\wedge(\theta_2=i_2)\cap\cdots\cap(\theta_n=i_n)\big)=\frac{k!}{\langle n\rangle_k}}$$

does ~~not~~ depend on $i_1, i_2, \ldots, i_n$

in striking contrast to Maxwell-Boltzmann, where probability of this event sure does depend on $i_1, i_2, \ldots, i_n$

Easy Proof

Now that we have this result, you say "There must be a reasonable way of seeing that"
Now let's do it easy.
The easy way of seeing that, most people don't get the first time.

Write down all the permutations:

$\{1, 2, \ldots, k\}$

$$k!\left\{\begin{array}{l} 1\ 2\ 3\ 4\,5\ldots k \\ 2\ 1\ 3\ 4\,5\ldots k \\ 2\ 3\ 1\ 4\,5\ldots k \\ \quad\vdots \end{array}\right.$$

You are given $i_1, i_2, \ldots, i_n$
Count the first i_1 and then I put in a stopper (|).
Then I count the next i_2 and then I put in a stopper, etc
I do this in exactly the same position for all of these.

$$k!\left\{\begin{array}{c|c|c|c} \overbrace{}^{i_1} & \overbrace{}^{i_2} & & \overbrace{}^{i_n} \\ 1\ 2 & 3\ 4\ 5 & \ldots & k \\ 2\ 1 & 3\ 4\ 5 & \ldots & k \\ 2\ 3 & 1\ 4\ 5 & \ldots & k \\ & \vdots & & \end{array}\right.$$

When I've done this, I've done all possible dispositions of balls into boxes using the given occupation numbers. Because they are lined up inside the stoppers, as pieces of all possible permutations.
This says that the number of possible ways of disposing the balls into the boxes with given occupation numbers is $k!$.
Because you know where to put the stoppers.

Easy proof: write down all permutations then put in the stoppers. This gives all possible dispositions.

Each of $k!$ dispositions has probability $\frac{1}{\langle n\rangle_k}$
so $P\big((\theta_1=i_1)\cap\ldots(\theta_n=i_n)\big)=\frac{k!}{\langle n\rangle_k}$

- Let's now go into the 3rd sample space in the theory of distribution and occupancy.

<u>Bose – Einstein statistics</u>

• • • . . . • k balls B

— — — . . . — n boxes U

In this case, the balls are <u>indistinguishable</u>

That means that you can visualize a sample point by simply putting checks (✓) corresponding to the number of balls in each box.
All that matters is how many balls are in each box.

In other words, the assignment of the occupation numbers to each box <u>completely determines</u> the placement of indistinguishable balls into the boxes.

All you know is how many balls you have.
So there is a check (✓)

provided that the number of checks (✓) add up to k

Before we interpret this, let's ask the question:

How many such placements are there of indistinguishable balls into boxes?

We've already done that. Look, that's why we did <u>dispositions</u>.
Look at this result

$$P\big((\theta_1=i_1)\cap(\theta_2=i_2)\cap\ldots\cap(\theta_n=i_n)\big) = \frac{k!}{\langle n\rangle_k}$$

for dispositions

That means that all possible distributions of occupation numbers are equally likely in the space of dispositions.

Therefore, how many possible distributions of occupation numbers are there?

One over that, i.e.,

$$\frac{1}{\dfrac{k!}{\langle n\rangle_k}} = \frac{\langle n\rangle_k}{k!}$$

Because they are all equally likely

From this result, we get that the number of possible $i_1, i_2, \dots, i_n$ such that $i_1 + i_2 + \dots + i_n = k$ is:

$$\frac{\langle n \rangle_k}{k!}$$

But this is exactly the number of ways of putting indistinguishable balls into boxes.
Because putting indistinguishable balls into boxes is the same as assigning numbers $i_1, i_2, \dots, i_n$ such that $i_1 + i_2 + \dots + i_n = k$,
And we just saw this.

Ω_{BE} has $\frac{\langle n \rangle_k}{k!}$ sample points $\qquad |\Omega_{BE}| = \frac{\langle n \rangle_k}{k!}$

↑ Bose-Einstein

$$\frac{\langle n \rangle_k}{k!} = \left\langle {n \atop k} \right\rangle$$

is called the <u>multiset coefficient</u>

$$P_{BE}(\omega) = \frac{1}{\left\langle {n \atop k} \right\rangle}$$

Let's prove this result by an entirely different method.
Re-prove it.

Another method to determine $|\Omega_{BE}|$

I will write down $n+k-1$ boxes

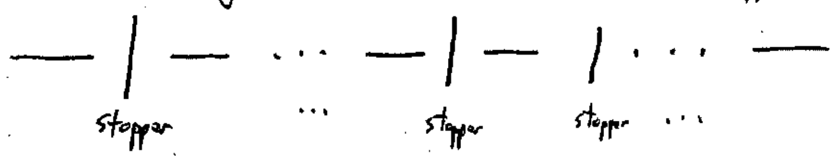
$$\underbrace{\_\_ \quad \_\_ \quad \dots \quad \_\_ \quad \_\_ \quad \dots \quad \_\_ \quad \_\_}_{n+k-1}$$

Now I do the following. Out of these $n+k-1$ boxes, I choose $n-1$ boxes.
How many ways can I do that?

There are $\binom{n+k-1}{n-1}$ subsets of $n-1$ boxes

So I pick from this set of $n+k-1$ boxes, a subset of $n-1$ boxes.
For each of the $n-1$ boxes I pick, what do I do to the boxes?

I do this. If I pick this box, I turn it 90°.
And I stop calling it a box – I call it a stopper.



So I pick the n-1 boxes and turn them into stoppers.
That way, I am left with: $n+k-1$

$$\frac{n+k-1 \quad - (n-1)}{k} \text{ boxes}$$

Each set of boxes is between stoppers.
So that the process of determining how many boxes are between stoppers is the same as the process of checking (✓) boxes.
Therefore, this is the same as placing k indistinguishable balls into n boxes.

But, now you say "the stoppers are completely different."
Well let's see.

$$\binom{n+k-1}{n-1} = \binom{n+k-1}{k} \longleftarrow \binom{n+k-1}{n-1} \; (n+k-1) \to (n-1),\ (k) \; \binom{n+k-1}{k}$$

Identity (choosing the complementary set)

$$= \frac{(n+k-1)_k}{k!}$$

$$= \frac{\overbrace{(n+k-1)(n+k-2)\ldots(n)}^{k \text{ terms}}}{k!} \longleftarrow \left(\begin{array}{l}\text{the numerator is just} \\ \langle n\rangle_k \text{ w/ terms in} \\ \text{the opposite order}\end{array}\right)$$

$$|\Omega_{BE}| = \frac{\langle n\rangle_k}{k!} = \left\langle {n \atop k} \right\rangle$$

So we got the same results.

Let's see a third proof of $|\Omega_{BE}| = \frac{\langle n\rangle_k}{k!}$

This is very close to the argument we've been using. But there is a variant and this is an important result.
I have the indistinguishable balls and I put them into boxes.
I do this in 2 steps:

1) I dispose the balls into the boxes $\langle n\rangle_k$
2) then I divide by $k!$, as the balls are indistinguishable and there are $k!$ permutations of each of these positions.

$$\Bigg\} \quad |\Omega_{BE}| = \frac{\langle n\rangle_k}{k!}$$

Step 2 identifies any dispositions which have the <u>same occupation numbers</u>.
But 2 dispositions have the same occupation numbers if they differ by a <u>permutation</u>.

It is an interesting exercise you should do when you go home to see why this <u>doesn't</u> work w/ Maxwell-Boltzmann statistics. In other words, why can't you just:

1) place the balls into the boxes by Maxwell-Boltzman and then
2) divide by $k!$

to get k indistinguishable balls in n boxes.

I won't tell you why this doesn't work, except you don't get integers. But, convince yourself of the reason why it doesn't work. Do it as an example w/ 2 balls + 3 boxes, and you'll see.

What about an <u>interpretation</u>?

I don't want to be very sophisticated.
Let's compute <u>occupation numbers</u>.

$$P_{BE}(\theta_1 = i) = \left\langle \begin{matrix} n-1 \\ k-i \end{matrix} \right\rangle \cdot \frac{1}{\left\langle \begin{matrix} n \\ k \end{matrix} \right\rangle}$$

$P_{BE}(\omega)$ – probability of sample point in BE

I pick a subset of i balls. Since the balls are <u>indistinguishable</u>, that can be done in only <u>one way</u>.

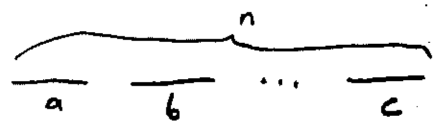
The remaining $k-i$ balls (all indistinguish... are disposed in $n-1$ boxes.

Research Problem
Find evident simplifications of this expression that I can put into my book.
This expression can be simplified in many ways.
Find something elegant.

This result is very important.
When particle physicists first started their game, they thought that all particles were Maxwell-Boltzmann. It's absurd to think that two particles were indistinguishable. <u>But</u>, then they computed the occupation numbers of states and they agreed in the formula.
<u>Not</u> w/ Maxwell-Boltzman formula, but w/ Bose-Einstein formula.
So they were relectantly lead to the conclusion that the particles are indistinguishable from this statistical fact.

Concrete Interpretation of Bose-Einstein statistics

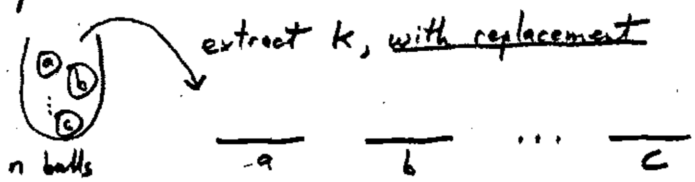
An occupancy interpretation:



The interpretation of the sample points is this.
You have a big urn in which you have n balls labelled a, b, ..., c
You extract a ball,
register the letter,
put the ball back.

Sampling with replacement

You extract k balls with replacement,
Then you add up how many a's you got, how many b's, etc.
Irrespective of the order.



count up how many of each kind you got.

This is called a multiset.
↑ Like a set, except elements may appear w/ a multiplicity.

It is a situation that occurs very often in applications.
(In the 19th Century, multisets were called combinations.)

A multiset is an element of the Bose-Einstein sample space.
Simply - a set of elements, where some element may occur more than once. The idea of a multiset carries the connotation that the same element may appear twice.

Just like there is an algebra of sets, there is also an algebra of multisets.
If you ask me, it's one of those evolutionary choices, that the foundations of mathematics have developed in terms of sets rather than in terms of multisets.

You could very easily conceive of another evolutionary tree where the idea of multisets occurs before the idea of sets.
Perfectly conceivable.
Then logic would have developed differently, etc.
The idea of multisets is just as natural as the idea of sets.

Sets – obey the laws of Boolean Algebra
Multisets – obey the laws of another algebra, which I don't have time to go into.

Last question:

$$P_{BE}\Big(\underbrace{(\theta_1=i_1)\cap(\theta_2=i_2)\cap\ldots\cap(\theta_n=i_n)}\Big)$$

How many sample points in this event? 1
One. Because the assignment of the occupation numbers completely determines the dispositions.
(assuming that $i_1+i_2+\ldots+i_n=k$, of course)

$$\therefore\ P_{BE}\Big((\theta_1=i_1)\cap(\theta_2=i_2)\cap\ldots\cap(\theta_n=i_n)\Big)=\frac{1}{\left\langle {n \atop k} \right\rangle}$$

← Read: "One over multiset n k"

And so, we are almost finished.
We are only left w/ Fermi-Dirac statistics, which we will deal with next time.

John Guidi
guidi@mit.edu

Distribution and Occupancy (cont'd)

There are 4 main sample spaces in distribution and occupancy.
So far, we have covered 3.
Let us cover the last one.

To review:

I) Ω_{MB} : k balls into n boxes

$$P_{MB}(\omega) = \frac{1}{n^k}$$

probability of a sample point - namely, the placement of the balls into the boxes.

$$P\big((\Theta_1 = i_1) \cap (\Theta_2 = i_2) \cap \cdots \cap (\Theta_n = i_n)\big) = \underbrace{\binom{k}{i_1, i_2, \ldots, i_n}}_{\text{multinomial coefficient}} \frac{1}{n^k}$$

probability of a given occupation number

And we computed the probability of several other events in this sample space.

II) Ω_D = dispositions

k balls into n boxes
<u>and</u> balls in each box are linearly ordered

Two models for this are:

1) balls are flags
boxes are poles } you place the flags on the poles. Clearly, the flags on each pole are linearly ordered.

Two different linear orders give you two different dispositions.

2) balls are cars
boxes are lanes at the entrance of a turnpike } It's not enough for cars to choose a lane, but the cars line up linearly in front of the entrance. This is the case for dispositions.

3) Further models?
I will give you a prize for further examples of dispositions.
I'd like some more examples.

$$P_D(\omega) = \frac{1}{\langle n \rangle_k}$$

In the case of dispositions, we have seen the remarkable fact that the distribution of occupation numbers is the same, no matter what the occupation numbers are.

$$P_D\big((\theta_1=i_1) \cap (\theta_2=i_2) \cap \ldots \cap (\theta_n=i_n)\big) = \frac{k!}{\langle n \rangle_k}$$

under the assumption that $i_1+i_2+\cdots+i_n = k$

$$= \left\langle {n \atop k} \right\rangle^{-1}$$

(otherwise, it is 0)

We have reasoned this out in a couple of ways [2/13/98. 4-7].

III) Ω_{BE} - Bose-Einstein sample space

the balls are indistinguishable

We have seen, by reasoning in terms of dispositions [2/13/98, 7-10], that:

$$P_{BE}(\omega) = \frac{1}{\left\langle {n \atop k} \right\rangle} \longleftarrow \text{multiset coefficient}$$

The interpretations of Bose-Einstein sample space are several. Let's review a couple.

distribution interpretation - views the balls as being put into the boxes

occupancy interpretation - views boxes as letters of the alphabet and the balls as places where we put the letters. Sample points are viewed as words.

$\left\{ \begin{array}{l} \text{I don't know a good occupancy interpretation for dispositions.} \\ \text{If you can think of one, I'd appreciate it.} \\ \text{I don't know a good way to visualize this} \end{array} \right\}$

Distribution interpretation of Bose - Einstein

You have the boxes, with checks in the boxes.
The number of checks adding up to k.

$\underline{\checkmark}\quad \underline{\checkmark\checkmark\checkmark}\quad \underline{\checkmark}\quad \cdots\quad \underline{\checkmark\checkmark}$

The balls being indistinguishable, all the data is how many balls are in each box.

Therefore, the distribution interpretation of BE can be interpreted as follows:

A <u>sample point ω</u> is simply the possible distribution of the b's, i.e., the occupation numbers.

From a distribution point of view, the elements of BE are simply possible occupation numbers.
That's how we obtained $P_{BE}(\omega) = \dfrac{1}{\left\langle {n \atop k} \right\rangle}$

Occupancy interpretation of Bose-Einstein

The occupancy interpretation is a little more visualizable.
View the boxes as letters:

$$\overbrace{\underset{a}{\underline{\quad}}\ \ \underset{b}{\underline{\quad}}\ \ \cdots\ \ \underset{c}{\underline{\quad}}}^{n}$$

n labelled balls are in an urn.
A sample point in Bose-Einstein sample space means:

do this k times $\left\{ \begin{array}{l} \text{pick out one ball from the urn,} \\ \text{register which ball it is,} \\ \text{put the ball back in the urn.} \end{array} \right.$

At the end, you have the data of how many times you have extracted ball "a", how many times ball "b", etc.
Altogether k extractions.

This is called, as we said last time, a <u>multiset</u>.

You can view the boxes as a set.
You are taking subsets of the the set, but a subset where some elements may be <u>repeated.</u>
This is a natural concept.

For example, suppose you use as letters $a, b, \ldots, c$ and you consider them as variables.

Then you ask: How many monomials of degree k can be made w/ these variables?

And finally:

IV) Ω_{FD} – Fermi-Dirac sample space

[Look at the illustrious names that are associated w/ these sample spaces, or statistics, as they are called.]

again k balls and n boxes
balls are indistinguishable
all occupation numbers $\theta_i = \begin{cases} 0 \\ 1 \end{cases}$

this is called the Pauli exclusion principle.

How many sample points are there?

$$\_\_ \quad \checkmark \quad \checkmark \quad \ldots \quad \_\_ \quad \checkmark$$

Each box either has a check ($\checkmark$) or not.
The balls are indistinguishable.
You check k of the n boxes.
Which means you are taking a subset of k boxes out of the n boxes.

So a sample point ω in Fermi-Dirac statistics is just a subset.
And physicists make a big deal of this.

A sample point ω is a k-subset of U.
Therefore:

$$P_{FD}(\omega) = \frac{1}{\binom{n}{k}}$$

Let's do, as an exercise, the following:

$$P_{FD}(\theta_1 = 0) = \frac{\binom{n-1}{k}}{\binom{n}{k}}$$

you take a subset of all the boxes and you leave out the first box (since $\theta_1 = 0$). Which means you pick a k-subset from n-1 boxes.

$$P_{FD}(\theta_1 = 1) = \frac{\binom{n-1}{k-1}}{\binom{n}{k}}$$

you pick one ball and put it in the first box. Since the balls are indistinguishable, it doesn't matter which, you just check the box. The rest (k-1) go into the rest of the boxes.

$$\underbrace{(\theta_1 = 0) \cup (\theta_1 = 1)}_{\text{independent events}} = \underbrace{\Omega_{FD}}_{\text{whole sample space}}$$

Then we have:

$$P(\theta_1 = 0) + P(\theta_1 = 1) = 1$$

$$\binom{n-1}{k} + \binom{n-1}{k-1} = \binom{n}{k}$$

This is a probabilistic proof of this identity.

Let's do similarly w/ Bose-Einstein:

You people don't ask enough questions. I'm not giving out chocolate bars. I never had a class full of geniuses before.

$$P_{BE}(\theta_1 = i) = \frac{\left\langle \begin{matrix} n-1 \\ k-i \end{matrix} \right\rangle}{\left\langle \begin{matrix} n \\ k \end{matrix} \right\rangle}$$

[2/13/98.10]

there's only one way of placing i indistinguishable balls in the 1st box. That means the remaining $k-i$ balls are placed, in a Bose-Einstein way, in the remaining n-1 boxes.

$$\underbrace{(\theta_1 = 0) \cup (\theta_1 = 1) \cup \ldots \cup (\theta_1 = k)}_{\text{all disjoint}} = \Omega_{BE}$$

since box 1 has some number, or none, of balls.

Taking probabilities:

$$P_{BE}(\theta_1 = 0) + P_{BE}(\theta_1 = 1) + \ldots + P_{BE}(\theta_1 = k) = 1$$

which gives:

$$\left\langle \begin{matrix} n-1 \\ k \end{matrix} \right\rangle + \left\langle \begin{matrix} n-1 \\ k-1 \end{matrix} \right\rangle + \left\langle \begin{matrix} n-1 \\ k-2 \end{matrix} \right\rangle + \ldots + \left\langle \begin{matrix} n-1 \\ 0 \end{matrix} \right\rangle = \left\langle \begin{matrix} n \\ k \end{matrix} \right\rangle$$

There is a strange analogy between the $\underbrace{\text{multiset coefficient } \left\langle {n \atop k} \right\rangle}_{\text{counts the number of multisets of size } k \text{ from a set of size } n.}$

and the $\underbrace{\text{binomial coefficient } \binom{n}{k}}_{\text{counts the number of subsets of size } k \text{ from a set of size } n.}$.

There is a very strange analogy between the two.
Which is one of the deepest problems of mathematics.

Negative sets are multisets.
Watch.

$$\binom{-n}{k} = \frac{\overbrace{(-n)(-n-1)\cdots(-n-k+1)}^{k \text{ terms}}}{k!}$$

$$= (-1)^k \frac{(n)(n+1)\cdots(n+k-1)}{k!}$$

$$= (-1)^k \frac{\langle n \rangle_k}{k!}$$

$$\boxed{\binom{-n}{k} = (-1)^k \left\langle {n \atop k} \right\rangle}$$

There is <u>no</u> satisfactory explanation of this identity.
If you get one, you get an A and you are excused from any further work in this course.

It's a very deep problem that no one has been able to figure out.
What is the interpretation of this identity?

Now - we want to do some deeper computations in those sample spaces.
And to this end, we must develop some more powerful techniques for computing probability.
At this point, I was going to just give you the Principle of Inclusion/Exclusion, then I thought - "Gee, they pay such high tuition. I can't give the something straight out of the book."
So, I will give something not in the book (not in any book), which is the general theory for establishing identities of probability.

Mr. Guidi, take this down carefully. ☺

Rényi's Principle

- Let's see some examples of identities holding in probability spaces.
The first example we have is:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \qquad [2/4/98.4\text{-}5]$$

This is an identity we established in the first day of class.

- Another identity is: Let's try to expand it and get a simple identity

$$P(A_1 \cup A_2 \cup A_3) = \underbrace{P(A_1 \cup (A_2 \cup A_3))}_{\text{applying identity for union of 2 sets}} \quad \text{associativity of } \cup \text{ of sets}$$

$$= P(A_1) + P(A_2 \cup A_3) - P(\underbrace{A_1 \cap (A_2 \cup A_3)}_{\text{distributive law}})$$

$$= P(A_1) + P(A_2) + P(A_3) - P(A_2 \cap A_3) - P((A_1 \cap A_2) \cup (A_1 \cap A_3))$$

$$= P(A_1) + P(A_2) + P(A_3) - P(A_2 \cap A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) + P(A_1 \cap A_2 \cap \cancel{A_1} \cap A_3)$$

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3)$$
$$- P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3)$$
$$+ P(A_1 \cap A_2 \cap A_3)$$

If you intersect twice w/ A_1, you might as well skip one of the $\cap$.

For the event $A_1 \cup A_2 \cup A_3$ we have the above inclusion/exclusion formula

- Now the inclusion/exclusion principle is the generalization of this formula to n events.
It can be proved brutally by plugging in. We could do that and there is a proof in the book.
But then, there are so many other formulas that come up and each time you have to prove them from scratch.

We are led to ask: Is there a super method that leads to all possible formulas pertaining to events and probabilities?

And it is an amazing fact that there is a super method which will check any formula.

By way of guidance, let's write the Inclusion-Exclusion Principle in general:

why they call it a "principle" beats me. It's just a formula.

Inclusion-Exclusion Principle

$A_1, A_2, \ldots, A_n$ = events in a sample space Ω, w/ probability P

Then the probability that at least one of the events occurs is given by:

which means the union

(**) $P(A_1 \cup A_2 \cup \ldots \cup A_n)$

$$= \underbrace{\sum_i P(A_i)}_{\text{sum of probability of each individual event}} - \underbrace{\sum_{i<j} P(A_i \cap A_j)}_{\text{subtract the probabilities of all the pairs.}} + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \ldots + \ldots \pm P(A_1 \cap A_2 \cap \ldots \cap A_n).$$

period

Except you have to be very careful to count each pair only once.
E.g., don't want to count both $A_1 \cap A_2$ and $A_2 \cap A_1$
All the pairs are counted only once.
You could take:

$$-\frac{1}{2}\sum_{\underline{all}\text{ pairs}} P(A_i \cup A_j)$$

but dividing by 2 gets you into trouble.
Instead, we limit the sum w/ $\sum_{i<j}$

It's a very powerful formula, as we are going to see.

Q: Can you generalize for infinite events?
A: Yes, but it's never been satisfactorily generalized.
I'll come back to that

The inclusion/exclusion formula is also an example of an identity holding in a probability space. And we'll soon see that there are other, more complicated, identities that have been developed to solve more complicated problems. There are oozes of identities.

There are also inequalities.
For example:

$$P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$$

which is immediate from the formula: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

So there are also complicated inequalities.

Now, we will see that there is one method for checking every identity and every inequality in a probability space.
Let us first state Réyni's Principle in a sort of philosophical way and later we will prove it.

I don't like the word involving, but I see no way out here.

~~Réyni's~~ Rényi's Principle

An identity or inequality involving n events $A_1, A_2, \ldots, A_n$ in a probability space holds for all $A_1, A_2, \ldots, A_n$

if and only if

it is true in the special case where $\underbrace{\text{each } A_i = \begin{cases} \phi \\ \Omega \end{cases}}_{\text{each } A_i \text{ is either the null set or the whole sample space}}$

This is extraordinary
In order to accept a formula, you only have to check it when the A_i's are the whole sample space and null.
If its true under these cases, then it's true forever.
So you can write down the most complicated inequality or identity and it's a cinch to check them.
That's Réyni's Principle.
It's a decision procedure to check all possible inequalities and identities.
We will prove it later. Now, let's apply it.

Application – A Mickey Mouse application

Let's prove:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

Let's pretend we don't know this. Let's apply Réyni's Principle to check it.

We only to check the following cases:

(1) $A_1 = \phi$, $A_2 = \phi$: $0 = 0 + 0 - 0$ ✓

(2) $A_1 = \phi$, $A_2 = \Omega$: $1 = 0 + 1 - 0$ ✓

(3) $A_1 = \Omega$, $A_2 = \phi$: $1 = 1 + 0 - 0$ ✓ (Symmetrical w/ (2))

(4) $A_1 = \Omega$, $A_2 = \Omega$: $1 = 1 + 1 - 1$ ✓

Therefore, identity is true for every event, says Réyni.
You got away w/ murder

- Let's check next the inclusion-exclusion formula by this Principle

Verification of (by Réyni's Principle):

$$(**)\quad P(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k)$$
$$- \ldots + \cdots \pm P(A_1 \cap A_2 \cap \ldots \cap A_n).$$

First, let me remind you that:

$$(1-1)^k = 0$$

In our 1st of 17 proofs of the binomial theorem [2/11/98.7-8], we showed that:

$$\sum_{i=0}^{k} \binom{k}{i} (n-1)^{k-i} = n^k$$

Let $n = 1-1$. This gives:

$$\sum_{i=0}^{k} \binom{k}{i} (-1)^{k-i} = 0$$

$$\pm \binom{k}{0} \mp \binom{k}{1} \pm \binom{k}{2} \mp \ldots + \binom{k}{k} = 0$$

sign dependent on whether k is even (+) or odd (-)

$$\binom{k}{0} = 1 \leftarrow \text{the empty set } \phi$$

Move 1st term to RHS.
Multiply both sides by -1, but <u>only if</u> k is even.
This gives:

$$\boxed{\binom{k}{1} - \binom{k}{2} + \cdots \pm \binom{k}{k} = 1}$$

We are going to use this.

By Réyni's Principle, in order to verify the identity $(**)$, all you have to do is choose some of the A_i's to be the whole sample space Ω and some to be ϕ.

By symmetry, we choose the 1st k A_i's to be Ω and the remaining $n-k$ A_i's to be ϕ. This is symmetric.

We need only check $(**)$ when:

$$A_1 = A_2 = \ldots = A_k = \Omega$$
$$A_{k+1} = \ldots = A_n = \phi$$

if we check it and it is true for <u>all</u> k, then it has to be true for <u>all</u> events.

$k=0$

$$A_1 = A_2 = \cdots = A_n = \phi$$

$$\underset{\checkmark}{0 = 0} \longleftarrow LHS = RHS \text{ of } (**)$$

$k>0$

$LHS = 1 \longleftarrow$ since <u>at least one</u> of the $A_i = \Omega$ (the whole space)

$$1 \overset{?}{=} \underbrace{\sum_i P(A_i)} - \underbrace{\sum_{i<j} P(A_i \cap A_j)} + \underbrace{\sum_{i<j<k} P(A_i \cap A_j \cap A_k)} - \cdots$$

k of A_i equal Ω. So k of $P(A_i)=1$. The other $n-k$ have $P(A_i)=0$. So this is k.

Take all pairs that are the whole space. Because if one of them is ϕ, then $\cap = \phi$. So you take <u>any</u> unordered <u>pair</u> of $A_1, A_2, \ldots, A_k$, since $A_1 = A_2 = \cdots = A_k = \Omega$. Any such pair has: $P(A_i \cap A_j) = P(\Omega) = 1$. How many unordered pairs are there out of a set of k elements? $\binom{k}{2}$

Here, you take all the <u>triples</u> of the k A_i's which have the whole space Ω. Otherwise $\cap = \phi$ and $P(\phi) = 0$. There are $\binom{k}{3}$ ways to choose such triples

$$= k - \binom{k}{2} + \binom{k}{3} - \cdots \pm \binom{k}{k}$$

On the previous page, we've shown that

$$\underbrace{\binom{k}{1}}_{=k} - \binom{k}{2} + \binom{k}{3} - \cdots \pm \binom{k}{k} = 1$$

Therefore, RHS above $=1$

$1 = 1$ ✓

Therefore, using Reýni's Principle, we have proved this fantastic inclusion/exclusion formula, With <u>no work!</u>

Actually, an application of In-Ex Principle.

Before going on to fancier formulas, let's have some applications of Réyni's Principle.
Beginning w/ the most famous one, which is the problem of the matching.

Application of Réyni's Principle – Matching Problem

I have 2 identical decks of cards.
I shuffle one.
Then I place them, one deck beside the other.
Then I turn over the ~~cards~~ top card in each deck simultaneously.
What is the probability that I get at least one match?

Or, equivalently, what's the probability I get no matches.

200 years ago, the Chevalier de Montmort, who is one of the earliest workers in probability, thought that this problem, together w/ the inclusion/exclusion principle, was "at the extreme limit of human reasoning." Now, we have it in 18.313 after 2 weeks.

What are we really talking about?
We have a set of n elements.

$$\underset{1}{\bullet}\ \underset{2}{\bullet}\ \bullet\ \bullet\ \cdots\ \underset{n}{\bullet}$$

And we take an arbitrary permutation

$$\underset{i_1}{\bullet}\ \underset{i_2}{\bullet}\ \underset{i_3}{\bullet}\ \underset{i_4}{\bullet}\ \cdots\ \underset{i_n}{\bullet}$$

We are asking for the probability that the random permutation will have a fixed point.

A = event that a random permutation has at least one fixed point.

There are many other ways of stating this.
For example:

The secretary collects grades from the students, but he mixes up the grades. What is the probability that a student will come out with his or her correct grade?

Or you have a party and a gentleman gives you a ticket for your top hat. (It's 50 years ago). And he mixes up the tags. What is the probability that the gentleman will give you back your own top hat?

Next time, we will apply the inclusion/exclusion principle to see how easy it is to solve these problems.

John Guidi
guidi@mit.edu

The inclusion-exclusion principle (cont'd)

Last time, we stated w/o proof Réyni's Principle.
We are going to see it again.
You can verify, from Réyni's Principle, one of the most important formulas in probability theory. Namely, the inclusion-exclusion principle.
The inclusion-exclusion principle says the following:

$A_1, A_2, \ldots, A_n$ are any events (in a probability space - that goes w/o saying)

The inclusion-exclusion principle allows you to express the probability of one of the events happening, in terms of the probability of certain subsets of the events.

(*) $$\underbrace{P(A_1 \cup A_2 \cup \ldots \cup A_n)}_{\text{probability that one of the events happens is the union of the events}} = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \ldots + \ldots + (-1)^{n+1} P(A_1 \cap A_2 \cap \ldots \cap A_n)$$

As I said, 200 years ago it was considered the utmost achievement - the limit of human reasoning.
It has extraordinarily vast applications for all sorts of situations.
So now let's see a few of them.
First, the problem of the matching.

Example: Matching

At commencement, every student gets a degree.
But the secretary mixed up the degrees, so that the president hands out the degrees in the wrong order.
What's the probability that at least one student will get his/her own degree?

$\Omega = \{\omega = \text{permutations of the set } \{1, 2, \ldots, n\}\}$

Ω = space of all sample points, which are permutations of the set $\{1, 2, \ldots, n\}$
A sample pt ω is a permutation

$$\underbrace{P(\omega)}_{\text{probability of a sample point}} = \frac{1}{n!}$$

Because in the absence of any further information, we give all permutations the same probability.

A = event that a random permutation has at least one fixpoint $(\underbrace{\omega(i)}_{} = i)$

ω is a permutation
$\omega(i)$ is the i^{th} element of the permutation ω

$P(A) = ?$

We compute this probability by the inclusion-exclusion principle.

A_1 = event that $\omega(1) = 1$ (i.e., that element 1 is a fixpoint)

A_2 = event that $\omega(2) = 2$ (" " " 2 " " ")

etc.

Then, of course:

$$A = A_1 \cup A_2 \cup \ldots \cup A_n$$

$P(A)$ is the probability that a permutation has at least one fixpoint.

↑ maybe more

So we compute $P(A)$ by the inclusion-exclusion principle, which I am about to cover.

$$\underbrace{P(A_1)}_{\text{probability that a permutation sends 1 into 1}} = \frac{(n-1)!}{n!}$$

← With 1 fixed, for 2 through n there are $(n-1)!$ permutations

← each permutation has probability $\frac{1}{n!}$

$$= \frac{1}{n}$$

Similarily:

$$P(A_i) = \frac{(n-1)!}{n!} = \frac{1}{n}$$

Now let's compute probability of A_1 and A_2:

$$P(\underbrace{A_1 \cap A_2}_{\text{want 1 to remain fixed and 2 to remain fixed in the permutations}}) = \frac{(n-2)!}{n!}$$

← you are producing the numbers 3 through n in an arbitrary way.

$$= \frac{1}{n(n-1)}$$

Similarily, of course:

$$P(A_i \cap A_j) = \frac{1}{n(n-1)} \quad , i<j$$

By the same reasoning.

It is clear that this kind of reasoning can be generalized:

$$\underbrace{P(A_i \cap A_j \cap A_k)}_{\substack{i^{\text{th}} \text{ point is fixed} \\ \text{and } j^{\text{th}} \text{ point is fixed} \\ \text{and } k^{\text{th}} \text{ point is fixed}}} = \frac{(n-3)!}{n!} \quad , i<j<k$$

you are permuting $n-3$ points arbitrarily.

$$= \frac{1}{n(n-1)(n-2)} \quad , i<j<k$$

With these data, we can now apply the inclusion-exclusion principle.

By $\overset{[2/18/98.1]}{(*)}$, we have:

$$P(A) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots + \cdots + (-1)^{n+1} P(A_1 \cap A_2 \cap \ldots \cap A_n)$$

$$= n \cdot \frac{1}{n} - \binom{n}{2}\frac{1}{n(n-1)} + \binom{n}{3}\frac{1}{n(n-1)(n-2)} - \cdots + \cdots + (-1)^{n+1}\frac{1}{n!}$$

n terms, each one equal to $\frac{1}{n}$.

all unordered pairs possible $\binom{n}{2}$, each one equal to $\frac{1}{n(n-1)}$.

all unordered triples possible $\binom{n}{3}$.

etc.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{(n)_k}{k!}$$

} different ways of simplifying the binomial coefficient

$$= 1 - \frac{\cancel{(n)_2}}{2!}\frac{1}{\cancel{n(n-1)}} + \frac{\cancel{(n)_3}}{3!}\frac{1}{\cancel{n(n-1)(n-2)}} - \cdots + \cdots \pm \frac{1}{n!}$$

$$P(A) = 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + \cdots \pm \frac{1}{n!}$$

That's our probability.
Is there a simple way of computing this probability?
Sure.

You recall from 18.01 :

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots - \cdots$$

set $x = -1$

$$e^{-1} = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \cdots - \cdots$$

It just so happens that the Taylor formula for e^x converges extremely fast. So that after 4 terms, you have a 4 digit approximation.

So our probability $P(A)$ is approximately equal to :

$$P(A) \simeq 1 - e^{-1}$$

$$\underbrace{P(A) = 1 - \frac{1}{e}} = .6321$$

For all practical purposes, this is the answer

We have a remarkable conclusion :

The probability of at least one match, in this matching problem, is independent of the number degrees/cards/etc.

This is paradoxical at first, but facts are facts.

Whether $n = 52$ or $n = 320$, the probability is practically the same.

Let's do another one:

Example Ω_{mB}

I want the probability that all the boxes are occupied.

event $\Big((\theta_1 > 0) \cap (\theta_2 > 0) \cap \ldots \cap (\theta_n > 0)\Big)$

The complement of this event, by De Morgan's Law :

$$(\theta_1 > 0)^c \cup (\theta_2 > 0)^c \cup \ldots \cup (\theta_n > 0)^c$$

where $(\theta_i > 0)^c = (\theta_i = 0)$ ← if you don't see that, I quit.

call this event A $\quad A = \underbrace{(\theta_1 = 0)}_{A_1} \cup \underbrace{(\theta_2 = 0)}_{A_2} \cup \ldots \cup \underbrace{(\theta_n = 0)}_{A_n}$

Let's compute $P(A)$ by the inclusion-exclusion principle.

In order to apply the inclusion-exclusion principle, we compute the probabilities of various intersections of these events.
Let's do that.

$$\underbrace{P(A_i)}_{\text{probability that box } i \text{ is zero}} = \frac{(n-1)^k}{n^k}$$

← box i empty means that all the k balls go into $n-1$ boxes

← each sample point has probability $\frac{1}{n^k}$

Next:

$$\underset{i<j}{P(A_i \cap A_j)} = \frac{(n-2)^k}{n^k}$$

← same idea. All the balls go into the other boxes.

etc.

Now we are ready to apply the inclusion-exclusion principle:

$$P(A) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \dots + \dots$$

$$= n\,\frac{(n-1)^k}{n^k} - \binom{n}{2}\frac{(n-2)^k}{n^k} + \binom{n}{3}\frac{(n-3)^k}{n^k} - \dots + \dots$$

Further simplification is possible, but not much.
In fact, the simplification of this expression is a cottage industry.
There are many ways of simplifying it.
It doesn't simplify like the previous one.
I'm tempted to simplify it, but we'd spend 15 minutes simplifying it and it's not worth it at this stage.

Finally, we want probability of the event that all the boxes are occupied.
This is the event A^c.
And we obtain this as:

$$P(A^c) = 1 - P(A)$$

Now let me pull a fast one on you.

Example Ω_{BE}

Let A be the same event, namely that at least one box is empty.

$$A = \underbrace{(\theta_1 = 0)}_{A_1} \cup \underbrace{(\theta_2 = 0)}_{A_2} \cup \ldots \cup \underbrace{(\theta_n = 0)}_{A_n}$$

what is $P_{BE}(A)$?

Do it:

$$\underbrace{P(A_i)}_{\text{probability that box } i \text{ is empty}} = \frac{\left\langle {n-1 \atop k} \right\rangle}{\left\langle {n \atop k} \right\rangle}$$

$\leftarrow$ all k indistinguishable balls go into $n-1$ boxes. The number of ways this can be done is the multiset coefficient $\left\langle {n-1 \atop k} \right\rangle$.

$$\underset{i<j}{P(A_i \cap A_j)} = \frac{\left\langle {n-2 \atop k} \right\rangle}{\left\langle {n \atop k} \right\rangle}$$

$\leftarrow$ number of ways of placing the k indistinguishable balls into $n-2$ boxes.

etc.

Therefore, by the mutual exclusion principle:

$$P(A) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \ldots + \ldots$$

event that at least one box is empty.

$$= n \frac{\left\langle {n-1 \atop k} \right\rangle}{\left\langle {n \atop k} \right\rangle} - \binom{n}{2} \frac{\left\langle {n-2 \atop k} \right\rangle}{\left\langle {n \atop k} \right\rangle} + \binom{n}{3} \frac{\left\langle {n-3 \atop k} \right\rangle}{\left\langle {n \atop k} \right\rangle} - \ldots + \ldots$$

That seems to be the answer.
Now – where did I pull a fast one?
This is correct.

This is stupid. Because I could have done it this way.

The balls are indistinguishable. I pick n balls and put each one in one box. They don't know.
That's the fundamental principle. Balls do not think.
So this is very deep actually.
We often think that they think.

We take any n of the balls and put each in a separate box.
How many ways can that be done? <u>One.</u>
Because the balls are <u>indistinguishable</u>. You put a check (✓) on it.
We are left with $k-n$ balls, which we put in any way in the boxes.

A = event <u>at least one</u> box is empty

A^c = event all the boxes are occupied

$$P(A^c) = \frac{\left\langle {n \atop k-n} \right\rangle}{\left\langle {n \atop k} \right\rangle}$$

That means I take the remaining $k-n$ balls and put them into n boxes. (You have to have more balls than boxes $k \geq n$)

Therefore:

$$P(A) = 1 - P(A^c)$$

$$= 1 - \frac{\left\langle {n \atop k-n} \right\rangle}{\left\langle {n \atop k} \right\rangle}$$

Taking the series we computed for $P(A)$ on the previous page and combining it with the above — we have proved a very interesting <u>identity</u>, with which you can tease your friends.

$$n \frac{\left\langle {n-1 \atop k} \right\rangle}{\left\langle {n \atop k} \right\rangle} - \binom{n}{2} \frac{\left\langle {n-2 \atop k} \right\rangle}{\left\langle {n \atop k} \right\rangle} + \binom{n}{3} \frac{\left\langle {n-3 \atop k} \right\rangle}{\left\langle {n \atop k} \right\rangle} - \cdots + \cdots = 1 - \frac{\left\langle {n \atop k-n} \right\rangle}{\left\langle {n \atop k} \right\rangle}$$

$\binom{n}{1} = n$

$$\left\langle {n \atop k} \right\rangle = \left\langle {n \atop k-n} \right\rangle + \sum_{i=1}^{n} (-1)^{n+1} \binom{n}{i} \left\langle {n-i \atop k} \right\rangle$$

How do you prove that? If you don't know the secret, it's a little tough.
There is an apocryphal story about von Neumann being given this problem (or a variant of it) and immediately he came up with this. But then one of his friends gave him this and he says "How did you write that that fast?"

- Now let's go back to Réyni's Principle and get some deeper identities for probabilities.
Let's jaze this up.

Again, given events $A_1, A_2, \dots, A_n$, we can ask other questions about these events.

The inclusion-exclusion principle gives you the probability that at least one of the events happens.

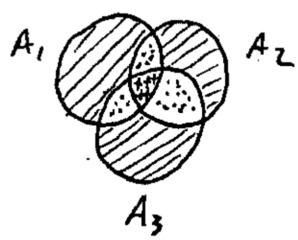
What if we need to compute the probability that exactly one of the events happens?

When you have a problem like that, what is step 1? Give it a name.

Let W_1 = event that exactly one of the A_i's happens (the letter "W")
W_2 = event that exactly two of the A_i's happens
etc.

Let's derive a formula for the probabilities of these events. Let's start w/ W_1, which will give us a good orientation.

W_1, in set theoretic terms, is the set of all sample points that belongs to exactly one of the A_i. We have the following picture



/// = W_1 - exactly 1 event
:: = W_2 - exactly 2 events
⁂ = W_3 - exactly 3 events

There are many situations where we need to compute this. Very practical situations. I'd be interested if you come up with some interesting practical examples. Life and death examples, for me to put in my book, so people buy the book.

$$P(W_1) = ?$$

I want to first define some simplifying summations. This is standard notation. Let me motivate it this way. Go back to the inclusion-exclusion principle. You give a name to the first sum, you give a name to the second sum, etc. It's useful to do that because the same sums occur in the expressions for the probabilities of the W_i.

$$S_0 = 1$$

$$S_1 = \sum_{i=1}^{n} P(A_i)$$

$$S_2 = \sum_{i<j} P(A_i \cap A_j)$$ S_2 is the sum of probabilities of all pairs of intersections.

etc.

Observe that $S_1, S_2, \ldots$ are not the probabilities of any events. They are just numbers.

In terms of S_i, the inclusion-exclusion principle can be written as follows:

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = S_1 - S_2 + S_3 - \ldots + \ldots$$

Another way of writing the inclusion-exclusion principle.
And this gives us a lead on how to compute the probability of W_1.

$$(*) \quad \underbrace{P(W_1)}_{\text{probability that exactly one of the events happens}} \overset{?}{=} S_1 - 2S_2 + 3S_3 - \ldots + \ldots$$

Let's check whether this is true or not.
Let's check this really quickly.
How do we check this?
We check it by Réyni's Principle
In other words, this formula is true if it is true when the events A_i are allowed to be only either the whole sample space (Ω) or the empty event (ϕ).
If it checks for these cases, we win.
Keep your fingers crossed.

By Réyni's Principle:

expression being symmetric,
set $A_1 = \Omega \quad A_2 = \ldots = A_n = \phi$

LHS of $(*)$ = 1 $\longleftarrow$ $P(W_1)$ = probability that exactly one of the A_i's occurs.
Every sample pt ω belongs to exactly one of the A_i's.
W_1 is the event that only one of the A_i's happens.
If A_1 is Ω and the rest of the A_i are ϕ, then for sure exactly one of the A_i happens.
So $P(W_1) = 1$

$S_1 = 1$, since $P(A_1) = 1$ and for $i = 2, \dots, n$, $P(A_i) = 0$.

What about S_2? S_2 involves the intersection of 2 events. But one of them will be the <u>empty set</u>. Therefore:

$$P(A_i \cap A_j) = 0, \quad \forall i,j \quad (i<j)$$

Same idea holds for $S_3, S_4, \dots$

$$P(W_1) \overset{?}{=} S_1 - 2S_2 + 3S_3 - \dots + \dots$$

$$1 \overset{\checkmark}{=} 1 - 0 + 0 - \overset{0}{\dots} + \overset{0}{\dots}$$

Good. It checks, at least for the first case.

Next:

set $A_1 = A_2 = \Omega \qquad A_3 = \dots = A_n = \phi$

Let's see what happens.
What is the probability that <u>exactly one</u> of the events A_i happens? <u>Zero</u>

$\uparrow A_1 = A_2 = \Omega$

$S_1 = 2$

$S_2 = \sum_{i<j} P(A_i \cap A_j) = 1$

$S_3 = S_4 = \dots = \phi \leftarrow$ since at least one of the events in each intersection is ϕ.

$$P(W_1) \overset{?}{=} S_1 - 2S_2 + 3S_3 - \dots + \dots$$

$$0 \overset{\checkmark}{=} 2 - 2 + 0 - \overset{0}{\dots} + \overset{0}{\dots}$$

It checks.

So now we jump to the general case:

$$\left.\begin{aligned} \text{set } A_1 &= A_2 = \dots = A_k = \Omega \\ A_{k+1} &= \dots = A_n = \phi \end{aligned}\right\} \text{ if it checks for this, for all } k \geq 3, \text{ then it's true by Réyni's Principle.}$$

$P(W_1) = 0 \leftarrow W_1$ is an impossible event, as more than one of the A_i's happens.

$$P(W_1) \stackrel{?}{=} S_1 - 2S_2 + 3S_3 - \dots + \dots$$

$$(**) \quad 0 \stackrel{?}{=} k - 2\underbrace{\binom{k}{2}} + 3\underbrace{\binom{k}{3}} - \dots + \dots$$

If you have k events w/ $A_i = \Omega$ then there are

$\binom{k}{2}$ terms $\underset{i<j}{A_i \cap A_j} = \Omega$, the rest will be empty

$\binom{k}{3}$ terms $\underset{i<j<k}{A_i \cap A_j \cap A_k} = \Omega$, the rest will be empty

etc

Is the RHS $= 0$?
It <u>is</u> 0 for the following reason.
I'm sorry, but I have to resort to a <u>non</u> probabilistic argument.
I could use a probabilistic argument, but it's too long.

$$(1+x)^k = 1 + \binom{k}{1}x + \binom{k}{2}x^2 + \binom{k}{3}x^3 + \binom{k}{4}x^4 + \dots$$

differentiate wrt x:

$$k(1+x)^{k-1} = 0 + \binom{k}{1} + 2\binom{k}{2}x + 3\binom{k}{3}x^2 + 4\binom{k}{4}x^3 + \dots$$

Set $x = -1$

$$0 = k - 2\binom{k}{2} + 3\binom{k}{3} - 4\binom{k}{4} + \dots$$

So the RHS of (**) is 0.
And we have:

$$0 \stackrel{\checkmark}{=} 0$$

By Réyni's Principle, it verifies the formula:

$$P(W_1) = S_1 - 2S_2 + 3S_3 + \dots - \dots$$

<u>Done.</u>

I leave it to you to prove the <u>general formula</u>:

$$P(W_k) = \sum_{i=0}^{n-k} (-1)^i \binom{k+i}{i} S_{i+k}$$

Verify by Réyni's Principle and you get a binomial coefficient identity, which reduces to the one I just covered, when you simplify.

Pretty cute.

You get these very fancy formula cheapo.
Next time, you ask – How does Réyni's Principle work? Why does it hold?

Because for your next problem set, I'm going to pull all sorts of inequalities out of the literature, which were proved by weirder methods, and you re-prove them by using Réyni's Principle.

So now you can prove any theorem. You can go to IEEE Transactions on Reliability Theory, which is about inclusion-exclusion formula, exclusively, and you can prove everything. You can trivialize all the papers there.

Q: When did Réyni come up w/ this?
A: He told it to me in 1969.

John Guidi
guidi@mit.edu

- Last time we used Réyni's Principle to verify the inclusion-exclusion principle. You are wondering, now, how to prove Réyni's Principle.
Well, I decided to write up the proof and give it out as a bulletin, so you can read it, understand it & me presenting it in class.
Why? Because it's boring. You can understand it better by watching the various steps in the proof.
The proof in writing is required material, but there's no point in my spending a half an hour writing the proof.

Let's go on to another extremely important chapter.

- Random Variables (integer)

for the moment, we will deal w/ one class of random variables, which we call integer random variables.
I will not keep saying integer. It will be understood that the random variables we deal w/ are integer random variables — until further notice.

First, the definition. And in this case, the definition conceals the meaning of the topic. Definitions conceal more than they reveal. You can not understand any thing by reading these definitions. Similarily, it's the only hope with which you can understand the meaning.

Official Definition:
A random variable X is a function from a probability space to the integers, with the property that, for every integer n,

$$\{\omega : X(\omega) = n\} = \underbrace{(X=n)} \text{ is an event}$$

The set of all sample points, where the random variable takes the value n, which is abbreviated this way.

The event that $X=n$.
This is the probabilistic meaning.
You just don't think in terms of sample points.

If X is a random variable, the events $(X=n)$ are disjoint and $\bigcup_n (X=n) = \Omega$

it is understood that this is an integer random variable.

This is intuitively obvious, because if X takes the value 3, it can not take the value 2.

This is just common sense.

The sequence $p_n = P(X=n)$ is called the probability distribution of the random variable X.

notice that we don't use 2 sets of parentheses and we ought to. We will only use one. This is, strictly speaking: $P((X=n))$ — the probability of an event (i.e., the event $(X=n)$).

It satisfies $p_n \geq 0$ and

$$\sum_n p_n = \sum_n P(X=n) = P\left(\bigcup_n (X=n)\right) = P(\Omega) = 1$$

The notion of a probability distribution is historically much older than the notion of a random variable.
For a long time, people could conceive of probability distributions, but they couldn't conceive of random variables.
The idea of random variables is a 20^{th} century idea. There were stalwarts who refused to accept it and said only distributions exist and not random variables.

This is the bare definition. In order to understand the definition, let's enrich it by typical examples. If they don't tell you the typical examples, then you don't know.

Let's see some typical examples. We already have seen them.

Example in Ω_{MB}

We talked about the number of balls in the 1^{st} box.
The occupancy number.
But what is the occupancy number?
The occupancy number is a random variable.
So, we were talking about random variables w/o realizing it.
θ_1 is a random variable

For once, let's use sample points:

$\theta_1(\omega)$

ω is a placement of the balls into the boxes.
So, for that placement, you count how many balls are in this 1^{st} box.
So it's a function.

There are various numbers of balls in the 1^{st} box for the placements $\omega \in \Omega$.
This function θ_1 is from the placement of balls into boxes to the number of balls in box 1.
It is a function.

* However, the spirit of probability is to use sample points as little as possible.
As the great mathematician John von Neumann used to say:

"The spirit of probability is to be pointless"

If you have to use sample points, you haven't really thought probability.
Of course, von Neumann had an ulterior motive. He was developing quantum probability. And in quantum probability, there aren't any points.
In order to understand quantum probability, you have to understand all of classical probability without ever mentioning the sample points.

What is the probability distribution of the random variable Θ_1?
We already have this.

$$p_i = P(\Theta_1 = i) = \frac{\binom{k}{i}(n-1)^{k-i}}{n^k}$$

And we did verify that: [2/11/98.7-8]

$$\sum_{i=0}^{n} p_i = \underbrace{\sum_{i=0}^{n} P(\Theta_1 = i)}_{} = \sum_{i=0}^{n} \frac{\binom{k}{i}(n-1)^{k-i}}{n^k} = 1$$

$$= P\left(\bigcup_{i=0}^{n}(\Theta_1 = i)\right) = P(\Omega)$$

Thus p_i is the probability distribution of the random variable Θ_1.

Now, what if I said at this point, "Let's do the same for Θ_2."
You would say "That's kind of stupid."
Why?
This stupidity conceals deep concepts.

So now I say: "Θ_2 is also a random variable"
And Θ_1 is different from Θ_2, since the number of balls in the 1st box is not the same as the number of balls in the 2nd box.
They are related, but that doesn't mean anything. Wishy-washy words.

$$\underbrace{P(\Theta_2 = i) = P(\Theta_1 = i)}_{} \longleftarrow$$ however, the probability of the event $(\Theta_2 = i)$ is the same as the event $(\Theta_1 = i)$.

It took people 200 years to unscramble what's the same and what's not the same here.
For a long time, people would say they are the same. ⟵ $(\Theta_1 = i)$ and $(\Theta_2 = i)$
But they are not the same.
They have the same distributions
But that does not entitle you to say they are the same random variables.
We reap the fruits of 200 years of misunderstandings.
We will, therefore, say that:

Two random variables are identically distributed when they have the same probability distributions.

Two random variables X and Y are said to be identically distributed when $P(X=n) = P(Y=n)$ for all n.

NB: being identically distributed is not the same as being the same.

Example: in Ω_{MB},

Θ_1 and Θ_2 are identically distributed,

since $P(\Theta_1 = i) = P(\Theta_2 = i)$ for all i

but random variables Θ_1 + Θ_2 are different

By the way, Θ_1 + Θ_2 are identically distributed in the other statistics (namely: Ω_D, Ω_{BE}, Ω_{FD}).

identically distributed
-vs-
same random variable

So it's an intuitive concept one has to distinguish.
This is strictly a probability concept.
It should guard you from using loosely the term "the same"

Next, we say:

Two random variables X and Y are independent when

$$P((X=n) \cap (Y=k)) = P(X=n)\,P(Y=k) \quad \text{for all integers } n \text{ and } k$$

In other words, the values of one variable are totally unrelated, to use sloppy language, to the values of the othe variable.

Similarily for 3 or more random variables.
NB: pairwise independence is not the same as independence
(similar to what we discussed wrt independence of events [2/16/98.4-5])
You must have that, for all possible subsets, the probability of the intersections of the random variable events = product of probabilities of the individual random variable events — for all values of the random variables.

For X, Y, Z to be independent random variables, we must have:

$$\left.\begin{aligned} P((X=n) \cap (Y=k)) &= P(X=n)P(Y=k) \\ P((X=n) \cap (Z=j)) &= P(X=n)P(Z=j) \\ P((Y=k) \cap (Z=j)) &= P(Y=k)P(Z=j) \end{aligned}\right\} \text{all pairs}$$

and

$$P((X=n) \cap (Y=k) \cap (Z=j)) = P(X=n)P(Y=k)P(Z=j) \quad \} \text{ and triple}$$

for all n, k, j } over all n, k, j

2/20/98.5

Research Problem:
Again, similar to independence of events, as discussed in [2/6/98.5]:
If you have a lot of random variables and you assume certain subsets to be independent, then you find that the other conditions are automatically satisfied.
No one has ever really figured it out. It's a nice research problem to work on.
If you give me \$1000, I'll work on it. I'll go off for a weekend.

- Example: θ_1 and θ_2 in MB are <u>not</u> independent random variables

$$P\big((\theta_1 = i) \cap (\theta_2 = j)\big) = \frac{\binom{k}{i}\binom{k-i}{j}(n-2)^{k-i-j}}{n^k}$$ ← We already discussed this [2/11/98.9]

$$\boxed{\frac{k!}{i!\,\cancel{(k-i)!}}\;\frac{\cancel{(k-i)!}}{j!\,(k-i-j)!} = \binom{k}{\underset{\text{1st box}}{i},\ \underset{\text{2nd box}}{j},\ \underset{\text{elsewhere}}{k-i-j}}} = \frac{\binom{k}{i,\,j,\,k-i-j}(n-2)^{k-i-j}}{n^k}$$

$\neq$

but

$$P(\theta_1 = i)\,P(\theta_2 = j) = \frac{\binom{k}{i}(n-1)^{k-i}}{n^k} \cdot \frac{\binom{k}{j}(n-1)^{k-j}}{n^k}$$

$$P\big((\theta_1 = i) \cap (\theta_2 = j)\big) \neq P(\theta_1 = i)\,P(\theta_2 = j)$$

Therefore, they are not independent.
But, <u>intuitively,</u> of course they <u>are not independent</u>
If you put more balls into the 1st box, then the 2nd box will have fewer balls.
So, we didn't have to do this.
It's just a matter of common sense that θ_1 + θ_2 are not independent.
I did it once just to show you that the math checks your common sense.

Let's do some examples of random variables that are independent.

Example — again in Maxwell-Boltzmann statistics

Let X_1 = position of the first ball

This is Maxwell-Boltzmann.
That means we label the boxes $1, 2, \dots, n$.
$U = \{1, 2, \dots, n\}$

Let X_2 = position of the second ball

etc.

I claim that X_1 and X_2 are independent.
That should be intuitively clear.
Balls do not think.
So, where the first ball goes has no influence on where the second ball goes.
Let's verify this mathematically.
Let's see that the formal notion of independence does coincide with our intuitive notion of independence.

The probability distribution:

$$\underbrace{P(X_1 = i)}_{\substack{\text{the event that}\\ \text{the 1st ball}\\ \text{goes into the}\\ \text{box labelled } i.}} = \frac{1}{n}$$

← all the boxes are equally possible.
Or we can compute this directly as:

$$P(X_1 = i) = \frac{n^{k-1}}{n^k} = \frac{1}{n}$$

Similarily:

$$P(X_2 = j) = \frac{1}{n}$$

Random variables X_1 and X_2 are identically distributed

$$P(X_1 = i) = P(X_2 = i) \quad \text{for all } i$$

"If you don't see that, I can not explain it." ←
Do you know how Leibniz (the famous philosopher) used to say ?
"That's because of the Principle of Sufficient Reason"

$$P((X_1 = i) \cap (X_2 = j)) = \frac{n^{k-2}}{n^k}$$

← the 1st ball goes into the ith box. Period.
the 2nd ball goes into the jth box. Period.
There is only one way to do that.
The remaining $k-2$ balls can go into any of the n boxes.

$$= \frac{1}{n^2}$$

$$= P(X_1 = i)\,P(X_2 = j) \quad \text{for all } i, j$$

Therefore random variables X_1 and X_2 are independent.

Similarily, random variables X_1, X_2, and X_3 are independent.

- Now we can backtrack.
  We defined the Maxwell-Boltzmann sample space by saying we have balls into boxes and so forth.
  But suppose we wanted a purely abstract definition of the Maxwell-Boltzmann sample space, as you might find in some probability books.

<u>Abstract Definition of Maxwell-Boltzmann Statistics</u>

You have k random variables $X_1, X_2, \dots, X_K$, which are independent, identically distributed, and $P(X_1 = i) = \frac{1}{n}$, $i = 1, 2, \dots, n$

So the Theory of Maxwell-Boltzmann statistics is the theory of k independent, random variables taking the distribution $p_i = \frac{1}{n}$.
This takes away some of the fun, because you don't visualize things.
It's a good, rigorous definition, but if you see a definition like this, you should immediately pick up a piece of paper and start drawing balls and boxes. Otherwise, you can't visualize it.

Such a definition, by the way, does not work in the case of the other statistics, where you have indistinguishable balls.

- Let's now see some fancier examples of random variables from the theory of the Bernoulli Process:

<u>Famous random variables in the Bernoulli Process</u>

$$\Omega = \left\{ \omega = (\omega_1, \omega_2, \dots \text{[redacted]}), \ \omega_n = \left\{ \begin{matrix} 1 \\ 0 \end{matrix} \right\} \right\}$$

all infinite sequences of 0's and 1's.

X_n = outcome of the n^{th} toss

Let's look at it from the <u>wrong</u> point of view – namely, as a function.
This says that:

$$X_n(\omega) = \omega_n \quad \longleftarrow \text{Sort of tautological}$$

ω is a sample point, which is an infinite sequence, you pick the n^{th} entry of it.

The probability distribution of the random variable X_n:

$$P(X_n = 1) = p$$

$$P(X_n = 0) = q = 1 - p$$

$$P(X_n = j) = 0 \quad \text{if } j \notin \{0, 1\}$$

We already verified, in different terms, that the random variables X_1 and X_2 are independent.
In other words, the outcome of the 1st toss is independent of the outcome of the 2nd toss.
We verified that explicitly at the beginning of the course.
We can now restate our verification in terms of random variables
It's much more efficient.

More generally, any finite set $X_i, X_j, \ldots, X_k$ is independent
So, if you take any distinct set of tosses, they are independent random variables.

Now, let's take this random variable:

$$X_1 + X_2 + \ldots + X_n = S_n$$

↑ this random variable is given the unfortunate name S_n, which we used in a different sense last time.
It's not my fault. All the books do it.
So you better get used to the ambiguity.

Random variables being functions, you can add them.
And multiply them (not divide them)

What does S_n measure? The number of Heads (1's) in the first n tosses.

Let's find the probability distribution of S_n:

This is Fermi-Dirac statistics. We have n boxes, which are the tosses, and k indistinguishable balls, which are the k Heads.
the win

$$P(S_n = k) = \underbrace{\binom{n}{k}}_{\substack{\binom{n}{k} \text{ ways of choosing } k \\ \text{Heads.} \\ \binom{n}{k} \text{ wins}}} \underbrace{p^k q^{n-k}}_{\substack{\text{For each win (i.e., } k \text{ heads, } n-k \text{ tails),} \\ \text{the probability is:} \\ p^k q^{n-k}}}$$

$$\overbrace{P(S_n=k)}^{\text{probability distribution of random variable } S_n} = \binom{n}{k} p^k q^{n-k}$$ is called the <u>binomial distribution</u>

Since:

$$\underbrace{\bigcup_k (S_n = k)}_{\substack{\text{Each of these}\\ \text{events is disjoint.}\\ (n \text{ is fixed})}} = \Omega \Rightarrow P\Big(\bigcup_k (S_n=k)\Big) = P(\Omega) = 1$$

it follows that:

$$\sum_k P(S_n=k) = P\Big(\bigcup_k (S_n=k)\Big)$$

$$\sum_k \binom{n}{k} p^k q^{n-k} = 1$$

and since $1 = p+q$, we have:

$$\boxed{\sum_k \binom{n}{k} p^k q^{n-k} = (p+q)^n}$$

The 3^{rd} proof of the <u>binomial theorem</u>
hardly a "proof" since one could put $(p+q)'$ on RHS too

Now, let's jazz this up a little.
These random variables are kind of boring.
Now let's look at random variables that are the hard core of probability.
What <u>random variables</u> are the <u>hard core of probability?</u>
It's not a secret.
They are:
Waiting Times

Most probability problems boil down to computing waiting times.
You will eventually learn.

Let's define a random variable connected w/ this Bernoulli Process:

W_1 = waiting time for the first Head (1)

Notice how artificial it would be to think of this as a function: $\cancel{W_1(\omega)}$

a sample pt ω is a sequence of 1's and 0's. You look for where the first 1 occurs in ω. Don't look at it this way.

Look at it pointlessly

What is $P(W_1 = n)$?
How does the event $(W_1 = n)$ occur?
It can occur in only one way:

The first $n-1$ tosses must be Tails (0)
the n^{th} toss must be Heads (1)

$$P(W_1 = n) = q^{n-1}p$$

Let's note that:

$$P(W_1 = 1) = p$$

Since $\underbrace{\bigcup_n (W_1 = n) = \Omega}$

That's a lie.
Where is the lie? Why did I glibbly write that?
Because, except for an event with probability 0, the above equality is true.

True, except for the sequence which is all 0's.

$$P(\underbrace{0_1 \cap 0_2 \cap \cdots}_{\text{the event where all of the infinite tosses are 0.}}) = q^{\infty}$$

Assuming $0 < p < 1$

$1 > \underbrace{1-p}_{q} > 0$

$\Rightarrow q^{\infty} = 0$

$$P(0_1 \cap 0_2 \cap \cdots) = 0$$

This event, all tosses are 0, has probability 0.

We will often write equalities this way:

$$\bigcup_n (W_1 = n) \overset{?}{=} \Omega$$

meaning that the two events are equal and they differ, at most by an event of probability 0.

This is a very useful distinction from the ordinary notion of equality.

Q/A: The probability of any sample point is 0 in the Bernoulli Process, if $0 < p < 1$.

$$P(\omega) = 0$$

$\uparrow$ single sample pt

An infinite product of numbers between 0 and 1 is 0.

Q: Can you remove a countable number of events?
A: We'll do that. That's a good idea.
A countable number of events with probability 0.
You see things like this, for example, when we deal w/ distances bt natural numbers.

With this notion of equality, from this identity:

$$\bigcup_n \underbrace{(W_1 = n)}_{\text{disjoint events}} = \Omega$$

we infer, by taking probabilities of both sides:

$$\sum_n P(W_1 = n) = 1$$

where, again, $P(W_1 = n) = q^{n-1} p$

and we have proved the following identity:

$$\boxed{p + pq + pq^2 + \cdots = 1}$$

You can now, of course, a posteriori, realize that this is:

$$p(1 + q + q^2 + \cdots) = p \frac{1}{1-q}$$
$$= \frac{p}{p}$$
$$= 1$$

So, the <u>probability distribution</u> of the <u>waiting time for the first success</u> is called the <u>geometric distribution</u>.

$$P(W_1 = n) = q^{n-1} p \quad \text{is called the } \underline{\text{geometric distribution}}$$

Don't let this go to your head.
These are <u>purely historical</u> names
(in this case, because of the geometric series, you see.)

Let's jazz this up, even more.
Let's now consider the <u>waiting time</u> for the <u>k^{th} success (k^{th} head)</u>

A reasonable thing to ask.
You'll see we jazz this up to the point that you won't believe it.
I've got something up my sleeve to give you in about 3 weeks, which has to do w/ Heads and Tails, that you wouldn't believe.
Absolutely, one of the deepest computations ever made. You will see a <u>paradox</u>. We are working up to it.
One of the most extraordinary paradoxes in probability.

$$W_k = \text{waiting time for } k^{\text{th}}\ 1$$

Let's find the probability distribution for this random variable:

$$P(W_k = n) = \begin{cases} 0 & \text{if } n < k \\ ? \end{cases}$$

it doesn't come out. You have to see this.

here are the tosses $\underbrace{\overbrace{\cdots\cdots\cdots}^{k-1 \text{ 1's in } n-1 \text{ tosses}}\ 1}_{n}$

We have n tosses.
The last one must be a 1, since its a waiting time for the k^{th} 1.

In the preceding $n-1$ tosses, we must have $k-1$ 1's.
That's kind of <u>Fermi-Diracish</u>.

So this gives:

$$\underbrace{\binom{n-1}{k-1}}_{\substack{\text{from } n-1 \text{ tosses,} \\ \text{we must have} \\ k-1 \text{ 1's.}}} p\, p^{k-1} q^{n-k}$$

the last toss has to be a 1.

You have $k-1$ 1's somewhere in between

the rest have to be 0

Therefore, we can rewrite this as:

$$P(W_k = n) = \binom{n-1}{k-1} p^k q^{n-k} \quad , \; n \geq k$$

This is called the Negative Binomial Distribution.

And we have proved, by probabilistic methods:

$$\sum_{n \geq k} \binom{n-1}{k-1} p^k q^{n-k} = 1$$

$$\left\{ \begin{array}{l} \bigcup_{n \geq k} \cancel{P}(W_k = n) = \Omega \\ \Downarrow \\ \sum_{n \geq k} P(W_k = n) = 1 \end{array} \right.$$

This is a non-trivial identity.

These are the easy ones.
These are absolutely universal distributions.
Even the notation is universally adopted – by everyone.
Next time, we'll do some non-trivial ones.

The easy probability distributions of random variables for the Bernoulli Process:

$$P(\underbrace{S_n = k}_{\#\text{ of Heads in } 1^{\text{st}}\ n \text{ tosses}}) = \binom{n}{k} p^k q^{n-k}$$ Binomial Distribution

$$P(\underbrace{W_1 = n}_{\#\text{ tosses until } 1^{\text{st}} \text{ Head}}) = q^{n-1} p$$ Geometric Distribution

$$P(\underbrace{W_k = n}_{\text{waiting time for } k^{\text{th}} \text{ head}}) = \binom{n-1}{k-1} p^k q^{n-k}$$ Negative Binomial Distribution

Random Variables (cont'd) (integer)

Last time, we saw the first examples of random variables.
Let's discuss some more.
But let's review the concept in light of the examples we have covered.

Again, I remind you, we are dealing w/ a special kind of random variable, which is integer random variable.
Meaning they are integer value functions.
Functions that take on integer values.
We consider these in advance, because it is easy to develop the theory.
Later we will consider the general case.

(it is understood that we are talking about integer random variables)

• Given a probability space Ω, P, a random variable X on Ω is a function from Ω to the integers, with the property that for every integer n, $\underbrace{\{\omega : X(\omega) = n\}}_{\text{The set of all sample points } \omega \text{ for which } X(\omega)=n. \text{ This is the last time I'm going to write it this way.}} = \underbrace{(X=n)}_{\text{From now on, we use this notation. The event that } X=n. \text{ Pointlessly}}$ is an event.

I want to fill in some details that I glossed over last time.

• $\bigcup_n (X=n) = \Omega$ $\underbrace{\text{except, possibly, for an event of null probability}}_{\text{I didn't tell you this last time, because I didn't want to overdo it}}$

What does this mean?
It means that the random variable need not be defined on the entire sample space.
It may be undefined on an event of null probability.

For example, last time we considered the waiting time for the first head in the Bernoulli Process. [2/20/98.10-12]
Strictly speaking, the waiting time is infinity for the one sample point that consists only of all tails.
The waiting time for the first head is not very well defined.
* This usually happens on infinite probability spaces.

- The sequence $p_n = P(X=n)$ is the probability distribution of the random variable X.

- X and Y are independent random variables when:

  intuitively, when every event definable in terms of X is independent of every event definable in terms of Y.

  This is the intuitive idea.
  It is rendered by saying:

  $$P((X=n)\cap(Y=k)) = P(X=n)P(Y=k) \quad \text{for all integers } k,n$$

- X and Y are identically distributed (i.d.) when:

  universal notation, used by everybody

  $$P(X=n) = P(Y=n) \quad \text{for all } n$$

  We've seen examples of random variables that are not the same, but are identically distributed.

- In the literature, you will see the following notation extremely frequently:
  i.i.d. = independent and identically distributed

Examples:

in Ω_{MB} occupation numbers θ_1 and θ_3 are i.d. but not i.i.d. [2/20/98.5]

in Ω_B X_1 and X_7 (outcome of 1^{st} and 7^{th} tosses, respectively) are i.i.d.

Bernoulli

In fact, the whole sequence $X_1, X_2, \ldots$ is i.i.d.
(X_n being the outcome of the n^{th} toss)

Q: Why can an infinite set of events be independent?
A: Good point. You get a chocolate bar.
Every finite subset of events $X_1, X_2, \ldots$ is independent.

Let's see some more examples of random variables.

Further examples

- in Ω_B, W_1 = waiting time for first Head (1)

  Last time, we saw the probability distribution for W_1 [2/20/98.10-12]:

$$P(W_1 = n) = q^{n-1}p = \text{geometric distribution}$$

Let's consider the problem of the following random variable:

T_2 = the gap between the first and second Heads (1) {exclude toss of first Heads; include toss of second Heads}

(you'll see the reason for the subscript 2 in a minute)

All right. Let's see if we can compute the probability distribution of T_2.

$$P(T_2 = n) = ?$$

Well, we have:

$$\bigcup_k (W_1 = k) = \Omega \longleftarrow \text{except for an event of probability } 0$$

Therefore, the event $(T_2 = n)$ can be written as:

$$(T_2 = n) = (T_2 = n) \cap \Omega$$

$$= (T_2 = n) \cap \left(\bigcup_k (W_1 = k) \right)$$

now use the distributive law for events:

$$(T_2 = n) = \bigcup_k \Big(\underbrace{(T_2 = n) \cap (W_1 = k)} \Big)$$

the union of these is disjoint.
Because if $W_1 = 31$ then W_1 can not be 32.

Therefore, since $P(LHS) = P(RHS)$, and $P(RHS)$ = sum of probabilities:

$$(*) \quad P(T_2 = n) = \sum_k P\big((T_2 = n) \cap (W_1 = k) \big)$$

So now we are confronted w/ computing the probability:

$$P\big((T_2=n)\cap(W_1=k)\big) = ?$$

That's not too hard.
Let's draw a picture. We have our tosses:

$$\overbrace{0\ 0\ 0\ 0\ 0}^{\text{all } 0}\ \underset{k^{\text{th}}}{1}\ \underbrace{\overbrace{0\ 0\ 0\ 0\ 0}^{\text{all } 0}\ 1}_{n}$$

the event $(W_1=k)$

the event $(T_2=n)$
The next 1 after the first 1 takes n tosses.

The event $\big((T_2=n)\cap(W_1=k)\big)$ can occur only in this way.
So we have here:

$$P\big((T_2=n)\cap(W_1=k)\big) = q^{k-1}p\,q^{n-1}p$$

Now that we have this, we can rewrite $(*)$:

$$P(T_2=n) = \sum_k q^{k-1}p\,q^{n-1}p$$

$$= q^{n-1}p \underbrace{\cancel{\sum_k q^{k-1}p}}_{\substack{\text{This is the sum of}\\ \text{the geometric distribution}}} \to 1$$

This is the sum of the <u>geometric distribution</u> $P(W_1=k) = q^{k-1}p$
W_1 = waiting time for 1st Head

and we showed [2/20/98.11-12] that

$$\sum_k q^{k-1}p = 1$$

Therefore:

$$P(T_2=n) = q^{n-1}p$$

We have obtained the following result:

The probability distribution of the <u>gaps between the 1st and the 2nd Heads</u> is <u>identically distributed</u> with that of the <u>waiting time for the 1st Heads</u>

T_2 and W_1 are i.d.

But you have no doubt noticed that the preceeding argument proves a lot more.
Ha. Ha. Ha.
What we have really proved is this.
By the same argument:

$$\underbrace{P\big((W_1=k)\cap(T_2=n)\big)}_{q^{k-1}p\,q^{n-1}p} = \underbrace{P(W_1=k)}_{q^{k-1}p}\,\underbrace{P(T_2=n)}_{q^{n-1}p}$$

obviously. If you don't see it, I can't explain it.

That means the random variables W_1 and T_2 are <u>independent</u>
W_1 and T_2 are i.i.d.
Now we readjust our intuition.
We say "Of course they have to be independent. No matter where the first Head occurs, we just start all over again."
A hand waving argument which is best rendered by the preceeding mathematical argument.

Similarily, we now rename $\underbrace{W_1 \Rightarrow T_1}_{\text{we give } W_1 \text{ two names}}$ and

T_3 = gap between the 2nd + 3rd Heads (1)
T_4 = " " " 3rd + 4th " "
etc.

We define all the gaps:

$T_1, T_2, T_3, \ldots$ = gaps between successive Heads (1's)

An extension of the above argument, which is pretty routine, shows that the random variables $T_1, T_2, \ldots$ are <u>independent</u> and <u>identically distributed</u>

$T_1, T_2, \ldots$ are i.i.d.

{ identically distributed with the geometric distribution }

We will see later that we can reconstruct the Bernoulli Process with these random variables alone.
The entire Bernoulli Process can be reconstructed from these data alone.

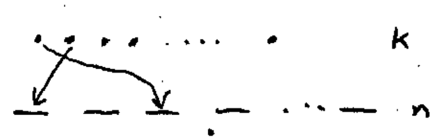
OK, Let's see some more random variables.

- Sampling

That's a major business, in statistics anyway.
This is time to have the exact, clear definition of what sampling is about.

There are a number of basic sample spaces, not all of them finite.

Sampling is just balls into boxes



the balls are called samples

set of boxes is called a population

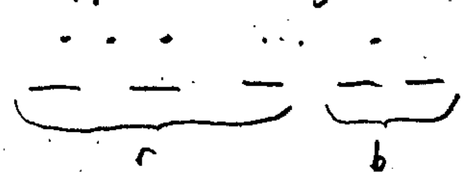
Intuitively, the function from the balls into the boxes means that.
at various times, you sample a member of the population.
Then you have a choice of 2 things. Either:

1) you throw it away ⟵ sampling w/o replacement

-or-

2) you put it back ⟵ sampling w/ replacement

- Sampling with replacement is just Ω_{MB}, except for one feature that comes in in sampling.
What makes sampling slightly different from ordinary Maxwell-Boltzmann statistics is that the population is usually of 2 or more colors:



Not only are the boxes distinguishable, but they are also endowed w/ certain features (e.g., red or black).
You sampling has to do w/ determining how many reds or blacks there are and also computations pertaining thereto.

So sampling is more than just Maxwell-Boltzmann, in that you have a partition of the population into several characteristics.
We will just take a partition of 2, since it's easier to compute.
But you can easily imagine any number of characteristics.

Sampling w/o replacement

Once you sample a member of the population, you can not replace it.
For example, you sample fruit flies and they die, because you kill them. That's $\underbrace{\text{sampling w/o replacement.}}_{\text{very often, you have to do that}}$

The sample space changes.

The sample space is a variant of Maxwell-Boltzmann where the occupation numbers can only be 0 or 1.

We have seen this (occupation numbers only 0 or 1) in Fermi-Dirac.
Now, I'm going to show you something very remarkable that is going to happen.

w/o replacement:

* Both samples and population are distinguishable, but occupation numbers are 0 or 1.

Q/A: Population is distinguishable. Different and distinct members.
You can not have an indistinguishable population.
Ex: There are people in this room, some of whom wear red socks some of whom wear black.
They are still distinct.
In real life, we don't have indistinguishable things.
(You are, in part, spoiling my act)
Population is distinguishable, until further notice.

What is indistinguishable — you have to prove that.
Things in the real world are always distinguishable.

There is a famous paper by Galileo where he is trying to convince someone that when you toss 2 dice, they are not indistinguishable. They are distinguishable. The probability has to be computed on the basis of the dice being distinguishable.
This is one of the dialogues by Galileo. It takes 13 pages.

Q: Could you give an example where the samples are distinguishable?
A: They are distinguishable. At a time, the first is picked out. Then the second is picked out. Etc.
Or John picks out the first. Mary picks out the second.
There is always some criteria that you can find to make them distinguishable, in real life.

So now you can ask: "Where does indistinguishable happen? Does it happen?"
That's where you talk about subsets.
When you pick a subset – thats where you get indistinguishable.
But we are not picking subsets here.
We are sampling.
And yet, watch what happens. Now this is interesting.

Let's consider now, in each of these 2 cases, the random variable

R = # of red members in a sample

Let's compute, in each case, what the probability distribution of R is.

case 1 Ω_{MB} (Sampling w/ replacement)

$$\underbrace{P(R=i)}_{\text{we have } i \text{ red}} = \frac{\binom{k}{i}\, r^i b^{k-i}}{n^k}$$

numerator is # of ways you can form i reds

denominator is # of sample points.

from the sample of k, we choose i red. There are $\binom{k}{i}$ ways to do this.

For each $\binom{k}{i}$, there are $r^i b^{k-i}$ ways to put i balls in red boxes and $k-i$ balls in black boxes.

What have we done?
We've given the 5th proof of the binomial theorem.

$$\underbrace{\bigcup_i (R=i)}_{\text{all disjoint}} = \Omega$$

$$\sum_i P(R=i) = 1$$

$$\sum_i \frac{\binom{k}{i} r^i b^{k-i}}{n^k} = 1$$

and since $r+b=n$

$$\boxed{\sum_i \binom{k}{i} r^i b^{k-i} = (r+b)^k}$$

That's the binomial theorem.

Now let's do it w/ feeling.
We kill the samples. Cruely.

case 2 (Sampling w/o replacement)

I know a biologist once who worked w/ mice all the time.
I won't say who it is, but it's someone very close to me.
And if she wasn't careful, she would kill the mouse.
She was so used to breaking their neck.
She sampled mice. And when she was asked to do samples with replacement, she had to be extremely careful. Because normally, she would do sampling w/o replacement.

This is a new sample space.
It doesn't have a name — no standard naming.

$$\Omega_{\underbrace{MBwoR}}$$

Maxwell – Boltzmann without Replacement

How many sample points are there?
You have k distinguishable balls going into n boxes.
And at most one ball per box.
We've done that.

$$|\Omega_{MBwoR}| = (n)_k$$ ← good old lower factorial

We now consider the random variable R.

$$P_{MBwoR}(R=i) = \frac{\binom{k}{i}(r)_i\,(b)_{k-i}}{(n)_k}$$

out of the k balls, I pick i balls

I place the i balls into the red boxes. How many ways can I place i balls into the red boxes w/ occupation numbers 0 or 1?
$(r)_i$

The remaining k-i go into black boxes w/ occupation numbers 0 or 1.

2/23/98.10

Now, let's do it the WRONG way.
Many books do it the wrong way.
They imagine that the balls are indistinguishable
They say, "Let's do this with Fermi-Dirac statistics"

In Ω_{FD}, we can ask the same question.
We have r red boxes and b black boxes.

In Ω_{FD}, the random variable R has the probability distribution:

$$P_{FD}(R=i) = \frac{\binom{r}{i}\binom{b}{k-i}}{\binom{n}{k}}$$

the balls are now indistinguishable

i balls go into the r red boxes.
That means I pick a subset of i elements from the set of red boxes.
This can be done in $\binom{r}{i}$ ways.

The remaining $k-i$ balls go into the b black boxes.

Remember that $\binom{n}{k} = \frac{(n)_k}{k!}$

$$= \frac{\frac{(r)_i}{i!} \cdot \frac{(b)_{k-i}}{(k-i)!}}{\frac{(n)_k}{k!}}$$

$$= \frac{\binom{k}{i}(r)_i (b)_{k-i}}{(n)_k}$$

Thus:

$$\left. P_{MBwoR}(R=i) = P_{FD}(R=i) \right\} \text{why?}$$

Now, my question to you is: Why?
I thought about this last night before going to bed for a half an hour - how best to say it.
I couldn't find a good explanation.
I give you 30 points if you find a good way of justifying it, that is completely convincing.
I can give you a mathematical interpretation.
But I'd like an intuitive, convincing, justification.

<u>Warning</u> - most of the books assume a priori that the balls are <u>indistinguishable</u>.
Which they have no right to do!

This distribution is called the <u>hypergeometric distribution</u>, usually written in the form:

$$P(R=i) = \frac{\binom{r}{i}\binom{b}{k-i}}{\binom{n}{k}}$$

<u>Research Problem</u>

Now develop an analogous coincidence for Bose-Einstein.
Something similar happens.
If you relate Bose-Einstein to Dispositions, you get again, if the balls go into either red or black, you get a similar coincidence, whether the balls are distinguishable or indistinguishable.
And <u>explain</u>. It's easy to work out. But it would be nice to have an explaination. So it's understandable to the public. Other than math.
Relate similarly Dispositions and Bose-Einstein.

So these are some examples of sampling.
Let's introduce a 3rd sample space associated w/ sampling.

Ω_S
↑ sampling

Now, you remember when we discussed the random variables associated w/ Bose-Einstein statistics, you could do it in 2 ways:

1) you could associate it w/ placing balls into boxes

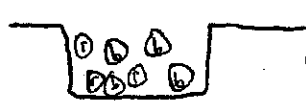
-or-

2) by taking the random variable X_1 to be the position of the 1st ball
X_2 " " " " " " 2nd ball
etc.

In this way, we get random variables that are independent

Now let's do this again, with an infinite number of balls.

The sample space is this.
I have an infinite number of balls, which are called samples. 1 2 3 ...
The way you represent <u>the population</u> is <u>one big urn.</u>
The boxes become balls.



boxes become balls, which are red or black

Urn

This is an occupancy situation
I pick a ball and throw it in here.
The balls are the letters of the alphabet.
Remember, there are distribution and occupancy points of view.
This is the occupancy point of view:
Pull the letter of the alphabet and
place it in it's place

Pull an infinite number of samples with replacement

- pull out a ball
- write the color of the ball
- put back the ball

For example: r, b, b, r, r, b, r, b, ...

$$X_1 = \begin{cases} 1 = \text{red} & \text{with probability } \frac{r}{r+b} \\ 0 = \text{black} & \text{" " } \frac{b}{r+b} \end{cases}$$

We have that $X_1, X_2, \ldots$ are independent random variables

This is equivalent to the Bernoulli Process except that we could have 3 or more colors and then we would have a slight generalization of the Bernoulli Process.

Q: Within each color are the balls distinguishable?
A: Yes. They are distinguishable.
In real life, they are distinguishable.
But you have paradoxes where the answer comes out the same.
However, you have to be careful. There are some goofy mistakes made all the time.

These are the basic sampling spaces.

Armed w/ these examples, we can go back to the theory and develop further the theory of random variables.
We introduce the notion of joint distribution and expectation.

You see, it's all balls into boxes when all is said and done.
Most problems in probability can be reduced to balls into boxes.

Joint Distribution

Joint Distribution of two $\overbrace{\text{random variables}}^{\text{integer}}$ X and Y is the probability:

$$P\big((X=n) \cap (Y=k)\big)$$

In common parlance, this is called a correlation.
It tells you how much the occurrence of the event $(X=n)$ affects the occurrence of the event $(Y=k)$.

When the random variables X and Y are independent, the probability is the product $\big(P(X=n)P(Y=k)\big)$.
In general, it's not the product, as the random variables are generally not independent.

This probability is indicated by a double indexed sequence

$$P\big((X=n) \cap (Y=k)\big) = p_{nk}$$

What are the properties of this doubly indexed sequence?
Very simple, Two properties.

Properties

1) $p_{nk} \geq 0$

2) $\sum_k p_{nk} = \underbrace{P(X=n)}_{\text{namely the probability distribution of } X.}$, $\sum_n p_{nk} = \underbrace{P(Y=k)}_{\text{namely the probability distribution of } Y.}$

Let's verify this and then we quit.

$$\sum_k p_{nk} = \sum_k P((X=n) \cap (Y=k))$$

$$= P\Big(\bigcup_k \underbrace{((X=n) \cap (Y=k))}_{\text{these are disjoint sets for each value } k.}\Big)$$

using the distributive law for events:

$$= P\Big((X=n) \cap \underbrace{\Big(\bigcup_k (Y=k)\Big)}_{\text{this is the whole space } \Omega}\Big)$$

$$\sum_k p_{nk} = P(X=n)$$

As desired.
It's all the distributive law and Boolean algebra that does this trick.

By the way – this is where quantum probability parts company with classical probability.

In quantum probability the distributive law of events does not hold.
Therefore, having joint distributions are a pain in the neck in quantum probability.
So far, everything I've said works in quantum probability too.
But here, that's where it stops.
That's where you start having headaches like the uncertainty principle.
We'll talk about this later.

Random Variables: joint distribution and expectation

X and Y are $\overbrace{\text{random variables}}^{\text{integer}}$

Associated w/ each random variable, we have a probability distribution:

$$P(X=n) = p_n \qquad\qquad P(Y=k) = r_k$$

where: | where:

$$p_n \geq 0 \qquad\qquad r_k \geq 0$$

$$\sum_n p_n = 1 \qquad\qquad \sum_k r_k = 1$$

At the end of last time, we defined the joint distribution of two random variables:

$$P\big((X=n) \cap (Y=k)\big) = p_{nk} \longleftarrow \text{sequence w/ 2 indices}$$

And, thereafter, we showed that:

$$\sum_k p_{nk} = p_n$$

$$\sum_n p_{nk} = r_k$$

The standard language is that when you are given a joint distribution (p_{nk}), and you compute from the joint distribution, the single distribution, the single distributions are called: the marginals

The sequence p_n, as well as the sequence r_k, are the 2 marginals of the double sequence p_{nk}.

It's just the way people speak.

The joint distribution of 2 random variables completely determines the statistical behavior of those random variables.
If you know the joint distribution, then you know also the individual distributions, by these formulas.

$$p_{nk} \begin{cases} \sum_k p_{nk} = p_n \\ \sum_n p_{nk} = r_k \end{cases}$$

In principle, knowing the joint distribution is knowing everything.

In the same way, of course, you can define the joint distribution of 3 random variables.

Similarily,

$$P\big((X=n)\cap(Y=k)\cap(Z=t)\big) = p_{nkt}$$ ← sequence w/ 3 indices

Completely determines the statistical behavior of 3 random variables

And you have similar identities for the marginals.
For example:

$$\underbrace{\sum_k p_{nkt}}_{\text{you sum all the values of } (Y=k)} = p_{nt} = P\big((X=n)\cap(Z=t)\big)$$

Intuitively, this is clear. You make the value of Y irrelevant by summing over all possible values of Y.

And you have these fancy identities. They look fancy, but they are really trivial.

- If X and Y are independent, then $p_{nk} = p_n r_k$ and conversely. This is just another restatement of the definition of independence in fancy language. Technically speaking, 2 random variables are independent when their joint distribution is the product of the marginals.
  which makes it sound very high floutent.

Since we are on this definition kick, let's define a stochastic process:

- A stochastic process is a probability space Ω, P, together with a family of random variables and the joint distributions given of any finite subset.

Needless to say, these joint distributions must be consistent with one another. In a sense that I will not even write. Because you know perfectly well. If you write up the marginal probabilities, automatically they are consistent with the random variables defined on a probability space.

For example, the Bernoulli Process is a stochastic process.
Because it is a probability space, together with the random variables $X_1, X_2, \ldots$, which are the outcomes of the tosses.

So a stochastic process is very general.
It can be almost anything.

Now let's see some examples of joint distributions.
And this is a typical situation, if I may be allowed to philosophize here, where we see just how negative definitions are.
The moment you see the example, you see that there is something missing from that definition where we can not really render it clear.

- Example Ω_{MB} ← a fairly familiar object by now

Let's take the joint distributions of occupation numbers
We've already computed 3 or 4 times the distributions of the occupation numbers.

$$P(\theta_1 = i) = \frac{\binom{k}{i}(n-1)^{k-i}}{n^k}$$

distribution of the first occupation number

It's also the distribution of any occupation number, as all the occupation are identically distributed (i.d.), for reasons that I can not explain, if you don't see that. [2/20/98.5]

(To miss one of those completely obvious things, and there are many in probability, sometimes you just have to click - you have to see that two occupation numbers are i.d. Imagine someone who can not see that. What can we do about it?

I'm saying that because I have something up my sleeve where it's not quite easy to see that.)

Now, lets compute the joint distribution of the 1st and 2nd occupation numbers:

$$P((\theta_1 = i) \cap (\theta_2 = j)) = \frac{\binom{k}{i, j, k-i-j}(n-2)^{k-i-j}}{n^k}$$

We've done this already twice.
Let's do it this time w/ multinomial coefficients

k balls, you pick i and put them in box 1.
From the remaining, you pick j and put them in box 2.

$$\binom{k}{i}\binom{k-i}{j} = \frac{k!}{i!\,\cancel{(k-i)!}}\,\frac{\cancel{(k-i)!}}{j!\,(k-i-j)!}$$

$$= \binom{k}{i,j,k-i-j} \leftarrow \text{multinomial coefficient}$$

The remaining balls go elsewhere (i.e., neither box 1 nor 2):

$$(n-2)^{k-i-j}$$

Now comes the fun.
From this we obtain, using the marginal identity $\sum_j p_{ij} = p_i$:

$$\overbrace{\sum_j \frac{\binom{k}{i,j,k-i-j}(n-2)^{k-i-j}}{n^k}}^{\sum_j p_{ij}} = \overbrace{\frac{\binom{k}{i}(n-1)^{k-i}}{n^k}}^{p_i}$$

The distribution of θ_1

$$\underbrace{\sum_j p((\theta_1=i)\cap(\theta_2=j))}_{\text{Summing over all possible values of } \theta_2} = P(\theta_1 = i)$$

The identity $\sum_j \frac{\binom{k}{i,j,k-i-j}(n-2)^{k-i-j}}{n^k} = \frac{\binom{k}{i}(n-1)^{k-i}}{n^k}$
is a special case of the identity $\sum_j p_{ij} = p_i$.
The innocent little marginal identity becomes this.
That is what the definition does not tell you.

So, in principle, even though we could have computed the distribution of $\underset{p_i}{\theta_1}$ by $\underbrace{\text{summing the joint distributions (over all values for } \theta_2)}_{\sum_j p_{ij}}$, in practice, it's easier to compute it directly and get a non-trivial binomial identity.
And that's easily the case.

Let's jazz it up,
Let's take the <u>joint distribution</u> of <u>all</u> occupation numbers.

$$P\left((\theta_1=i)\cap(\theta_2=j)\cap\ldots\cap(\theta_n=r)\right)=\frac{\binom{k}{i,j,\ldots,r}\cdot 1}{n^k}$$ ← they're all placed

That's it. That's the joint distribution of all the occupation numbers.
And now, by adding over $j,\ldots,r$, this sum of the joint distribution is the marginal p_i:

$$\sum_{j,\ldots,r} p_{ij\ldots r} = p_i$$

$$\underbrace{\sum_{j,\ldots,r}\frac{\binom{k}{i,j,\ldots,r}\cdot 1}{n^k}}_{\text{We've added this over all the occupation numbers except the } i^{th}} = \frac{\binom{k}{i}(n-1)^{k-i}}{n^k}$$

} Try to prove this directly! You can, but it's work.

So you see, that definition <u>conceals</u> more than it <u>displays</u>

Ok, Let's do another one.
Let's do a new kind of distribution.
Namely:
Example – <u>Multivariate hypergeometric distribution</u> ← A nice mouthful. You can impress your friends when they hear you say this. When you apply for a job.

Now what's the most rational way of doing this?
What do you think this is?
<u>Balls into boxes.</u>

We have a population of n members. These are the boxes.
And we take a sample of k elements.
The population is a population of people, like in this room, and on your neck is branded a letter of the alphabet. Some people branded $a, b, \ldots, c$.
They are <u>distinct</u> people.
Therefore, you have the <u>partition of the population</u>, called boxes:

$\bullet\ \bullet\ \bullet\ \ldots\ \bullet$ k balls

$\underbrace{\_\ \_\ \_}_{a}\ \underbrace{\_\ \_\ \_}_{b}\ \ldots\ \underbrace{\_\ \_}_{c}$ n members population

$a, b, \ldots, c$ are a number of letters (or colors, etc). {There are always n boxes + today they are called population}
Altogether $a+b+\ldots+c=n$

We pick a sample and then we ask "How many people branded a are in the sample?" "How many branded b?" etc.
This generalizes the hypergeometric distribution we discussed last time [2/23/98.9-11]

Let A = # of a-people in a sample of k balls into n boxes

B = # of b-people " " " " " " " " "

etc.

Random variables.
Let's find their distribution.

$$P(A=i) = \frac{\binom{a}{i}\binom{n-a}{k-i}}{\binom{n}{k}}$$

pick i that are branded a.

from branding other than a $(n-a)$, pick remaining $k-i$.

This is the hypergeometric distribution we discussed last time.
No multivariate yet.

Now, let's take the joint distribution of A and B:

$$P((A=i)\cap(B=j)) = \frac{\binom{a}{i}\binom{b}{j}\binom{n-a-b}{k-i-j}}{\binom{n}{k}}$$

From a, you pick i
From b, you pick j
From the rest $(n-a-b)$, you pick $k-i-j$

Again, as we discussed last time, this is the same as if the balls were indistinguishable.
(I hope some of you have come up w/ a nice explanation).

This is a bivariate hypergeometric distribution

You may ask why do we call this a bivariate hypergeometric distribution.
The answer will be given later, when we do randomization.
Because I want to give you the real answer, not some phony one.

The marginal identity gives us:

$$\sum_j p_{ij} = p_i$$

$$\sum_j \frac{\binom{a}{i}\binom{b}{j}\binom{n-a-b}{k-i-j}}{\binom{n}{k}} = \frac{\binom{a}{i}\binom{n-a}{k-i}}{\binom{n}{k}}$$

We've proved this.
We don't need to prove it in a complicated way

By the way, this is not as impressive as it looks.
A more interesting identity is:

Since $\sum_i p_i = 1$

$$\sum_i P(A=i) = 1$$

$$\sum_i \frac{\binom{a}{i}\binom{n-a}{k-i}}{\binom{n}{k}} = 1$$

$$\boxed{\sum_i \binom{a}{i}\binom{n-a}{k-i} = \binom{n}{k}}$$ Vandermonde's identity

We've proved this.
It doesn't need any proof

Now let's go all the way.
Let's take the joint distribution of __all__ the brands (colors, etc)

$$P\big((A=i) \cap (B=j) \cap \ldots \cap (C=r)\big) = \frac{\binom{a}{i}\binom{b}{j}\cdots\binom{c}{r}}{\binom{n}{k}}$$ Multivariate hypergeometric distribution

automatically $i+j+\ldots+r = n$, the number of boxes

From this, you can get fantastic identities by using marginals.
Statisticians would spend 2 or 3 weeks just writing down all these identities individually.

$\left\{\begin{array}{l}\text{I worked out also the Bose-Einstein.} \\ \text{You can do the same stuff w/ Bose-Einstein.} \\ \text{Big deal.}\end{array}\right\}$

Let's go into expectation.

2/25/98.8

Expectation - a fundamental concept

X = integer random variable

$E(X)$ = average value it takes

expectation of X

How do we write down correctly the average value it takes?
This is simple.

$$E(X) = \sum_n n\, P(X=n)$$

That's the average value.
You weigh each value of X (i.e., n) by the probability that X takes that value (i.e., $P(X=n)$).

This is provided that the series converges.
If the series doesn't converge, then we say that the random variable does not have an expectation.
It only has an infinite expectation if all the terms $nP(X=n) > 0$.

Now comes the one fact in probability that has kept probability in business.
As a matter of fact, if not for this, probability would not be in business.
It's true, even for quantum probability.
It's really a god given fact.

Namely - the expectation of a sum of random variables is always equal to the sum of the expectations.
Even when the random variables are dependent.
Always

Basic fact

For any two (or more) random variables X and Y, we have

$$\boxed{E(X+Y) = E(X) + E(Y)}$$

This is a very powerful identity because in many circumstances, it is the only way we have to work with dependent random variables.
Because in many cases, computing the joint distribution is a pain in the neck.
So you have to use other techniques. And this is what you use.

[I hope you have started the problem set, by the way.
Yesterday I spent 6 hours on the next problem set.
In vicious glee.]

Proof

Let's take the random variable $X+Y$
Let's take the event $(X+Y=n)$

Can we simplify this event? Well, yes.
Intuitively, it's like this:

$$(X+Y=n) = \bigcup_{i+j=n} \left((X=i)\cap(Y=j)\right)$$

The event $(X+Y=n)$ if and only if X takes some value i and Y takes some value j and $i+j=n$. This is the only way $(X+Y=n)$.
And these are <u>disjoint</u> ways.

We have the above set theoretic identity.

By definition, the expectation of the random variable $X+Y$ is:

$$E(X+Y) = \sum_n n\, P(X+Y=n)$$

$$= \sum_n n \left[\bigcup_{i+j=n} \left((X=i)\cap(Y=j)\right)\right]$$ using above identity

$$= \sum_n n \sum_{i+j=n} P\big(\underbrace{(X=i)\cap(Y=j)}\big)$$

probability of disjoint events equals sum of probabilities

you can't simplify this because the random variables X and Y may not be independent

Combine these sums into one double sum:

($\sum_{i+j=n}$ over all n is the same as the sum over all i and over all j independently)

$$= \sum_{i,j} \overbrace{(i+j)}^{n}\, P\left((X=i)\cap(Y=j)\right)$$

$$= \sum_{i,j} i\, P\left((X=i)\cap(Y=j)\right) + \sum_{i,j} j\, P\left((X=i)\cap(Y=j)\right)$$

$$= \sum_i i \underbrace{\sum_j P\left((X=i)\cap(Y=j)\right)}_{\text{this is the marginal identity}} + \sum_j j \underbrace{\sum_i P\left((X=i)\cap(Y=j)\right)}_{\text{ditto}}$$

(Taking joint probabilities and adding over one variable gives the marginal)

$$= \sum_i i\, P(X=i) + \sum_j j\, P(Y=j)$$

$$= E(X) + E(Y)$$

$$E(X+Y) = E(X) + E(Y)$$

When we went over marginal identities, I was training you to see this.
Expectation is always additive.

Now, let's use this fact.
What shall we compute next?
The expectation of the occupation number in Maxwell-Boltzmann.
We always start w/ that.

- Example in Ω_{MB}

$$E(\theta_1) = \sum_i i\, P(\theta_1 = i) \quad \text{by definition}$$

We've computed $P(\theta_1 = i)$ several times.
It is:

$$= \underbrace{\sum_i i \frac{\binom{k}{i}(n-1)^{k-i}}{n^k}}_{} = ?$$

What is this equal to?
It looks like a mess.

Therefore, we ring the alarm bell and say, there must be an easier way of doing this.
Actually, if you think carefully, I've already done this - in a different notation. You have it in your notes.

Now let's try to figure this out in some clever way.

$\theta_1, \theta_2, \ldots, \theta_n$ are i.d. (identically distributed)

Hence,

$$E(\theta_1) = E(\theta_2) = \ldots = E(\theta_n)$$

Obviously, since they all have the same probability distribution $P(\theta_i = i)$

$$E(\theta_1 + \theta_2 + \ldots + \theta_n) = k$$

I take the random variable $\theta_1 + \theta_2 + \ldots + \theta_n$.
What's that?
The number of balls (k).
Furthermore So this random variable takes the value k with probability 1.

$$P(\theta_1 + \theta_2 + \ldots + \theta_n = k) = 1$$

Therefore, its expectation is:

$$E(\theta_1+\theta_2+\ldots+\theta_n) = \sum_i i\, P(\theta_1+\theta_2+\cdots+\theta_n = i)$$

$$= k\cdot 1$$

$$E(\theta_1+\theta_2+\ldots+\theta_n) = k$$

By the main theorem:

$$E(\theta_1+\theta_2+\ldots+\theta_n) = \underbrace{E(\theta_1)+E(\theta_2)+\ldots+E(\theta_n)}_{\text{since } E(\theta_1)=E(\theta_2)=\ldots=E(\theta_n)}$$

$$= nE(\theta_1)$$

Therefore,

$$k = n\,E(\theta_1)$$

$$\Longrightarrow E(\theta_1) = \frac{k}{n}$$

Which also gives us the identity:

$$E(\theta_1) = \sum_i i\,\frac{\binom{k}{i}(n-1)^{k-i}}{n^k} = \frac{k}{n}$$

Another fancy binomial identity proved cheapo.

Let's do another one
A real fancy one.

- Example Ω_{hyper}
  multivariate hypergeometric distribution
  with $a, b, \dots, c$ colors (brands)

What is the expectation of A?

$$E(A) = ?$$

So we have a population of n people whose necks are branded $a, b, \dots, c$.
And we take a sample of k people.
What is the expected number of people who are branded a?

$$\bullet\ \bullet\ \bullet\ \cdots\ \bullet \quad k$$

$$\underbrace{\_\ \_\ \_}_{a}\ \ \underbrace{\_\ \_}_{b}\ \ \cdots\ \ \underbrace{\_\ \_}_{c} \qquad n$$

$$E(A) = \sum_i i\,P(A=i)$$

$$= \underbrace{\sum_i i \frac{\binom{a}{i}\binom{n-a}{k-i}}{\binom{n}{k}}}$$

This looks like a mess.
You can do it.
Let's see if we can do it
another way.

We look at the problem from the following point of view:

$$\bullet\ \bullet\ \bullet\ \ \cdots\ \bullet \quad k$$

$$\underbrace{\_\ \_\ \_}_{a}\ \underbrace{\_\ \_}_{b}\ \cdots\ \underbrace{\_\ \_}_{c} \qquad n$$

We are picking a sample of k.
Let's look at the 1^{st} box (i.e., the 1^{st} person Mr. Torres).
Either the 1^{st} box (i.e., Mr. Torres) is picked or is not picked.

What is the probability Mr. Torres gets picked?
It's $\frac{k}{n}$. ← If you don't see that, I can't explain it.

Let's introduce a random variable Z_1 for the 1^{st} box:

$$Z_1 = \begin{cases} 1 & \text{if } 1^{st} \text{ box is picked} \\ 0 & \text{if } 1^{st} \text{ box is not picked} \end{cases}$$

$Prob(Z_1 = 0) =$
$\frac{n-1}{\cancel{n}} \cdot \frac{\cancel{n-2}}{\cancel{n-1}} \cdot \frac{n-3}{\cancel{n-2}} \cdots \frac{n-k}{n-k+1} = \frac{n-k}{n}$
so $Prob(Z_1 = 1) = 1 - \frac{n-k}{n} = \frac{k}{n}$

The <u>probability distribution</u> of Z_1 is:

$$P(Z_1 = 1) = \frac{k}{n}$$

$$P(Z_1 = 0) = 1 - \frac{k}{n}$$

The number of elements picked out of people who are branded a is simply:

$$\underbrace{\_\ \_\ \_}_{a}\quad \underbrace{\_\ \_}_{b}\quad \cdots\quad \underbrace{\_\ \_}_{c}$$

$$Z_1 + Z_2 + \ldots + Z_a = A$$

What's the problem?
The random variables Z_i are <u>not</u> independent.
However, the <u>expectation is additive</u> - <u>even</u> if they are <u>not independent</u>.
Therefore, we can still compute the expectation,

$$E(A) = E(Z_1 + Z_2 + \ldots + Z_a)$$
$$= E(Z_1) + E(Z_2) + \ldots + E(Z_a)$$

← {whether the Z_i are independent or not.}

and since $Z_1, Z_2, \ldots, Z_n$ are i.d.

$$E(A) = a\frac{k}{n}$$ ← a result that takes people 3 pages to work out, if they evaluate the sum.

And we have another binomial identity:

$$\sum_i i \frac{\binom{a}{i}\binom{n-a}{k-i}}{\binom{n}{k}} = a\frac{k}{n}$$

Now let's do a couple of examples from the Bernoulli Process.
Ah, I forgot.
There's a nice trick to compute the expectation of the waiting time for the 1st Heads.
Done in a clever way.
I'll do it the hard way.

- Example Bernoulli

$$W_1 = T_1 = \text{waiting time for the } 1^{st} \text{ Heads}$$

$$E(T_1) = \sum_{n>0} n\, P(T_1 = n)$$

I'm ashamed of what I am about to do.
You don't remember how it goes?
There's a really simple trick. ← You could just write down the answer and call it a day.

$$= \sum_{n\geq 1} n\, q^{n-1} p \qquad \longleftarrow \quad P(T_1 = n) = \underbrace{q^{n-1}}_{n-1 \text{ Tails}} \underset{\text{Head}}{p}$$

$$= \frac{d}{dq} \underbrace{\sum_{n\geq 0} q^n}\, p$$

telescoping sum: $1 + q + q^2 + \dots = \frac{1}{1-q}$, $q < 1$

$$= \frac{d}{dq}\left(\frac{1}{1-q}\right) p$$

$$= \frac{1}{(1-q)^2}\, p$$

$$= \frac{1}{1-(1-p))^2}\, p$$

$$E(T_1) = \frac{1}{p}$$

This is not the right way to do it.
But it works.

The point I was coming to is this:
The waiting time for the k^{th} success (k^{th} Heads) is:

$$W_k = \underbrace{T_1}_{\substack{\text{waiting time} \\ \text{for } 1^{\text{st}} \text{ Heads} \\ \text{(success)}}} + \underbrace{T_2}_{2^{\text{nd}}} + \underbrace{T_3}_{3^{\text{rd}}} + \underbrace{\cdots}_{\cdots} + \underbrace{T_k}_{k^{\text{th}}}$$

As we've seen [2/23/98.3-5], $T_1, T_2, \ldots, T_k$ are identically distributed (i.d.)

$$E(W_k) = E(T_1 + T_2 + \cdots + T_k)$$
$$= E(T_1) + E(T_2) + \cdots + E(T_k)$$
$$= k\frac{1}{p}$$
$$E(W_k) = \frac{k}{p}$$

Doing it the <u>hard way</u> gives you:

$$E(W_k) = \sum_n n \, \underbrace{P(W_k = n)}$$

negative binomial distribution [2/20/98.12-13]

$$= \sum_n n \overbrace{\binom{n-1}{k-1} q^{n-k} p^k}$$

We have yet another binomial identity:

$$\sum_n n \binom{n-1}{k-1} q^{n-k} p^k = \frac{k}{p}$$

[Next time I will figure out a better way to compute $E(W_k)$]

The moral is that expectation is directly computed from the sum of expectations of very simple random variables, which are <u>not</u> always independent.

Conditional Probability

The basic technique for computing probabilities

What is conditional probability?
Let's take a sample space Ω, with sample point ω.
On this sample space, you have events which have a probability.

The idea is this.
Suppose that you know that event B has occurred.
Then the probability of an event is altered by your knowledge that the event B has occurred.
How?

Intuitively, it's easy to answer "How?"
You have an event A, which has some probability.
You know that event B has happened.
So only the part of A that is inside B can happen.

Therefore, the probability of A, given that B has happened will depend only on the intersection of A and B $(A \cap B)$.

(Q: When you say that "B has happened," do you know that nothing other than B has happened?
A: Yes. Perhaps something more than B has happened, but my knowledge of it stops there. The probabilities all say something about what we don't know. If we knew everything, we'd do this by the sample points.)

Let's repeat this in set theoretic terms.
Instead of taking the whole set Ω, we take only the set of sample points that are in the set B.
Then we want the probability of sample pts of A that are in the set B.

This probability, for events in the set B, has to be computed in terms of the given probability of Ω

So the temptation is to say:

If we know that the event B with $P(B) > 0$ has happened, then the "new" probability of any event A will depend only on $A \cap B$
(e.g., the "new" probability of the whole sample space Ω will depend only on $\Omega \cap B = B$)

However, the "new" probability can not be $P(A \cap B)$. The "new" probability of an event can not be $P(A \cap B)$. Because then we would have:

$$\underbrace{P(\Omega \cap B)}_{\text{"new" probability of whole sample space is not } 1} = P(B) \neq 1$$

If we have to renormalize our probability in the knowledge that the event B has happened, then after renormalization, we have to end up w/ a probability. Therefore, the only way we have to renormalize a probability of A in the knowledge that B has happened is to set:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

probability of event A, given that event B has happened.

read: Conditional Probability of A given B

For example:

$$P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)}$$

(probability of whole sample space given event B has happened.)

$$= \frac{P(B)}{P(B)}$$

$$= 1$$

That kind of checks.

This notation, of universal use, is <u>somewhat misleading.</u>
Because of the following:
We have a sample space with events and we are given a probability.
With conditional probability, we <u>keep</u> the same events, but we <u>change the probability.</u>

Ω, P
$\uparrow$
P_B

We also write: $\underbrace{P_B}(A) = P(A|B)$

to indicate that the conditional probability is a new probability.

- P_B satisfies the three probability axioms:

1) $P_B(\Omega) = 1$

2) $P_B(A^c) = 1 - P_B(A)$

3) if $A_1, A_2, \ldots$ is a disjoint sequence of events then $P_B(A_1 \cup A_2 \cup \ldots) = P_B(A_1) + P_B(A_2) + \ldots$

In other words, P_B has full rights to be called a probability.

There is no reason to assume that, given a family of events, that there should be a unique probability for that family.
There are lots of probabilities,
Now we see some of them. – the conditional probabilities

Now the verification of these 3 axioms: you just write down the definition and it checks. I leave it to you,

(1) $P_B(\Omega) = \dfrac{P(\Omega \cap B)}{P(B)}$

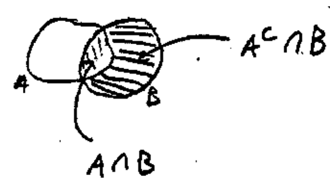
$= \dfrac{P(B)}{P(B)}$

$P_B(\Omega) = 1$ ✓

(2) $P_B(A^c) \overset{?}{=} 1 - P_B(A)$

$$\frac{P(A^c \cap B)}{\cancel{P(B)}} \overset{?}{=} \frac{P(B)}{\cancel{P(B)}} - \frac{P(A\cap B)}{\cancel{P(B)}}$$

$$P(B) \overset{?}{=} P(A^c \cap B) + P(A \cap B)$$



Since B is the union of disjoint sets:

$$B = (A^c \cap B) \cup (A \cap B)$$

$$\therefore\ P(B) = P(A^c \cap B) + P(A \cap B)$$

We can, thus, erase the "?" and we have:

$$P_B(A^c) = 1 - P_B(A) \quad ✓$$

(3) $P_B(A_1 \cup A_2 \cup \dots) = \dfrac{P\big((A_1 \cup A_2 \cup \dots) \cap B\big)}{P(B)}$

each of the A_i are disjoint, thus each of $(A_i \cap B)$ are too.

$= \dfrac{P\big((A_1\cap B) \cup (A_2 \cap B) \cup \dots\big)}{P(B)}$

$= \dfrac{P(A_1 \cap B) + P(A_2 \cap B) + \dots}{P(B)}$

$P_B(A_1 + A_2 + \dots) = P_B(A_1) + P_B(A_2) + \dots$ ✓

What's interesting is to understand what <u>conditional probability</u> <u>really means.</u> In practice. When you see what it really means, you will realize that this is the <u>greatest discovery</u> in the theory of probability.

Because, as we will see from the examples, it really renders our intuitive notion of how the probability of an event A changes, in the knowledge that some other event B has happened.

So let's see now the significant examples.
As with all of this, you don't understand anything w/o significant examples.
It's always best to start w/ the most Mickey Mouse possible example.
Because those Mickey Mouse examples are likely to give away the secrets more than the fancy ones.
So let's start w/ a super Mickey Mouse example.

Super Mickey Mouse.
What do you think I'm going to do?
<u>Balls into boxes</u>
Which one? Maxwell-Boltzmann — ordinary balls into ordinary boxes

<u>Example</u> in Ω_{MB}
Let's compute the following conditional probability.
What's the probability that the 2^{nd} box has j balls <u>given</u> that the 1^{st} box has no balls?

$$P(\theta_2 = j \mid \theta_1 = 0)$$

Now you can work it out <u>intuitively.</u> Because intuitively, if you assume the 1^{st} box has no balls, it is as though you had $n-1$ boxes. } If you don't see that, I'm sorry.

Let's make this come out automatically from the definition of conditional probability.
Let's pretend we don't know what our intuition tells us and let's see what happens.
Use the definition:

$$P(\theta_2 = j / \theta_1 = 0) = \frac{P((\theta_2 = j) \cap (\theta_1 = 0))}{P(\theta_1 = 0)} \Leftarrow \text{The scriptures say}$$

That's what the scriptures say. Let's compute this.
We've computed it umpteen times. Till you're sick of it.
What's the <u>numerator</u>?
Well, it's the joint distribution of 2 random variables

The <u>numerator</u> is:
We take j balls out of K and place them in box 2.
None of them go into box 1.
You're left w/ $K-j$ balls which go into $n-2$ boxes (i.e., neither box 1 nor 2)

$$P\left(\left(\theta_2=j\right)\cap\left(\theta_1=0\right)\right) = \frac{\binom{k}{j}(n-2)^{k-j}}{n^k}$$

The <u>denominator</u>:

Thats easy. All k balls go into the boxes other than box 1.
This can happen $(n-1)^k$ ways.

$$P(\theta_1=0) = \frac{(n-1)^k}{n^k}$$

So we have:

$$P(\theta_2=j \mid \theta_1=0) = \frac{\dfrac{\binom{k}{j}(n-2)^{k-j}}{\cancel{n^k}}}{\dfrac{(n-1)^k}{\cancel{n^k}}}$$

$$= \frac{\binom{k}{j}(n-2)^{k-j}}{(n-1)^k}$$

This is precisely what we got from the intuitive idea.
So this begins to convince you that maybe there is something to this.
Contrary to your initial skepticism.

OK - now let's do a non-intuitive example.

- Example - Bernoulli Process Ω_B

Q/A: Is there a bias.
No. Coin is fair

Let event $A = \underbrace{(X_2 = 1)}$

namely, the 2nd toss is a Heads (1)

event B = at least one of X_1, X_2 equals Heads (1)

Let's compute:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$A \cap B$ = event that: $\underbrace{(X_2 = 1)}_{\text{2nd toss is 1}}$ and $\underbrace{\text{at least one of } X_1, X_2 \text{ equals } 1}$

So this consists of the following events:

$$A \cap B = \underbrace{((X_1 = 1) \cap (X_2 = 1))} \cup \underbrace{((X_1 = 0) \cap (X_2 = 1))}$$

disjoint events

$$P(A \cap B) = P((X_1 = 1) \cap (X_2 = 1)) + P((X_1 = 0) \cap (X_2 = 1))$$
$$= \frac{1}{4} + \frac{1}{4}$$

$$P(A \cap B) = \frac{1}{2}$$

Note that $P(B) = \frac{3}{4}$

So:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$
$$= \frac{\frac{1}{2}}{\frac{3}{4}}$$
$$P(A|B) = 2/3$$

It's not what we expected

Q/A: Is this the Monte Hall problem?

You can't go very far w/ just the definition.
You now have to develope the <u>basic techniques</u> for working with conditional probabilities, And we'll do very fancy problems.

<u>Basic Techniques for working w/ conditional probabilities</u>

There are 3 basic rules, which will be used over and over again and which can solve extraordinary problems, if properly used.

<u>Rule 1</u> - Rule of <u>successive probabilities</u>

Remember that $P(A|B) = P_B(A)$

Therefore, we can consider:

$P_B(A|C)$ because P_B is a probability

Lets work it out:

$$P_B(A|C) = \frac{P_B(A \cap C)}{P_B(C)}$$

and this simplifies to:

$$= \frac{\dfrac{P(A \cap C \cap B)}{\cancel{P(B)}}}{\dfrac{P(C \cap B)}{\cancel{P(B)}}}$$

since $P(B)$ is assumed to be > 0.

$$= \frac{P(A \cap C \cap B)}{P(C \cap B)}$$

This can be viewed as <u>another conditional probability</u>

{This is the rule of successive probabilities} $= P(A | C \cap B) = \underbrace{P((A|C)|B)}_{\text{we rewrite } P_B(A|C) \text{ in standard terms}}$

That's Rule 1:
iterated conditioning

Rule 2 is the <u>most important</u>:

<u>Rule 2</u> – Law of alternatives

a partition is a <u>set</u> of sets

You must grasp the notion of this on a <u>partition.</u>

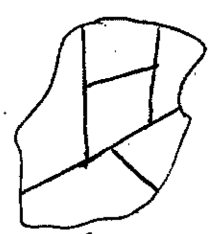
A partition π of a probability space Ω, P is a family of events $B \in \pi$ with the following properties:

(1) if $B \in \pi$ then $\underbrace{P(B) > 0}_{P(B) \text{ is positive}}$

(2) if $\overbrace{B \neq B' \in \pi}^{B, B' \in \pi,\ B \neq B'}$ then $B \cap B' = \phi$ except for an event of zero probability
(any 2 members of π are <u>disjoint</u>)

(3) $\bigcup_{B \in \pi} B = \Omega$ (except for an event of zero probability)

The intuitive picture of a partition is this:



Ω

You have your sample space Ω
And you cut it up.

The members of the partion are called <u>blocks</u>
if $B \in \pi$, B is called a block of π

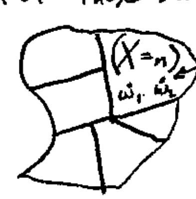
So it's an intuitive concept – cutting up the sample space into blocks.

<u>Example of a partition</u>

we've only seen integer random variables to this point

if X is any (integer) random variable, the events $(X=n)$ for which $\underbrace{P(X=n) > 0}_{\text{positive}}$ are a partition of the sample space Ω.

In other words, every <u>block</u> consists of all sample points where the random variable takes the value n.
Two sample points ω_1, ω_2 belong to the same block iff X of those sample points is the same (i.e., $X(\omega_1) = X(\omega_2)$).
This is an important concept.
So to every random variable, you can associate a partition.



$X(\omega_1) = X(\omega_2) =$

Most partitions come from random variables in this way.
Not all, but most.

There is a beautiful algebra of partitions, which is analogous to the algebra of sets, which we might go over in the super-class.
Partitions are duals to sets in a deep philosophical sense.
A partition is a basic notion for classification.
It is the dual notion of the event.

Later on, we will see that, just as you assign probabilities to events, there is also a number you assign to partitions, which is called the entropy. The analogy is Entropy is to partitions what probability is to events.

Q/A: Note that property (2) says that the intersection of 2 distinct blocks B and B' is null, except for an event of zero probability.

$$B \cap B' = \phi \cup \text{possibly an event of zero probability}$$

Property (3) says that:

$$\bigcup_{B \in \pi} B = \Omega - \text{possibly an event of zero probability}$$

The intersection of two distinct blocks is disjoint (subject to the above proviso).
They can't overlap.
In other words, partitions are classifications. You can view the random variable as the classification criteria.

Q/A: It is much more elegant to state properties (2) + (3) as:

$$\left.\begin{array}{ll} (2) & P(B \cap B') = 0 \\ (3) & P(\Omega - \bigcup_{B \in \pi} B) = 0 \end{array}\right\}$$

this way you avoid having to mention the proviso "possibly an event of zero probability".

By the way, if you want to be really fancy, you can prove that in a partition, there is at most a countable number of blocks.
This follows automatically.

Q: Does having a countable number of blocks then imply that there be some random variable X that forms the partition?
A: You can always invent a random variable.
You can always find a random variable to choose them with. Many of them.
The point is that the partition abstracts the blocks of the random variable that determines the partition.
In other words, the blocks of a partition are distinct, but they are not distinguished by specifics because of the random variable.
I leave this for next hour.

We now return to the Law of Alternatives:

if π is a partition of $\underbrace{\Omega, P}_{\text{sample space } \Omega \text{ with probability } P}$

then

$$P(A) = \underbrace{\sum_{B \in \pi}}_{\text{sum over all blocks in the partition}} P(A|B)P(B)$$

← That's the Law of Alternatives

Proof

$$RHS = \sum_{B \in \pi} \frac{P(A \cap B)}{\cancel{P(B)}} \cancel{P(B)}$$

$$= \sum_{B \in \pi} P(A \cap B)$$

The blocks B are disjoint.
Therefore $A \cap B$ are even more disjoint.

$$= P\left(\bigcup_{B \in \pi} (B \cap A)\right)$$

Now, we use the distributive law of $\cup$, $\cap$.

$$= P\left(\underbrace{\left(\bigcup_{B \in \pi} B\right)}_{\Omega \text{ (from property (3))}} \cap A\right)$$

$$= P(\Omega \cap A)$$

$$P(A) = P(A)$$

as desired

By the way, this is what failed in quantum probability, because in quantum probability, the distributive law doesn't hold. So the whole idea of conditional probability still exists, but it gets much more sophisticated, as Mr. Feynman showed when he was taking this course in 1939. Why don't you show this? He could do it in 1939. You can do it too. Actually it was 1938. At that time the course was taught by Norbert Wiener. Sorry.

Q: If we use the restated property (3) $P(\Omega - (\bigcup_{B \in \pi} B)) = 0$, how do we carry out this proof?

A: Instead of "=" say "=, except for an event of probability zero"

Rule 3 – Law of successive conditioning

Last night, just as I went to bed I thought "Gee, the examples are too trivial." I thought – maybe we can do the problem of the matching w/ conditional probability. And since I couldn't sleep, I worked out the problem of the matching using conditional probability. That's one of the examples coming up. It's quite interesting.

If events B_1 and B_2 are independent, then

$$P(B_1 | B_2) = P(B_1)$$

The idea is that event B_2 doesn't influence B_1. Therefore it shouldn't change.

and conversely.

This is immediately checked:

$$P(B_1|B_2) = \frac{P(B_1 \cap B_2)}{P(B_2)}$$

Since B_1 + B_2 are independent, $P(B_1 \cap B_2) = P(B_1)\,P(B_2)$

$$= \frac{P(B_1)\,\cancel{P(B_2)}}{\cancel{P(B_2)}}$$

$$= P(B_1) \checkmark$$

The law of successive conditioning is based on the following imaginary reasoning:

If B_1 and B_2 are independent, then $P(B_1 \cap B_2) = P(B_1)P(B_2)$

Suppose B_1 and B_2 are not independent. Can we fix this so we have something like it? Yes.

In general:

$$P(B_1 \cap B_2) = P(B_1)\,P(B_2|B_1)$$

Trivial. From the definition.

So this gives you a new way of looking at it – Conditional Probability. You can compute the probability of the intersection of any two events as the probability of the 1st event times the probability of the 2nd event, given that the 1st has happened.
This is kind of nice.

So nice, that you can generalize to <u>any number of events.</u>

Let's see.

$$P(B_1 \cap B_2 \cap B_3) = P(B_1)\,P(B_2|B_1)\,P(B_3|B_1 \cap B_2)$$

$$= \cancel{P(B_1)}\,\frac{\cancel{P(B_2 \cap B_1)}}{\cancel{P(B_1)}}\,\frac{P(B_3 \cap B_1 \cap B_2)}{\cancel{P(B_1 \cap B_2)}}$$

$$= P(B_1 \cap B_2 \cap B_3) \checkmark$$

It's <u>very</u> nice.
It tells us how to immediately generalize this for n events.
Exactly the same argument works for n events.

$$P(B_1 \cap B_2 \cap \ldots \cap B_n) = P(B_1)P(B_2|B_1)P(B_3|B_1 \cap B_2) \ldots P(B_n|B_1 \cap B_2 \cap \ldots \cap B_{n-1})$$

That's the <u>law of successive conditioning.</u>

Now you are wondering: "This looks so abstruse, it can not have any practical application."
Next time we'll spend the whole just doing applications of these laws.
And you'll see how extraordinarily they simplify computations of probability.

John Guidi
guidi@mit.edu

Conditional Probability (cont'd) Probability trees

We continue our discussion of conditional probability, which, as we said, is the most important technique for the computation of probability.
Let's review some fundamental facts:

In a sample space Ω, the probability of A, given that B has happened, is defined as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \Big\} \; P(A \cap B) \text{ normalized by } P(B).$$

This, of course, requires that the probability of the hypothesis B is positive:

$$P(B) > 0$$

We have seen already some examples (and we will keep convincing ourselves) that this corresponds indeed to the intuitive notion of the event B happening. How the probability of A changes in the knowledge that the event B happens.

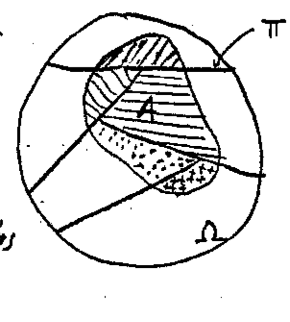
The fundamental properties of traditional probability are 3, but for the purposes of today's lecture, let's recall 2.

1) Law of alternatives

You are given Π = a partition of Ω
As you know, Π is a family of disjoint events called blocks, which cover the entire sample space, except for probability zero.
Each block has positive probability.
Blocks of probability 0 are not blocks.

We saw that every integer random variable defines a partition.
Let me write a formula that we didn't see last time:

$$(*) \qquad P(A) = \underbrace{\sum_{B \in \Pi}}_{\substack{B \text{ ranges over} \\ \text{all the blocks} \\ \text{of the partition} \\ \Pi.}} P(A \cap B)$$



This is proved by the picture:
The probability of A is the sum of the probabilities of the intersection of A with the blocks of the partition.
That's obvious.

From (*) we get the <u>Law of Alternatives</u>, which is a more sophisticated version, which states:

$$P(A) = \sum_{B \in \pi} P(A|B)\,P(B) \qquad \text{The Law of Alternatives}$$

The Law of Alternatives follows from formula (*) immediately, since:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \implies P(A \cap B) = P(A|B)\,P(B)$$

So the two formulas are really equivalent, although they have different interpretations.

2) <u>Law of successive conditioning</u>

$B_1, B_2, \ldots, B_n$ - finite set of events

The probability of <u>all these events happening</u> - if they were independent, that would be the product of the probabilities. They are <u>not</u> independent, so we fix that.
How do we fix it?
Using conditional probabilities.

$$P(B_1 \cap B_2 \cap \ldots \cap B_n) = P(B_1)\,P(B_2|B_1)\,P(B_3|B_1 \cap B_2) \cdots P(B_n|B_1 \cap B_2 \cap \ldots \cap B$$

Important <u>special cases</u>:

- if A and B are independent then $\underbrace{P(A|B) = P(A)}$

  This follows immediately from the definition, but intuitively, this confirms our interpretation of conditional probability.
  Because this formula says that whether or not B happens does <u>not</u> influence the probability of A.

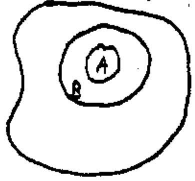
  $$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)\cancel{P(B)}}{\cancel{P(B)}} = P(A)$$

- if $\overbrace{A \subseteq B}^{A \text{ contained in } B}$ then $P(A|B) = \dfrac{P(A)}{P(B)}$

  $$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} \qquad A \cap B = A$$

  This further confirms our intutive interpretation. Let's draw a picture:

If you know that B has happened, then probability of B must be normalized to 1.



so we renormalize the probability of A.

- if $B_1 \supseteq B_2 \supseteq \ldots \supseteq B_n$ then the Law of Successive conditioning becomes:

$$P(B_1 \cap B_2 \cap \ldots \cap B_n) = P(B_1)P(B_2|B_1)\underbrace{P(B_3|B_2)}_{P(B_3|\underbrace{B_1 \cap B_2})}P(B_4|B_3)\cdots P(B_n|B_{n-1})$$

since $B_1 \supseteq B_2$,
$B_1 \cap B_2 = B_2$.
Thus:
$$P(B_3|B_1 \cap B_2) = P(B_3|B_2)$$

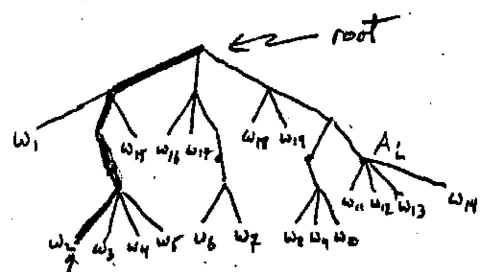
This is trivial.
None the less, the formula is useful, as you will see.

Now we want to get a really good feeling for these formulas.
So we want to look at them from different perspectives.
These are the crucial formulas for the computation of probabilities.

I can give you the theory first and the examples later, or the examples first.
I'll give you the theory first.
Slightly more sadistic.

- A very important type of sample space is a <u>sample space</u> which is constructed on <u>trees.</u>
We want to discuss them in great detail because they occur in practice.

<u>Sample spaces on trees</u> (none of this is in the book, by the way)



The sample spaces that are associated w/ trees have for a <u>sample point</u>, a <u>complete branch</u> of the tree.

this is a sample point ω_2.
We mark it just at the end,
but we really mean the unique branch starting at the root and ending here is the sample point.

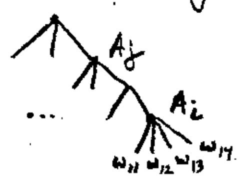
A probability tree means that you give probabilities to each of these sample points that add up to 1.

$$\sum_{i=1}^{19} P(\omega_i) = 1$$

But watch the following:

Every vertex of the probability tree is an event.

Take, for example, the vertex labelled A_i.
<u>A_i is the event containing all the sample points passing through A_i.</u>

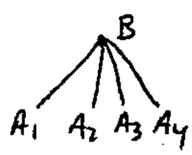


if vertex j is closer to the root than vertex i,

$$\underbrace{A_j \supseteq A_i}_{\text{event } A_j \text{ contains } A_i}$$

In the example above, the event A_i contains the 4 sample points $\omega_{11}, \omega_{12}, \omega_{13}, \omega_{14}$.
The event A_j contains all these sample points, plus extras.
The farther <u>up</u> the tree you go, the bigger the event.

If you go <u>down</u>, from the root of the tree, you get a set of smaller events, until you get to the sample points.



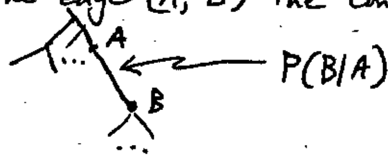
if A_i are all vertices immediately below B (say), then the events A_i are a <u>partition</u> of B.

This is <u>obvious.</u>

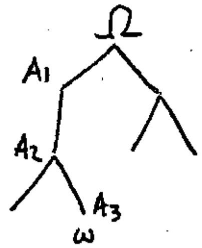
So you say "So what?"
Now you can do a wonderful thing.
We can give an <u>interpretation</u> to the <u>edges</u> of the tree
By <u>conditional probability.</u>
Watch.

<u>Basic fact</u> – The sample points are the <u>complete</u> path
Everything else follows from this basic fact.

If we have A immediately above B, so that $A \supseteq B$, then we assign to the edge (A, B) the conditional probability $P(B|A)$.



Why is that good?
Watch – from the <u>Law of successive conditioning</u>:



← I want the probability of this sample point.
$P(\omega)$

We can express this in terms of <u>conditional probability.</u>

$$\overbrace{P(\omega)}^{P(A_1 \cap A_2 \cap \dots)} = \underbrace{P(A_1|\Omega)}_{\text{this is the same as } P(A_1)} P(A_2|A_1)P(A_3|A_2) \dots$$

$$P(\omega) = P(A_1)P(A_2|A_1)P(A_3|A_2) \dots$$

Notice that:

$$A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots$$

So we are in the special case of the law of successive conditioning where the events are contained in one another [3/2/98.3]

The point is, in real life, you are <u>not</u> given the probabilities of the sample points. You are only given the conditional probabilities.
And you <u>build</u> the probabilities of the sample points <u>from</u> the conditional probabilities.

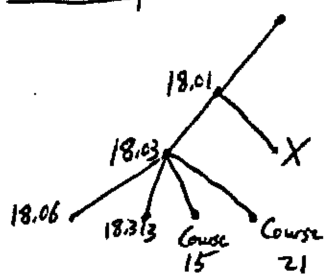
That's called <u>Bayesian Theory</u>

If you are given the <u>probabilities on the path</u>, then you can determine the probabilities of all the vertices.
The <u>vertices</u> correspond to <u>events.</u>
The <u>edges</u> correspond to <u>conditional probabilities.</u>

Conversely:
If you are given <u>conditional probabilities to the edges</u>, which are consistent, then you have a sample space! And you define the probabilities on all the paths.

This is the way a lot of sample spaces are defined.
Just like this.
Even <u>infinite</u> sample spaces.

- <u>For example</u>



You are admitted to MIT and you take 18.01.
If you pass, you take 18.03
If you flunk, you flunk out.

If you pass 18.03, you take either 18.06 or 18.313.
If you flunk it, you switch to either Course 15 or Course 21.

All you are given are the <u>conditional probabilities</u> (e.g., that you pass 18.313, given that you've passed 18.03.
That's the <u>natural data.</u>
You are <u>not</u> given the <u>probability of the sample points</u> (i.e., your history at MIT).

The event that you graduate is a union of sample points.
All the sample points which end up w/ you graduating.

Flunking out is another event.

So you do computations this way.

- <u>Example</u> The test problem – a classical example

In a hospital, people have a disease (the black plague) with probability:

$$p \text{ for plague} \longrightarrow p = .1$$
$$p^c = 1 - .1 = .9$$

There is a test for the plague that is 80% accurate:

accurate $a = .8$

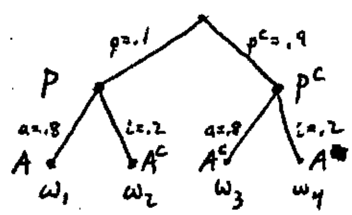
inaccurate $i = 1 - .8 = .2$

You go to the hospital to get tested and the test turns out positive.
What is the probability that you have the plague?

Let's build a tree.
Either you have the plague or you don't have the plague.
If you <u>have the plague</u> and the test is <u>accurate</u>, then it will come out as your having the plague. Or the test is <u>inaccurate</u>, then it will come out as your not having the plague.

If you <u>don't have the plague</u> and the test is <u>accurate</u>, test result is you don't have the plague. Or the test is <u>inaccurate</u>, then it will come out as your having the plague

Let P = event that you have the plague
A = event that test result says you have the plague



We want the conditional probability of the event that you have the plague, given that the test result says you have the plague.
Namely:

$$\text{What is } P(P|A)?$$

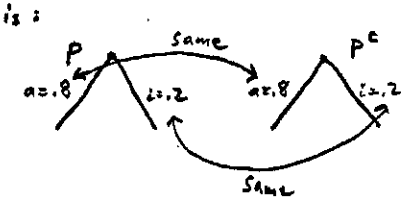
$$P(P|A) = \frac{P(P \cap A)}{P(A)}$$

(1) The event A = union of the two sample points ω_1 and ω_4

$$A = \{\omega_1\} \cup \{\omega_4\}$$

Thus $P(A) = P(\omega_1) + P(\omega_4)$

(2) In this first approximation, let's assume that the <u>accuracy/inaccuracy</u> of the test results are <u>independent</u> of whether you have the <u>plague or not</u>.
That is:



$$P(P|A) = \frac{P(P \cap A)}{P(\omega_1) + P(\omega_4)} \leftarrow \text{from (1)}$$

$$= \frac{P(\omega_1)}{P(\omega_1) + P(\omega_4)} \leftarrow P \cap A = \{\omega_1, \omega_2\} \cap \{\omega_1, \omega_4\} = \{\omega_1\}$$

$$= \frac{(.1)(.8)}{(.1)(.8) + (.9)(.2)} = .308$$

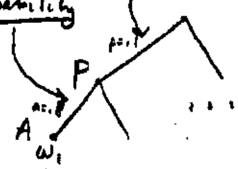
How comforting.

By the way, now let's assume the test accuracy $a = .9$ and inaccuracy $i = .1$.

Then $P(P|A) = \dfrac{(.1)(.9)}{(.1)(.9) + (.9)(.1)}$

$P(P|A) = .5 \leftarrow$ The probability that you have the plague, given that your test result says you have it is a <u>fair coin toss</u>.

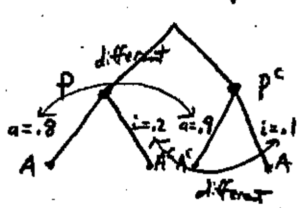
The probability of the sample point ω_1 is the <u>conditional probability</u> times the <u>conditional probability</u>



$$P(\omega_1) = P(P \cap A)$$
$$= P(P|\Omega)\, P(A|P)$$
$$= P(P)\, P(A|P)$$
$$= (.1)(.8)$$

Doctors make mistakes 80% of the time on this and have to hire statisticians at \$100,000/year. Because they can't do this computation.

- How can we jazz it up?
  We can say the test has <u>different accuracy</u> according to whether you administer it to a patient <u>with the plague, or w/o the plague.</u>
  How would we set up the tree in that case?
  We set up the tree in much the same way.
  In this case, we no longer have independent events.



Setting it up as a tree helps a lot.

Just as the probability of your flunking out can be set up as a probability tree, with all the conditional probabilities.
You can compute, for example, "What's the probability that I flunk out, given that I've taken 18.313?"
Or I can ask "Given a student that has flunked out, what is the probability he has taken 18.313?"
It's all in the tree. You can work it out. Once you've set up the tree, you can answer all these questions.

Now what do you think comes next?
Balls into boxes.

We jazz up Maxwell-Boltzmann this way:

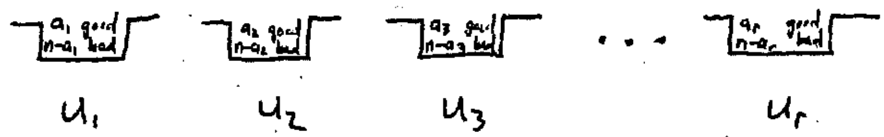
I have an indefinite number of boxes. It might even be infinite.

Each box is an urn.
Every box (urn) has the same number of balls (n).

The 1^{st} box has a_1 good balls, $n-a_1$ bad balls
" 2^{nd} " " a_2 " " , $n-a_2$ " "
etc.

I would be greatful for a good interpretation of this.
A griping interpretation.

Let's restrict ourselves to a finite number of boxes/urns (r):



all boxes have n balls

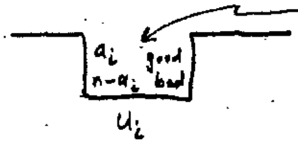
Each urn has a different number of good balls.
My strategy is the following:

1) pick an urn at random

then 2) pick k balls out of that urn

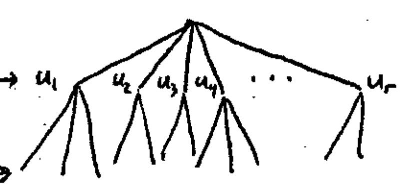
Let random variable A = number of good balls I pick.
What is the probability distribution of A?

Strictly speaking, I'm deviating from my usual terminology.



This should be a population.
The balls are members of a population.
The balls are in the form of boxes in Maxwell-Boltzmann.
The balls are really boxes.

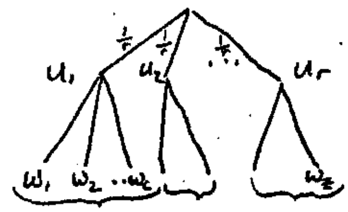
Let's set up a tree for the sample space.
We have one edge for each urn. So, you choose one of $U_1, \ldots, U_r$.



Having chosen an urn, we are in the ordinary Maxwell-Boltzmann situation, We put k balls into n boxes.
The boxes are labelled good or not good

You can do this w/ replacement or w/o replacement.
Let's do both cases.
First:

With replacement



The sample points $\omega_1, \dots, \omega_\ell$ correspond to all the possible sets of k distinguishable balls that can be extracted from urn U_1

Similary, for each subtree under urn U_i, the sample points ω correspond to all the possible sets of k distinguishable balls that can be extracted from that urn.

The sample points ω of our sample space Ω are these sets.
This is very important - there are different sample spaces Ω_{MB}, since we pick urns first.

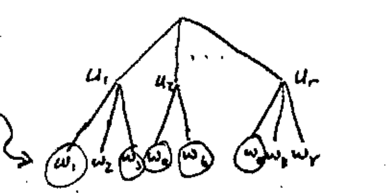
In the absence of any further information, we give the choice of urns the same probability $(\frac{1}{r})$.
We will see later what happens if you alter these probabilities - if you give an uneven probability for the choice of urns.

We want the probability distribution $P(A=i)$.
This can be computed in one, and only one, way - with the law of alternatives.

$$P(A=i) = \sum_{j=1}^{r} \underbrace{P(A=i \mid U_j)}_{\substack{\text{probability of the event}\\ \text{that we extract } i\\ \text{good balls, given that}\\ \text{we choose urn } j.}} P(U_j)$$

$\left\{ \begin{array}{l} U_j = \text{event that we choose urn } j \\ P(U_j) = \frac{1}{r} \end{array} \right\}$

We are adding all the sample points, as per theory. We didn't even need to know the law of alternatives!



O = sample points w/ good sets of K extractions that have i good balls.

$$= \frac{1}{r} \sum_{j=1}^{r} P(A=i \mid U_j)$$

$$P(A=i) = \frac{1}{r} \sum_{j=1}^{r} \binom{k}{i} \left(\frac{a_j}{n}\right)^{i} \left(\frac{n-a_j}{n}\right)^{k-i}$$

probability ball is a good one.

probability ball is not a good one.

You mentally check that this is a probability distribution:

$$\sum_{i=0}^{k} P(A=i) = 1$$

$$\sum_{i=0}^{k} \frac{1}{r} \sum_{j=1}^{r} \binom{k}{i} \left(\frac{a_j}{n}\right)^i \left(\frac{n-a_j}{n}\right)^{k-i} = 1$$

$$\sum_{i=0}^{k} \binom{k}{i} \underbrace{\frac{1}{r} \sum_{j=1}^{r} \left(\frac{a_j}{n}\right)^i \left(\frac{n-a_j}{n}\right)^{k-i}} = 1$$

Since, from the binomial theorem:

$$\sum_{i=0}^{k} \binom{k}{i} \overbrace{p^i q^{k-i}} = 1 \quad , \text{where } p+q=1$$

- That's pretty easy.
  Now, in the same sample space, let's ask another question.
  What is the probability I've chosen the 1st urn, give that I have extracted i good balls?

$$P(U_1 | A=i) = \frac{P(U_1 \cap (A=i))}{P(A=i)}$$

We can rewrite the numerator as:

$$P(U_1 \cap (A=i)) = P((A=i) \cap U_1)$$
$$= P((A=i) | U_1)\, P(U_1)$$

by definition of conditional probability.

$$= \frac{P((A=i)|U_1)\, P(U_1)}{P(A=i)}$$

We know these data:

$$P((A=i)|U_1) = \binom{k}{i} \left(\frac{a_1}{n}\right)^i \left(\frac{n-a_1}{n}\right)^{k-i}$$

$$P(U_1) = \frac{1}{r}$$

$$P(A=i) = \frac{1}{r} \sum_{j=1}^{r} \binom{k}{i} \left(\frac{a_j}{n}\right)^i \left(\frac{n-a_j}{n}\right)^{k-i}$$

$$= \frac{\cancel{\binom{k}{i}}\left(\frac{a_1}{n}\right)^i\left(\frac{n-a_1}{n}\right)^{k-i}\cdot\cancel{\frac{1}{r}}}{\cancel{\frac{1}{r}}\sum_{j=1}^{r}\cancel{\binom{k}{i}}\left(\frac{a_j}{n}\right)^i\left(\frac{n-a_j}{n}\right)^{k-i}}$$

$$P(U_1|A=i) = \frac{\left(\frac{a_1}{n}\right)^i\left(\frac{n-a_1}{n}\right)^{k-i}}{\sum_{j=1}^{r}\left(\frac{a_j}{n}\right)^i\left(\frac{n-a_j}{n}\right)^{k-i}}$$

The idea is this. You don't know which urn the balls are extracted from. And you want an estimate of the probability that you've picked a given urn on the basis of how many good balls you have.
That estimate is $P(U_1 | A = i)$.

↳ the probability that you've picked urn 1, given that you've extracted i good balls.

This is called the <u>Bayesian Probability</u>

We will discuss, next time, Bayes' Law.
All this nonsense about Bayes Law disappears when you visualize everything in terms of trees.

Quiz 1 average = 31
median ≈ 60

Conditional Probability (cont'd)

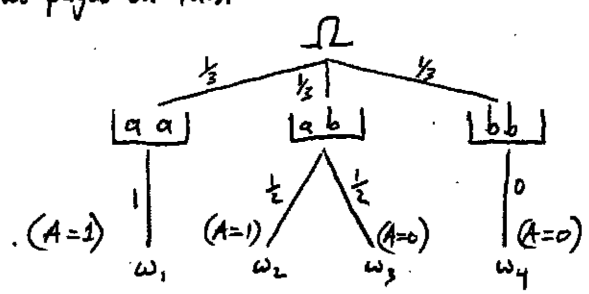
The main technique for computing probabilities.

$$P(A|B)$$

We're going to examine this from all sorts of different viewpoints.
Let's start w/ a couple of easy problems.

An urn has 2 balls, colored a or b. The balls are distinguishable.
An a-ball is extracted (a ball is extracted from the urn + it turns out to be an a-ball).
What is the probability that both balls are a-balls?

This is a typical problem of the kind that we have begun to name last time as Bayesian, after the writer Sir John Bayes, who, by accident, wrote two pages on this.



We don't know what the distribution of the balls in an urn is.
Our sample space consists of first imagining 3 possibilities for the urn.

Then the extraction.

Let event A = number of a-balls extracted from the single extraction
event U = number of a-balls in the urn before extraction

All together, you have 4 sample points $(\omega_1, \omega_2, \omega_3, \omega_4)$

In the absence of any further information, we assume we choose one of the 3 urn configurations with equal probability.
This is something we will discuss later.
Under what circumstances should we make a different assumption?
This is something that is extremely hard to justify.
That is, why is it, in the absence of any further information, that you assume all the possibilities have equal probability.

Q: By symmetry?
A: No. That's circular.

Reams of words have been written on this.
There are people in the philosophy department who are working full-time on this.

Let's first consider $P(A=1)$.

Evaluated by the <u>law of alternatives</u>:

$$P(A=1) = \sum_{(U=i)\in\Pi} P(A=1 \mid U=i)\, P(U=i)$$

$$= \underbrace{\overset{0}{\cancel{P(A=1 \mid U=0)}}}\, P(U=0) + P(A=1 \mid U=1)\, P(U=1) + P(A=1 \mid U=2)\, P(U=2)$$

This is 0. Since the urn has no a-ball, the probability of extracting an a-ball is 0.

The event $(A=1 / U=1)$ has the sample point ω_2

$$P(A=1 \mid U=1) = \tfrac{1}{2}$$

The event $(A=1 / U=2)$ has the sample point ω_1

$$P(A=1 \mid U=2) = 1$$

$$P(A=1) = \tfrac{1}{2}\cdot\tfrac{1}{3} + 1\cdot\tfrac{1}{3}$$

$$= \tfrac{1}{2}$$

Now to the question being asked.
What's being asked is $\underbrace{P(U=2 \mid A=1)}$.

if you extract one a-ball, what's the probability that the urn had two a-balls?

We have the whole sample space on our tree diagram.
So you can compute this w/ conditional probability.

$$P(U=2 \mid A=1) = \frac{P((U=2)\cap(A=1))}{P(A=1)}$$ by definition of conditional probability.

This is one of the most important tricks in probability.
What we are given are the conditional probabilities coming <u>down</u> the tree.

So we rewrite the numerator:

$$P((U=2)\cap(A=1)) = P((A=1)\cap(U=2))$$
$$= P(A=1\mid U=2)\,P(U=2)$$

$$= \frac{P(A=1\mid U=2)\,P(U=2)}{P(A=1)}$$

Note that: $\underbrace{P(A=1\mid U=2)}_{\text{Because this is \underline{certain}}} = 1$

$$P(U=2) = \frac{1}{3}$$

$$P(A=1) = \frac{|\{\omega_1,\omega_2\}|}{|\Omega|} = \frac{2}{4} = \frac{1}{2}$$

$$P(U=2\mid A=1) = \frac{1\cdot\frac{1}{3}}{\frac{1}{2}}$$
$$= \frac{2}{3}$$

Now you think you all know that.
Therefore I pose a harder problem.
<u>Now we do this all over with feeling</u>

↑ do you know where this expression comes from?
It comes from a New Yorker cartoon of the '40's.
Which you have never seen, which portrays the scene of a film being filmed in Hollywood. A love scene – a woman + a man on a sofa. The camera around, the director. They just finished shooting the love scene and the director says to the two actors:
"Now let's do again all over with, with feeling."

This is a tough problem. 3/6/98.4

- Example – Bernoulli Process

Let A = event that a run of h ones occurs before a run of t zeros. (the first run, of course)

Let's compute $P(A)$.

The way it's done in the book is lousy.

For the sake of this argument, say $h=4$, $t=3$

The method will be completely apparent from this example.

Consider $P(A \mid (X_1=0) \cap (X_2=1))$

This is the conditional probability that A occurs (namely, a run of 4 heads occurs before a run of 3 tails) given that the 1st toss is a tail and the 2nd toss is a head.

I claim that this $(X_1=0)$ is irrelevant.

Namely that:

$$P(A \mid (X_1=0) \cap (X_2=1)) = P(A \mid X_2=1)$$

The probability of event A, given that $(X_2=1)$ is independent of $(X_1=0)$.

Lets write this in a pedantic way:

$$P(A \mid (X_1=0) \cap (X_2=1)) = P((A \mid X_1=0) \mid X_2=1)$$

I'm just rubbing it in.

$$= P_{(X_2=1)}(A \mid X_1=0)$$

we write the subscript here to stress the fact that conditional probability is a probability.

If you have tosses 0, 1, then a number of tosses and then the event A occurs:

(Note: if instead we considered the case $X_1=1$, then for event A, $(X_1=1)$ would be relevant.)

$$\to 0\ 1\ \ldots\ A$$

(This 0 can not play any role in the event A occurring. The event that the 1st toss is 0 is independent of the event A occurring. If you don't see that, I can't explain it any more.)

(It doesn't influence the event A ever occurring. It's shit. So, $P_{(X_2=1)}(A \mid X_1=0) = P_{(X_2=1)}(A)$ The occurrence of $(X_2=1)$ makes $(X_1=0)$ irrelevant. If $X_1=0$, it can not play any role. It's independent.)

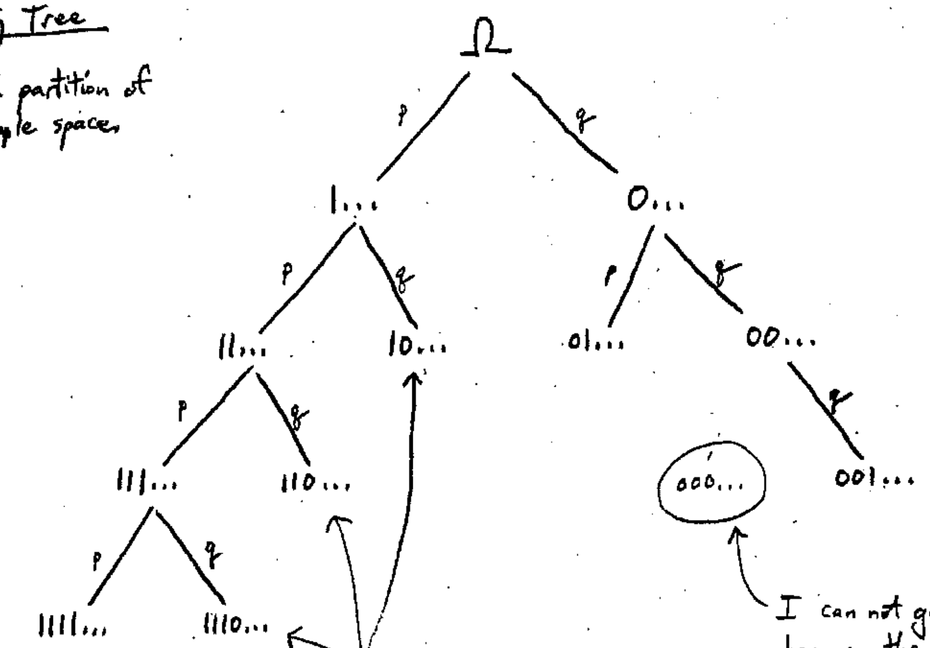
$$= P_{(X_2=1)}(A)$$

$$P(A \mid (X_1=0) \cap (X_2=1)) = P(A \mid X_2=1)$$

As claimed.

Probability Tree

This is a partition of the sample spaces



Here, we win. All the sample points stating w/ this are in our event A.

why don't I continue these? We'll discuss it further in a moment, but it's for the reason just mentioned.

I can not get 000..., because then I lose. Event A can not happen.

$$P(A) = P(\underbrace{1111\ldots}_{\text{event } A\ldots}) + P(\underbrace{1110\ldots}_{\text{event } 1110\ldots\, A\ldots}) + P(110\ldots) + P(10\ldots) + P(01\ldots) + P(001\ldots)$$

The tree displays it right away.

Now you say, whats $P(10\ldots)$?
This is the probability that the event A occurs, given that $(X_1 = 1)$ and $(X_2 = 0)$.

$$P(A \mid X_1 = 1, X_2 = 0) = P(A \mid X_2 = 0)$$

sometimes we write commas instead of $\cap$.

for exactly the same argument at end of previous page. The leading 1 has no influence on the occurrence of event A.

$\cancel{1}0\ldots\ A$

$$= P(A \mid X_1 = 0)$$

Since what happened before is independent, probability of A occurring after 2nd toss is 0 is the same as the probability of A occuring given the 1st toss is 0.

Similarily, $P(A \mid X_1 = 1, X_2 = 1, X_3 = 0) = P(A \mid X_3 = 0)$

It's only X_3 that matters.
X_1 and X_2 can not make A occur at all.
The event A is independent of $(X_1 = 1)$ and $(X_2 = 1)$, if you have $(X_3 = 0)$

Similarly, $P(A|X_1=1, X_2=1, X_3=1, X_4=0) = P(A|X_1=0)$

Exactly the <u>same argument</u> holds.
None of X_1, X_2, X_3 can influence event A.
Events $(X_1=1), (X_2=1), (X_3=1)$ are <u>independent</u> of A, <u>given</u> that $(X_4=0)$

Same idea gives:

$$P(A|X_1=0, X_2=1) = P(A|X_2=1) = P(A|X_1=1)$$

The event $(X_1=0)$ is <u>independent</u> of A, <u>given</u> that $(X_2=1)$
So you can cross out $X_1=0$ and say probability of A, given that 1 occurs after the second toss.
But this is the same as probability of A, given that 1 occurs after the 1st toss.

Similarly,

$$P(A|X_1=0, X_2=0, X_3=1) = P(A|X_1=1)$$

$(X_1=0)$ and $(X_2=0)$ are <u>independent</u> of A, <u>given</u> that $(X_3=1)$
So you might as well start at the 3rd toss w/ a 1 and say "What's the probability of A occurring after the 3rd toss, given that the 3rd toss is a 1?"
But $\underbrace{P(A|X_3=1)}$ does not know that two tosses have already gone off.
probability that A occurs after the 3rd toss, given that the 3rd toss is 1.

So you <u>relabel</u> the tosses, as if the 3rd was the 1st.

relabel tosses
$$\left\{\begin{array}{l} P(A|X_3=1) \Rightarrow \underset{1^{st}}{?} \quad \underset{2^{nd}}{?} \quad \underset{3^{rd}}{1} \quad \ldots \quad A \\ \\ P(A|X_1=1) \Rightarrow \underset{1^{st}}{1} \quad \ldots \quad A \end{array}\right.$$

Now we are done.

$$P(1111\ldots) = P((X_1=1) \cap (X_2=1) \cap (X_3=1) \cap (X_4=1))$$

$$= p \quad * \quad p \quad * \quad p \quad * \quad p$$

$$\boxed{P(1111\ldots) = p^4}$$

$$P(1110\ldots) = P\left(A \cap (X_1=1) \cap (X_2=1) \cap (X_3=1) \cap (X_4=0)\right)$$

from definition of conditional probability:

$$= \underbrace{P(A \mid X_1=1, X_2=1, X_3=1, X_4=0)}_{\substack{\text{we've already shown that} \\ \text{this is: } P(A \mid X_1=0)}} \underbrace{P\left((X_1=1) \cap (X_2=1) \cap (X_3=1) \cap (X_4=0)\right)}_{\text{expand using law of successive conditioning}}$$

$$= P(A \mid X_1=0)\, P(X_1=1)\, P(X_2=1 \mid X_1=1)\, P(X_3=1 \mid X_1=1, X_2=1)\, P(X_4=0 \mid X_1=1, X_2=1, X_3=1)$$

Note that the coin tosses are independent

$$P(X_i = 0/1 \mid \text{any other tosses}) = P(X_i = 0/1)$$

$$= P(A \mid X_1=0)\, P(X_1=1)\, P(X_2=1)\, P(X_3=1)\, P(X_4=0)$$

$$\boxed{P(1110\ldots) = P(A \mid X_1=0)\, p^3 q}$$

$$P(110\ldots) = P\left(A \cap (X_1=1) \cap (X_2=1) \cap (X_3=0)\right)$$

$$= P(A \mid X_1=1, X_2=1, X_3=0)\, P\left((X_1=1) \cap (X_2=1) \cap (X_3=0)\right)$$

$$\boxed{P(110\ldots) = P(A \mid X_1=0)\, p^2 q}$$

$$P(10\ldots) = P\left(A \cap (X_1=1) \cap (X_2=0)\right)$$

$$= P(A \mid X_1=1, X_2=0)\, P\left((X_1=1) \cap (X_2=0)\right)$$

$$\boxed{P(10\ldots) = P(A \mid X_1=0)\, pq}$$

$$P(01\ldots) = P\left(A \cap (X_1=0) \cap (X_2=1)\right)$$

$$= P(A \mid X_1=0, X_2=1)\, P\left((X_1=0) \cap (X_2=1)\right)$$

$$\boxed{P(01\ldots) = P(A \mid X_1=1)\, qp}$$

$$P(001\ldots) = P(A \cap (X_1=0) \cap (X_2=0) \cap (X_3=1))$$
$$= P(A \mid X_1=0, X_2=0, X_3=1)\, P((X_1=0) \cap (X_2=0) \cap (X_3=1))$$
$$\boxed{P(001\ldots) = P(A \mid X_1=1)\, q^2 p}$$

Q/A: You don't include $P(000\ldots)$, as this means a run of 3 tails occurs <u>before</u> any heads – thus the event A (a run of 4 heads before a run of 3 tails) can <u>not</u> happen. $P(000\ldots) = 0$

Therefore:

$$P(A) = P(1111\ldots) + P(1110\ldots) + P(110\ldots) + P(10\ldots) + P(01\ldots) + P(001\ldots)$$
$$P(A) = p^4 + P(A \mid X_1=0)p^3q + P(A \mid X_1=0)p^2q + P(A \mid X_1=0)pq + P(A \mid X_1=1)qp + P(A \mid X_1=1)q^2p$$

And now we are behind the 8-ball.
Because, where are we going to get probabilities $P(A \mid X_1=0)$ and $P(A \mid X_1=1)$?

Where did I go wrong? (On purpose, I must say)

Consider $P(A \mid X_1=1)$

$$P(A \mid X_1=1) = \frac{P(A \cap (X_1=1))}{P(X_1=1)}$$ counts only events under $(X_1=1)$, that is the entire left subtree on [3/6/98.5], normalized by $P(X_1=1)$.
That's what conditional probability is all about.

$$P(A \mid X_1=1) = \frac{P(1111\ldots) + P(1110\ldots) + P(110\ldots) + P(10\ldots)}{P(X_1=1)}$$

We've already computed the terms in the numerator.
For the denominator, we know that:

$$P(X_1=1) = p$$

$$\Downarrow$$

$$P(A \mid X_1=1) = p^3 + P(A \mid X_1=0)p^2q + P(A \mid X_1=0)pq + P(A \mid X_1=0)q$$

$$= (q + pq + p^2 q)\, P(A|X_1=0) + p^3$$

$$= q(1 + p + p^2)\, P(A|X_1=0) + p^3$$

$$= \cancel{q}\left(\frac{1-p^3}{\cancel{1-p}}\right) P(A|X_1=0) + p^3$$

$1-p=q$

$$\boxed{P(A|X_1=1) = (1-p^3)\, P(A|X_1=0) + p^3}$$

Why is this good?
Because this expresses $P(A|X_1=1)$ in terms of $P(A|X_1=0)$.
We do the same on the right subtree and we get an expression where $P(A|X_1=0)$ is in terms of $P(A|X_1=1)$.
We end up w/ 2 linear equations with 2 unknowns.
We can solve this.

Now consider $P(A|X_1=0)$

Now we consider the right subtree, counting events under $(X_1=0)$, that is the entire right subtree on [3/6/98.5], normalized by $P(X_1=0)$.

$$P(A|X_1=0) = \frac{P(A \cap (X_1=0))}{P(X_1=0)}$$

$$P(A|X_1=0) = \frac{P(01\ldots) + P(001\ldots)}{P(X_1=0)}$$ ← computed in previous pages

$$= P(A|X_1=1)\,p + P(A|X_1=1)\,qp$$

$$= p(1+q)\, P(A|X_1=1)$$

$$= \cancel{p}\left(\frac{1-q^2}{\cancel{1-q}}\right) P(A|X_1=1)$$

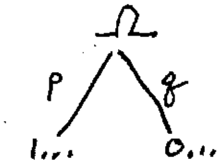
$1-q=p$

$$\boxed{P(A|X_1=0) = (1-q^2)\, P(A|X_1=1)}$$

Now, We can solve the two simultaneous linear equations and obtain:

$P(A|X_1=1)$
$P(A|X_1=0)$

Consider again our probability tree of [3/6/98.5] – only now just the first branching.



The <u>Law of Alternatives</u>: gives:

$$P(A) = \underbrace{P(A|X_1=1)} P(X_1=1) + \underbrace{P(A|X_1=0)} P(X_1=0)$$

we just showed how to compute these.

$$\boxed{P(A) = P(A|X_1=1)\, p + P(A|X_1=0)\, q}$$

If you can find a better way, I'll give you 100 points.
The answer is given in the book.

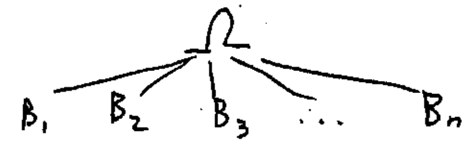
In all these probability problems, you have to get a recursion for the probability that you are trying to solve.
In essence, what we got is a recurrence difference equation for the probability that we solved.

You have 2 problems on this, for next week, which are fairly tough.

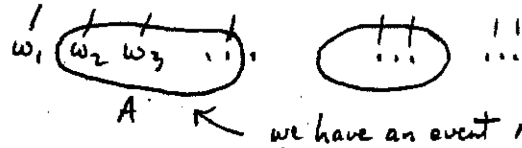
By this method we can solve any problem, whatsoever, like:
What's the probability that the pattern 011100 occurs before pattern 100011 (the complement)?

Just draw the tree and do the possibilities.
It comes out.

Summarizing what we've observed



We have a sample space comprised of a number of events (B_i).



At the end of the branches, we have the sample points.

we have an event A here.

Now you can compute the probability of A by conditioning on the probabilities of the B_i's (the Law of Alternatives):

$$P(A) = \sum_i P(A|B_i)\,P(B_i)$$

However, these B_i are used in another sense, as well.
For example - you are given several B_i's, one of which is the good one.
You perform experiment A.
You find that $A=32$.
On the basis of these data, what's the probability that the first B_i (i.e., B_1) is the good one?

$$P(B_1|A=32) = ?$$

where A = random variable
B_1 = event that B_1 is good

$$P(B_1|A=32) = \frac{P(B_1 \cap (A=32))}{P(A=32)}$$

definition of conditional probability

Bayes' Law is sometimes written like this, w/ the denominator expanded by the Law of Alternatives:

$$P(B_1|A=32) = \frac{P(A=32|B_1)\,P(B_1)}{P(A=32)} = \frac{P(A=32|B_1)\,P(B_1)}{\sum_i P(A=32|B_i)\,P(B_i)}$$

This is called Bayes' Law.

You decide among several hypotheses $(B_1, \ldots, B_n)$ on the basis of performing an experiment. The result of the experiment determines the probability of each of the possible hypotheses.
We are going to see several more examples of this, because it's very important.

Conditional Probability and Bayes' Law (cont'd)

The Law of Alternatives:

$$P(A) = \sum_{B \in \pi} P(A|B)\, P(B)$$

where π is a partition of the sample space Ω.
I.e., π is a family of events such that:

(1) for $B \in \pi$, $P(B) > 0$

(2) if $B, B' \in \pi$, $B \neq B'$, then $B \cap B' = \phi$

(3) $P(\bigcup_{B \in \pi} B) = 1$
probability of union of all the blocks of π = 1.

The Law of Alternatives, along with the Law of Successive Conditioning, is the main means for computing probabilities.
We are going to go over some more examples. There are so many angles to it.
There are so many ways to do things.
You never finish learning.
When you think you've learned all about it, you discover some new twist, whereby you apply things in a new way.

Here's a new twist. Suppose we compute the conditional probability of A, given event C.

$$P(A|C) = P_C(A)$$

Now, we've said many times that $P(A|C)$ is itself a probability.
Namely, defining a conditional probability is taking the same event and renormalizing the given probability — replacing it with another probability $P_C(A)$, which has all the same properties of all probabilities.

Therefore, the Law of Alternatives for the conditional probability P_C reads as follows:

$$P_C(A) = \sum_{B \in \pi} P_C(A|B)\, P_C(B)$$

The Law of Alternatives is valid for every probability.
Therefore, it is valid for the probability P_C.

Now let's unscramble this and write it in the "|" (given) notation.
Let's see what it looks like.
Sounds like childs play. Watch.

$$P(A|C) = \sum_{B\in\pi} P((A|B)|C)\,P(B|C)$$

From the Law of Successive Probabilities, we have [2/27/98.7]:

$$P((A|B)|C) = P(A|B\cap C)$$

$$P(A|C) = \sum_{B\in\pi} P(A|B\cap C)\,P(B|C)$$

This is the Law of Alternatives, for conditioning relative to C.

It's the same Law, but when applied, you have to be very careful about the conditioning event.

- Let's do the same problem I did last time w/ probability trees [3/6/98.4-10], but this time, <u>w/o probability trees</u>.

<u>Example</u> - Bernoulli Process

Let A = event that the first run of h one's precedes the first run of t zero's.

Let's just fool around a little bit.

What is $P(A|(X_1=0)\cap(X_2=0)\cap\ldots\cap(X_t=0))$?

$$P(A|(X_1=0)\cap(X_2=0)\cap\ldots\cap(X_t=0)) = 0$$

Because if the first t tosses are 0, then the first run of t zeros occurs before any ones. A run of t zeros occurs <u>certainly</u> before a run of h ones.
Therefore, this probability is 0.
If you don't see it, I can't explain it.

Let's jazz this up a little bit.

What is $P(A|(X_1=0)\cap(X_2=1))$?

This is the probability of the event A taken on all sequences of zeros and ones that start with 01.
You just restrict to that sample space of all sequences that start w/ 01.

I claim that:

$$P(A|(X_1=0)\cap(X_2=1)) = P(A|(X_1=1))$$

Because the 1st zero can not possibly influence the occurrence of A – in view of the fact $(X_2=1)$, so long as $h, t \geq 2$.

This again – if you don't see it, I can't explain it.

$P(A|(X_1=0)\cap(X_2=1))$ is independent of $(X_1=0)$, because the event in between, i.e., $(X_2=1)$ breaks any sequence that could satisfy A.

It's a crucial fact.

This is a very simple application of the Law of Alternatives.

Let's take as a partition $\pi = \{\underbrace{(X_1=1)}_{B_1}, \underbrace{(X_1=0)}_{B_2}\}$

$$P(A) = \sum_{B\in\pi} P(A|B)\,P(B) \Longleftarrow$$

B_1 + B_2 are disjoint blocks
Their union is the entire sample space
$P(\bigcup_{B\in\pi} B) = 1$.
Therefore, the Law of Alternatives applies.

$$= P(A|(X_1=1))\underbrace{P(X_1=1)}_{p} + P(A|(X_1=0))\underbrace{P(X_1=0)}_{q}$$

$$(**)\quad \boxed{P(A) = P(A|X_1=1)\,p + P(A|X_1=0)\,q}$$

Let W^0 = waiting time for the first zero

To solve P(A), I need to compute $P(A|X_1=1)$ and $P(A|X_1=0)$.

First, we compute $P_{(X_1=1)}(A)$, by conditioning on ~~the following~~ partition:

$$\pi = \{(W^0 \leq h), (W^0 > h)\}$$

1) These two events are disjoint (i.e., blocks)
2) Their union is the whole sample space.

Therefore, we can apply the Law of Alternatives

We apply the Law of Alternatives, but w/ a kick.

We apply it for the probability $P_{(X_1=1)}$:

$$P_{(X_1=1)}(A) = \sum_{B \in \pi} P_{(X_1=1)}(A|B)\, P_{(X_1=1)}(B)$$

$$= P_{(X_1=1)}(A|W^0 \le h) P_{(X_1=1)}(W^0 \le h) + P_{(X_1=1)}(A|W^0 > h) P_{(X_1=1)}(W^0 > h)$$

Let's try to simplify this.
First consider:

$$P_{(X_1=1)}(A|W^0 > h)$$

This is the probability of event A, given that the waiting time for the first zero is strictly greater than h.
That means at least the first h tosses are ones.
Thus event A happens absolutely.
The first h tosses are one.
Therefore:

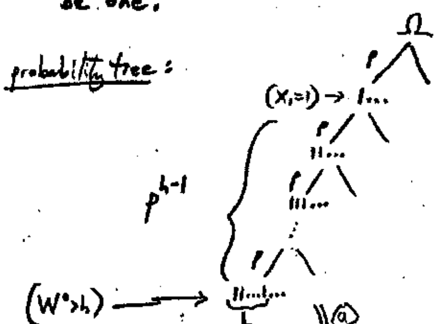
$$P_{(X_1=1)}(A|W^0 > h) = 1$$

Next, we consider $P_{(X_1=1)}(W^0 > h)$:

$(X_1=1)$ means the first toss is one.
$(W^0 > h)$ means that at least the first h tosses are one.
One of them, the first has already happened. So the next h-1 tosses have to be one.

{ Q/A: Note that getting a run of more than h heads includes getting a run of h heads. }

(a) by probability tree:



(b) Alternatively – by the Law of Alternatives:

$$P_{(X_1=1)}(W^0>h) = \underbrace{P_{(X_1=1)}\left(W^0>h \,\middle|\, \substack{\text{tosses } 2\ldots h \\ \text{all } 1}\right)}_{\substack{(X=1) \text{ and } \left(\substack{\text{tosses } 2\ldots h\\ \text{all } 1}\right) \text{ mean}\\ 1^{\text{st}}\ h \text{ tosses } 1. \text{ Thus } (W^0>h) \\ \text{is a certainty: } 1}} \underbrace{P_{(X_1=1)}\left(\substack{\text{tosses } 2\ldots h \\ \text{all } 1}\right)}_{p^{h-1}}$$

$$+ \underbrace{P_{(X_1=1)}\left(W^0>h \,\middle|\, \left(\substack{\text{tosses } 2\ldots h \\ \text{all } 1}\right)^c\right)}_{\substack{\left(\substack{\text{tosses } 2\ldots h \\ \text{all } 1}\right)^c \text{ means one of the first}\\ h \text{ tosses is } 0. \text{ Thus } (W^0>h) \text{ can}\\ \text{never happen: } 0}} P_{(X_1=1)}\left(\left(\substack{\text{tosses } 2\ldots h \\ \text{all } 1}\right)^c\right)$$

(b)

$$P_{(X_1=1)}(W^0 > h) = p^{h-1}$$

Now let's evaluate the first term.
First, we have:

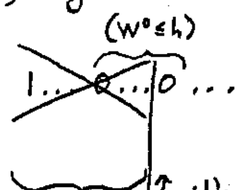
$$P_{(X_1=1)}(W^0 \le h) = 1 - P_{(X_1=1)}(W^0 > h)$$

$$= 1 - p^{h-1}$$

Finally, we are left with the toughest part:

$$P_{(X_1=1)}(A/W^0 \le h)$$

Let's think of what this is.
This is the probability $P_{(X_1=1)}$ that some zeros occur $(W^0 \le h)$ before you have a run of h ones.
So the zeros associated with the event $(W^0 \le h)$ break the run.
Therefore, any run that has occurred before is useless.



it's as if you were starting with a zero.
Because everything up to this is irrelevant.

The only thing that matters is that given event $(W^0 \le h)$, there are some zeros that break any sequence, therefore you might as well assume you start with a zero.

$$P_{(X_1=1)}(A/W^0 \le h) = P(A/X_1=0)$$

So we therefore obtain, using these terms to rewrite the above equation:

$P_{(X_1=1)}(A)$

$$P(A/X_1=1) = P(A/X_1=0)(1-p^{h-1}) + 1 \cdot p^{h-1}$$

$$(*) \quad \boxed{P(A/X_1=1) = P(A/X_1=0)(1-p^{h-1}) + p^{h-1}}$$

This is a linear equation linking the conditional probability of the event given $(X_1=1)$ with the conditional probability of the event, given $(X_1=0)$. Notice we get both conditional probabilities.
We need another linear equation linking $P(A/X_1=0)$ to $P(A/X_1=1)$.
Let's do that.

Let W' = waiting time for <u>first one</u>

Consider the partition:

$$\pi = \{(W' > t), (W' \le t)\}$$

These two events form a partition
They are disjoint.
And their union is the whole sample space.

We therefore condition, by the Law of Alternatives:

$$P_{(X_1=0)}(A) = \sum_{B \in \pi} P_{(X_1=0)}(A|B)\, P_{(X_1=0)}(B)$$

$$= P_{(X_1=0)}(A|W'>t)P_{(X_1=0)}(W'>t) + P_{(X_1=0)}(A|W'\le t)P_{(X_1=0)}(W'\le t)$$

Let's simplify this.
Consider the first term:

$$P_{(X_1=0)}(A|W'>t)$$

The waiting time for the first one is greater than t.
This means that <u>at least</u> the first t tosses are zeros.
Therefore you've lost.
You've lost completely, if this happens.
Event A (event that first run of h ones precedes first run of t zeros) can <u>never</u> happen.
Since, if $(W'>t)$, A can not happen

$$P_{(X_1=0)}(A|W'>t) = 0$$

Now let's do the second term:

$$P_{(X_1=0)}(A|W'\le t)$$

$(W'\le t)$ means that the first 1 occurs <u>before</u> t zeros occur.

$$\overbrace{\cancel{0 \ldots 0}}^{<t}\,|\,1 \ldots$$

therefore, it's as if you were <u>starting with a one</u>

$$P_{(X_1=0)}(A|W'\le t) = P(A|X_1=1)$$

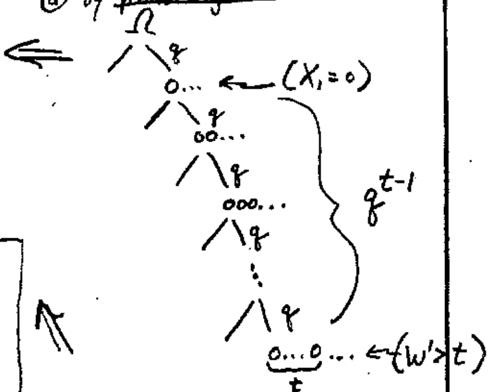
Finally:

$$P_{(X_1=0)}(W' \le t) = 1 - P_{(X_1=0)}(W' > t)$$

$$= 1 - q^{t-1}$$

$P_{(X_1=0)}(W' > t)$

ⓐ by probability tree:



ⓑ by Law of Alternatives:

$$P_{(X_1=0)}(W'>t) = \underbrace{P_{(X_1=0)}\left(W'>t \,\middle|\, \begin{smallmatrix}\text{tosses } 2\ldots t\\ \text{all } 0\end{smallmatrix}\right)}_{1 \text{ - since first 1 after } t \text{ tosses. } \underbrace{0\ldots0}_{t}\ldots} \underbrace{P_{(X_1=0)}\left(\begin{smallmatrix}\text{tosses } 2\ldots\\ \text{all } 0\end{smallmatrix}\right)}_{q^{t-1}}$$

$$+ \underbrace{P_{(X_1=0)}\left(W'>t \,\middle|\, \left(\begin{smallmatrix}\text{tosses } 2\ldots t\\ \text{all } 0\end{smallmatrix}\right)^c\right)}_{0 \text{ - since at least one of tosses } 2\ldots t \text{ is } 1} P_{(X_1=0)}\left(\left(\begin{smallmatrix}\text{tosses } 2\\ \text{all } 0\end{smallmatrix}\right)\ldots\right)$$

We therefore obtain, using these terms to rewrite the above equation:

$$P_{(X_1=0)}(A) = 0 + P(A|X_1=1)\,(1-q^{t-1})$$

(**)
$$\boxed{P(A|X_1=0) = P(A|X_1=1)\,(1-q^{t-1})}$$

Solve (*) and (**) for $P(A|X_1=0)$ and $P(A|X_1=1)$

Then substitute into (***) to obtain the desired $P(A)$.

$$P(A) = \frac{p^{h-1}(1-q^t)}{p^{h-1} + q^{t-1} - p^{h-1}q^{t-1}}$$

I confess that I do not see this result intuitively.
It's not exactly intuitively clear.

<u>One remark</u>

This reasoning is exactly the same as the reasoning we went through before [3/6/98.4-10].
These are 2 ways of saying the same thing.
I see it more clearly by drawing trees.
I would never have guessed this conditioning if I hadn't drawn the tree.

You have 3 problems in the problem set that are hard, where you really have to draw the tree.
I suggest you do the last problems in the problem set first, because they are very easy.

BTW - You might compare this treatment of the problem w/ that in the book.
Notice that the book has a mistake.
Although the result works out.

Example – this typical Bayes' Law

This is the Problem of Scientific Induction

You have:

Population of n members, n = known; a = good members, a = unknown

We know that the population has n members, but we don't know how many good guys there are.

We are looking at a sample of size k w/o replacement.

Each sample guy is killed.

On the basis of this sample, we have the random variable A:
Let random variable A = number of a-guys in the sample

The Problem of Scientific Induction, or Bayes' Law, is this.
On the basis of the number of a-guys in the sample of size k, what's the most probable inference of the number of a-guys in the population?

This is a problem of universal occurrence.
I state it in these Mickey Mouse terms.
But you can state it in terms that are very highfalutin.
Life and death situations where people will pay millions of dollars.
I leave it to you to be applied mathematicians, which is to recognize when you walk into a problem that is just this.
I can't teach you that.

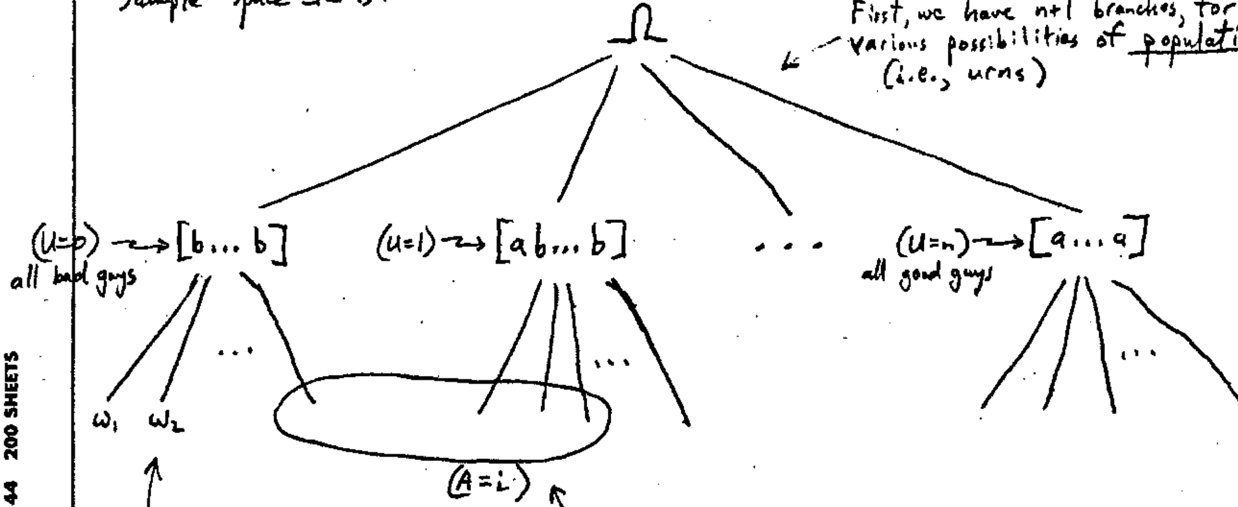
Let's set it up.
We want to set up a sample space where all possible distributions of numbers of a-guys (good guys) and bad guys exist.
And then we sample one of the population

We set up $n+1$ urns.
The urn $(U=j)$ contains $\begin{cases} j & \text{a-guys (good guys)} \\ n-j & \text{bad guys} \end{cases}$

We set up the experiment as follows:
We first pick one of the urns.
Then from that urn, we sample k members.
Then we look for a-guys.

That's our sample space.

Sample space Ω is:



First, we have n+1 branches, for various possibilities of <u>populations</u> (i.e., urns)

Then, each urn has various sample points ω_i

The event $(A=i)$ is the event that in our sample of size k, there are i a-guys (good guys).
The event $(A=i)$ includes the appropriate sample points.

What's missing?
In order for this to be a probability space, we have to assign <u>probabilities</u>.

The <u>sample points</u> are the <u>branches of the tree.</u> {NB: <u>branches</u>, not just the leaves}
You have to assign probabilities in such a way that the probabilities added up over all the branches add up to 1.

Let's take <u>urn j.</u>
Without knowing the sample space, we can compute compute the conditional probability:

$$\underbrace{P(A=i \mid U=j)}_{\text{probability of event } (A=i)\text{, given that you've chosen the population w/ } j \text{ a-guys (good guys).}} = \frac{\binom{j}{i}\binom{n-j}{k-i}}{\binom{n}{k}}$$

← {From the j good guys, pick i; From the $n-j$ bad guys, pick $k-i$}

← because you have a sample of size k

We've done this before [2/23/98.8–11]

That is, the <u>hypergeometric distribution</u>

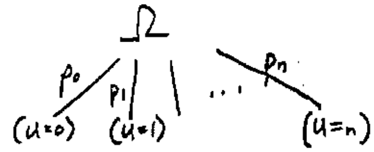
<u>Remember</u> – there was something very funny happening here (The population is the urns. The sample is the balls).
We obtained the <u>same result</u> whether the balls are distinguishable or indistinguishable.
This is an important fact–physically.
Namely, if you <u>sample w/o replacement</u>, it doesn't matter whether the balls are distinguishable or not.

If we know the probability of choosing urn j $P(U=j)$, then we could determine the probability distribution of A on this entire sample space, immediately, from the Law of Alternatives.

$$P(A=i) = \sum_{i=0}^{n} P(A=i \mid U=j)\, \underbrace{P(U=j)}$$

Unfortunately, we have no information as to which distribution of populations $P(U=j)$ is more popular than which.

So we have $p_j = P(U=j)$:



where $p_0 + p_1 + \cdots + p_n = 1$

Q: If you have absolutely no way of knowing which urn is more probable than another, why can't you just say the probability of each of the $n+1$ urns is just $\frac{1}{n+1}$?

$$p_i = \frac{1}{n+1}$$

A: You're spoiling my act.
Someone may have slipped us some information:
You can't have more than 15 good guys
-or-
Good guys come in multiples of three.

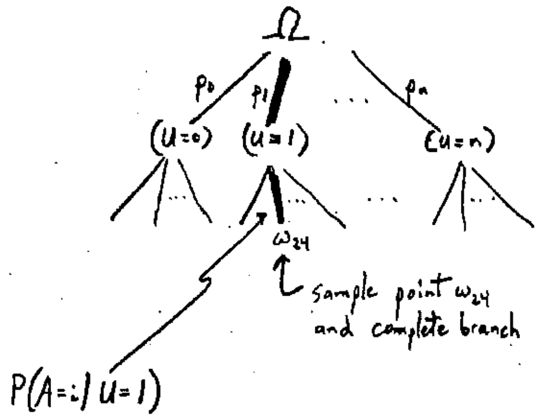
But in the absolutely most general case, this is what we get.

Therefore, the probability distribution of the random variable A is the following:

$$P(A=i) = \sum_{j=0}^{n} P(A=i \mid U=j)\, P(U=j) \longleftarrow$$ Very standard application of the <u>Law of Alternatives</u>

$$= \sum_{j=0}^{n} \frac{\binom{j}{i}\binom{n-j}{k-i}}{\binom{n}{k}}\, p_j$$

Now the sample space is completely defined.
The <u>sample point</u> is the <u>complete branch</u>.



The edges, as we've said, are the conditional probabilities.
Since this edge ends at ω_{24}, we have:

$$P(\omega_{24}) = \underbrace{P(U=1)}_{p_i} P(A=i \mid U=1)$$

So now we can compute <u>everything</u> about this sample space.
For example, we <u>can compute</u>:

$$P(U=j \mid A=i)$$

This is the process of <u>induction</u>.
If you draw a sample of k guys and i of them are good (i.e., the event $(A=i)$), which $(U=j)$ is the most probable distribution?

$$P(U=j \mid A=i) = \frac{P((U=j)\cap(A=i))}{P(A=i)}$$ definition of conditional probability

$$(*) \qquad P(U=j \mid A=i) = \frac{P(A=i \mid U=j)\,\overbrace{P(U=j)}^{p_j}}{P(A=i)}$$ rewrite numerator

Now some <u>jargon</u>. In expression (*):

$p_j = P(U=j)$ is called the <u>prior</u>

$\dfrac{P(A=i \mid U=j)}{P(A=i)}$ is called the <u>likelihood</u>

$P(U=j \mid A=i)$ is called the <u>posterior</u>

These are weird terms, but everybody uses them. They were invented by business people, so they are really weird.

Begin Digression

Some philosophical remarks.
This is something I should have told you before, but I missed it.

Point 1

Expectation of a random variable:

If X is an integer random variable,
then
$$E(X) = \underbrace{\sum_n}_{\text{sum over all integers } n} n\, P(X=n)$$

Then, on the basis of this definition, we verified [2/25/98.8-11]:
$$E(X+Y) = E(X) + E(Y)$$
by an argument that is not at all trivial.

You should have asked, at this point, where does this come from?
Let's digress for a minute and motivate the definition.

Suppose sample space Ω is finite.
And X = random variable.
I want to define the average value of that random variable.
What would be the most natural definition of average value for that random variable?
The most natural definition of average would be:

Intuitive Point of View

$$E(X) = \underbrace{\sum_{\omega\in\Omega}}_{\text{sum over all sample points}} X(\omega)\, P(\omega)$$

times probability of ω

value of the random variable X at the sample point ω.

This is as natural a definition of the average of a random variable as one can have.
Intuitive.
Add up the values, each one weighted by its probability.
In fact, if you use this definition, then the main formula $[E(X+Y) = E(X)+E(Y)]$ is obvious. Completely obvious.
Because you have:
$$E(X+Y) = \sum_{\omega\in\Omega} \big(X(\omega)+Y(\omega)\big)P(\omega)$$
which is obviously:
$$= \sum_{\omega\in\Omega} X(\omega)P(\omega) + \sum_{\omega\in\Omega} Y(\omega)P(\omega)$$
$$= E(X) + E(Y)$$

From this point of view, the linearity of expectation is something completely trivial.

Of course, we assume that our sample space Ω is finite.

Now let's see that the intuitive definition reduces to the definition we gave before:

$$E(X)=\sum_{\omega\in\Omega} X(\omega)P(\omega) \qquad E(X)=\sum_n n\,P(X=n)$$

$$\underbrace{\sum_{\omega\in\Omega}}\, X(\omega)P(\omega) = \sum_n \sum_{\{\omega : X(\omega)=n\}} X(\omega)P(\omega)$$

split this sum into sums, one for each value of $X(\omega)$

First, I sum over all ω for which $X(\omega)=n$.
Then, I sum over all n.
That gives me the sum over all Ω.

the whole sample space partitions according to the value of the random variable.

Now we have the Principle of Who is Buried in Grant's Tomb.
What is:

$$\sum_{\{\omega : X(\omega)=n\}} X(\omega)P(\omega) \quad ?$$

Since $X(\omega)=n$ then, then you have n here.
If you don't see that, I leave.

$$= \sum_n \sum_{\{\omega : X(\omega)=n\}} n\,P(\omega)$$

$$= \sum_n n \underbrace{\sum_{\{\omega : X(\omega)=n\}} P(\omega)}$$

$$= \sum_n n\, \overbrace{P(X=n)}$$

{And this is precisely the other formula.}

This is another of "Who's buried in Grant's Tomb" argument.
$\sum_{\{\omega : X(\omega)=n\}} P(\omega)$, sum of $P(\omega)$ where $X(\omega)=n$
is the probability of the event that $(X=n)$, namely:
$P(X=n)$. Again, if you don't see that, I quit.

$$\sum_{\omega\in\Omega} X(\omega)P(\omega) = \sum_n n\,P(X=n)$$

intuitive definition — unintuitive definition

Therefore, for a finite sample space, the intuitive definition is an equal to the unintuitive one.

Unfortunately, most sample spaces of this world are not finite.
And, therefore, the intuitive definition can not be used.
However, the unintuitive definition still works.

And so, in the knowledge that the intuitive and the unintuitive are equivalent in the simple case of the finite sample space, we introduce the unintuitive definition and we pawn that off as the expectation value.

That's the motivation that I didn't give you before.

This is one of the greatest arguments of mathematics – don't kid yourselves.
It took hundreds of years for this to be discovered.
Now we work it out.
It took a long time for people to realize this very simple fact.

So that's the computation of expectation.
Intuitively, when we are given :

$$\sum_{\omega \in \Omega} X(\omega)\, P(\omega)$$

we think of the sum of infintesimal probabilities over $X(\omega) P(\omega)$.
That can be done by integrating the Lebesque integral.
But we don't want to do that now.
Since we don't want to do the fancy integration, we use the unintuitive definition :

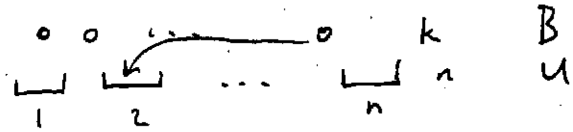
$$= \sum_n n\, P(X=n)$$

That's the first point.

Point 2

Another philosophical point:

Two descriptions of the classical statistics.
The classical statistics, which are the statistics of <u>k balls</u> into <u>n boxes</u>.
The sample points are <u>functions</u> from a set B to a set U
(or something akin to functions, like equivalence classes, when the balls are indistinguishable.)



We said that there are two points of view, whereby you can visualize the stochastic processes.

1) <u>Distribution</u> point of view
where you imagine placing the balls into the boxes.
Mostly, you consider the occupation number

B = balls
U = boxes

the other is: 2) <u>Occupancy</u> point of view
U = population
B = sample of that population

The formulas are the <u>same</u>, but sometimes the logic becomes clear if you look at it one way, or the other.

Let's make this distinction more rigorous.

<u>Distribution:</u>

We define the random variable capital theta Θ:

$$\Theta_j = \text{number of balls in the first } j \text{ boxes}$$
(we assume the boxes are numbered)
$$= \underbrace{\theta_1 + \theta_2 + \cdots + \theta_j}_{\text{sum of occupation numbers for boxes } 1 \text{ to } j}$$

The sequence of random variables $\Theta_1, \Theta_2, \ldots, \Theta_n$ completely determines any phenomenon pertaining to the occupation numbers of the boxes.

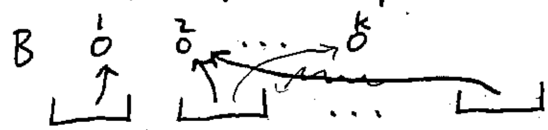
For example, if you want the occupation number θ_j, you take:

$$\theta_j = \Theta_j - \Theta_{j-1}$$

I am introducing this term to motivate the next discussion.

Occupancy: This point of view is more sophisticated.
Lets do first Ω_{MB}

From an occupancy point of view, that's the situation where we: [2/13/98.2]
sample, put it in position 1,
sample, put it in position 2,
⋮
sample, put it in position k.



random variable X_1 = first sample

For the occupancy point of view, the stochastic process is completely described by giving k independent random variables $X_1, X_2, \dots, X_k$ that are independent and identically distributed (i.i.d.) and with the probability that:

$$P(X_1 = j) = \frac{1}{n}$$ } Because the probability of a ball coming from any box is $\frac{1}{n}$.

The Maxwell-Boltzmann process can be completely described by saying I have a sequence of k i.i.d. random variables with this distribution.

The occupancy description carries more information than the distribution description. Because, for example, in the distribution description, you only count how many balls are in each box. You can not say whether ball 3 has gone into box 5. Θ_j doesn't provide this information.

Whereas the occupancy description provides the entire information.
For example, $(X_3 = 5)$ is the event that the 3^{rd} sample was picked from the 5^{th} box.

Let's do the occupancy description for the other stochastic processes.
Namely, for Fermi-Dirac and Bose-Einstein.
As you see, this is purely philosophical.

Let's take Fermi-Dirac [2/17/98.4]
We'd like to give an occupancy description in terms of a sequence of random variables.
But that's impossible, because the balls are indistinguishable.

What do we do?
We take another sample space, which is Maxwell-Boltzmann conditioned on the boxes having 0 or 1 ball.

$$\Omega'_{FD} = \Omega_{MB} \text{ conditioned on } \Theta_j = \begin{cases} 0 \\ 1 \end{cases} \text{ for all } j$$

Then divide by $k!$ and you get Fermi-Dirac.

Q: Why not use $\Omega_{BE} \leftarrow$ Bose-Einstein?
A: Because I want the balls to be distinguishable.
In Bose-Einstein, the balls are indistinguishable, so I can not talk about X_1 = first sample, X_2 = second sample, etc.

In both Fermi-Dirac and Bose-Einstein, what you are doing is:
distribution description: placing marks on boxes

— ✓ ✓ ... ✓

-or-

occupancy description: taking a sample and forgetting in which order you took it in.
This is why we divide by $k!$

If you want the order of the sample – every hour on the hour, we sample one, then they are distinguishable because we can tell which one has been sampled which hour.

So, we are forced to use Maxwell-Boltzmann. We condition it this way and divide by $k!$ to get Fermi-Dirac.

$X_1, X_2, \ldots, X_k$ = successive balls extracted (the samples)

These are no longer independent

$$P(X_1 = j) = \frac{1}{n}$$

What is $P(X_2 = j)$?
Let's work it out.

$$P(X_2=j) = \sum_{B\in\pi} P(X_2=j \mid B)\, P(B)$$

Law of Alternatives.
For the blocks B, choose the disjoint $(X_1=i)$, $i=\{1,\ldots,n\}$

$$= \sum_{i=1}^{n} \underbrace{P(X_2=j \mid X_1=i)}_{\substack{\text{given that the } 1^{st} \text{ sample chosen is } i\\ (X_1=i)\text{, the } 2^{nd} \text{ sample can not be}\\ i\text{, since this is Fermi-Dirac (Pauli}\\ \text{exclusion). So instead of } n\text{, there are } n-1\\ \text{possible samples for the } 2^{nd} \text{ choice. Thus:}}} \underbrace{P(X_1=i)}_{\frac{1}{n}}$$

$$P(X_2=j \mid X_1=i) = \frac{1}{n-1}$$

Finally, note that $P(X_2=j \mid X_1=j)=0$, as the 2^{nd} choice can <u>not</u> choose the same as the 1^{st} choice.

$$= \sum_{\substack{i=1\\ i\neq j}}^{n} P(X_2=j \mid X_1=i)\, P(X_1=i)$$

$$= (n-1)\left(\frac{1}{n-1}\right)\left(\frac{1}{n}\right)$$

$$P(X_2=j) = \frac{1}{n}$$

$X_1, X_2, \ldots, X_K$ are <u>identically distributed</u> (i.d.)

What about the <u>joint distribution</u>?

$$P((X_1=i)\cap(X_2=j)) = \underbrace{P(X_2=j \mid X_1=i)}_{\frac{1}{n-1}}\ \underbrace{P(X_1=i)}_{\frac{1}{n}}$$

as discussed above.

$$P((X_1=i)\cap(X_2=j)) = \left(\frac{1}{n-1}\right)\left(\frac{1}{n}\right) \neq \underbrace{P(X_1=i)P(X_2=j)}_{\left(\frac{1}{n}\right)\left(\frac{1}{n}\right)}$$

$$P((X_1=i)\cap(X_2=j)) \neq P(X_1=i)P(X_2=j)$$

Therefore X_1 and X_2 are <u>not independent</u>

$X_1, X_2, \ldots, X_K$ are <u>not independent</u>

So for this Fermi-Dirac case, you have these identically distributed random variables, which are <u>not independent</u>.

This sequence of random variables, together w/ their <u>joint distributions</u>, completely determines the process.

In fact, instead of the joint distribution, you might think of describing the process by the conditional probability.

{For Bose-Einstein, the occupancy description is a similar situation, I leave it to you as an exercise.}

[Remember for Disposition (Flags), we discussed a distribution interpretation, but we did not discuss an occupancy interpretation. That's very easy to work out. I leave it to you.]

~~3rd + Last Point~~

Bayes' Law

A = event

π = partition of sample space Ω

The Law of Alternatives gives:

$$P(A) = \sum_{B \in \pi} P(A|B)\, P(B)$$

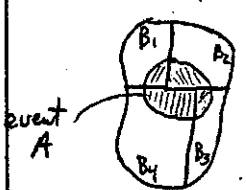
Bayes' Law is a jazzing up of the above.

Bayes' Law:

$$P(B|A) = \frac{P(A|B)\, P(B)}{P(A)} = \frac{P(A|B)\, P(B)}{\sum_{B\in\pi} P(A|B) P(B)}$$

We are beginning to see some applications of this and we're coming to this in a minute.

This is interpreted as (let me do this in terms of sets):



You have the partition π with blocks B_i.
Then you have the event A.

Then you want to estimate the probability of each block B_i of the partition on the basis of the intersection of A with that block.

In order to estimate $P(B|A)$, you need to know conditional probabilities $P(A|B)$.

$$\underbrace{P(B|A)}_{\text{posterior}} = \frac{P(A|B)\, \overbrace{P(B)}^{\text{prior}}}{\sum_{B\in\pi} P(A|B) P(B)}$$

likelihood

It is impossible to understand this w/o examples.

End: Digression

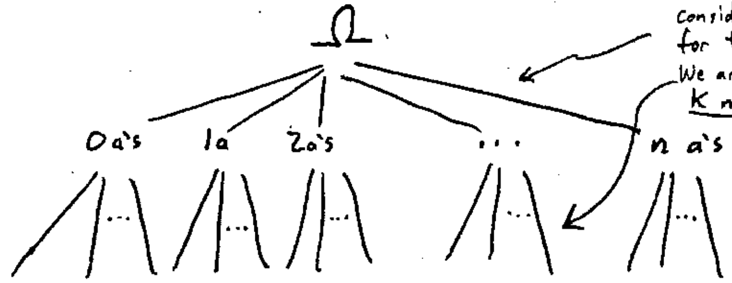
Example 1

Given a population of n members, a of which are good. a = unknown
We are allowed to sample, w/o replacement, K numbers
And we find A = number of good members in the sample.
clearly A is a random variable

What is the best estimate of the $\overbrace{\text{number of good guys in the population}}^{a}$,
given the $\underbrace{\text{number of good guys in a sample of size } k}_{A}$?

That's the base problem,
We want the solution to the problem: Best guess of a based on random variable A.

Set up a sample space Ω:



consider the various number of possibilities for the number of a's in the population.
We are sampling, from one of these possibilities, K members. But we don't know from which possibility

How do we set up the sample space?
We need to add probabilities for each branch.

Let U = number of a's in the population
another random variable

$$P(A=i/U=j) = \frac{\binom{j}{i}\binom{n-j}{k-i}}{\binom{n}{k}}$$

We've computed this a few times.
For example, see [3/9/98.9].
From j, choose i good guys.
From remaining $n-j$, choose $k-i$ bad guys.

the hypergeometric distribution

Watch it, because I am preparing a trap!

This seems like something elementary, but you should be on guard.

From the Law of Alternatives:

$$P(A=i) = \sum_{j=0}^{n} P(A=i \mid U=j)\, P(U=j)$$

By Bayes' Law:

$$P(U=j \mid A=i) = \frac{P(A=i \mid U=j)\, P(U=j)}{\sum_{j=0}^{n} P(A=i \mid U=j)\, P(U=j)}$$

These are different j. Strictly speaking, this is correct, as in denominator, j refers to summation variable. Although it is slightly deceiving.

(*) $$P(U=j \mid A=i) = \frac{\binom{j}{i}\binom{n-j}{k-i} P(U=j)}{\sum_{j=0}^{n} \binom{j}{i}\binom{n-j}{k-i} P(U=j)}$$

{the $\binom{n}{k}$ in the denominator of the hypergeometric distributions cancel out.}

We can not determine this until we have the probability $P(U=j)$.
We have to chose a prior probability, namely:

$$P(U=j)$$

It comes naturally to say that if we know absolutely nothing, then we assume that all the possibilities have the same probability.

But we will see that that is not the most natural thing to assume.

Case 1 - Uniform prior (each of $\overbrace{0a\text{'s}, 1a, 2a\text{'s}, \ldots, na\text{'s}}^{n+1}$ is equally probable)

$$P(U=j) = \frac{1}{n+1}, \text{ for all } j$$

Then (*) simplifies as follows:

$$P(U=j \mid A=i) = \frac{\binom{j}{i}\binom{n-j}{k-i} \cancel{\frac{1}{n+1}}}{\sum_{j=0}^{n} \binom{j}{i}\binom{n-j}{k-i} \cancel{\frac{1}{n+1}}}$$

$$= \frac{\binom{j}{i}\binom{n-j}{k-i}}{\sum_{j=0}^{n} \binom{j}{i}\binom{n-j}{k-i}}$$

We do not have a good answer, unless we are able to give a closed form for the denominator.

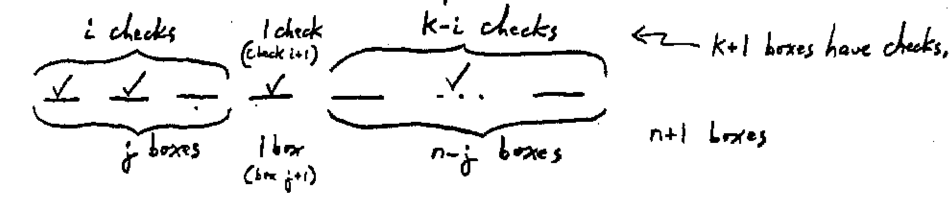
Simplify the denominator as follows:
We want a closed form for:

$$\sum_{j=0}^{n} \binom{j}{i}\binom{n-j}{k-i}$$

It's not that easy.
You can do it directly.
But we'll do the following side argument:

In Fermi-Dirac statistics, take sample space Ω_{FD} with $k+1$ balls, $n+1$ boxes.

Let random variable C_{i+1} = the position of the $i+1^{st}$ box checked.



$$P(C_{i+1}=j+1) = \frac{\binom{j}{i}\,1\,\binom{n-j}{k-i}}{\binom{n+1}{k+1}}$$

There are i checks in the first j boxes.

total number of possibilities.

$(C_{i+1}=j+1)$ means the $j+1^{st}$ box has a check, by definition.

The remaining $k-i$ checks go into the following $n-j$ boxes.

$$P(C_{i+1}=j+1) = \frac{\binom{j}{i}\binom{n-j}{k-i}}{\binom{n+1}{k+1}}$$

Since $P(C_{i+1} = j+1)$ is a probability distribution, its sum over all j must equal 1.

$$\sum_{j=0}^{n} \left(C_{i+1} = j+1\right) = 1 \quad \Big\} \text{ by definition of probability distribution}$$

$$\sum_{j=0}^{n} \frac{\binom{j}{i}\binom{n-j}{k-i}}{\binom{n+1}{k+1}} = 1$$

Which gives the identity:

$$\boxed{\sum_{j=0}^{n} \binom{j}{i}\binom{n-j}{k-i} = \binom{n+1}{k+1}}$$

This completes the side argument.

Now we can reduce $P(U=j \mid A=i)$ to:

$$P(U=j \mid A=i) = \frac{\binom{j}{i}\binom{n-j}{k-i}}{\sum_{j=0}^{n}\binom{j}{i}\binom{n-j}{k-i}}$$

substitute the above identity.

$$\boxed{P(U=j \mid A=i) = \frac{\binom{j}{i}\binom{n-j}{k-i}}{\binom{n+1}{k+1}}}$$

(tape unavailable)

Bayesian Theory (cont'd) – conjugate priors

This is slightly digressive.
Last time, we discussed an interesting binomial identity, namely:

$$\sum_j \underbrace{\binom{j}{i}\binom{n-j}{k-i}}_{} = \binom{n+1}{k+1}$$

Note that i appears only on the LHS, but neither as the sum index nor on the RHS.

Thus there are an infinite number of such binomial identities, one for each i.

Let's consider some special cases of this binomial identity:

Case $i=0$

We want to verify that the binomial identity holds for $i=0$:

$$\sum_j \binom{j}{i}\underbrace{\binom{n-j}{k-i}}_{=\binom{n-j}{k}} \stackrel{?}{=} \binom{n+1}{k+1}$$

$$\binom{j}{0} = \frac{j!}{0!(j-0)!} = 1$$ = only one way to choose the empty set from j elements

$$\sum_j \binom{n-j}{k} \stackrel{?}{=} \binom{n+1}{k+1}$$

$$\underbrace{\binom{n}{k} + \binom{n-1}{k} + \cdots}_{} \stackrel{?}{=} \binom{n+1}{k+1}$$

number of ways to choose k subsets out of n
+ " " " " " " " " $n-1$
+ ⋮

This is an interesting binomial identity.
Let's verify this binomial identity with a probabilistic model.

<u>Model</u> - Fermi-Dirac with k+1 checks, n+1 boxes

k+1 checks, at most one check per box

$$\underset{1}{\_\_}\quad \underset{2}{\_\_}\quad \underset{\substack{3\\ \text{positions}}}{\checkmark}\quad \underset{4}{\checkmark}\quad \cdots\quad \underset{n+1}{\checkmark}$$

boxes are numbered by position

Let random variable C_1 = position of 1^{st} check

Ex: In the above, $C_1 = 3$

Let's consider the probability distribution of the event $(C_1 = j+1)$, i.e., that the first check occurs at position $j+1$:

$$\overbrace{\underbrace{\_\_\quad \_\_}_{j\text{ positions}}}^{0\text{ checks}}\quad \overbrace{\underbrace{\underset{j+1}{\checkmark}}_{1\text{ position}}}^{1\text{ check}}\quad \overbrace{\underbrace{\checkmark\quad \cdots\quad \underset{n+1}{\checkmark}}_{n-j\text{ positions}}}^{k\text{ checks}}$$

$\leftarrow$ k+1 checks

$\leftarrow$ n+1 boxes

$$P(C_1 = j+1) = \frac{\cancel{\binom{j}{0}}^{1}\, 1 \binom{n-j}{k}}{\binom{n+1}{k+1}}$$

$$P(C_1 = j+1) = \frac{\binom{n-j}{k}}{\binom{n+1}{k+1}}$$

Since C_1 is a random variable, the sum over all j of its probability distribution must equal 1. We have:

$$\sum_j P(C_1 = j+1) = 1 \quad \Big\}\ \text{by definition}$$

$$\sum_j \frac{\binom{n-j}{k}}{\binom{n+1}{k+1}} = 1$$

$$\sum_j \binom{n-j}{k} = \binom{n+1}{k+1}$$

$$\binom{n}{k} + \binom{n-1}{k} + \cdots = \binom{n+1}{k+1}$$

which verifies the case for $i=0$, as desired.

Case $i=1$

We want to show that the binomial identity holds for $i=1$:

$$\sum_j \binom{j}{i}\binom{n-j}{k-i} \stackrel{?}{=} \binom{n+1}{k+1}$$

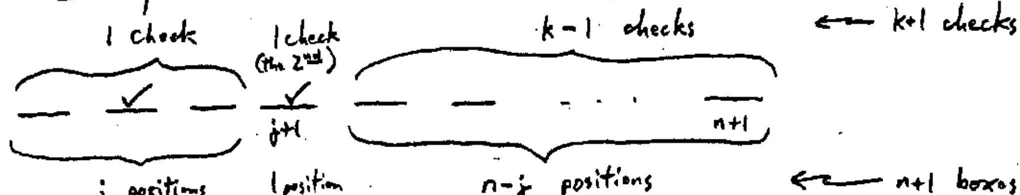
$$\binom{j}{1} = \frac{j!}{1!(j-1)!} = j = \text{number of ways to pick one element from } j$$

$$\sum_j j\binom{n-j}{k-1} \stackrel{?}{=} \binom{n+1}{k+1}$$

$$\binom{n-1}{k-1} + 2\binom{n-2}{k-1} + 3\binom{n-3}{k-1} + \cdots \stackrel{?}{=} \binom{n+1}{k+1}$$

Another interesting binomial identity.
We use the same Fermi-Dirac probabilistic model as before ($k+1$ checks, $n+1$ boxes), but now we consider the random variable:

$$C_2 = \text{position of the } 2^{\text{nd}} \text{ check}$$



The probability distribution for C_2 is:

$$P(C_2 = j+1) = \frac{\cancel{\binom{j}{1}}^{j}\,1\,\binom{n-j}{k-1}}{\binom{n+1}{k+1}}$$

$$P(C_2 = j+1) = \frac{j\binom{n-j}{k-1}}{\binom{n+1}{k+1}}$$

And since the sum over all possible values of a probability distribution equals one, we have:

$$\sum_j P(C_2 = j+1) = 1$$

$$\sum_j \frac{j\binom{n-j}{k-1}}{\binom{n+1}{k+1}} = 1$$

$$\sum_j j\binom{n-j}{k-1} = \binom{n+1}{k+1}$$

$$\binom{n-1}{k-1} + 2\binom{n-2}{k-1} + 3\binom{n-3}{k-1} + \dots = \binom{n+1}{k+1}$$

which verifies the case for $i=1$, as desired.

- You can get additional interesting binomial identity proofs, in exactly the same way, for different i.

- Almost every binomial identity has a probabilistic model proof. Experience, for example taking 18.313, will teach you how to prove these.

- Note: instead of binomial identities, we can get multiset identities in the same fashion, but instead of Fermi-Dirac statistics, we use Bose-Einstein statistics for the probabilistic model proof.

identity	probabilistic model proof
binomial	Fermi-Dirac statistics
multiset	Bose-Einstein statistics

(i.e., w/ feeling. We kill the sampled members.) 3/13/98.5

- Bayes' Theorem for Sampling w/o Replacement

population : n members (not random)
sample : k

Let's define the random variables :

U = number of good guys (in the population)
A = number of good guys in the sample

Bayes' Law, compliments of Reverend Sir Thomas Bayes, tells us:

$$\underbrace{P(U=j|A=i)}_{\uparrow} = \frac{P(A=i|U=j)\,P(U=j)}{P(A=i)}$$

The meaning of this is that we perform experiment A. Given that we observe $(A=i)$, $P(U=j|A=i)$ gives the probability that the population has j good guys.
We are estimating the number of good guys in the population, based on the observed number of good guys in the sample.

$$= \frac{P(A=i|U=j)\,P(U=j)}{\sum_j P(A=i|U=j)\,P(U=j)}$$

rewrite $P(A_i)$ by Law of Alternatives

As we discussed last time [3/11/98.10], $P(A=i|U=j)$ is the hypergeometric distribution :

$$P(A=i|U=j) = \frac{\binom{j}{i}\binom{n-j}{k-i}}{\binom{n}{k}}$$

(*) $$\boxed{P(U=j|A=i) = \frac{\binom{j}{i}\binom{n-j}{k-i}P(U=j)}{\sum_j \binom{j}{i}\binom{n-j}{k-i}P(U=j)\}}}$$

prior probability

denominator

Typically, we are provided with the conditional probabilities $P(A=i|U=j)$ and $P(A=i)$.
That is, we can readily compute the likelihood probability.

In the absence of any further information, what do we assign to the prior?

- Case 1 – Uniform Prior

This summarizes what we discussed last time [3/11/98.10-13].

Assume $P(U=j) = \frac{1}{n+1}$, for all j

Then equation (*) simplifies to:

$$P(U=j|A=i) = \frac{\binom{j}{i}\binom{n-j}{k-i}\cancel{\frac{1}{n+1}}}{\underbrace{\sum_j \binom{j}{i}\binom{n-j}{k-i}}\cancel{\frac{1}{n+1}}}$$

$$P(U=j/A=i) = \frac{\binom{j}{i}\binom{n-j}{k-i}}{\binom{n+1}{k+1}}$$

from binomial identity proved probabilistically [3/11/98.12-13]

Now we ask "Is the assumption of a uniform prior sensible?"
Is there another prior that should be considered?

Surprisingly, the answer is yes.
The conjugate prior

- Case 2 – Conjugate Prior

members of the population are Bernoulli random variables ← (each equal to 0 or 1)

$$\underbrace{\underline{X_1}\quad \underline{X_2}\quad \underline{X_3}\quad \cdots \quad \underline{X_n}}_{n \text{ members}} \quad \text{population}$$

As we recall from [2/20/98.8]:

$$P(X_i=1) = p$$
$$P(X_i=0) = q = 1-p$$

random variable U = number of $\overbrace{\text{good guys}}^{1\text{'s}}$ in the sample

$$= X_1 + X_2 + \cdots + X_n$$

We've shown the probability distribution of U is the binomial distribution:

$$\underbrace{P(U=j)}_{\text{this is the conjugate prior}} = \binom{n}{j} p^j q^{n-j}$$

With this prior $P(U=j)$, equation $(*)$ gives the posterior probability as:

$$P(U=j \mid A=i) = \frac{\dfrac{\binom{j}{i}\binom{n-j}{k-i}}{\binom{n}{k}} P(U=j)}{\sum_j \dfrac{\binom{j}{i}\binom{n-j}{k-i}}{\binom{n}{k}} P(U=j)}$$

$$= \frac{\dfrac{\binom{j}{i}\binom{n-j}{k-i}}{\binom{n}{k}} \binom{n}{j} p^j q^{n-j}}{\sum_j \dfrac{\binom{j}{i}\binom{n-j}{k-i}}{\binom{n}{k}} \binom{n}{j} p^j q^{n-j}} \quad \Big\} \text{ A mess!}$$

the denominator works out as follows.
Trust me, I worked it out.
You, too, are welcome to.

$$\binom{k}{i} p^i q^{k-i}$$

$$= \frac{\dfrac{\binom{j}{i}\binom{n-j}{k-i}}{\binom{n}{k}} \binom{n}{j} p^j q^{n-j}}{\binom{k}{i} p^i q^{k-i}}$$

$$P(U=j \mid A=i) = \binom{n-k}{j-i} p^{j-i} q^{n-k-j+i}$$

Alternatively

But I didn't have to calculate this mess.
I can reason as follows:

- Since all members of the population are random, choosing the first k random members as the sample is the same as choosing randomly k members.

So, for a sample of size k, we might as well choose $X_1, X_2, \ldots, X_k$.

$$\underline{X_1}\ \underline{X_2}\ \underline{X_3} \ldots \underline{X_k} \ldots \underline{X_n} \;=\; \underline{X_1}\ \underline{X_2}\ \underline{X}\ \underline{X}\ \underline{X}\ \underline{X} \ldots \underline{X_n}$$

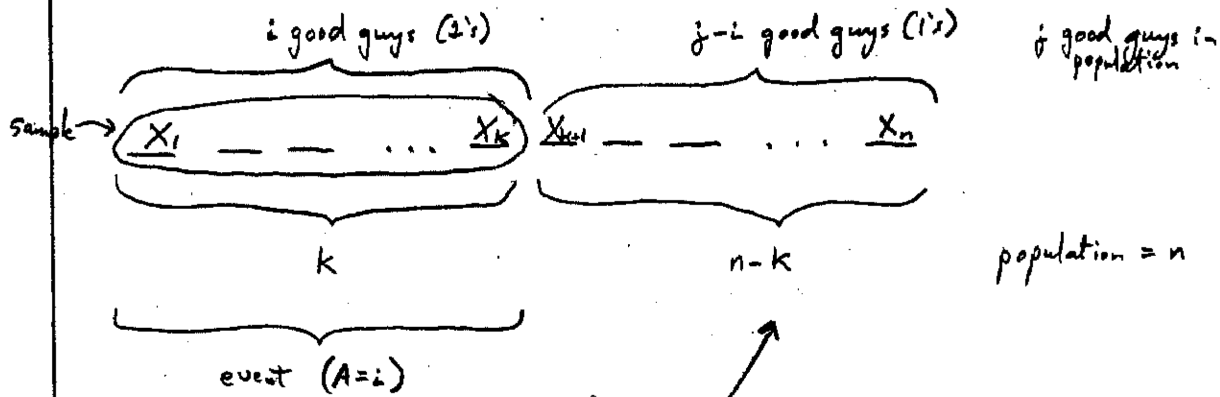
1^{st} k = sample of size k

random sample of size k

Therefore, the random variable A is defined as:

$$A = \text{number of } \overbrace{\text{good guys}}^{1\text{'s}} \text{ in the sample}$$
$$= X_1 + X_2 + \cdots + X_k \quad \longleftarrow \text{first } k \text{ random variables}$$

We next consider the conditional probability $P(U=j \mid A=i)$ directly, from the Bernoulli Process point of view.



$P(U=j \mid A=i)$ is just the <u>binomial distribution</u>:

$$P(U=j \mid A=i) = \binom{n-k}{j-i} p^{j-i} q^{n-k-j+i}$$

given event $(A=i)$, number of ways of choosing remaining $j-i$ good guys from remaining $n-k$ random variables,

probability of good guys

probability of bad guys

As desired, there is agreement in the posterior probability $P(U=j \mid A=i)$, whether calculated analytically (i.e., evaluation of the mess on [3/13/98.7]) or viewed from the Bernoulli process point of view.

Note that with the prior given as a binomial distribution, the posterior was also a binomial distribution:

<u>prior</u> $\quad P(U=j) = \binom{n}{j} p^j q^{n-j}$

<u>posterior</u> $\quad P(U=j \mid A=i) = \binom{n-k}{j-i} p^{j-i} q^{n-k-j+i}$

both <u>binomial distributions</u>

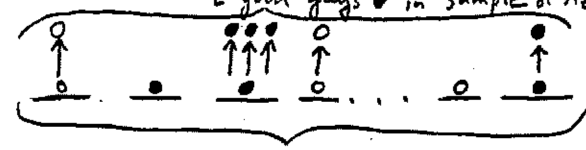
This is an example of <u>Conjugate Prior Theory.</u>

In practice, the prior used depends on the application. Many different priors are plausible (we've shown two). A statistician's sense and experience are involved in choosing the appropriate prior for a particular problem.

Now, let's do the same w/o feeling – sampling w/ replacement:

Bayes' Theorem for Sampling W/ Replacement

i good guys ● in sample of size k



random sample of size k with replacement

n = population size
j = number of good guys ● in population

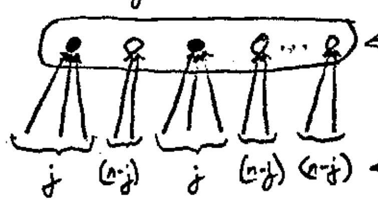
Sample of size k with replacement

Let U = number of good guys (in the population)
A = number of good guys in a random sample of k with repetition

Computation of the conditional probability $P(A=i \mid U=j)$ is straightforward (for example, see [2/23/98.8]):

$$P(A=i \mid U=j) = \frac{\binom{k}{i} j^i (n-j)^{k-i}}{n^k}$$

In a sample of size k, there are $\binom{k}{i}$ ways to have i good guys,
Each of the i good guys can come from any of the j good members of the population $\Rightarrow j^i$
Each of the $k-i$ bad guys can come from any of the $n-j$ bad members of the population $\Rightarrow (n-j)^{k-i}$



given one of the $\binom{k}{i}$ ways for sample to have i good guys $k-i$ bad guys

there are: $j^i (n-j)^{k-i}$ possibilities

$$P(A=i \mid U=j) = \binom{k}{i}\left(\frac{j}{n}\right)^i \left(\frac{n-j}{n}\right)^{k-i} = \binom{k}{i}\left(\frac{j}{n}\right)^i \left(1-\frac{j}{n}\right)^{k-i}$$

Since the above is a probability distribution for the random variable A, it satisfies:

$$\sum_i P_{(U=j)}(A=i) = 1 \quad , \text{ for all } j$$

which gives the binomial identity (our friend again, the binomial theorem):

$$\sum_i \binom{k}{i}\left(\frac{j}{n}\right)^i \left(\frac{n-j}{n}\right)^{k-i} = 1 \quad , \text{ for all } j$$

Bayes' Law gives:

$$P(U=j|A=i) = \frac{P(A=i|U=j)\,P(U=j)}{P(A=i)}$$

$$= \frac{P(A=i|U=j)\,P(U=j)}{\sum_j P(A=i|U=j)\,P(U=j)}$$

Law of Alternatives

Substituting our recently derived

$$P(A=i|U=j) = \binom{k}{i}\left(\frac{j}{n}\right)^i\left(\frac{n-j}{n}\right)^{k-i}$$

$$= \frac{\cancel{\binom{k}{i}}\left(\frac{j}{n}\right)^i\left(\frac{n-j}{n}\right)^{k-i} P(U=j)}{\sum_j \cancel{\binom{k}{i}}\left(\frac{j}{n}\right)^i\left(\frac{n-j}{n}\right)^{k-i} P(U=j)}$$

$$P(U=j|A=i) = \frac{\left(\frac{j}{n}\right)^i\left(\frac{n-j}{n}\right)^{k-i} P(U=j)}{\sum_j \left(\frac{j}{n}\right)^i\left(\frac{n-j}{n}\right)^{k-i} P(U=j)}$$

Let's consider the posterior probability with different prior probabilities:

(1) Uniform Prior (Sampling w/ Replacement)

Assume $P(U=j) = \frac{1}{n+1}$

which gives the posterior as:

$$P(U=j|A=i) = \frac{\left(\frac{j}{n}\right)^i\left(\frac{n-j}{n}\right)^{k-i}\cancel{\left(\frac{1}{n+1}\right)}}{\underbrace{\sum_j \left(\frac{j}{n}\right)^i\left(\frac{n-j}{n}\right)^{k-i}}\cancel{\left(\frac{1}{n+1}\right)}}$$

this is the discrete beta function, denoted as:

$$C_{i,k,n}$$

$$\boxed{P(U=j \mid A=i) = C_{i,k,n}^{-1} \left(\frac{j}{n}\right)^{i} \left(\frac{n-j}{n}\right)^{k-i}}$$

Coverage of: Sampling w/ Replacement

Case 1 – uniform prior

Case 2 – conjugate prior

Continues on AfterNotes [AN 3/16/98.6–9]
Lecture 16

John Guidi
guidi@mit.edu

Conjugate Priors

Sampling w/o Replacement

population : n
sample : k

Random Variables:

U = number of good in population
A = number of good in sample

$$P(A=i \mid U=j) = \frac{\binom{j}{i}\binom{n-j}{k-i}}{\binom{n}{k}}$$

from the j good, we choose i.

from the $n-j$ bad, we choose $k-i$.

number of points in sample space,

$$P(U=j \mid A=i) = \frac{P(A=i \mid U=j)\, P(U=j)}{P(A=i)}$$

$$= \frac{P(A=i \mid U=j)\, P(U=j)}{\sum_j P(A=i \mid U=j)\, P(U=j)}$$ ← Law of Alternatives

$$= \frac{\dfrac{\binom{j}{i}\binom{n-j}{k-i}}{\cancel{\binom{n}{k}}}\, P(U=j)}{\sum_j \dfrac{\binom{j}{i}\binom{n-j}{k-i}}{\cancel{\binom{n}{k}}}\, P(U=j)}$$

$$\underbrace{P(U=j \mid A=i)}_{\text{this is the posterior}} = \frac{\binom{j}{i}\binom{n-j}{k-i}\, P(U=j)}{\sum_j \binom{j}{i}\binom{n-j}{k-i}\, P(U=j)}$$

given arbitrary priors

Now we have the following cases:

- Case 1 – Uniform prior

$$\text{prior} \quad P(U=j) = \frac{1}{n+1}$$

This gives:

$$P(U=j \mid A=i) = \frac{\binom{j}{i}\binom{n-j}{k-i}}{\sum_j \binom{j}{i}\binom{n-j}{k-i}}$$

we proved binomial identity probabilistically [3/11/98.12-13]:

$$\sum_j \binom{j}{i}\binom{n-j}{k-i} = \binom{n+1}{k+1}$$

$$P(U=j \mid A=i) = \frac{\binom{j}{i}\binom{n-j}{k-i}}{\binom{n+1}{k+1}}$$

- Case 2 – Conjugate Prior

$$\text{prior} \quad P(U=j) = \binom{n}{j} p^j q^{n-j}$$

$$P(A=i) = \sum_j P(A=i \mid U=j)\, P(U=j)$$

$$= \sum_j \frac{\binom{j}{i}\binom{n-j}{k-i}}{\binom{n}{k}} \binom{n}{j} p^j q^{n-j}$$

$$= \sum_j \frac{k!\,\boxed{(n-k)!}}{\cancel{n!}} \frac{\cancel{j!}}{i!\,\boxed{(j-i)!}} \frac{\cancel{(n-j)!}}{(k-i)!\,\boxed{(n-j-k+i)!}} \frac{\cancel{n!}}{\cancel{j!}\cancel{(n-j)!}} p^j q^{n-j}$$

$$= \sum_j \frac{k!}{i!\,(k-i)!} \frac{(n-k)!}{(j-i)!\,(n-k-j+i)!} p^j q^{n-j}$$

$$= \sum_j \binom{k}{i}\binom{n-k}{j-i} p^j q^{n-j}$$

$$= \binom{k}{i} \sum_j \binom{n-k}{j-i} p^j q^{n-j}$$

Let $t = j - i$

$$= \binom{k}{i} \sum_t \binom{n-k}{t} \underbrace{p^i p^t}_{p^j} \underbrace{q^{k-i} q^{n-k-t}}_{q^{n-j}}$$

$$= \binom{k}{i} p^i q^{k-i} \underbrace{\sum_t \binom{n-k}{t} p^t q^{n-k-t}}_{1\text{, by the binomial theorem}}$$

$$\boxed{P(A=i) = \binom{k}{i} p^i q^{k-i}}$$

Now we figure our posterior probability:

$$P(U=j \mid A=i) = \frac{P(A=i \mid U=j)\, P(U=j)}{P(A=i)}$$

$$= \frac{\dfrac{\binom{j}{i}\binom{n-j}{k-i}}{\binom{n}{k}} \binom{n}{j} p^j q^{n-j}}{\binom{k}{i} p^i q^{k-i}}$$

$$= \frac{\cancel{k!}(n-k)!}{\cancel{n!}} \frac{\cancel{i!(k-i)!}}{\cancel{k!}} \frac{\cancel{j!}}{\cancel{i!}(j-i)!} \frac{\cancel{(n-j)!}}{\cancel{(k-i)!}(n-j-k+i)!} \frac{\cancel{n!}}{\cancel{j!(n-j)!}} p^{j-i} q^{n-j-k+i}$$

$$= \frac{(n-k)!}{(j-i)!\,(n-k-j+i)!} p^{j-i} q^{n-j-k+i}$$

$$\boxed{P(U=j \mid A=i) = \binom{n-k}{j-i} p^{j-i} q^{n-j-k+i}}$$

$$P_{(A=i)}(U=j) = \binom{n-k}{j-i} p^{j-i} q^{n-j-k+i}$$

- Since this is a probability, if you fix i and add up over j, the result must equal 1:

$$\sum_j P_{(A=i)}(U=j) = 1$$

$$\sum_j \binom{n-k}{j-i} p^{j-i} q^{n-j-k+i} = 1$$

More binomial identities!

{close to, but not quite, a binomial distribution}

Conjugate Prior	Posterior
$P(U=j) = \binom{n}{j} p^j q^{n-j}$	$P_{(A=i)}(U=j) = \binom{n-k}{j-i} p^{j-i} q^{n-j-k+i}$
$\sum_j \binom{n}{j} p^j q^{n-j} = 1$	$\sum_j \binom{n-k}{j-i} p^{j-i} q^{n-j-k+i} = 1$

Q/A: We start w/ the conjugate prior distribution, which is concentrated from 0 to n.
Then we end up with a posterior distribution, which is concentrated from i to n (since event $(A=i)$ is given).
So the posterior distribution is the distribution for some other family.
So our family for this posterior distribution is not closed under sampling, which you want from conjugate priors — unless you define a broader family in general.
But if you define a broader family, you will need to do this computation for all representatives of the broader family.
It's unclear what the broader family should be like.
It is not a well-defined concept, this notion of conjugate priors.

In this case, the conjugate prior and posterior are roughly something similar, but not quite the same.

In class, we discussed a direct probabilistic interpretation to find the conditional probability $P(U=j|A=i)$ [3/13/98.7-8].

Uniform Prior

$$P(U=j) = \frac{1}{n+1}$$

Posterior

$$P_{(A=i)}(U=j) = \frac{\binom{j}{i}\binom{n-j}{k-i}}{\binom{n+1}{k+1}}$$

Note how very different the prior + the posterior are in this case. The posterior is very close to a hypergeometric distribution.
The conjugate prior and its posterior are much closer than the uniform prior and its posterior.

Q: In looking for prior probabilities that give posteriors in the same family, what do we look for?
A: This happens mostly when you take samples w/ replacement.
Then it's no problem.
But if you take samples w/o replacement, the combinatorics get in the way.

- I don't think anybody worked out the Pólya case.
I just put it in.
I have no idea what it is.

- I have also not seen anywhere worked out (in this sense) the posterior multivariate hypergeometric distribution.

I think for the multivariate geometric, it is something very similar to this.
For Pólya, I really don't know.

- In practice, most people use sampling w/ replacement.
But that's not very good w/ small samples.

- I suspect that except for this posterior from conjugate prior, multivariate hypergeometric, and Pólya, these are the only 3 that have this weird so called prior, which is computable.
No other family has this computable so called conjugate

$$\underbrace{P(U=j \mid A=i)}_{\text{Posteriors}} = \frac{P(A=i \mid U=j)\,P(U=j)}{P(A=i)} \Longrightarrow \text{Conjugate Priors}$$

Posteriors

1) approx binomial
2) multivariate hypergeometric
3) Pólya

Posterior + Conjugate Prior similar distributions

- The problem is that for a long time, people didn't believe in Bayes' Law. And now, there is a big turn towards Bayesian statistics, If you look at all the classical statistics books, the word Bayes is <u>forbidden</u>. And all this stuff is left out.

- I would love to have an argument that gives <u>the</u> binomial distribution as posterior.

- <u>Sampling w/ Replacement</u>

$$P(A=i|U=j) = \frac{\underbrace{\binom{k}{i}}_{\text{from } k\text{, pick } i \text{ good ones}} \overbrace{j^i}^{\text{number of ways to have } i \text{ good guys}} \overbrace{(n-j)^{k-i}}^{\text{number of ways to have } i \text{ bad guys}}}{\underbrace{n^k}_{\text{number of sample points}}}$$

given one of $\binom{k}{i}$ ways to pick i good:

$$P(U=j|A=i) = \frac{P(A=i|U=j)\,P(U=j)}{P(A=i)}$$

$$= \frac{P(A=i|U=j)\,P(U=j)}{\sum_j P(A=i|U=j)\,P(U=j)}$$

$$= \frac{\frac{\cancel{\binom{k}{i}} j^i (n-j)^{k-i}}{n^k} P(U=j)}{\sum_j \frac{\cancel{\binom{k}{i}} j^i (n-j)^{k-i}}{n^k} P(U=j)}$$

$$\boxed{P(U=j|A=i) = \frac{\left(\frac{j}{n}\right)^i \left(\frac{n-j}{n}\right)^{k-i} P(U=j)}{\sum_j \left(\frac{j}{n}\right)^i \left(\frac{n-j}{n}\right)^{k-i} P(U=j)}}$$

Just need to plug in the priors.

Case 1 - Uniform prior

Assume $P(U=j)$ = same, for all j

So,

$$P(U=j|A=i) = \frac{\left(\frac{j}{n}\right)^i \left(\frac{n-j}{n}\right)^{k-i} \cancel{P(U=j)}}{\sum_j \left(\frac{j}{n}\right)^i \left(\frac{n-j}{n}\right)^{k-i} \cancel{P(U=j)}}$$

$$P(U=j|A=i) = \frac{\left(\frac{j}{n}\right)^i \left(\frac{n-j}{n}\right)^{k-i}}{\underbrace{\sum_j \left(\frac{j}{n}\right)^i \left(\frac{n-j}{n}\right)^{k-i}}}$$

The denominator is the discrete beta function
You can evaluate it, but it's pointless.
You sum over j.
Note: if you look in the books, they tend to do the integral, because the beta function integral has been well described.

We refer to the discrete beta function, which depends on i, k, n, as:

$$C_{i,k,n}$$

$$\boxed{P(U=j|A=i) = C_{i,k,n}^{-1} \left(\frac{j}{n}\right)^i \left(\frac{n-j}{n}\right)^{k-i}}$$

Posterior for uniform prior with replacement

We know that $\sum_j P_{(A=i)}(U=j) = 1$

Thus it is understood that the $C_{i,k,n}$ constant will be such that the above sum equals 1.
Statisticians often ignore such constants, with the understanding that probabilities add up to 1, and write such probabilities as:

$$P(U=j|A=i) \propto \left(\frac{j}{n}\right)^i \left(\frac{n-j}{n}\right)^{k-i}$$

proportional

Case 2 - Conjugate Prior

Assume $P(U=j) = C^{-1}_{t,u,n} \left(\frac{j}{n}\right)^t \left(\frac{n-j}{n}\right)^{u-t}$

So,

$$P(U=j|A=i) = \frac{\left(\frac{j}{n}\right)^i \left(\frac{n-j}{n}\right)^{k-i} \cancel{C^{-1}_{t,u,n}} \left(\frac{j}{n}\right)^t \left(\frac{n-j}{n}\right)^{u-t}}{\sum_j \left(\frac{j}{n}\right)^i \left(\frac{n-j}{n}\right)^{k-i} \cancel{C^{-1}_{t,u,n}} \left(\frac{j}{n}\right)^t \left(\frac{n-j}{n}\right)^{u-t}}$$

$$= \frac{\left(\frac{j}{n}\right)^{i+t} \left(\frac{n-j}{n}\right)^{k+u-i-t}}{\underbrace{\sum_j \left(\frac{j}{n}\right)^{i+t} \left(\frac{n-j}{n}\right)^{k+u-i-t}}}$$

This denominator is the discrete beta function denoted as:

$$C_{i+t,k+u,n}$$

$$\boxed{P(U=j|A=i) = C^{-1}_{i+t,k+u,n} \left(\frac{j}{n}\right)^{i+t} \left(\frac{n-j}{n}\right)^{k+u-i-t}}$$

$$P(U=j|A=i) \propto \left(\frac{j}{n}\right)^{i+t} \left(\frac{n-j}{n}\right)^{k+u-i-t}$$

Conjugate Prior

$$P(U=j) = C^{-1}_{t,u,n} \left(\frac{j}{n}\right)^t \left(\frac{n-j}{n}\right)^{u-t}$$

Posterior

$$P_{(A=i)}(U=j) = C^{-1}_{i+t,k+u,n} \left(\frac{j}{n}\right)^{i+t} \left(\frac{n-j}{n}\right)^{k+u-i-t}$$

Since these are probabilities, then have to add up to 1.
So you also have the discrete beta function identity series:

$$\sum_j P(U=j) = 1 \qquad\qquad \sum_j P_{(A=i)}(U=j) = 1$$

$$\Rightarrow \sum_j \left(\frac{j}{n}\right)^t \left(\frac{n-j}{n}\right)^{u-t} = C_{t,u,n} \qquad \Rightarrow \sum_j \left(\frac{j}{n}\right)^{i+t} \left(\frac{n-j}{n}\right)^{k+u-i-t} = C_{i+t,k+u,n}$$

what does it mean to take $P(U=j) = C^{-1}_{t,u,n} \left(\frac{j}{n}\right)^t \left(\frac{n-j}{n}\right)^{u-t}$ as prior?

We already did this experiment and got the result $P(U=j \mid A=t)$.
We did the experiment with a sample of size u and got t good.
Then, having done this experiment, we use that as prior for a second experiment, pretending that we haven't done the first experiment.
So, for the 2^{nd} experiment, the prior is really:

$$\underbrace{P(U=j)}_{\text{prior for } 2^{nd} \text{ experiment}} = \underbrace{P(U=j \mid A=t)}_{\text{result from } 1^{st} \text{ experiment}}$$

$$= C^{-1}_{t,u,n} \left(\frac{j}{n}\right)^t \left(\frac{n-j}{n}\right)^{u-t}$$

This means we have previously done the experiment.
We use that information as our prior.
Plug it in and we get the posterior.

Q/A: Start the experiment and take a sample of size t.
Then I pretend I have not done the experiment.
I say "Let's use the result of this experiment as prior."

Then I do the actual experiment. Taking a sample of k, I find i good.
The result of this experiment is as if you had a single experiment with a sample size of $\underbrace{u}_{1^{st}\text{ pretend}}+\underbrace{k}_{\text{actual experiment}}$ and found $\underbrace{t}_{1^{st}\text{ pretend}}+\underbrace{i}_{\text{actual experiment}}$ good.

$$\underbrace{P(U=j \mid A=i)}_{\text{posterior}} = C_{i+t,k+u,n} \left(\frac{j}{n}\right)^{i+t} \left(\frac{n-j}{n}\right)^{k+u-i-t}$$

- There is a fancy book on this by Sir Harold Jeffreys entitled "Theory of Probability"
  Some statisticians have made an entire career rewriting chapters of that book.
  700 pages. No table of contents. No index.
  It's all stream of consciousness — you have to know where you are.
  This is sort of the bible of Bayesian Probability.
  I'm almost sure this is in Sir Harold Jeffreys.
  The sampling w/o replacement examples are not.

- I did all this because I want ~~to~~ get to DeFinetti's Theorem.

  one of the few honest theorems in this business.

Given an infinite sequence of exchangable random variables that take values 0 + 1.
Thats the same as:

1) Taking a number p between 0 and 1 with your favorite distribution.

2) After you do that, you play Bernoulli Trials.

So the prior can be any infinite sequence of 0 and 1, which is exchangable, i.e., position independent.

This means that there is a posterior of the Bernoulli Process with an arbitrary prior on the probability of the bias of the coin.

It's a really nice theorem.
I'll do it in class.

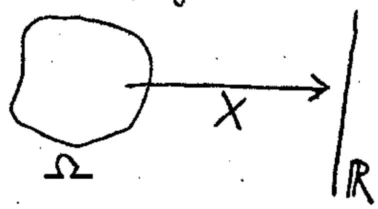
In some sense you can not have infinite sequences of exchangable 0-1 random variables. They are very close to coin tosses.

John Guidi
guidi@mit.edu

Continuous Random Variables (Beginning)

Let me skip the motivation and just start with the definitions, then go into the theory, then we'll give motivational examples.

So far, we have considered random variables that are integer valued. A random variable X on the probability space Ω, P that takes integer values. Guess what's coming next?



Random variable X, whose values are arbitrary real numbers.

General Definition of a Random Variable ← no restrictions

A random variable X on a probability space is a function from Ω to the real numbers such that for every real number a and b,

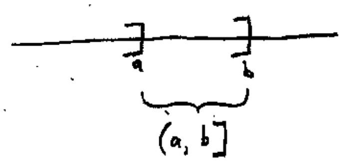
$$\underbrace{\{\omega : a < X(\omega) \le b\}}_{\text{the set of all } \omega \text{ for which } X(\omega) \text{ lies in } (a,b]} = \underbrace{(a < X \le b)}_{\text{which we call, from now on, the \underline{event} that } X \text{ takes values in } (a,b]} \quad \text{is an event.}$$

This is an almost vacuous definition.

To this end, it suffices to assume that $\overbrace{(X \le t)}^{\text{event } X \le t}$ is an event for all real numbers t, since

$$\underbrace{(a < X \le b)}_{\text{The event that } X \text{ is in } (a,b]} = \underbrace{(X \le b)}_{\text{The event that } X \le b} - \underbrace{(X \le a)}_{\text{The event that } X \le a}$$

And now you realize why we say $(a < X \le b)$ instead of $(a \le X \le b)$
↖ strictly.



- The function
$$F(t) = \overbrace{P(X \leq t)}^{\text{probability of the event } X \leq t}$$
is called the <u>cumulative distribution function</u> of the random variable X.

<u>Example:</u>
$$\underbrace{P(a < X \leq b)}_{\substack{\text{Probability that } X \text{ lies} \\ \text{in } (a, b]}} = \underbrace{F(b)}_{P(X \leq b)} - \underbrace{F(a)}_{P(X \leq a)}$$

- The random variable X <u>takes values between a and b</u> when
$$P(a \leq X \leq b) = 1$$
{here, we want $\leq$
Because X could take value a w/ positive probability}

i.e., $P(X > b) = 0$
and $P(X < a) = 0$

- Two random variables X and Y are <u>independent</u> when:
$$P\big((a < X \leq b) \cap (c < Y \leq d)\big) = P(a < X \leq b)\, P(c < Y \leq d) \text{ for all } a, b, c, d$$

- The cumulative probability distribution $F(t)$, also written as $F_X(t)$, satisfies:

{when you want to identify which random variable this is the cumulative probability distribution of.}

$$\lim_{t \to \infty} F(t) = 1$$

$$\lim_{t \to -\infty} F(t) = 0$$

and $F(t) \leq F(t')$ if $t < t'$

In other words, $F(t)$ is <u>increasing</u>, and it tends to zero at $-\infty$ and 1 at $+\infty$.
As follows immediately, from the definition:
$$F(t) = P(X \leq t)$$

- If the random variable X takes the values between a and b, we say that X is a <u>continuous</u> random variable when there is a function $f(t)$ for which:

$$P(c < X \leq d) = \int_c^d f(t)\,dt \quad , \text{ for } a \leq c \leq d < b$$

Probability of the event that X takes values in $(c, d]$

This is equivalent to:

$$F(d) - F(c) = \int_c^d f(t)\,dt$$

$$\left\{ P(c < X \leq d) = F(d) - F(c) \right\}$$

and too:

$$F(t) = \int_{-\infty}^t f(s)\,ds$$

$$\left(\begin{aligned} P(-\infty < X \leq t) &= \int_{-\infty}^t f(s)\,ds \\ F(t) - \underbrace{F(-\infty)}_{0} &= \int_{-\infty}^t f(s)\,ds \\ F(t) &= \int_{-\infty}^t f(s)\,ds \end{aligned} \right)$$

and too:

$$f(t) = \frac{d}{dt} F(t) = \frac{d}{dt} P(X \leq t)$$

for t between a and b

- Sometimes one writes:

$$\text{dens}(X = t) = f(t) \longleftarrow \left\{ \text{dens}(X = t) \text{ instead of } f(t) \right\}$$

and calls it the <u>density</u> of the continuous random variable X (density is 0 outside the interval $[a, b]$).

- Density is always non-negative

$$\text{dens}(X = t) = f(t) \geq 0$$

This is the continuous extension of the notion of random variables that we have been dealing with.
But now, shifting from the integers to the real numbers.
And we are going to see that a lot of the phenomena that we've already studied (e.g., balls into boxes) have <u>continuous analogues</u>.

Q: We have that: $F(t) = \int_{-\infty}^{t} f(s)\,ds$

and $f(t) = \frac{d}{dt} F(t)$

What if $f(t)$ is <u>not a continuous function</u>? In which case $F(t)$ will not be differentiable everywhere.
For example, what if $f(t)$ is a step function?

A: Then X is not a <u>continuous</u> random variable.
It's not true that every random variable is either integer or continuous.
There are other random variables, which we will consider later.
Continuous random variables are a special kind of random variable, just as integer random variables are a special kind of random variable.

This is the <u>basic definition.</u> It is vacuous, like all definitions.
But, gradually you will see what it means.

Q/A: if a random variable is continuous, we say that these integrals

$\int_{c}^{d} f(t)\,dt$, for $a \le c \le d \le b$

and $\int_{-\infty}^{t} f(s)\,ds$

<u>automatically</u> converge.

Let me write another formula:

$$\underbrace{\int_{-\infty}^{\infty} f(t)\,dt}_{P(-\infty < X \le \infty)} = 1$$

[Q/A: $f(t)$ has to be a <u>function</u>]

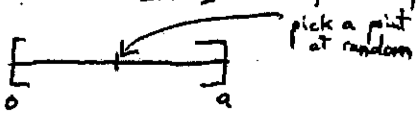
Now lets see examples.
As usual, the examples are not what one expects.
You are used to seeing functions like functions and integers like integers.
Now you see that this relationship between the cumulative distribution of a random variable and the density of a random variable, which are the probabilistic analogues of the fundamental theorem of the Calculus

By the way, I did not say what it means for two random variables to be identically distributed. It's obvious what it means. They have the same cumulative probability distribution function.

Now the density of a random variable is not the probability. It's something like the mass density.

Example 1 The uniform or Dirichlet process

Take an interval $[0, a]$ and pick a point at random.



} The simplest possible random variable you can think of.

The sample space is the interval $[0, a]$ $\quad \Omega = [0, a]$

sample point $\omega \in [0, a]$

The events are all intervals $(c, d]$ and any events that can be obtained on these intervals by taking union, intersection, complement, and infinite countable unions of disjoint [redacted] intervals.

Q: How do you get an event of 0?
A: You can easily get, from half open intervals, closed intervals, using these operations. So you can easily get the closed interval $[c, d]$. Thus this is also an event.
Then you can subtract the two:

$$[c, d] - (c, d] = c \longleftarrow \text{to get a point}$$

So any sensible set is in the interval $[0, a]$.
These are sometimes called Borel sets.

[redacted]

We define the probability as:

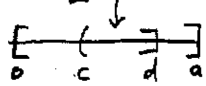
$$P((c, d]) = \frac{d-c}{a}$$

Then, by a theorem that we stated, but did not prove, this probability extends, in a well-defined way, to $\underset{\uparrow \text{ all Borel sets}}{\underline{\text{all events}}}$

By the fundamental theorem we stated at the beginning of the course, this gives a well-defined probability on all events.

Let $\underset{\uparrow \text{ a random variable}}{X}$ = random point in $[0, a]$

$$\underbrace{P(c < X \leq d)}_{\substack{\text{probability that } X \\ \text{takes a value in} \\ (c,d]}} = \frac{d-c}{a} \quad \leftarrow \text{We define it that way}$$



$$\underbrace{\text{dens}(X=t)}_{\text{density of this random variable}} = \begin{cases} \frac{1}{a} & \text{if } 0 \leq t \leq a \\ 0 & \text{elsewhere} \end{cases}$$

X is a continuous random variable on the interval $[0, a]$

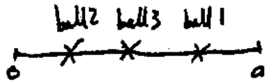
$$P(c < X \leq d) = \int_c^d \underbrace{\frac{1}{a}}_{\text{dens}(X=t) = f(t) = \frac{1}{a}} dt$$

- Let's pick n points <u>independently</u>.

<u>Dirichelet Process</u> – Take $X_1, X_2, \dots, X_n$ <u>independent</u> random variables, <u>identically distributed</u> with X.

we'll define the distribution in a moment.

What is the Dirichelet Process really?
Really, what we are doing is – guess what?
We take the interval $[0, a]$ and we drop <u>n balls</u> on that interval
So it's balls into boxes.
Except the <u>boxes</u> are now <u>continuous.</u>



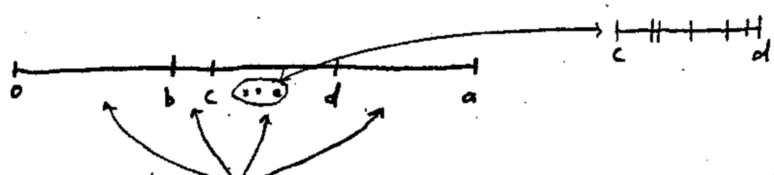
Now we can ask questions. The gist of the questions are those that we asked for Maxwell-Boltzmann statistics. Because these are <u>distinguishable balls.</u>

Q/A: Unfortunately in the book, I write X_n all over the place.
If we kept the original notation (i.e., k balls into n boxes), we'd instead have independent random variables $X_1, X_2, \dots, X_k$ for the k balls and the interval (i.e., the boxes) would be $[0, n]$.

I debated last night whether to change it everywhere.
Let's keep it as written.

So let's ask a suitable question.
<u>Occupation Numbers</u>

We take the interval $[0, a]$ and we divide it into intervals:

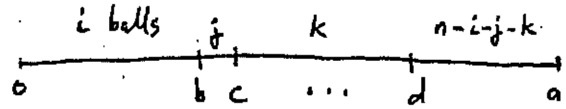


Now, we have <u>genuine boxes</u>, except they are of <u>different sizes</u>.

So the probability of a ball going into a box is <u>different</u> from the probability of a ball going into another box, because it could be <u>bigger or smaller.</u>

Let's compute the probability that the intervals have the following occupation numbers:

$A =$



Let's introduce a convenient and efficient notation for this type of computation.
From now on I talk about <u>balls</u>. The <u>X_i refer to balls.</u>

Let $N(t)$ = number of X_i in the interval $[0,t)$
(i.e., number of balls falling in $[0,t)$)

$$P(N(t)=k) = \binom{n}{k}\underbrace{\left(\frac{t}{a}\right)^k\left(\frac{a-t}{a}\right)^{n-k}}$$

from the n balls, you pick k.

The X_i are <u>independent</u>, so their probabilities multiply.
$\left(\frac{t}{a}\right)^k$ is the probability that they drop in $[0,t)$
$\left(\frac{a-t}{a}\right)^{n-k}$ is the probability that they drop outside $[0,t)$

$P(N(t)=k)$ = probability that there are k balls in $[0,t)$

Now let's go back to the event illustrated at the top of the page:

Event:

$$A = (N(b)=i) \cap (N(c)-N(b)=j) \cap (N(d)-N(c)=k) \cap (N(a)-N(d)=n-i-j-k)$$

We use <u>multinomial coefficients</u> to compute $P(A)$.
We have n balls, of which i go into the first interval
j " " " second "
k " " " third "
n-i-j-k " " " remaining interval

$\binom{n}{i,j,k,n-i-j-k}$ read "multinomial n $i,j,k,n-i-j-k$"

This is what multinomial coefficients are for.

$$P(A) = \underbrace{\binom{n}{i,j,k,n-i-j-k}}\left(\frac{b}{a}\right)^i\left(\frac{c-b}{a}\right)^j\left(\frac{d-c}{a}\right)^k\left(\frac{a-d}{a}\right)^{n-i-j-k}$$

you pick the balls that go in each box.

Now they go <u>independently</u>, so the probabilities multiply.
i go independently into the first interval
j " " " " second "
etc.

And that's the way you compute probabilities.

Let's repeat what we just did in highfalutin terms. Really high class.
What we've done is compute the <u>joint distribution of a random function.</u>
Watch.

What is $N(t)$, really?

$N(t)$ is infinitely many random variables, one for each t.
And these random variables are <u>not</u> independent. ← (As the formula for $P(A)$ clearly indicates.)

In some sense <u>$N(t)$</u> tells you <u>everything</u> about <u>occupation numbers.</u>

$N(t)$ is a random variable for each t

One way of saying this (i.e., that you have a random variable for each t) is to say you have a <u>random function.</u>

$N(t)$ is a random function ← {The word function is somewhat misleading here, but not too much.}

The random function is a function of t, which is a random variable for each t.
The random function $N(t)$, associated w/ the Dirichlet Process, contains all the information about occupation numbers for the Dirichlet Process.

<u>Anything</u> about the balls dropping in the interval $[0,a]$, you can infer by studying <u>joint distributions of the random function.</u>

You think that's easy? Just wait.

• <u>Example</u> You can view the dropping of the balls on the interval $[0,a]$ as putting checks on the interval.
Once the ball is dropped, you can forget about which ball is which.
We'll talk later about indistinguishable balls.

The probability that two balls will fall to the same place is 0.
So we may disregard it from now on.
We can ask the question:

When you drop all n balls, how close to the origin is the ball that is closest to the origin?

There is a <u>standard name</u> for that, used by everybody.

not my fault ⟶ $$X_{(1)} = \min(X_1, X_2, \ldots, X_n)$$

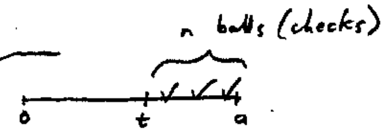
read: "X order 1"

What is the probability distribution of $\widetilde{X_{(1)}}$?

$$P(X_{(1)} > t) = ?$$

The event $(X_{(1)} > t)$ is the same as the event $(N(t) = 0)$.
If you don't see that, I can't explain it.

$$(X_{(1)} > t) = (N(t) = 0)$$



$$P(X_{(1)} > t) = P(N(t) = 0) = \binom{n}{0}\left(\frac{t}{a}\right)^0\left(\frac{a-t}{a}\right)^n = \left(\frac{a-t}{a}\right)^n$$

Therefore,

$$P(X_{(1)} > t) = \left(\frac{a-t}{a}\right)^n$$

And therefore,

the <u>cumulative distribution</u> of the random variable $X_{(1)}$ is:

$$\begin{aligned} F(t) &= P(X_{(1)} \le t) \\ &= 1 - P((X_{(1)} \le t)^c) \\ &= 1 - P(X_{(1)} > t) \\ &= 1 - \left(\frac{a-t}{a}\right)^n \end{aligned}$$

$$\boxed{F(t) = 1 - \frac{(a-t)^n}{a^n}}$$

Next, we compute the <u>density</u> of the continuous random variable.
Here we have the first non-trivial example of a continuous random variable, whose density is <u>not obvious</u>.

We have that:

$$\int_{-\infty}^{t} f(s)\,ds = F(t) \implies f(t) = \frac{d}{dt}F(t)$$

So the density $f(t)$ is:

$$f(t) = \text{dens}(X_{(1)} = t)$$

$$= \frac{d}{dt}\left(1 - \frac{(a-t)^n}{a^n}\right)$$

$$\boxed{f(t) = \frac{n(a-t)^{n-1}}{a^n}}$$

Observe that:

$$\boxed{\int_0^a \frac{n(a-t)^{n-1}}{a^n}\,dt = 1}$$

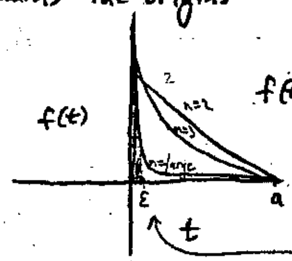
Automatically!

Since the integral of the density is always 1.

$$1 = \int_{-\infty}^{\infty} f(t)\,dt = \int_{-\infty}^{\infty} \frac{n(a-t)^{n-1}}{a^n}\,dt$$

$$= \int_0^a \frac{n(a-t)^{n-1}}{a^n}\,dt$$

Why is this result significant?
If you plot these curves, you'll find they are more and more condensed towards the origin.



$$f(t) = \frac{n(a-t)^{n-1}}{a^n}$$

graph of $\text{dens}(X_{(1)} = t)$

$\left\{\text{as } n \text{ increases (i.e., more balls dropped), more mass is concentrated to the origin. That is, for any arbitrarily small } \varepsilon,\ \lim_{n\to\infty} \int_0^{\varepsilon} f(t)\,dt \to 1\right\}$

You get:

$$P(X_{(1)} \leq \varepsilon) = F(\varepsilon) \quad \text{cumulative distribution}$$

$$= \int_{-\infty}^{\varepsilon} f(t)\,dt$$

$$\boxed{P(X_{(1)} \leq \varepsilon) = \int_0^{\varepsilon} \frac{n(a-t)^{n-1}}{a^n}\,dt}$$

{for fixed ε, the larger n is, the more the mass of the density curve moves towards the origin.}

So the more balls you drop, the bigger the chance that the closest ball will go to the origin!
This gives the exact expression, by integrating the density.

- So you have to get used to the fact that besides probability, there is this other measure, which is density.

$$\text{density} : \text{probability} :: \underbrace{\text{mass density}} : \text{mass}$$

mass density means nothing in itself.
It means something only when integrated.
The same happens now with density.
Probability density means nothing in itself.
It means something when you integrate it over an interval, in which case, it is the probability of a continuous random variable taking values within that interval.

- The probability of a continuous random variable taking one value is 0.

$$P(X = x) = \lim_{h \to 0} P(x-h < X \leq x)$$

$$= \lim_{h \to 0} \int_{x-h}^{h} f(t)\,dt$$

$$P(X = x) = 0$$

This is what the calculus is all about.
Your chickens are coming to roost now, if you don't know calculus.
From now on, we use calculus.

- Next comes the question:

$$\underbrace{X_{(2)}}_{\text{read: "X order 2"}} \text{ is the next ball after } X_{(1)}$$

Then you have:

$$X_{(1)} < X_{(2)} < X_{(3)} < \dots < X_{(n)}$$

We are going to compute the densities of these random variables, which are not independent, of course.

The order in which you find balls as you go down the interval.
These are called the:

Order Statistics

Tremendously important in statistics.
With these, you can answer questions such as how big these intervals are.
To do so, you need to know the facts: gaps
biggest
smallest
median

[On the problem set that I gave you, there are 2 problems that are unsolved.
You are MIT students, you're not students at Oshkosh College.
It would be insulting to you if I gave you only solved problems all the time.]

John Guidi
guidi@mit.edu

Continuous Random Variables (cont'd)

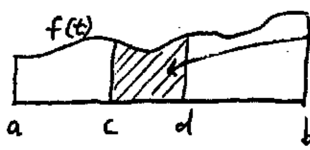
Last time we saw the definition of Random Variable in general.

if X is a random variable and $P(a < X \le b) = 1$
then X is said to be continuous when it has a density $f(t) = \text{dens}(X=t)$,
i.e., when:

$$\underbrace{P(c < X \le d)}_{\text{probability that } X \text{ is in the interval } (c,d]} = \int_c^d f(t)\,dt = \int_c^d \text{dens}(X=t)\,dt \qquad a \le c \le d \le b$$

This is the continuous analogue of what we have been doing so far.
Given a random variable taking its values in the interval $(a,b]$, this random variable is said to be continuous when there is a density function

(the probability that a random variable takes a given value is 0. $(P(X=x)=0)$.
All we can talk about is that the random variable takes a value between c and d.



That probability is the area under the density

Exactly the analogue of what we call calculus.
That's what this is all about.

A density is characterized by:

(1) $f(t) \ge 0$ for t between a and b

(2) $\int_a^b f(t)\,dt = 1$

since $\int_a^b f(t)\,dt = \underbrace{P(a < X \le b) = 1}_{\text{by assumption (see above)}}$

This is the continuous analogue of the probability distribution of an integer random variable. Integer random variables have probability distributions, which are sequences and if you sum all the sequences, they add up to 1.

Here we have functions integrating to 1. It's exactly parallel.

In addition to the density, we have the <u>cumulative distribution</u>.
We saw last time that cumulative distribution is defined for <u>all</u> random variables, not just continuous random variables.
We will see later examples of random variables that are neither continuous nor integer.
You can immediately imagine one — take the sum of a continuous random variable and an integer random variable. The <u>resulting random variable</u> is <u>neither continuous nor random.</u>

$$\underbrace{X}_{\substack{\text{continuous}\\ \text{random variable}}} + \underbrace{Y}_{\substack{\text{integer}\\ \text{random variable}}}$$

<u>Cumulative Distribution</u>

$$F(t) = F_X(t) = P(X \le t)$$

we use this notation when we want to be explicit about which random variable we are talking about.

For a continuous random variable, you have certain important <u>relationships</u> between the <u>cumulative distribution</u> and the <u>density</u>, which are essential to manipulate.

Let's review them:

$$P(c < X \le d) = F(d) - F(c)$$

By an obvious argument, which we made last time.

Also:

$$F(t) = \int_{\underbrace{-\infty}}^{t} \text{dens}(X=s)\, ds$$

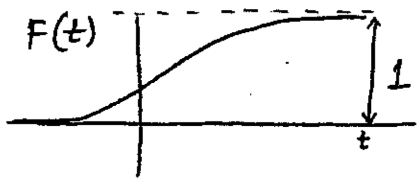
I put $-\infty$ instead of a, since $\text{dens}(X=s)$ will be 0 when $s < a$.
So it doesn't matter.

And from this,
by the Fundamental Theorem of the Calculus, you obtain the fundamental relation:

$$\frac{d}{dt} F(t) = \text{dens}(X=t) = f(t)$$

The cumulative distribution is the integral of the density.
The density is the derivative of the cumulative distribution.

What does the cumulative distribution look like?



$F(t)$ is increasing. Why? Because of the definition:

$$F(t) = P(X \le t)$$

Since probability is non-zero, if you increase t, then the probability of the event $(X \le t)$ either stays the same, or increases for the larger event.

$$\lim_{t \to -\infty} F(t) = 0$$

$$\lim_{t \to \infty} F(t) = 1$$

If the random variable is continuous then, of course, the cumulative distribution will also be continuous - in fact, differentiable.

Let's add a few tidbits to what we said last time:

if X is continuous, then there is a number m, for which $F(m) = \frac{1}{2}$

Because $F(m)$ goes from 0 to 1 continuously.
So there is one number where $F(m) = \frac{1}{2}$.

This number m is the median of the random variable X

Similarity, you can define the quartiles, as they say in statistics.

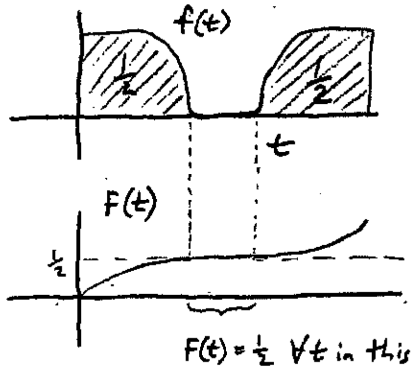
$m_{1/4}$ with the propeity that $F(m_{1/4}) = \frac{1}{4}$ ← that's called the first quartile

$m_{3/4}$ " " " " $F(m_{3/4}) = \frac{3}{4}$

This is the precise definition of the common sense, everywhere used, notion.

You can interpret these numbers explicitly in terms of probabilities.
Let's redefine the median w/o using the cumulative distribution.

Q: Is there only one median m?
A: No. Assuming we are talking about continuous random variables. If it's not continuous, all bets are off.
Consider the following density:



$$F(t) = \int_{-\infty}^{t} f(s)\,ds$$

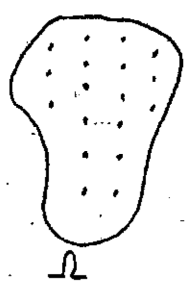
$F(t) = \frac{1}{2}\ \forall t$ in this interval
$m \in$ this interval
So there can be many medians m

If $f(t) > 0$ (strictly), then there can be only one median m

$$P(X < m) = \frac{1}{2} = P(X \geq m)$$

This is just repeating the definition.
This is for continuous random variables.

Let's add one more concept.
That's the expectation.



For the integer random variables, the idea of expectation was clear, if you had a finite sample space.
There, you took the value of the random variable at each point, multiplied it by the probability at that point and added it all over:

$$E(X) = \sum_{n} n\, P(X = n)$$

Now, we are in a similar situation — even worse, since we have a continuum.
Points have probability 0.

We define the expectation for a continuous random variable X:

$$E(X) = \int_a^b t \underbrace{f(t)}_{\text{density}} dt$$

The expectation is the average of the density, which is the continuous analogue of the notion of expectation that we had for integer random variables.

And we'll see later that the expectation of the sum of continuous random variables is the sum of the expectations, even though the random variables may not be independent.

$$E(X_1 + X_2 + \ldots + X_n) = E(X_1) + E(X_2) + \ldots + E(X_n)$$

Example Uniform (Dirichlet) Distribution

This is one of the fundamental stochastic processes.
To put it in really blunt terms:

Take n measurements of some quantity.
Then you want all the statistical facts about these n measurements

Really rock bottom phenomena.
A measurement of the quantity that is known to lie between 0 and a
Beyond which, we know nothing whatsoever.

The uniform process is characterized by being given n random variables $X_1, \ldots, X_n$ which are independent and identically distributed.

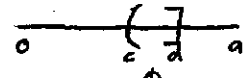
The probability distribution:

$$P(c < X_1 \leq d) = \int_c^d \frac{1}{a}\, dt \qquad \text{if } 0 \leq c \leq d \leq a$$

$$= \frac{d-c}{a}$$

In other words, if I have an interval $[0, a]$ and I pick a point at random, I can not tell anything about the probability that this point is some specific point — that probability is 0.

$$P(X_1 = x) = 0$$



$P(c < X_1 \leq d) = \frac{d-c}{a}$

But we <u>can</u> tell the probability that the picked point lies in an <u>interval</u>.
For example:

$$P(x-\varepsilon < X_1 \leq x+\varepsilon) = \int_{x-\varepsilon}^{x+\varepsilon} \frac{1}{a}\,dt$$

$$= \frac{2\varepsilon}{a}$$

The density $f(t) = \frac{1}{a}$
The density is <u>not</u> the probability.
It's just like mass density and mass. Same analogy. It's extremely similar.
As a matter of fact there is a concept relevant to math + physics that, as you will see, is very clearly visualized in probability.

The variables are all independent and they all have this obvious $\underbrace{\text{uniform distribution.}}_{P(c < X_1 \leq d)}$

What can we say about them?

We associate w/ the Dirichlet Process the integer random variable:

$U(t)$ = number of X_i's that are $\leq t$

We computed the probability distribution of this last time [3/16/98.8]:

$$P(U(t) = k) = \binom{n}{k}\left(\frac{t}{a}\right)^k\left(\frac{a-t}{a}\right)^{n-k}$$

We saw that $U(t)$ is a random variable for each t.
And this situation is called <u>random functions.</u>
A random function is a random variable for each value of t. Observe that to get a random function is <u>not</u> just to describe one random variable for each t. <u>All</u> the <u>joint distributions</u> for <u>any</u> <u>finite set</u> of t, are needed to completely describe the random function.

> On Friday, Problem Set w/ 20 problems. A regular problem set and training problems for the quiz.
> There is a very nice collection of problems made up by a British <u>Reverend Whitworth</u> who include some beautiful problems. I'll crib from those. They're all stated in British terms. Some Bayes problems that are really super.
> By the way, if you look up in the literature on Bayesian stuff, the philosophers have written many books w/ names such as "Paradoxes of Rationality and Cooperation" Several hundred pages.

That's what we did last time. We subdivided the interval into a number of boxes and then we computed the probability that 1st box had i X's, 2nd j X's, etc. [3/16/98.7-9]
This gives you joint distributions of the integer random variable $U(t)$.

$U(t)$ is an integer random variable associated w/ the Dirichlet Process.

On the other hand, the order statistics, which we began to discuss last time, are continuous random variables.
These are probably the most important single sequence of random variables there is.
Perhaps 20 books have been written on the subject of order statistics.

$$\left.\begin{aligned}
\underbrace{X_{(1)}}_{\text{read: "X order 1"}} &= \min(X_1, X_2, \dots, X_n) \\
X_{(2)} &= \text{next after the minimum} \\
\vdots \qquad & \qquad \vdots \\
X_{(n)} &= \max(X_1, X_2, \dots, X_n)
\end{aligned}\right\} \text{These are the order statistics}$$

The order statistics are the X_i rearranged in increasing order.
In statistical situations, in problems of distribution, that's what happens.
We will later consider problems of occupancy.

So let's now study the order statistics
These are not independent random variables.

I have to curb my tendencies to philosophize.
But, philosophically, the order statistics correspond to indistinguishable balls.
I'm not going to explain that.
If you don't see it, forget it.

Shifting from $X_1, X_2, \dots, X_n$ to $X_{(1)}, X_{(2)}, \dots, X_{(n)}$
means that you are shifting from distinguishable balls to indistinguishable.

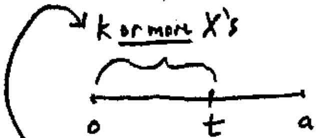
Q: Can we assume that no two X_i are the same?
A: Always. Probability is 0. These are continuous random variables.
We don't even mention it anymore.

Our first job is to compute the probability distribution of the order statistics.
We'll do it two ways.
First we do it brutal. Then we do it using calculus.

$$P(X_{(k)} \le t) = ?$$

The cumulative distribution
of the k^{th} order statistic

Well, what do you do to compute the probability?
Let's write the event $(X_{(k)} \le t)$ in terms of events that we can compute.



What does it mean for $(X_{(k)} \le t)$?
At least k X's fall in interval $(0, t]$.

{This kind of trick is the only trick there is. Once you've mastered that trick, you can do everything}

So we have:

$$(X_{(k)} \le t) = \big((U(t)=k) \cup (U(t)=k+1) \cup \dots \cup (U(t)=n)\big)$$

Therefore, these events being disjoint, we are saved from doing mutual inclusion-exclusion.

$$P(X_{(k)} \le t) = P(U(t)=k) + P(U(t)=k+1) + \dots + P(U(t)=n)$$

$$= \binom{n}{k}\frac{t^k (a-t)^{n-k}}{a^n} + \binom{n}{k+1}\frac{t^{k+1}(a-t)^{n-k-1}}{a^n} + \dots + \underbrace{\binom{n}{n}\frac{t^n \cancel{(a-t)^0}}{a^n}}_{\text{i.e., all of them fall in the interval } (0,t]} \to \frac{t^n}{a^n}.$$

Now, I have an unpleasant announcement.
Let's compute the density of this random variable.
That's the derivative of the mess above.

1st – Density – The Hard Way

$$\text{dens}(X_{(k)} = t) = \frac{d}{dt}F(t)$$

$$= \frac{d}{dt}P(X_{(k)} \le t)$$

Miracles of cancellation occur!
$(fg)' = fg' + gf'$

$$= \binom{n}{k}\frac{k t^{k-1}(a-t)^{n-k}}{a^n} - \cancel{\binom{n}{k}\frac{(n-k)t^k (a-t)^{n-k-1}}{a^n}} \quad \Big\} \text{ derivative of 1st term}$$

$$+ \cancel{\binom{n}{k+1}\frac{(k+1)t^k (a-t)^{n-k-1}}{a^n}} - \cancel{\binom{n}{k+1}\frac{(n-k-1)t^{k+1}(a-t)^{n-k-2}}{a^n}} \quad \Big\} \text{ of 2nd term}$$

$$+ \cancel{\dots} - \cancel{\dots} \quad \Big\} \quad \vdots$$

$$+ \cancel{\frac{n t^{n-1}}{a^n}} \quad \Big\} \text{ of last term}$$

Miracles of cancellation occur because alternating −/+ terms have the same factors for t and $(a-t)$ and the coefficients cancel:

$$-\binom{n}{k+i}(n-k-i) + \binom{n}{k+i+1}(k+i+1) \stackrel{?}{=} 0$$

$$-\frac{\cancel{n!}\,(n-k-i)}{(k+i)!\,(n-k-i)!} + \frac{\cancel{n!}\,(k+i+1)}{(k+i+1)!\,(n-k-i-1)!} \stackrel{?}{=} 0$$

$$-\frac{1}{(k+i)!\,(n-k-i-1)!} + \frac{1}{(k+i)!\,(n-k-i-1)!} = 0 \quad ✓$$

So only the first term remains and we have:

$$(*) \quad \boxed{dens(X_{(k)}=t) = \binom{n}{k}\frac{k\,t^{k-1}(a-t)^{n-k}}{a^n}}$$

2nd - Density - The Easy Way

Now let's rederive this result directly using calculus, without going through the cumulative distribution.

Again, with feeling:

By definition, we have:

$$P(t < X_{(k)} \le t+dt) = \int_t^{t+dt} dens(X_{(k)}=s)\,ds$$

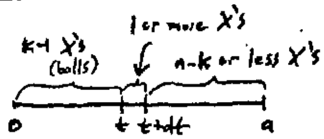
$$\lim_{dt\to 0} P(t < X_{(k)} \le t+dt) = \lim_{dt\to 0}\int_t^{t+dt} dens(X_{(k)}=s)\,ds$$

$$= \lim_{dt\to 0} dens(X_{(k)}=t)\,dt$$

which gives:

$$\lim_{dt\to 0}\frac{P(t < X_{(k)} \le t+dt)}{dt} = dens(X_{(k)}=t)$$

This is calculus.



The event that the k^{th} order statistic is in the interval $(t, t+dt]$.

$(t < X_{(k)} \le t+dt)$

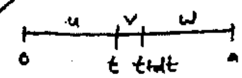
multinomial coefficient

probability of all events where more than 1 X (ball) falls in $(t, t+dt]$

$$P(t < X_{(k)} \le t+dt) = \underbrace{\binom{n}{k-1,\,1,\,n-k}\left(\frac{t}{a}\right)^{k-1}\left(\frac{dt}{a}\right)\left(\frac{(a-t-dt)}{a}\right)^{n-k}} + \overbrace{O\left((dt)^2\right)}$$

probability of event that k-1 X's (balls) are in $(0,t]$
exactly 1 in $(t,t+dt]$
exactly n-k in $(t+dt, a]$

$$\sum_{\substack{\forall\, u<k\le u+v \\ \text{where: } u \ne k-1 \\ v \ne 1}} \binom{n}{u,\,v,\,w}\left(\frac{t}{a}\right)^{u}\left(\frac{dt}{a}\right)^{v}\left(\frac{a-t-dt}{a}\right)^{w}$$



Note that $v \ge 2$

$$\lim_{dt\to 0}\frac{P(t<X_{(k)}\le t+dt)}{dt} = \lim_{dt\to 0}\left\{\binom{n}{k-1,\,1,\,n-k}\left(\frac{t}{a}\right)^{k-1}\left(\frac{1}{a}\right)\left(\frac{a-t-\cancel{dt}^{\,0}}{a}\right)^{n-k} + O(\cancel{dt}^{\,0})\right\}$$

$$= \binom{n}{k-1,\,1,\,n-k}\left(\frac{t}{a}\right)^{k-1}\left(\frac{1}{a}\right)\left(\frac{a-t}{a}\right)^{n-k}$$

which gives:

$$\boxed{\text{dens}(X_{(k)}=t) = \binom{n}{k-1,\,1,\,n-k}\frac{t^{k-1}(a-t)^{n-k}}{a^n}}$$

Done. Directly. No computation necessary.
This is the same as equation (*), obtained from differentiating the cumulative density as:

$$\binom{n}{k}k = \binom{n}{k-1,\,1,\,n-k}$$

$$\frac{n!}{k!(n-k)!}k = \frac{n!}{(k-1)!\,1!\,(n-k)!}$$

$$= \binom{n}{k-1,\,1,\,n-k} \checkmark$$

Now let's reap something from our labors.
As a consequence, we obtain the following:

$$\int_0^a \mathrm{dens}(X_{(k)} = t)\,dt = 1$$ ← {We know that the integral of the density is 1}

This gives us an extraordinary integral:

$$\int_0^a \binom{n}{k} k \frac{t^{k-1}(a-t)^{n-k}}{a^n}\,dt = 1$$

Automatically, as they say. Give that to your friends to evaluate directly.

$$\int_0^a \frac{n!}{(k-1)!(n-k)!} \frac{t^{k-1}(a-t)^{n-k}}{a^n}\,dt = 1$$

Let's rewrite this in a way that masks where it comes from.

Namely:

$$\boxed{\int_0^a t^{k-1}(a-t)^{n-k}\,dt = \frac{a^n (k-1)!(n-k)!}{n!}}$$

This is the Beta Function Integral

The distribution of the k^{th} order statistic is called the Beta density distribution:

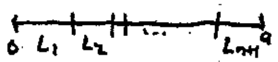
$$P(X_{(k)} \le t)$$

Not bad.
Let's do a little more.

This is one of the cutest arguments in this business.
Really cute. Super cute.
You have to really use probability on this one.
Watch.

We now have the probability distribution of all the order statistics.
Where the 1st point goes, the 2nd point, etc.
Now we want the gaps between successive points.

If you drop n points on the interval $[0,a]$, there are $n+1$ gaps



$n+1$ gaps

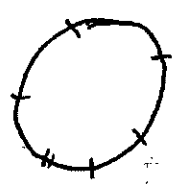
It is our duty to compute the probability densities of these gaps.
How far apart are the points?

Here we have the main result.
These gaps are identically distributed. They are not independent, of course.

$$L_i \text{ is the gap between } X_{(i-1)} \text{ and } X_{(i)} \quad \longleftarrow \text{ where } X_{(0)} = 0, \quad X_{(n+1)} = a$$

I claim that random variables $L_1, \dots, L_{n+1}$ are identically distributed.
It doesn't seem obvious until I show you the trick,
which is the following:

Instead of dropping n points on the interval $(0,a]$, let's take a circle w/ circumference a and drop, on this circumference, $n+1$ points.



$n+1$ points

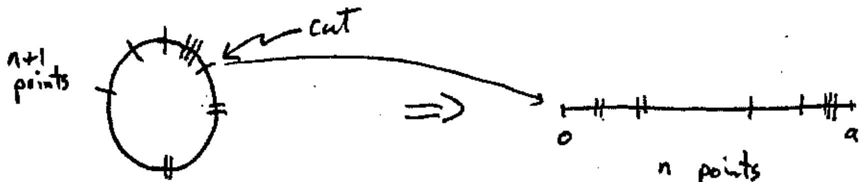
circumference a

If I drop $n+1$ points at random, then you have $n+1$ gaps on this circumference.
Now I say – in the case of a circumference of the circle, any 2 gaps are identically distributed.
If you don't see that, I can not explain it.
In fact, I can not prove it.
By the Principle of Sufficient Reason.

So by dropping $n+1$ points on the circumference of a circle, it is obvious that the gaps are i.d.

Now, instead of dropping n points on the interval $(0,a]$, I do the following:

I drop $n+1$ points on the circumference. Then I see where the first one lands. Then with a scissors, I cut the circle. The result is the same as dropping n points on the interval $(0,a]$.



Circumference a
gaps i.d.

n points
on interval $(0,a]$ gaps i.d.

$\Uparrow$

$\left\{ \text{Therefore, the gaps are identically distributed.} \right\}$

That's a purely probabilistic argument.

In other words, the experiment of dropping n points is exactly the same as dropping $n+1$ points on the circle.
This gives you an intuitive grasp that the gaps are identically distributed.
We will later prove it again, using conditional probability.

This is fairly convincing.
And this leads immediately to a way of computing the expectation.

$$E(L_1) = E(L_2) = \dots = E(L_{n+1})$$ ← Because these random variables are identically distributed.

Note that: $L_1 + L_2 + \dots + L_{n+1} = a$

$$E(L_1 + L_2 + \dots + L_{n+1}) = a$$

So,
$$E(L_1) + E(L_2) + \dots + E(L_{n+1}) = a$$

$\left\{ \text{Here we anticipate that the expectation of the sum is the sum of the expectations, which we have not yet proved.} \right\}$

This gives:

$$E(L_1) = \frac{a}{n+1}$$

No computation necessary.

What's more, we even get the expectation of the order statistics.

$$X_{(k)} = L_1 + L_2 + \dots + L_k$$

We easily get the expectation:

$$E(X_{(k)}) = E(L_1 + L_2 + \dots + L_k)$$
$$= \underbrace{E(L_1) + E(L_2) + \dots + E(L_k)}_{k}$$

and since $E(L_1) = E(L_2) = \dots = E(L_k)$

$$E(X_{(k)}) = \frac{ka}{n+1}$$

You can also compute this expectation by evaluating the integral [3/18/98, 5]:

$$E(X_{(k)}) = \int_{-\infty}^{\infty} t \underbrace{f(t)}_{\text{we determined the density earlier [3/18/98, 8-10]}} dt$$
$$= \int_0^a t \overbrace{\binom{n}{k} \frac{k\, t^{k-1} (a-t)^{n-k}}{a^n}} dt$$

Same result. So we also have another identity:

$$\int_0^a \binom{n}{k} \frac{k\, t^k (a-t)^{n-k}}{a^n} dt = \frac{ka}{n+1}$$

This is the easy stuff.
The hard stuff is to find the maximum length of a gap.
To solve this, we have to use all the machinery of conditional probability, in the continuous case.

Problem 6b of PS #6 is an open question.
In the Pólya Urn process, no conjugate prior is known for Bayes Theory.

Continuous Random Variables (cont'd)

I want to really rub in order statistics

Order statistics (cont'd)

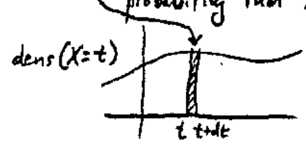
Let X = continuous random variable

Then:

$$\underbrace{P(a < X \le b)}_{\substack{\text{probability that } X \\ \text{lies bt any two} \\ \text{numbers } a + b}} = \overbrace{\int_a^b \underbrace{\mathrm{dens}(X=t)}_{}\, dt}^{\text{integrating the density gives the probability}}$$

density itself doesn't mean anything.
It is not a probability.

$\mathrm{dens}(X=t)\,dt$ is, in a manner of speaking, the probability that X takes values between t and $t+dt$.

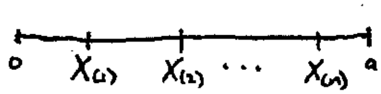


This is meaningless.
This discussion went on for 200 years after the discovery of the calculus.

People would say this is meaningless.
And it is meaningless.

Nowadays there is something called non-standard analysis, whereby you can make $\mathrm{dens}(X=t)$ meaningful if you talk about infinitesimal probability.
It comes at a cost of enlarging the real numbers, adding infinitesimals, which can be done in a very simple way.
It hasn't taken on yet, but eventually, people will catch on.

Meanwhile, we have to get used to the density.
The order statistics we saw comes from the Dirichlet process, where we drop n points on the interval $(0,a]$, by looking at the points in the order in which they appear on the interval



the closest point to 0, after dropping n points, is
$X_{(1)}$ (read, "X order 1"),
next is $X_{(2)}$, and so on,

The <u>uniform process</u>, together with the <u>order statistics</u>, is the basic statisical gadget we work with.

We take n measurements in a given interval.
We want to really study what probability can say about this basic process, taking n measurements.

We saw, last time, how to compute the density of the order statistics:

$$\text{dens}\left(X_{(k)} = t\right)$$

Let's do it again – with <u>less feeling</u>.
We've done this twice already.
Once – by computing the <u>cummulative distribution</u>

Remember, the cummulative distribution is:

$$\underbrace{P\left(X_{(k)} \le t\right) = F_{X_{(k)}}(t)}_{\text{This is a genuine probability}}$$

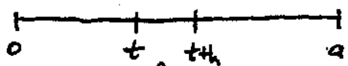
From this, we get the <u>density</u> by taking the derivative of the cummulative distribution:

$$\text{dens}\left(X_{(k)} = t\right) = \frac{d}{dt} F_{X_{(k)}}(t)$$

Which, by the Fundamental Theorem of the calculus, is the same as saying:

$$F_{X_{(k)}}(t) = \int_{-\infty}^{t} \text{dens}\left(X_{(k)} = s\right) ds$$

We tend to poo-poo this, but it's highly non-trivial.
This is the whole calculus. 200 years ago, people would shout at each other about this.
We really see the calculus in action when dealing with density.
So, I want to do it again to really rub it in.
Unless you unanimously rise and say "No."



$X_{(k)}$ lies in $(t, t+h]$

Let $U(t)$ = number of X_i in $[0, a)$

Computations w/ the Dirichlet Process are performed by <u>reducing</u> it to the <u>random function</u> that is associated w/ the Dirichlet Process.

event $(t < X_{(k)} \le t+h)$ =

$$\left.\begin{aligned} &(U(t) = k-1) \cap ((U(t+h) = k) \cup (U(t+h) = k+1) \cup \dots \cup (U(t+h) = n)) \\ \cup\ &(U(t) = k-2) \cap ((U(t+h) = k) \cup (U(t+h) = k+1) \cup \dots \cup (U(t+h) = n)) \\ \cup\ &\vdots \\ \cup\ &(U(t) = 0) \cap ((U(t+h) = k) \cup (U(t+h) = k+1) \cup \dots \cup (U(t+h) = n)) \end{aligned}\right\}$$

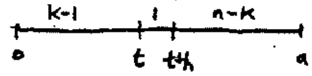
We can simplify by the distributive law:

$$\left.\begin{aligned} &\boxed{((U(t) = k-1) \cap (U(t+h) - U(t) = 1))} \cup ((U(t) = k-1) \cap (U(t+h) - U(t) = 2)) \cup \dots \\ \cup\ &((U(t) = k-2) \cap (U(t+h) - U(t) = 2)) \cup ((U(t) = k-2) \cap (U(t+h) - U(t) = 3)) \cup \dots \\ \cup\ &\vdots \\ \cup\ &((U(t) = 0) \cap (U(t+h) - U(t) = k)) \cup ((U(t) = 0) \cap (U(t+h) - U(t) = k+1)) \cup \dots \end{aligned}\right\}$$

So we have that:

$$\text{event } (t < X_{(k)} \le t+h) = \underbrace{((U(t) = k-1) \cap (U(t+h) - U(t) = 1))}_{} \cup \text{ disjoint events that have 2 or more points in } (t, t+h]$$

{i.e., all but first term (boxed) disjunction above}



Therefore, the probability is:

$$P(t < X_{(k)} \le t+h) = \binom{n}{k-1,\, 1,\, n-k} \left(\frac{t}{a}\right)^{k-1} \left(\frac{h}{a}\right)^{1} \left(\frac{a-t-h}{a}\right)^{n-k} + \underbrace{O\left(\left(\frac{h}{a}\right)^2\right)}_{\text{from at least 2 points in } (t,t+h]}$$

In terms of cummulative distributions, we have:

$$P(t < X_{(k)} \le t+h) = \underbrace{F_{X_{(k)}}(t+h)}_{P(X_{(k)} \le t+h)} - \underbrace{F_{X_{(k)}}(t)}_{P(X_{(k)} \le t)}$$

which gives:

$$F_{X_{(k)}}(t+h) - F_{X_{(k)}}(t) = \binom{n}{k-1,\ 1,\ n-k}\left(\frac{t}{a}\right)^{k-1}\left(\frac{h}{a}\right)\left(\frac{a-t-h}{a}\right)^{n-k} + O\left(\left(\frac{h}{a}\right)^2\right)$$

divide by h, and take limit as $h \to 0$:

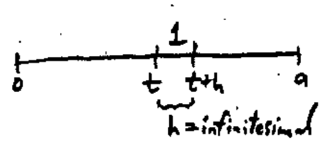
$$\underbrace{\lim_{h\to 0} \frac{F_{X_{(k)}}(t+h) - F_{X_{(k)}}(t)}{h}} = \binom{n}{k-1,\ 1,\ n-k}\left(\frac{t}{a}\right)^{k-1}\left(\frac{1}{a}\right)\left(\frac{a-t}{a}\right)^{n-k} + \underbrace{\cancel{O(h)}}_{} \to 0$$

This is precisely the derivative of the cummulative distribution, which is exactly the density.

{as they said 200 years ago, you drop higher order infinitesimals.}

$$\text{dens}(X_{(k)} = t) = \underbrace{\binom{n}{k-1,\ 1,\ n-k}}_{\frac{n!}{(k-1)!(n-k)!}}\left(\frac{t}{a}\right)^{k-1}\left(\frac{1}{a}\right)\left(\frac{a-t}{a}\right)^{n-k}$$

- Note that we could have argued, from the start, that no more than 1 point could fall in $(t, t+h]$, as h is infinitesimal:



From now on, we're going to argue that way.
I'm not going to go through this again.
The whole idea of the calculus is to simplify.

- Plotting the curves of the density, $\text{dens}(X_{(k)} = t)$, tells you how the probability is distributed.
Is the density lumped on one side? Then it's more probable that event occurs on that side.
You understand the probability of a continuous random variable by looking at the density, not by looking at the cummulative distribution.
If you plot the density, you see the lumps.
Ex: unimodal density — one lump
bimodal " — two lumps

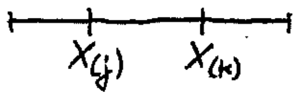
{It's very instructive to plot $\text{dens}(X_{(k)} = t)$ to see how various order statistics fit.}

Some interesting random variables for order statistics: 3/20/98.5

- Range $\triangleq X_{(n)} - X_{(1)}$
- Midpoint, when n is odd $\triangleq X_{\left(\frac{n-1}{2}\right)}$

You can go on forever w/ some very delicate computations of this kind.

- Let's take now two order statistics, $X_{(j)}$ and $X_{(k)}$:



We ask: How are $X_{(j)}$ and $X_{(k)}$ correlated?

These are clearly not independent random variables.
Therefore, we have to develope the analogue for the joint distribution of integer random variables.
Now we need the continuous analogue of the joint distribution.

Then we have to do a complete computation of the correlation of all the order statistics.
All possible ones.
Sorry – that's the way it is.

This requires that we first develope the theory of joint distribution of continuous random variables. This is farraginous.
Don't let yourself be distracted by this.
(see comments at end of [3/20/98.7]

{consisting of a confused mixture. A hodgepodge}

- Joint distribution of 2 continuous random variables

First, the bad news: you have to use partial derivatives, ∂ and multiple integrals, $\iint$ } The definitions involve these.

The good news – in practice, there are all these tricks to avoid them.

Let X, Y be two continuous random variables.

$$\underbrace{P((X\le t)\cap(Y\le s))}_{\text{a genuine, garden variety probability.}} = \int_{-\infty}^{s}\int_{-\infty}^{t} \overbrace{\text{dens}(X=u, Y=v)}^{\text{joint density}}\, du\, dv$$

NB: this is not the intersection of 2 sets
These are not events

joint cummulative distribution

Random variables $X + Y$ are jointly continuous when their joint cummulative distribution function is obtained by integrating a function of these variables that is called the joint density.

Q/A: everything is non-negative, so the order of integration doesn't matter.

$$\iint \text{dens}(X=u, Y=v)\, du\, dv = \iint \text{dens}(X=u, Y=v)\, dv\, du$$

$\text{dens}(X=t, Y=s)$ is called the joint density of X and Y

Let's fool around a little bit.

- The marginal density $\text{dens}(X=t)$ is obtained by integrating out the Y variable:

$$\text{dens}(X=t) = \int_{-\infty}^{\infty} \text{dens}(X=t, Y=s)\, ds$$

and similarily:

$$\text{dens}(Y=s) = \int_{-\infty}^{\infty} \text{dens}(X=t, Y=s)\, dt$$

- The probability $P(X \le t)$ is:

$$P(X \le t) = \int_{-\infty}^{t} \int_{-\infty}^{\infty} \text{dens}(X=u, Y=v)\, dv\, du$$

the Y variable is integrated out

and then integrating w/ these limits gives the cummulative distribution $P(X \le t)$.

- The basic formula to remember is:

$$P((X \le t) \cap (Y \le s)) = \int_{-\infty}^{s} \int_{-\infty}^{t} \text{dens}(X=u, Y=v)\, du\, dv$$

Rewriting this in a slightly different form:

$$P\big((a < X \le b) \cap (c < Y \le d)\big) = \int_{t=a}^{b} \int_{s=c}^{d} dens(X=t, Y=s)\, ds\, dt$$

useful notation to keep track of limits and variables of integration.

This tells you that <u>any event</u> involving two continuous random variables can be computed, and the <u>probability</u> of any such event can be computed by <u>integrating</u> the <u>joint density</u>.

- Lastly, by the Fundamental Theorem of the Calculus, the partial derivative of the joint cummulative distribution is the joint density:

$$\underbrace{dens(X=t, Y=s)}_{\text{joint density}} = \frac{\partial^2}{\partial t\, \partial s} \underbrace{P((X \le t, Y \le s))}_{\text{joint cummulative distribution}}$$

- You can go on forever w/ these formulas and in some statistics books, you have 20 pages doing this stuff w/ $i = 2, \ldots, n$ random variables.
  Your eyes can't see anymore.
  It's clear that once you understand this for $i=2$ that you can see what happens for $i=3, \ldots,$ any number of random variables.

John Guidi
guidi@mit.edu

18.313

Lecture 20 (Sara Billey)
3/30/98.1

Today : finish Chapt 3
joint distributions/densities
expectation (again)

Wed : in-class review

Fri : Quiz

it doesn't matter if they are continuous, integer, etc.

• **Def** The joint distribution of any two random variables X, Y is

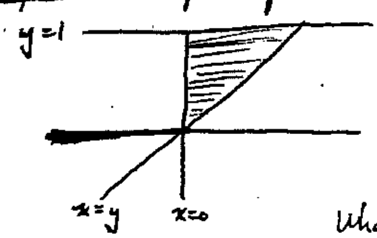
$$\underbrace{P(X \le x, Y \le y)}$$

Caution: In the literature, it's a totally standard technique to use
lower-case x, y for the variables
upper-case X, Y for the random variables
This does get people into trouble, as you're using a letter in two ways.

• **Def** If X, Y are continuous random variables, then their joint density is:

$$\text{dens}(X=x, Y=y) = f(x,y) = \underbrace{\frac{\partial}{\partial x}\frac{\partial}{\partial y} P(X \le x, Y \le y)}_{\text{derivative of the probability function}}$$

• **Example** – Drop a point uniformly on triangle enclosed by $x=0, y=1, x=y$:



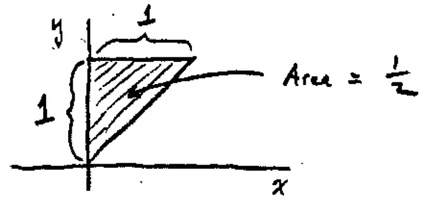
Let continuous r.v. X, Y be the coordinates of the drop point.

What does it mean to say that a point is dropped uniformly on this triangle?
It's a statement about what the density function should look like.

uniformly implies density function is a constant $\Rightarrow$ $\text{dens}(X=x, Y=y) = c$ ← a constant

We want to figure out what this constant c is.

To get this joint density function, we know that if we integrate it over the whole space, we get 1.
The whole space here is over a certain area.
What's the area?



We know that:

$$\underbrace{P(X \le 1, Y \le 1)}_{\text{joint distribution}} = 1$$

$$P(X \le 1, Y \le 1) = \int_y \int_x f(x,y)\, dx\, dy$$

We've just argued that:

$$f(x,y) = dens(X = x, Y = y) = c \quad \text{when } 0 \le x \le y \le 1$$

$$= \begin{cases} \int_0^1 \int_0^y c \; dx\, dy & \text{when } 0 \le x \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

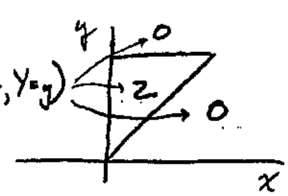
$$= \begin{cases} c * \overbrace{\int_0^1 \int_0^y dx\, dy}^{\text{area of triangle} = \frac{1}{2}} & \text{when } 0 \le x \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$P(X \le 1, Y \le 1) = \begin{cases} \frac{c}{2} & \text{when } 0 \le x \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

This gives:

$$1 = \frac{c}{2} \implies c = 2$$

dens(X=x, Y=y)



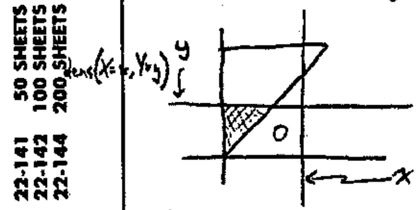
$$\text{joint density} = \begin{cases} 2 & 0 \le x \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{dens}(X=x, Y=y) = \begin{cases} 2 & 0 \le x \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Once you have the density function, we can get anything we want.
From this, we can compute $P(X \le x, Y \le y)$

$$P(X \le x, Y \le y)$$

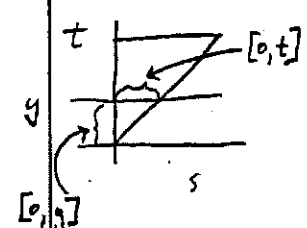
- Case 1 – $0 \le y \le x \le 1$



$y \le x$

$$P(X \le x, Y \le y) = \iint_{y\ x} \overbrace{\text{dens}(X=x, Y=y)}^{\text{we've just shown this is 2 within the triangle}}\, dx\, dy$$

To keep our sanity, lets switch the variables of integration:
$x \to s$
$y \to t$



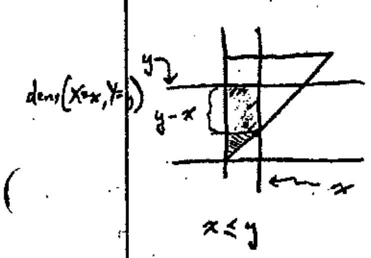
$$= \iint_{t\ s} 2\ ds\, dt$$

$$= \int_0^y \underbrace{\int_0^t 2\ ds}\, dt$$

$$= 2\int_0^y t\, dt$$

$$P(X \le x, Y \le y) = y^2 \quad \text{when} \quad 0 \le y \le x \le 1$$

- Case 2 – $0 \le x \le y \le 1$



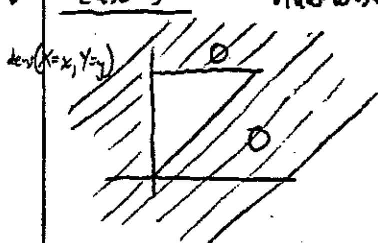
$$P(X \le x, Y \le y) = \iint_{y\ x} \overbrace{\text{dens}(X=x, Y=y)}^{\text{2 everywhere w/in triangle}}\, dx\, dy$$

$$= (2 * \text{area triangle}) + (2 * \text{area rectangle})$$

$$= 2\left(\frac{x^2}{2}\right) \qquad + 2(y-x)x$$

$$P(X \le x, Y \le y) = y^2 + 2x(y-x) \quad \text{when} \quad 0 \le x \le y \le 1$$

- case 3 - otherwise



$$P(X \le x, Y \le y) = \iint_{y\,x} \overbrace{dens(X=x, Y=y)}^{0}\, dx\,dy$$

$$P(X \le x, Y \le y) = 0$$

And we have:

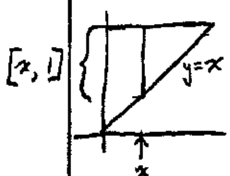
$$P(X \le x, Y \le y) = \begin{cases} y^2 + 2x(y-x) & \text{when } 0 \le x \le y \le 1 \\ y^2 & 0 \le y \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

- Are the continuous random variables X and Y of this example uniformly distributed?

We look at the marginal density functions, which are obtained from the joint density function.
If the marginal density functions are constants, then X and Y are uniformly distributed.

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$

integrate away the y component.
Integrate over all values of y allowed by a particular x.



$$= \int_x^1 2\,dy$$

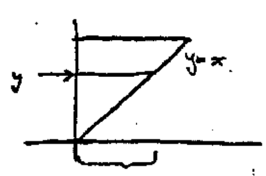
{we need only concern ourselves w/ limits of integration $[x,1]$, because outside of that region the density function is 0. $(f(x,y)=0)$}

$$f_X(x) = 2(1-x)$$

X is not uniformly distributed, as $f_X(x)$ is not constant.

Similarly:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$



$$= \int_0^y 2\,dx$$

$[0, y]$ all values of x allowed by a particular y

$$f_Y(y) = 2y$$

Y is not uniformly distributed, as $f_Y(y)$ is not constant.

- Are the continuous random variables X and Y of this example <u>independent</u>?

  How can we test for this?

  Test 1 : $P(X \le x, Y \le y) \stackrel{?}{=} P(X \le x)\,P(Y \le y)$

  alternatively:

  Test 2 : $\text{dens}(X = x, Y = y) \stackrel{?}{=}$ product of the marginals $f_X(x)\,f_Y(y)$

These tests and independence are all equivalent.

In this example:

$$\text{dens}(X=x, Y=y) \stackrel{?}{=} f_X(x)\,f_Y(y)$$

$$2 \stackrel{?}{=} 2(1-x)\,2y$$

<u>No</u>!

Product of marginals does not equal 2.

Therefore X and Y are <u>not</u> independent r.v.

Let's do a fancier example: 3/30/98.6

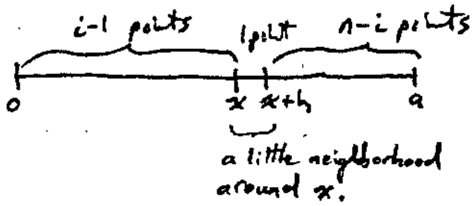
- Joint densities and joint distributions for the order statistics for the uniform process.

Uniform process – dropping n points on interval $[0, a]$

$X_{(i)}$ = position of the i^{th} largest point

Recall the density function for the i^{th} order statistic:

$$\text{dens}(X_{(i)} = x) = \binom{n}{i-1,\ 1,\ n-i}\left(\frac{x}{a}\right)^{i-1}\left(\frac{1}{a}\right)\left(\frac{a-x}{a}\right)^{n-i}$$

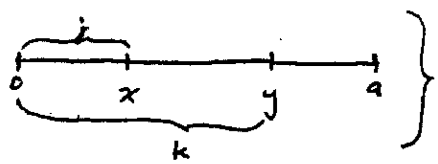


$$\frac{n!}{(i-1)!\,1!\,(n-i)!} = n\frac{(n-1)!}{(i-1)!\,(n-1-i+1)!} = n\binom{n-1}{i-1}$$

$$\text{dens}(X_{(i)} = x) = n\binom{n-1}{i-1}\left(\frac{x}{a}\right)^{i-1}\left(\frac{1}{a}\right)\left(\frac{a-x}{a}\right)^{n-i} \quad \text{where } 0 \le x \le a$$

Let $X_{(j)}$ and $X_{(k)}$ be random variables such that $j < k$.
We want to compute the joint distribution:

$$P(X_{(j)} \le x,\ X_{(k)} \le y) = \binom{n}{j,\ k-j,\ n-k}\left(\frac{x}{a}\right)^{j}\left(\frac{y-x}{a}\right)^{k-j}\left(\frac{a-y}{a}\right)^{n-k}$$



exactly j fall in $(0, x]$
exactly $k-j$ fall in $(x, y]$
exactly $n-k$ fall in $(y, a]$

$$+ \underbrace{\quad \cdots \quad}$$

all terms where more than j fall in $(0, x]$
and where more than k fall in $(0, y]$

So $P(X_{(j)} \le x, X_{(k)} \le y)$ is a big expression, with lots of terms.

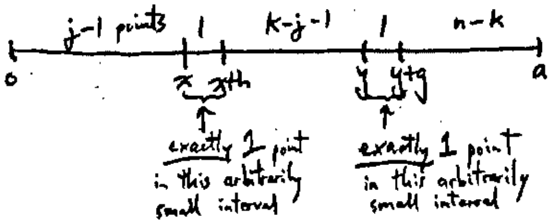
Recall that the density is:

$$\text{dens}(X_{(j)} = x,\ X_{(k)} = y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y} P(X_{(j)} \le x,\ X_{(k)} \le y)$$

to get the density, we take the derivative of the probability.

It turns out that when we take the derivative, that the ... terms for $P(X_{(j)} \le x, X_{(k)} \le k)$ are zero. So we could simply take the derivative.

Alternatively, what's the probability of dropping points with arbitrarily small intervals $[x, x+h]$, $[y, y+g]$?



Let $A_{h,g}$ = event that $X_{(j)}$ (i.e., the j^{th} order statistic) falls in $[x, x+h]$ and $X_{(k)}$ falls in $[y, y+g]$

$$\text{dens}\left(X_{(j)} = x, X_{(k)} = y\right) = \lim_{h\to 0} \lim_{g \to 0} \frac{P\left(A_{h,g}\right)}{hg}$$

probability of event $A_{h,g}$ is not hard to figure out, as we have exact occupation numbers in each interval.

$$= \lim_{h\to 0} \lim_{g\to 0} \frac{\binom{n}{j-1,1,k-j-1,1,n-k}\left(\frac{x}{a}\right)^{j-1}\left(\frac{\cancel{h}}{a}\right)\left(\frac{y-\cancel{x-h}}{a}\right)^{k-j-1}\left(\frac{\cancel{g}}{a}\right)\left(\frac{a-\cancel{y-g}}{a}\right)}{\cancel{hg}}$$

$$\text{dens}\left(X_{(j)} = x, X_{(k)} = y\right) = \begin{cases} \binom{n}{j-1,1,k-j-1,1,n-k}\left(\frac{x}{a}\right)^{j-1}\left(\frac{1}{a}\right)\left(\frac{y-x}{a}\right)^{k-j-1}\left(\frac{1}{a}\right)\left(\frac{a-y}{a}\right)^{n-k} & \text{when } 0 \le x \le y \le a \\ 0 & \text{otherwise} \end{cases}$$

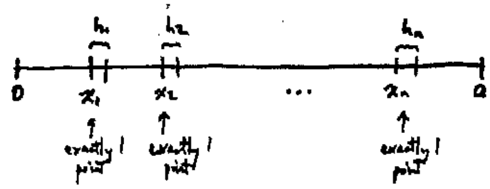
Warning: Note that $\text{dens}(X_{(j)} = x, X_{(k)} = y)$ looks a lot like $P(X_{(j)} \le x, X_{(k)} \le y)$ for this problem.
Do not get confused by this. Density function and probability distribution function are different beasts.

Remember: Density can be bigger than 1, unlike probability distribution function. There's no upper bound on density function.
(density does have to be non-negative)

- what if want the joint density for <u>all</u> of the order statistics?

What is $\text{dens}\left(X_{(1)} = x_1, \ldots, X_{(n)} = x_n\right)$?

Divide the interval $[0, a]$ into intervals which include arbitrarily small intervals about the order statistics:



$$\text{dens}\left(X_{(1)}=x_1, \ldots, X_{(n)}=x_n\right) = \lim_{h_1, h_2, \ldots, h_n \to 0} \frac{P\left(\bigcap_{i=1}^{n} \left(x_i < X_{(i)} \le x_i + h_i\right)\right)}{h_1 \cdot h_2 \cdot \ldots \cdot h_n}$$

$$= \lim_{h_1, h_2, \ldots, h_n \to 0} \frac{\binom{n}{1, 1, \ldots, 1} \prod_{i=1}^{n} \left(\frac{\cancel{h_i}}{a}\right)}{\prod_{i=1}^{n} \cancel{h_i}}$$

$$\binom{n}{\underbrace{1, 1, \ldots 1}_{n}} = \frac{n!}{1!1! \ldots 1!} = n!$$

$$\text{dens}\left(X_{(1)}=x_1, \ldots, X_{(n)}=x_n\right) = \frac{n!}{a^n} \quad \text{when } 0 < x_1 < x_2 < \ldots < x_n < a$$

The <u>joint density function</u> for <u>all</u> the order statistics.

There are whole books on order statistics.
For example, there are a lot of nice tests for determining whether the random number generator on a computer actually provides <u>random</u> numbers. You can do this by checking the order statistics.
If the <u>distribution</u> for the <u>order statistics</u> of your <u>random number generator</u> <u>doesn't come out right</u>, perhaps you have a <u>bug in your code</u>.

- Integration of the density function over the whole sample space gives 1.

$$\iint \cdots \int dens(X_{(1)}=x_1, \dots, X_{(n)}=x_n)\, dx_1\, dx_2 \dots dx_n = 1$$

$$\int_0^a \int_0^{x_{n-1}} \cdots \int_0^{x_2} \frac{n!}{a^n}\, dx_1\, dx_2 \dots dx_n = 1$$

the reason the integration limits are like this is because, for example,
the 1^{st} order statistic lies between 0 and x_2
the 2^{nd} " " " " 0 and x_3
etc.

Which gives the cool integral identity (try proving this directly from calculus):

$$\int_0^a \int_0^{x_{n-1}} \cdots \int_0^{x_2} dx_1\, dx_2 \dots dx_n = \frac{a^n}{n!}, \quad 0<x_1<x_2<\dots<x_n<a$$

- <u>Expectation</u>

For a continuous random variable, we have:

$$E(X) = \int_{-\infty}^{\infty} x \; dens(X=x)\, dx$$

Back to our example [3/30/98.4]:

given the density function : $f_Y(t) = 2t, \quad 0 \le t \le 1$
the expected value of the random variable Y is:

$$\left.\begin{aligned} E(Y) &= \int_{-\infty}^{\infty} t\, f_Y(t)\, dt \\ &= \int_0^1 t\,(2t)\, dt \\ &= \int_0^1 2t^2\, dt \\ &= \frac{2t^3}{3}\Big|_0^1 = \frac{2}{3} \end{aligned}\right\} E(Y) = \frac{2}{3}$$

- The main trick w/ expectations is the following <u>basic fact</u>:

For any continuous random variables X, Y,

$$E(X+Y) = E(X) + E(Y)$$

<u>Proof</u>

The proof practically jumps right out from the definition:

$$E(X+Y) = \iint (x+y) \underbrace{f(x,y)}_{\text{joint density}} dx\,dy$$

By linearity of integrals, this can be separated into:

$$= \iint x\, f(x,y)\, dx\,dy + \iint y\, f(x,y)\, dx\,dy$$

$$= \int x \underbrace{\int f(x,y)\, dy}_{f_X(x)}\, dx + \int y \underbrace{\int f(x,y)\, dx}_{f_Y(y)}\, dy$$

$\longleftarrow$ marginals $\longrightarrow$

$$= \underbrace{\int x f_X(x)\, dx}_{E(X)} + \underbrace{\int y f_Y(y)\, dy}_{E(Y)}$$

$$E(X+Y) = E(X) + E(Y) \quad \checkmark$$

As a nice example of this:
Expectation of order statistics

Anytime someone says "expectation," think Basic Fact $\Rightarrow E(X+Y) = E(X) + E(Y)$
How can I break it up into the expectation of a sum of random variables?
This should be a knee jerk reaction.

We want $E(X_{(i)})$

The way we do this is to look at the gaps

$$\left.\begin{aligned} &\underset{\text{the first gap}}{L_1} = X_{(1)} \\ &L_i = X_{(i)} - X_{(i-1)} \end{aligned}\right\} \text{ gaps in a uniform process}$$

If we can figure out the expectation for the gaps $E(L_i)$, then we can clearly figure out $E(X_{(i)})$

$E(L_1 + \dots + L_{n+1}) = a \longleftarrow$ The expectation of the sum of all the gaps – we already know this, as the sum of all the gaps must cover the whole interval $[0, a]$.

Also, by the basic fact of expectations:

$$E(L_1 + \dots + L_{n+1}) = E(L_1) + \dots + E(L_{n+1})$$

Previously ([3/18/98.12-13]), we argued that the gaps are all identically distributed (equidistributed).

The intuitive circle argument.

$$E(L_1 + \dots + L_{n+1}) = (n+1)E(L_1)$$

Thus: $E(L_1) = \dfrac{a}{n+1}$

From the definition of gap, we have: $L_i = X_{(i)} - X_{(i-1)}$

$$\Rightarrow X_{(i)} = L_i + X_{(i-1)}$$

Successive applications of gap definitions gives:

$$X_{(i)} = L_i + \underbrace{X_{(i-1)}}_{L_{i-1} + \underbrace{X_{(i-2)}}_{L_{i-2} + \underbrace{X_{(i-3)}}_{\vdots}}}$$

$$+ L_1$$

$$X_{(i)} = L_1 + L_2 + \dots + L_i$$

Taking expectations:

$$\left.\begin{aligned} E(X_{(i)}) &= E(L_1 + L_2 + \dots + L_i) \\ &= E(L_1) + E(L_2) + \dots + E(L_i) \end{aligned}\right\} \text{from basic fact of expectations}$$

We just showed that:

$$E(L_1) = E(L_2) = \dots = E(L_i) = \frac{a}{n+1}$$

$$= i\,E(L_1)$$

$$E(X_{(i)}) = i\,\frac{a}{n+1}$$

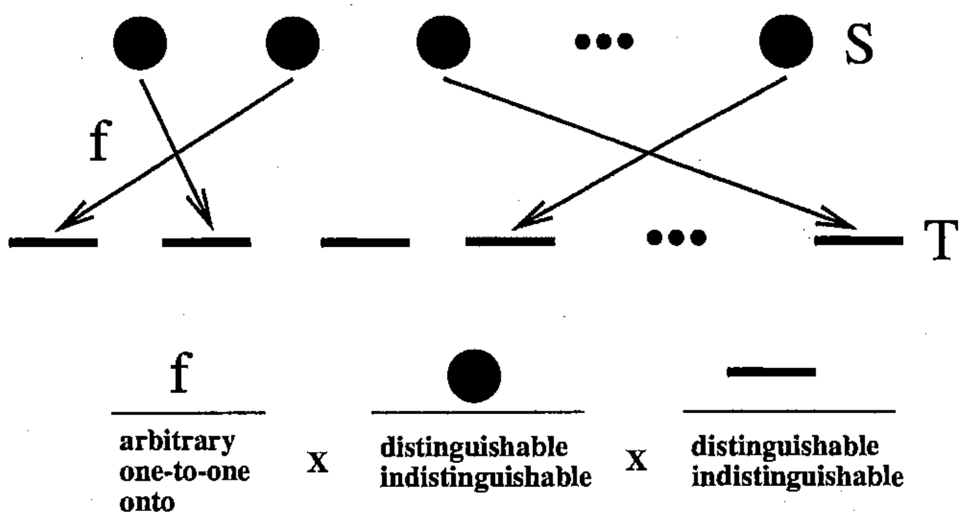
I did not attend this lecture and no tape of the lecture is available.

The lecture was an <u>in-class review</u> of the material to be tested by the Quiz on Friday.

Within the text, pagination is of the form *lecture_date.page*. Note that a different convention is used in both the table of contents and the index, where pagination is of the form *lecture_number.page*. The table of contents provides the mapping between *lecture_number* and *lecture_date*. For example, page 1.1 in the table of contents and index corresponds to page 2/4/98.1 in the text, where the topic anomaly detection is discussed.

Index

Probability



John N. Guidi
Lecture Notes - Spring 1998
MIT Course 18.313
Professor Gian-Carlo Rota
Volume 2 of 2
(Lectures 22–35, Super Classes 1–4)

John N. Guidi
Visiting Scholar in Applied Mathematics
Department of Mathematics
Room 2-392
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139
Email: guidi@mit.edu

These *Lecture Notes* originated from the lectures presented by Gian-Carlo Rota, Professor of Applied Mathematics, for course 18.313 - Probability, which he taught at MIT, during the Spring 1998 semester. The *Lecture Notes* were produced from notes I made during class, audio recordings I made of lectures, as well as clarifications and expansions I made of the material presented, after the fact. Although no attempt was made to provide a literal transcription of the lectures, an effort was made to present the material here in the same spirit and I hope to have introduced only a limited number of errors. The *Lecture Notes* are comprised of a two volume set. Volume 1 contains lectures 1–22, volume 2 contains lectures 22–35 and super class lectures 1–4.

Sara Billey, Assistant Professor of Applied Mathematics, who was the course director and whose recitations I attended, presented lectures 20, 21, and 27. The other recitation instructors were Alex Perlin and Ioana Popescu. Their efforts clearly contributed to the success of this course.

I was absent from lecture 25 and my thanks go to Jeff Lieberman, who taped the lecture for me and provided his notes, as well as to Ira Gerhardt and Rick Monte, who also provided their notes of this class.

During the course, I found that preparation of these *Lecture Notes* was a particularly useful way to really understand the material (or, as Professor Rota is fond of saying, "*to really rub it in*"). My gratitude goes to him for imparting his enthusiasm and sharing his profound knowledge of mathematics, as well as for presenting these superb lectures on probability.

John N. Guidi
October 15, 1998

Within the text, pagination is of the form *lecture_date.page*. Note that a different convention is used in both the table of contents and the index, where pagination is of the form *lecture_number.page*. The table of contents provides the mapping between *lecture_number* and *lecture_date*. For example, page 1.1 in the table of contents and index corresponds to page 2/4/98.1 in the text, where the topic anomaly detection is discussed.

Contents

Lecture 3 - [2/9/98]

Lecture 4 - [2/11/98]

Lecture 5 - [2/13/98]

Lecture 9 - [2/23/98]

Lecture 10 - [2/25/98]

Lecture 15 - [3/11/98]

Lecture 16 - [3/13/98]

After Notes Lecture AN16 - [3/16/98]

Lecture 17 - [3/16/98]

Lecture 18 - [3/18/98]

Lecture 19 - [3/20/98]

Lecture 27 (Sara Billey) - [4/17/98]

Lecture 28 - [4/22/98]

Lecture 29 - [4/24/98]

Lecture 33 - [5/6/98]

Lecture 34 - [5/8/98]

Lecture 35 - [5/11/98]

Super Class Lecture SC1 - [2/27/98]

Super Class Lecture SC2 - [3/13/98]

Super Class Lecture SC3 - [4/10/98]

Super Class Lecture SC4 - [4/24/98]

Index

Within the text, pagination is of the form *lecture_date.page*. Note that a different convention is used in both the table of contents and the index, where pagination is of the form *lecture_number.page*. The table of contents provides the mapping between *lecture_number* and *lecture_date*. For example, page 1.1 in the table of contents and index corresponds to page 2/4/98.1 in the text, where the topic anomaly detection is discussed.

Research Problems

John Guidi
guidi@mit.edu

We're going to change the policy on problem sets from now on, because the material is getting very tough.
I'll give you problem sets that are very closely related to the material given in class. Instead of giving fancy, unsolved problems.
From now on, we're getting to the absolute heart of probability, as well as some of the toughest chapters of probability.
This is what you really have to learn. So far, balls into boxes and all that stuff you can visualize it.
When it comes to continuous probability, you really have to work on it to get it — to develope the concepts behind it.
I'm going to give you problems that are really lecture based from now on.
No fancy stuff, no unsolved problems. Well, maybe unsolved problems.

The ideal thing is that at the end of this course, each one of you does a UROP with me and pulls out one of my unsolved problems and solves it + we write a joint paper.
You are young, you have lot's of time.
Help me out.

Continuous conditional probability (beginning)

Now we are doing a very tough concept.

Let X = continuous random variable

Then we know that:

$$P(a < X \leq b) = \int_a^b dens(X=t)\, dt$$

In particular, we have that:

$$\underbrace{P(X \leq t) = \int_{-\infty}^{t} dens(X=s)\, ds}_{\text{This is called the cummulative distribution.}}$$

The density in itself doesn't have a meaning as a probability.
It has a meaning only when integrated.
The density is always a non-negative quantity.

Our objective is to learn to work w/ these densities, even though they do not have a direct meaning.

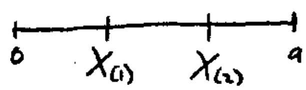
- The toughest thing we are facing is to find the right definition, as well as interpretation, of this formula:

$$\underbrace{P(A \mid X=t)}_{} = ?$$

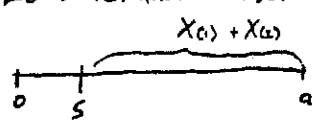
probability of event A, given that $X=t$, where X is a continuous random variable.
So it's a conditioning relative to a density.
This is what we have to work on.
Let's work up to it gradually, by increasing difficulty + working up to that.

- Let's start the whole discussion w/ one of the simplest examples we have seen.
Let's take the interval $[0,a]$ and pick two points.

Dirichlet Process with $n=2$ $\Big\}$ so we have two order statistics, $X_{(1)}$ and $X_{(2)}$



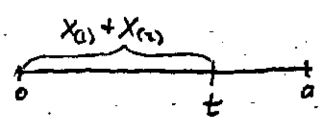
Let's remind ourselves that:



$$P(X_{(1)} > s) = \binom{2}{0,2}\left(\frac{a-s}{a}\right)\left(\frac{a-s}{a}\right)$$ ← both marbles fall between s and a, i.e., $(s,a]$.

$$= \frac{(a-s)^2}{a^2}$$

and, as well:



$$P(X_{(2)} \le t) = \binom{2}{2,0}\left(\frac{t}{a}\right)^2$$ [redacted] ← both marbles fall between 0 and t, i.e., $(0,t]$.

$$= \frac{t^2}{a^2}$$

Since we know $P(X_{(1)} > s)$, the cummulative distribution for the first order statistic is:

$$P(X_{(1)} \le s) = 1 - P(X_{(1)} > s)$$

$$= 1 - \frac{(a-s)^2}{a^2}$$

- The densities are obtained by differentiating the cummulative distributions:

$$\text{dens}(X_{(1)} = s) = \frac{d}{ds} \underbrace{P(X_{(1)} \le s)}_{1 - \frac{(a-s)^2}{a^2}}$$

$$= \frac{2(a-s)}{a^2}$$

Similarily:

$$\text{dens}(X_{(2)} = t) = \frac{d}{dt} \underbrace{P(X_{(2)} \le t)}_{\frac{t^2}{a^2}}$$

$$= \frac{2t}{a^2}$$

- This thing is tough. Let's not kid ourselves.
  Inside the Dirichlet Process (Uniform Process), you find all of probability.
  There are unsolved questions all over the place.
  From such a simple thing as dropping n marbles, there is research going on.
  So this is tough.
  Genuinely tough.
  In my days, this would be a graduate course.

- In particular, we have:

$$\int_{-\infty}^{\infty} \text{dens}(X_{(1)} = s)\, ds = 1 \quad \text{, by definition} \quad \left\{ \begin{array}{c} \text{integral of density} \\ \text{is } 1 \end{array} \right\}$$

Thus:

$$\int_0^a \frac{2(a-s)}{a^2}\, ds = 1$$

} You can verify this by hand, but this is automatic. We get a definite integral cheap.

Similarily:

$$\int_{-\infty}^{\infty} dens\left(X_{(2)}=t\right) dt \quad = 1$$

gives the definite integral:

$$\int_0^a \frac{2t}{a^2}\, dt \quad = 1$$

- Let's jazz this up.
We now ask:

$$P\Big(\big(X_{(1)}>s\big) \cap \big(X_{(2)}>t\big)\Big) \quad = \underbrace{?}$$

I don't know.
Because to do a combinatorial count, I don't know how to fix it.
Hard to do directly.

(always assume $s<t$)

But I do know how to compute:

$$P\Big(\underbrace{\big(X_{(1)}>s\big)}_{\text{the first marble is } >s} \cap \underbrace{\big(X_{(2)}\le t\big)}_{\text{the second marble is } \le t}\Big) \longleftarrow \text{This is } \underline{\text{easy}}$$

Therefore both marbles fall in the interval $(s,t]$.
Two marbles falling in the same interval is easy because the marbles are independent.

$$P\Big(\big(X_{(1)}>s\big) \cap \big(X_{(2)}\le t\big)\Big) = \binom{2}{0,2,0}\left(\frac{t-s}{a}\right)^2$$

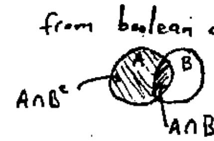
$$= \frac{(t-s)^2}{a^2}$$

Now we can use this probability to get the desired $P\big((X_{(1)}>s) \cap (X_{(2)}>t)\big)$.

By essential reasoning you've seen before, but don't mind seeing again:

$$\text{event } \big(X_{(1)}>s\big) \cap \big(X_{(2)}>t\big) = \big(X_{(1)}>s\big) - \Big(\big(X_{(1)}>s\big) \cap \big(X_{(2)}\le t\big)\Big)$$

from boolean algebra:



$$A \cap B = A - (A \cap B^c)$$

Therefore, because probabilities always subtract:

$$P\big((X_{(1)} > s) \cap (X_{(2)} > t)\big) = \underbrace{P\big(X_{(1)} > s\big)}_{\frac{(a-s)^2}{a^2}\ [4/6/98.2]} - \underbrace{P\big((X_{(1)} > s) \cap (X_{(2)} \le t)\big)}_{\frac{(t-s)^2}{a^2}\ \text{We just computed this.}}$$

$$P\big((X_{(1)} > s) \cap (X_{(2)} > t)\big) = \frac{(a-s)^2}{a^2} - \frac{(t-s)^2}{a^2} = \frac{2}{a^2}\left(\frac{a^2}{2} - sa - \frac{t^2}{2} + st\right)$$

I don't know of an intuitive justification of this.

Now, lets do the same the hard way.
By integrating densities. Painful, but good.

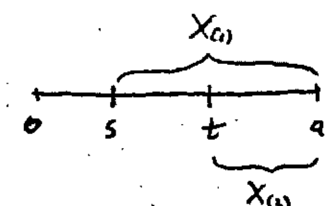
Let's find the joint density $X_{(1)}$ and $X_{(2)}$, namely: $\text{dens}(X_{(1)} = s, X_{(2)} = t)$.

We then obtain the joint distribution by integrating the joint density – just like the scriptures say:

NB: not the intersection of events

$$P\big((X_{(1)} > s) \cap (X_{(2)} > t)\big) = \iint_{\text{some set}} \text{dens}\big(X_{(1)} = u, X_{(2)} = v\big)\, du\, dv$$

Now, I've told you, in the strictist confidence, that this kind of computation is avoided most of the time. We just avoided it, because we did it directly. But for once, let's make it come out.



Roll up our sleeves and let's do it.
We have to find the limits of integration and do a double integral.

where: $s < t$
$s < X_{(1)} < X_{(2)}$
$t < X_{(2)} \le a$

What are the limits of integration?

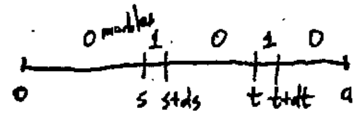
$$= \int_{v=t}^{a} \int_{u=s}^{v} \text{dens}\big(X_{(1)} = u, X_{(2)} = v\big)\, du\, dv$$

$X_{(2)} \le a$ $X_{(1)} < X_{(2)}$ $t < X_{(2)}$ $s < X_{(1)}$

- Next, we need to compute the <u>joint density</u>. We've done this several times.

$$\text{dens}\left(X_{(1)}=s, X_{(2)}=t\right) = ?$$

We use the <u>infinitesimal approach</u>



$$\text{dens}\left(X_{(1)}=s, X_{(2)}=t\right) = \lim_{\substack{ds\to 0\\ dt\to 0}} \frac{\binom{2}{0,1,0,1,0}\left(\frac{ds}{a}\right)\left(\frac{dt}{a}\right)}{ds\, dt}$$

This is, of course, an approximation argument by a limiting process.

$$\text{dens}\left(X_{(1)}=s, X_{(2)}=t\right) = \frac{2}{a^2}$$

Recall, from Dr. Billey's lecture [3/30/98.8], that the joint density function for n order statistics, derived in exactly the same fashion, is:

$$\text{dens}\left(X_{(1)}=x_1, \dots, X_{(n)}=x_n\right) = \frac{n!}{a^n}$$

- Now, with the limits of integration and the joint density, we can go back to our double integral:

$$P\left(\left(X_{(1)}>s\right) \cap \left(X_{(2)}>t\right)\right) = \int_{v=t}^{a} \underbrace{\int_{u=s}^{v} \overbrace{\frac{2}{a^2}}^{\text{dens}(X_{(1)}=u,\, X_{(2)}=v)} du}_{\frac{2(v-s)}{a^2}}\, dv$$

$$= \int_{v=t}^{a} \frac{2(v-s)}{a^2}\, dv$$

$$= \frac{2}{a^2}\left[\frac{v^2}{2} - sv\right]_{v=t}^{a}$$

$$= \frac{2}{a^2}\left(\frac{a^2}{2} - sa - \frac{t^2}{2} + st\right) \checkmark$$

$$= \frac{(a-s)^2}{a^2} - \frac{(t-s)^2}{a^2}$$

probability obtained by integration of joint density checks w/ directly obtained probability [4/6/98.4-5].

So it checks out.
In 90% of the cases, we avoid evaluating double integrals.

You think this is the end?
Sorry - this is just the beginning.

- Now let's compute the conditional probability:

$$\underbrace{P(X_{(2)} > t \,/\, X_{(1)} > s)}_{} = ?$$

Probability that $X_{(2)} > t$
given that $X_{(1)} > s$.

Let's do it 2 ways.
First, by the book.
Then, intuitively.
} This will justify our intuition, as often happens in life.

1) By the book:

$$P(X_{(2)} > t \,/\, X_{(1)} > s) = \frac{P((X_{(2)} > t) \cap (X_{(1)} > s))}{P(X_{(1)} > s)}$$ } by definition of conditional probability

{ The numerator we just computed two different ways.
The denominator we computed on [4/6/98.2]. }

$$= \frac{\frac{(a-s)^2}{\cancel{a^2}} - \frac{(t-s)^2}{\cancel{a^2}}}{\frac{(a-s)^2}{\cancel{a^2}}}$$

$$= 1 - \frac{(t-s)^2}{(a-s)^2}$$

2) By intuition:

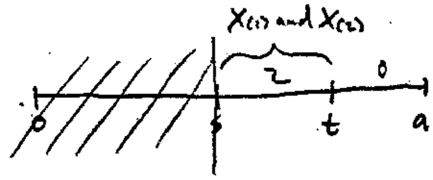
What is $P(X_{(2)} > t \,/\, X_{(1)} > s)$ telling you?

You know that no marble has fallen in the interval $(0, s]$.
Since it is given that $X_{(1)} > s$ (i.e., first order statistic strictly $> s$).

It's as if you were dealing w/ a shorter interval.
Namely, the interval from s to a.

Therefore, you recalculate this notion on the interval $[s, a]$.
This is what it's telling you. $P(X_{(2)} > t \,/\, X_{(1)} > s)$ is the probability distribution of the second marble, given that both marbles fall in the interval $[s, a]$.

Consider $P_{(X_{(1)}>s)}(X_{(2)} \le t)$



given that $X_{(1)}>s$, no balls drop in interval 0 to s, so we shorten interval to $[s, a]$.

$$P_{(X_{(1)}>s)}(X_{(2)} \le t) = \binom{2}{2,0}\left(\frac{t-s}{a-s}\right)^2$$

$$= \frac{(t-s)^2}{(a-s)^2}$$

And since we have:

$$P_{(X_{(1)}>s)}(X_{(2)}>t) = 1 - P_{(X_{(1)}>s)}(X_{(2)} \le t)$$

Our intuition is confirmed.

$$P(X_{(2)}>t \mid X_{(1)}>s) = 1 - \frac{(t-s)^2}{(a-s)^2} \quad ✓$$

Now we try another one that is a little tougher.

$$P(X_{(2)} > t \mid X_{(1)} \le s) = ?$$

By the definition of conditional probability:

$$P(X_{(2)} > t \mid X_{(1)} \le s) = \frac{P((X_{(2)} > t) \cap (X_{(1)} \le s))}{P(X_{(1)} \le s)}$$

First - the numerator:

Recall that we have a nice probability that is <u>almost</u> that of the complement of the numerator's event [4/6/98.4], namely:

$$P((X_{(2)} \le t) \cap (X_{(1)} > s)) = \frac{(t-s)^2}{a^2}$$

which gives:

$$P([(X_{(2)} \le t) \cap (X_{(1)} > s)]^c) = 1 - \underbrace{P((X_{(2)} \le t) \cap (X_{(1)} > s))}$$

$\Downarrow$ by DeMorgan's Law

$$P((X_{(2)} > t) \cup (X_{(1)} \le s)) = \boxed{1 - \frac{(t-s)^2}{a^2}}$$

this probability differs from the numerator in that this operation is a <u>union</u>, <u>not</u> an <u>intersection</u>.

Let's use the <u>inclusion-exclusion principle</u> on the events of the numerator:

$$\underbrace{P((X_{(2)} > t) \cup (X_{(1)} \le s))}_{\boxed{1 - \frac{(t-s)^2}{a^2}}} = \underbrace{P(X_{(2)} > t)}_{1 - \underbrace{P(X_{(2)} \le t)}_{\frac{t^2}{a^2}}} + \underbrace{P(X_{(1)} \le s)}_{1 - \frac{(a-s)^2}{a^2}} - \underbrace{P((X_{(2)} > t) \cap (X_{(1)} \le s))}_{\text{the numerator, what we desire}}$$

See [4/6/98.2]

See [4/6/98.3]

Rearranging and simplifying gives:

$$P\left((X_{(2)} > t) \cap (X_{(1)} \le s)\right) = 1 - \frac{t^2}{a^2} \cancel{+1} - \frac{(a-s)^2}{a^2} \cancel{-1} + \frac{(t-s)^2}{a^2}$$

$$= 1 - \left(\frac{t^2 - (t-s)^2 + (a-s)^2}{a^2}\right)$$

Second - the denominator:

We already know this, we've done it a hundred times [4/6/98.3].

$$P(X_{(1)} \le s) = 1 - \frac{(a-s)^2}{a^2}$$

Which gives the desired <u>conditional probability</u>:

$$P(X_{(2)} > t \mid X_{(1)} \le s) = \frac{1 - \left(\frac{t^2 - (t-s)^2 + (a-s)^2}{a^2}\right)}{1 - \frac{(a-s)^2}{a^2}}$$

$$= \frac{\dfrac{\cancel{a^2} - \cancel{t^2} - (\cancel{t^2} - 2st + \cancel{s^2}) + (\cancel{a^2} - 2as + \cancel{s^2})}{\cancel{a^2}}}{\dfrac{\cancel{a^2} - (\cancel{a^2} - 2as + s^2)}{\cancel{a^2}}}$$

$$= \frac{2as - 2st}{2as - s^2}$$

$$P(X_{(2)} > t \mid X_{(1)} \le s) = \frac{a-t}{a - \frac{s}{2}}$$

- Now we compute the <u>density</u>, <u>given an event</u>.

<u>Conditional Density</u>

$$\text{dens}\left(X_{(2)}=t \mid X_{(1)}>s\right) = \frac{d}{dt} P\left(X_{(2)} \le t \mid X_{(1)}>s\right)$$

$\left\{\begin{array}{l}\text{Take the conditional cummulative probability} \\ \text{and differentiate.}\end{array}\right\}$

$$= \frac{d}{dt} \frac{P\left(\left(X_{(2)} \le t\right) \cap \left(X_{(1)}>s\right)\right)}{P\left(X_{(1)}>s\right)}$$

← numerator from [4/6/98.4] probability <u>both</u> marbles fall in [s, t]

← denominator from [4/6/98.2]

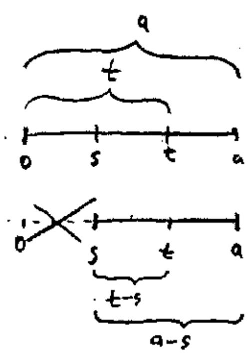
$$= \frac{d}{dt} \frac{\frac{(t-s)^2}{\cancel{a^2}}}{\frac{(a-s)^2}{\cancel{a^2}}}$$

$$= \frac{2(t-s)}{(a-s)^2}$$

- Now, let's justify this <u>intuitively</u>.
Note that:

$$\text{dens}\left(X_{(2)}=t\right) = \frac{2t}{a^2}$$

$$\text{dens}\left(X_{(2)}=t \mid X_{(1)}>s\right) = \frac{2(t-s)}{(a-s)^2}$$



The condition is that $X_{(1)}>s$.
What does this mean in ordinary language?
It means that both marbles fall in the interval $(s, a]$.
Therefore, we have a <u>uniform process</u> on a <u>shorter interval</u>.
And we take the density in the shorter interval.

$\text{dens}\left(X_{(2)}=t / X_{(1)}>s\right)$ is <u>exactly</u> the formula for the density on the <u>shorter interval</u>:

$$\text{dens}\left(X_{(2)}=t\right) = \frac{2t}{a^2}, \quad \text{where } \begin{array}{l} t \Rightarrow t-s \\ a \Rightarrow a-s \end{array}$$

- So <u>conditional densities</u> are computed just as other densities. Either:
  a) take the conditional cummulative distribution and differentiate
  -or- b) if you can, give an intuitive argument.

- Now comes the tough part.
  What is :

$$P\left(X_{(2)} \le t \,\middle|\, \underbrace{X_{(1)} = s}_{}\right) = ?$$

random variable $X_{(1)}$ takes <u>specific value</u> s.

This is not yet defined. Previously, we looked at probabilities and densities <u>conditioned on events</u> of continuous random variables, where the probability of the conditioning event was non-zero. Here, the conditioning event, that $X_{(1)}$ takes a <u>specific value</u> s, has probability:

$$P(X_{(1)} = s) = 0$$

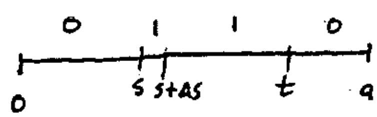
How will we define this?
1) Take the conditioning event that $X_{(1)}$ is in a small Δs neighborhood of s.
2) Then take the limit.

$$\lim_{\Delta s \to 0} P\left(X_{(2)} \le t \,\middle|\, \underbrace{s < X_{(1)} \le s+\Delta s}_{}\right) = \lim_{\Delta s \to 0} \frac{P\left((X_{(2)} \le t) \cap (s < X_{(1)} \le s+\Delta s)\right)}{P(s < X_{(1)} \le s+\Delta s)}$$

This is the event, of <u>non-zero</u> probability, that $X_{(1)}$ lies in the interval $(s, s+\Delta s]$. So we rewrite it using the rule of conditional probability to get the RHS.

Let's compute this with probabilities, rather than with densities.

<u>Numerator:</u>



Could have $X_{(2)}$ maybe fall in $(s, s+\Delta s]$, but in limit $\Delta s \to 0$, only $X_{(1)}$ will be in $(s, s+\Delta s]$.

$$P\left((X_{(2)} \le t) \cap (s < X_{(1)} \le s+\Delta s)\right) = \binom{2}{0,1,1,0}\left(\frac{\Delta s}{a}\right)\left(\frac{t-s-\Delta s}{a}\right) + \underbrace{(\Delta s)^2 \,(\text{stuff})}_{}$$

<u>Denominator:</u>

$$P(s < X_{(1)} \le s+\Delta s)$$

We could do, cheapo. Is there something more elegant?

$$\lim_{\Delta s \to 0} P(s < X_{(1)} \le s+\Delta s) = \lim_{\Delta s \to 0} \int_s^{s+\Delta s} \mathrm{dens}(X_{(1)} = u)\, du$$

$$= \lim_{\Delta s \to 0} \mathrm{dens}(X = s)\,\Delta s$$

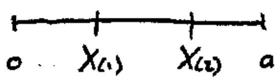
{more next time}

- Our motivation is the following:
  <u>Objective</u> : $\mathrm{dens}(X_{(i)} = s \mid X_{(j)} = t) = ?$

John Guidi
guidi@mit.edu

Continuous Conditional Probability (cont'd)

Let's continue from where we left off last time.
We take the Uniform or Dirichlet Process with $n=2$
So we drop 2 points at random in an interval.



We're trying to find a rigorous definition of the following term:

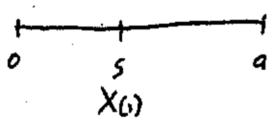
$$P(X_{(2)} > t \mid \underbrace{X_{(1)} = s}) = ?$$

in other words, the condition is an event of probability zero

$$P(X_{(1)} = s) = 0$$

We want to show that conditioning on an event of probability zero makes sense.

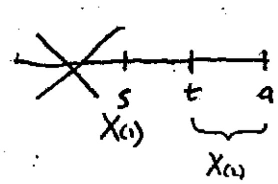
Let's first guess what it ought to be.
What should it be? Intuitively.



we know that the first marble is placed at position s, as the given conditioning event is $(X_{(1)} = s)$.
That's our knowledge.
The placement of $X_{(1)}$ is no longer a random event.
It's a certain event.

Given that we know that the first marble is placed at position s, what is the probability distribution for the second marble?

The answer intuitively is clear.
If the first marble is placed in position s, then you are left w/ a Dirichlet Process with only one marble, on a smaller interval.



$\{s < t \text{ always}\}$

So it ought to be that the interval of the first order statistic $X'_{(1)}$ $(= X_{(2)})$ is $(s, a]$. So, you take one point in this interval.
So it ought to be:

$$P(X_{(2)} > t \mid X_{(1)} = s) = \frac{a-t}{a-s}$$

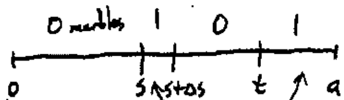
Now, let's confirm this by actual, rigorous computation.

The only possible definition of $P(X_{(2)} > t \mid X_{(1)} = s)$ is:

$$P(X_{(2)} > t \mid X_{(1)} = s) = \lim_{\Delta s \to 0} \underbrace{P(X_{(2)} > t \mid s < X_{(1)} \le s + \Delta s)}_{\text{this is a genuine probability}}$$

(*) $$= \lim_{\Delta s \to 0} \frac{P((X_{(2)} > t) \cap (s < X_{(1)} \le s + \Delta s))}{P(s < X_{(1)} \le s + \Delta s)}$$

The numerator is:

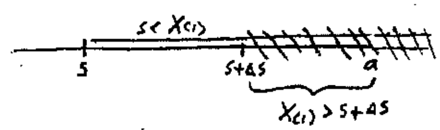


{In the limit, we need not concern ourselves w/ multiple balls in an interval.}

$$P(\overbrace{(X_{(2)} > t)} \cap \overbrace{(s < X_{(1)} \le s + \Delta s)}) = \cancel{\binom{2}{0,1,0,1}}\, 2 \left(\frac{\Delta s}{a}\right)\left(\frac{a-t}{a}\right)$$

The denominator is:

$$\underbrace{(s < X_{(1)} \le s + \Delta s)}_{\text{event}} = (s < X_{(1)}) - (X_{(1)} > s + \Delta s)$$



Pure logic.
The limits of rationality

Probabilities always subtract, thus:

$$P(s < X_{(1)} \le s + \Delta s) = \underbrace{P(s < X_{(1)})}_{\substack{\text{we've computed}\\ \text{this a hundred times}\\ [4/6/98.2]:\\ \frac{(a-s)^2}{a^2}}} - \underbrace{P(X_{(1)} > s + \Delta s)}_{\frac{(a-s-\Delta s)^2}{a^2}}$$

Therefore equation (*) is :

$$P\left(X_{(2)} > t \,\middle|\, X_{(1)} = s\right) = \lim_{\Delta s \to 0} \frac{2\left(\frac{\Delta s}{\cancel{a}}\right)\left(\frac{a-t}{\cancel{a}}\right)}{\frac{(a-s)^2}{\cancel{a^2}} - \frac{(a-s-\Delta s)^2}{\cancel{a^2}}}$$

$$= \lim_{\Delta s \to 0} \frac{2\,\Delta s\,(a-t)}{\left[\cancel{(a-s)} - (\cancel{a-s} - \Delta s)\right]\left[(a-s) + (a-s-\Delta s)\right]}$$

$$= \lim_{\Delta s \to 0} \frac{\cancel{2}\,\cancel{\Delta s}\,(a-t)}{\cancel{\Delta s}\left[\cancel{2}(a-s) - \overset{0}{\cancel{\Delta s}}\right]}$$

$$= \frac{a-t}{a-s} \quad \checkmark$$

Lo and Behold, this confirms our intuition.

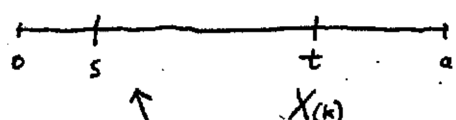
- Let's do a fancy one.

Example : Uniform Process, n arbitrary ← any number of marbles.

What is :

$$P\left(X_{(1)} > s \,\middle|\, X_{(k)} = t\right) = ?$$

Let's first guess what it ought to be.



↑ we are given the data that the k^{th} marble falls exactly at position t.

Given these data, what is the probability distribution that $(X_{(1)} > s)$?

Well – let's think.

If you are given these data, then the positions of all marbles from $X_{(k+1)}$ to $X_{(n)}$ are irrelevant.

It is as if you dropped $k-1$ marbles on the interval $(0,t)$.

So it ought to be:

$$P\left(X_{(1)} > s \,\middle|\, X_{(k)} = t\right) = \left(\frac{t-s}{t}\right)^{k-1}$$

- Now, we justify it the hard way.
(Last time we do it the hard way)

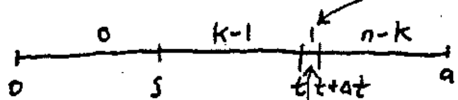
$$P(X_{(1)} > s \mid X_{(k)} = t) = \lim_{\Delta t \to 0} \underbrace{P(X_{(1)} > s \mid t < X_{(k)} \le t + \Delta t)}_{\text{a genuine probability, as we are conditioning by an actual event of non-zero probability, not an event of zero probability.}}$$

a genuine probability, as we are conditioning by an actual event of non-zero probability, <u>not</u> an event of zero probability.

$$(**) \qquad = \lim_{\Delta t \to 0} \frac{P\big((X_{(1)} > s) \cap (t < X_{(k)} \le t + \Delta t)\big)}{P(t < X_{(k)} \le t + \Delta t)}$$

The <u>numerator</u> is:

{we may assume WLOG that there is only 1 marble here, because Δt is small. We've gone through this crisis before.}

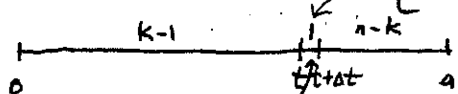


$$P\big((X_{(1)} > s) \cap (t < X_{(k)} \le t + \Delta t)\big) = \binom{n}{0, k-1, 1, n-k} \left(\frac{t-s}{a}\right)^{k-1} \left(\frac{\Delta t}{a}\right) \left(\frac{a-t-\Delta t}{a}\right)^{n-k}$$

Note: a good check is that power of denominator adds up to n.

The <u>denominator</u> is:

{again, WLOG, because of the calculus.}



$$P(t < X_{(k)} \le t + \Delta t) = \binom{n}{k-1, 1, n-k} \underbrace{\left(\frac{t}{a}\right)^{k-1}}_{\text{ex: that's how k-1 marbles fall independently in the interval } (0,t]} \left(\frac{\Delta t}{a}\right) \left(\frac{a-t-\Delta t}{a}\right)^{n-k}$$

ex: that's how k-1 marbles fall independently in the interval $(0,t]$.

Therefore, $(**)$ equals:

$$P(X_{(1)} > s \mid X_{(k)} = t) = \lim_{\Delta t \to 0} \frac{\cancel{\binom{n}{0, k-1, 1, n-k}} \left(\frac{t-s}{\cancel{a}}\right)^{k-1} \cancel{\left(\frac{\Delta t}{a}\right)} \cancel{\left(\frac{a-t-\Delta t}{a}\right)^{n-k}}}{\cancel{\binom{n}{k-1, 1, n-k}} \left(\frac{t}{\cancel{a}}\right)^{k-1} \cancel{\left(\frac{\Delta t}{a}\right)} \cancel{\left(\frac{a-t-\Delta t}{a}\right)^{n-k}}}$$

$$= \frac{(t-s)^{k-1}}{t^{k-1}} \quad \checkmark$$

It comes out. Our intuition is again confirmed.

- In *general*:

$$P(A \mid X=t) = \lim_{\Delta t \to 0} P(A \mid t < X \leq t + \Delta t)$$

probability of event A, given the event that continuous random variable X has value t

X is any continuous variable.
A is any event in the same sample space.

Now you say "So what?"
Now we come to the deepest, and most important, formula in probability.
This is what I was leading up to.

- **Continuous Version of the Law of Alternatives**

It is stated:

$$(*) \quad P(A) = \int_{-\infty}^{\infty} P(A \mid X=t)\, dens(X=t)\, dt$$

where X is a continuous random variable

This is the continuous analogue of the Law of Alternatives.
It is amazing that it works.

Let's roll up our sleeves and prove it.
We prove it by a limiting process, as in the calculus. What did you expect?

Proof

$$\ldots\ t_3\ t_2\ t_1\ t_0\ t_1\ t_2\ t_3\ \ldots$$

take events that X takes a value in the interval $(t_n, t_{n+1}]$.

event $(t_n < X \leq t_{n+1})$

These events are a partition of the sample space Ω

i) events disjoint
ii) events cover the sample space.

Hence, we can apply the Law of Alternatives. ← ordinary, garden variety.
Ordinary Law of Alternatives gives you:

$$(**) \quad P(A) = \sum_n P(A \mid t_n < X \leq t_{n+1})\, P(t_n < X \leq t_{n+1})$$

Now we close our eyes and take the limit as all the intervals between successive $(t_n, t_{n+1}]$ simultaneously tend to 0.
Let's see what happens.

Recall that:

$$P(t_n < X \leq t_{n+1}) = \int_{t_n}^{t_{n+1}} dens(X=t)\,dt$$

<u>Integral Mean Value Theorem</u>
An Intro. to Numerical Analysis
Kendall E. Atkinson
Let $w(x)$ be non-negative + integrable on $[a,b]$ and let $f(x)$ be <u>continuous</u> on $[a,b]$. Then
$$\int_a^b w(x) f(x)\,dx = f(\xi) \int_a^b w(x)\,dx$$
for some $\xi \in [a,b]$
In this case, $w(x)=1$, $f(x) = dens(X=x)$
$$\int_{t_n}^{t_{n+1}} dens(X=x)\,dx = dens(X=s_n)\, x \Big|_{t_n}^{t_{n+1}}$$
$s_n \in [t_n, t_{n+1}]$

Now we use something that you thought you'd never see again when you took 18.01:
The <u>Mean Value Theorem</u>
When you integrate a ~~non-negative~~ continuous quantity, it's the same as taking the length of the interval of integration times the value of the integrand at some intermediate value.

$$= (t_{n+1} - t_n)\, dens(X = s_n)$$

for some s_n between t_n and t_{n+1}

Therefore (**) can be simplified as follows:

$$P(A) = \sum_n P(A \mid t_n < X \leq t_{n+1})(t_{n+1} - t_n)\, dens(X = s_n), \quad s_n \in [t_n, t_{n+1}]$$

This is the way you define an <u>integral</u>, as limit of all successive intervals $(t_n, t_{n+1}] \to 0$.

$$P(A) = \lim_{(t_n, t_{n+1}] \to 0} \sum_n \underbrace{P(A \mid t_n < X \leq t_{n+1})}_{\text{in the limit, tends to } P(A|X=t)} dens(X=s_n)(t_{n+1} - t_n), \quad s_n \in [t_n, t_{n+1}]$$

$$= \int_{-\infty}^{\infty} P(A \mid X=t)\, dens(X=t)\,dt$$

For the next couple of lectures, we will learn much more of this.
We will now apply this all over the place.
It is very important.

- Lets draw a consequence of: $P(A) = \int_{-\infty}^{\infty} P(A|X=t)\, dens(X=t)\, dt$

Applying this to the event:

$$A = (Y \le s)$$

↑ Y is another continuous random variable

you get:

$$P(Y \le s) = \int_{-\infty}^{\infty} P(Y \le s | X = t)\, dens(X=t)\, dt$$

Now take the derivative, relative to s, on both sides.
Then you get the density.
The derivative of the cummulative distribution is the density.

$$\frac{d}{ds} P(Y \le s) = dens(Y=s)$$

$$\frac{d}{ds} P_{(X=t)}(Y \le s) = dens(Y=s|X=t)$$

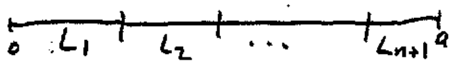
$$\frac{d}{ds} P(Y \le s) = \frac{d}{ds}\left[\int_{-\infty}^{\infty} P(Y \le s | X=t)\, dens(X=t)\, dt\right]$$

$$dens(Y=s) = \int_{-\infty}^{\infty} \frac{d}{ds} P(Y \le s | X=t)\, dens(X=t)\, dt$$

$$(***) \quad dens(Y=s) = \int_{-\infty}^{\infty} dens(Y=s|X=t)\, dens(X=t)\, dt$$

Let's start w/ some easy applications.

Example: Uniform Process



The gaps are $\underbrace{L_1, L_2, \dots, L_{n+1}}_{n+1 \text{ gaps}}$

By an <u>intuitive</u> argument, we've convinced ourselves that the gaps are identically distributed (i.d.), even though they are <u>not</u> <u>independent</u>.

Let's now verify this rigorously:

<u>Gaps in the uniform process are i.d.</u>

<u>Proof</u>

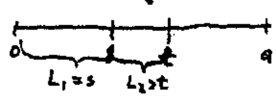
Let's find the probability distribution of the 2^{nd} gap (L_2).

Let $A = (L_2 > t)$

By the Continuous Law of Alternatives:

$$P(L_2 > t) = \int_{s=0}^{a-t} P(L_2 > t \mid L_1 = s)\, dens\,(L_1 = s)\, ds$$

Limits of integration:

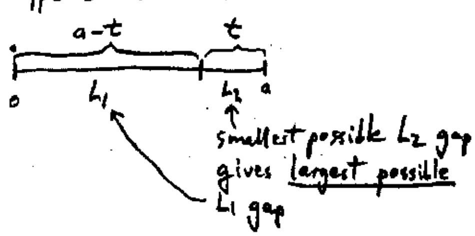


The first gap $L_1 = s$ ranges from $s = 0$ to $\underbrace{a-t}$,

because the 2^{nd} gap $L_2 > t$

You can <u>not</u> go beyond that because otherwise the 2^{nd} gap would <u>not be able to fit</u> in the interval $[0, a]$.

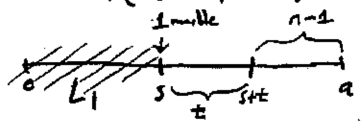
upper bound:



Therefore, integration limits are:

$$\int_{s=0}^{a-t}$$

- Now we do $P(L_2 > t \mid L_1 = s)$ intuitively:



$L_1 = s \Rightarrow 1^{st}$ gap = length s
That means that the first marble falls at s.
Therefore the second marble, which falls after the first marble, falls in a smaller interval.

Since $L_2 > t$, the 2^{nd} through n^{th} marble ($n-1$ total) fall on the smaller interval, as illustrated above. Therefore, we have the conditional probability:

$$P(L_2 > t \mid L_1 = s) = \left(\frac{a-s-t}{a-s}\right)^{n-1}$$

- Next, we need the density of the 1^{st} gap – $dens(L_1 = s)$.

Note that the density of the 1^{st} gap is the same as the density of the 1^{st} order statistic $X_{(1)}$. The position of the 1^{st} marble $X_{(1)}$ is the same as the gap of the 1^{st} marble, $L_1 = X_{(1)}$.
Thus:

$$dens(L_1 = s) = dens(X_{(1)} = s)$$

We've computed density of order statistics two different ways – see [3/18/98. 8–10].

$$dens(X_{(k)} = s) = \binom{n}{k} \frac{k\, s^{k-1} (a-s)^{n-k}}{a^n}$$

$$= \frac{n(a-s)^{n-1}}{a^n}$$

- Now that we have $P(L_2 > t \mid L_1 = s)$ and $dens(L_1 = s)$, we can continue computing the integral for $P(L_2 > t)$:

$$P(L_2 > t) = \int_{s=0}^{a-t} P(L_2 > t \mid L_1 = s)\, dens(L_1 = s)\, ds$$

$$= \int_{s=0}^{a-t} \left(\frac{a-s-t}{\cancel{a-s}}\right)^{n-1} \frac{n\,\cancel{(a-s)^{n-1}}}{a^n}\, ds$$

$$= \int_{s=0}^{a-t} \frac{n\,(a-s-t)^{n-1}}{a^n}\, ds$$

$$= \frac{n}{a^n} \int_{s=0}^{a-t} (a-s-t)^{n-1}\, ds$$

$$= \frac{1}{a^n} \Big[-(a-s-t)^n \Big]_{s=0}^{s=a-t}$$

$$P(L_2 > t) = \frac{(a-t)^n}{a^n}$$

Probability that the 2nd gap is greater than t is exactly the same as Probability that the 1st gap is greater than t.

$$P(L_2 > t) = P(L_1 > t) = \frac{(a-t)^n}{a^n}$$

This proves that L_1 & L_2 are identically distributed (i.d.)

What more do you want?

And similarily, you can show L_2 i.d. to L_3, etc.

John Guidi
guidi@mit.edu

Continuous Conditional Probability (cont'd)

We begin to see applications of what is perhaps the fundamental identity of probability theory. Namely:

The continuous analogue of the Law of Alternatives:

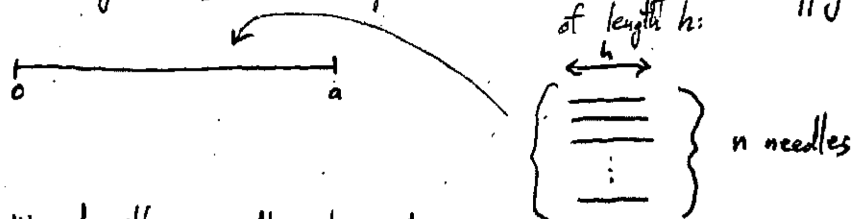
$$\underset{\text{probability of an event } A}{P(A)} = \int_{-\infty}^{\infty} P(A \mid X=t)\, \underbrace{\text{dens}(X=t)}_{\text{density of a continuous random variable, A concept which we took some time to define.}}\, dt$$

- We saw one application of this last time, when we showed that if we take the first 2 gaps in the Dirichlet Process, then the 2 gaps L_1 and L_2 are identically distributed [4/8/98.8-10]. And, of course, from the computations we did last time, it's easy to infer that the computations can be further pursued to show that any 2 gaps (L_i and L_j) are identically distributed.

Now let's attack one of the toughest problems of probability, which is the problem of needles on a stick.

- Example Problem of needles on a stick

We are given a stick of length a: We are given a supply of n identical needles of length h:



We drop the n needles at random on the stick.
The needle discipline is that the needles should fit entirely w/in the stick. They can not stick out.

So we drop the needles on the stick, subject to this needle discipline.

Problem: What is the probability that no 2 needles overlap?

Note – this problem is solved in the book, but there are several mistakes.

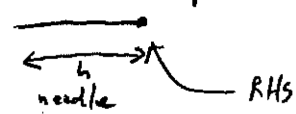
Let $B_{a,n}$ = event that $\underbrace{\text{no two needles}}_{\text{out of } n \text{ needles}}$ $\underbrace{\text{overlap}}_{\text{on interval of length } a}$

Our objective is to compute the probability of the event $B_{a,n}$:

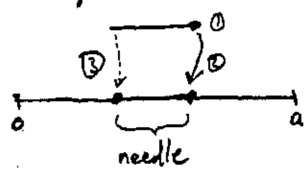
$$P(B_{a,n}) = \ ?$$

Step 1: rephrase the whole problem in terms of the Uniform Process
In terms of the uniform process, what have we done?
Let's decide what it means to drop a needle.
You may use several tricks, which are equivalent.
The one we will use is the following:

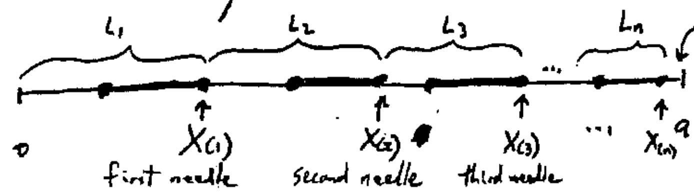
① We first pick the right hand endpoint of the needle.



② Then we drop the RHS onto the stick.
③ Then we put the needles left hand endpoint onto the stick.



If we understand the dropping of the needles from this point of view, then the problem can be immediately reformulated in terms of the uniform process.



The dropping of the RHS of the needle closest to 0 is the same as choosing point $X_{(1)}$ of the Uniform Process, and so forth.

And therefore, the event that the needles shall not overlap (i.e., $B_{a,n}$) is equivalent to saying that the first n gaps must be > h.
So, the event $B_{a,n}$ is the same as:

$$B_{a,n} = (L_1 > h) \cap (L_2 > h) \cap \ldots \cap (L_n > h)$$

Nothing is said about the last gap, L_{n+1}, for which we don't care.

So that's the event whose probability we have to compute.

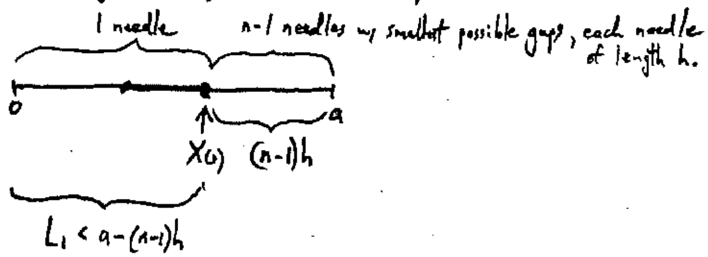
The way known to man of to compute this probability is by applying the Continuous Law of Alternatives, in a judicious way.

We condition relative to either a) the position of the 1st needle
or b) the position of the last (nth) needle

Let's see what happens if we condition relative to the position of the 1st needle.
By the Continuous Law of Alternatives:

$$P(B_{a,n}) = \int_{-\infty}^{\infty} P(B_{a,n} \mid L_1 = t)\, \mathrm{dens}(L_1 = t)\, dt$$

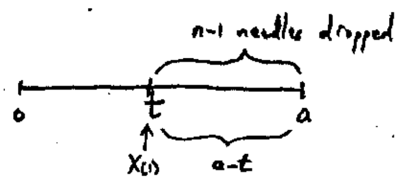
Now, let's put the limits of integration as carefully as we can.
Certainly, $L_1 > h$, otherwise the first needle doesn't fit,
in which case $P(B_{a,n} \mid L_1 = t) = 0$.
At the other limit, the largest 1st gap L_1 that is possible is:



$$= \int_{h}^{a-(n-1)h} \underbrace{P(B_{a,n} \mid L_1 = t)}\, \mathrm{dens}(L_1 = t)\, dt$$

Now we look at the conditional probability and say:
"What is this, really?"
Note that $L_1 = X_{(1)}$ ⟵ {the 1st gap and the 1st order statistic are the same.}

If $L_1 = X_{(1)} = t$ (i.e., the 1st order statistic is a t), then this conditioning tells you that you are dropping $n-1$ needles on a stick of length $a-t$.



So,

$$P(B_{a,n} \mid L_1 = t) = P(B_{a-t,\, n-1})$$

$$(*) \qquad \doteq \int_h^{a-(n-1)h} P(B_{a-t,n-1})\, \text{dens}(L_1 = t)\, dt$$

Next, we will evaluate this. First by handwaving, then by rigorous mathematics.

- By handwaving:

Since we have $P(B_{a,n}) = \int_h^{a-(n-1)h} \underbrace{P(B_{a-t,n-1})} \,\text{dens}(L_1 = t)\,dt$

we can use this equality to obtain an expression for $P(B_{\cdot,n-1})$ in terms of $P(B_{\cdot,n-2})$

We can repeatedly use this inequality and eventually get:

$P(B_{a,n})$ in terms of $P(B_{\cdot,1})$

The toilet trained way of saying this is:
By induction

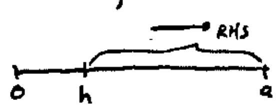
- By rigorous mathematics (i.e., by induction)

We want to show that:

$$P(B_{a,n}) = \frac{(a-nh)^n}{a^n}$$

At this point, I will assume that this probability has already been computed and shown to be correct up to $n-1$, which is the induction hypothesis.

Basis: for $n=1$, the solution is trivial. One needle has to fall inside the stick. $(L_1 = X_{(1)}) > h$



$$P(B_{a,1}) = \frac{a-h}{a} \checkmark$$

Inductive Step: By the inductive hypothesis, where I replace $\left\{ \begin{matrix} a \to a-t \\ n \to n-1 \end{matrix} \right\}$, we have:

$$P(B_{a-t,n-1}) = \frac{((a-t)-(n-1)h)^{n-1}}{(a-t)^{n-1}}$$

Then we have $P(B_{a,n})$ from equation $(*)$ by computing the integral:

$$P(B_{a,n}) = \int_h^{a-(n-1)h} \underbrace{\frac{((a-t)-(n-1)h)^{n-1}}{(a-t)^{n-1}}}_{\text{mutatis mutandis} = \text{Latin: You all would see, You all will have seen.}} \text{dens}(L_1 = t)\,dt$$

We've computed dens$(L_1=t)$ a hundred thousand times $\left\{\begin{matrix}[3/18/98.8-10]\\ [3/20/98.2-4]\end{matrix}\right\}$

$$\text{dens}(L_1=t) = \text{dens}(X_{(1)}=t)$$

$$= \frac{n(a-t)^{n-1}}{a^n}$$

mutatis mutandis

Thus, we have:

$$P(B_{a,n}) = \int_h^{a-(n-1)h} \frac{((a-t)-(n-1)h)^{n-1}}{\cancel{(a-t)^{n-1}}} \cdot \overbrace{\frac{n\cancel{(a-t)^{n-1}}}{a^n}}^{\text{mutatis mutandis}} \, dt$$

$$= \int_h^{a-(n-1)h} \frac{n(a-t-(n-1)h)^{n-1}}{a^n} \, dt$$

$$= \frac{n}{a^n} \underbrace{\int_h^{a-(n-1)h} (a-t-(n-1)h)^{n-1} \, dt}$$

Evaluate this integral by the substitution method (aka change of variables)

$$u = a-t-(n-1)h \qquad du = -dt$$

$$\int_h^{a-(n-1)h} (a-t-(n-1)h)^{n-1} dt = -\int_{t=h}^{a-(n-1)h} u^{n-1} du$$

$$= -\frac{u^n}{n} \Big|_{t=h}^{a-(n-1)h}$$

$$= -\frac{(a-t-(n-1)h)^n}{n} \Big|_{t=h}^{a-(n-1)h}$$

$$= 0 - \left(-\frac{(a-h-(n-1)h)^n}{n}\right)$$

$$\int_h^{a-(n-1)h} (a-t-(n-1)h)^{n-1} dt = \frac{(a-nh)^n}{n}$$

So we have:

$$= \frac{\cancel{n}}{a^n} \frac{(a-nh)^n}{\cancel{n}}$$

$$P(B_{a,n}) = \frac{(a-nh)^n}{a^n}$$, as desired

This completes the proof by induction of $P(B_{a,n})$, which was obtained using the Continuous Law of Alternatives.

That's it.
The Problem of needles on a stick – I told you about this the 1st day of class.
The 2-Dimensional analogue of this problem is <u>unsolved</u>.
Want to get a Nobel Prize? Work this out.
(You'll get the Nobel Prize in Physics).
80% of solid-state physics is this problem.
I am not joking.

Q/A: I should have given the following as an exercise.
Work out all the configurations where the needles can fall off the right of the stick, but not the left. Or they can fall off the left of the stick.

Q/A: Another way of stating the problem is to say:
The right end of each needle must fall into the interval 0 to a.

This is a very famous problem. There are about 1,000 physics reasearch papers written on this.

Now you get it in 10 minutes.
It's worth your tuition.
Some of you may help me do a UROP to do a search through the literature of all the <u>physics papers</u> that have been written on the <u>uniform process</u> <u>w/o acknowledging</u> it. That's called 1-dimensional physics.

Lets go on to bigger + better things.
In fact, let's go back to our origins — as they say in philosophy.
What's the fundamental problem of probability?
A: Balls into boxes? ← { No. You toss a coin. Even more fundamental than balls into boxes. }

A: Tossing a coin
↑ the fundamental problem of probability

So let's go back to the Bernoulli Process and apply the Continuous Law of Alternatives to the Bernoulli Process.

Example – Bernoulli Process

I toss the coin 17 times and I get 5 heads.
Is the coin fair?

How do we answer that question?
We have the means to answer it. You are now in a position to answer this fundamental question of life.
Is the coin fair based on these data?

You don't know if the coin is fair. You are being asked the bias of the coin. You don't know the coin's bias.
So what do you do when you don't know the coin's bias?
You say the bias is a random variable

$P =$ a continuous random variable w/ a density
$= $ bias

Now, let's compute the probability that in n tosses, we get k heads.
The binomial distribution.
Do you remember? I hope so,

By the Continuous Law of Alternatives:

$$\underbrace{P(S_n = k)}_{\substack{\text{Same old}\\ \text{notation [2/20/98.8-9]}}} = \int_{-\infty}^{\infty} P(S_n = k \mid P = t)\, \underbrace{\text{dens}(P=t)}_{\text{dens}(P=t)=0 \text{ for } t<0,\ t>1}\, dt$$

Random variable S_n is number of heads in n tosses.

$$= \int_0^1 P(S_n = k \mid P = t)\, \text{dens}(P = t)\, dt$$

continuous random variable } bias

$$\underbrace{P(S_n=k \mid P=t)}_{} = \binom{n}{k} t^k (1-t)^{n-k}$$

probability of k heads, given coin bias is t. [2/20/98.8] $\Big\}$ $\lim_{\Delta t \to 0} P(S_n=k \mid t < P \le t+\Delta t)$

$$P(S_n=k) = \int_0^1 \binom{n}{k} t^k (1-t)^{n-k} \, dens(P=t)\, dt$$

Q: Isn't this the Beta Function?
A: Not quite, because the dens $(P=t)$ is an <u>unknown</u> function. You don't know what the density is. It can be anything. It could be bimodal.
Recall the Beta Function [3/18/98.11], which rewritten is:

$$\int_0^a \binom{n}{k} t^{k-1} (1-t)^{n-k} \, dt = \frac{a^n}{k}$$

No, No. It's not Beta.

- So, in order to solve the problem of whether the coin is biased on the basis of given data, we have to develope a <u>continuous analogue</u> of <u>Bayes' Law.</u>

Remember Bayes' Law:

$$\underset{\text{posterior}}{P(A \mid B)} = \frac{\underset{\text{likelihood}}{P(B \mid A)} \; \overset{\text{prior}}{P(A)}}{P(B)} \quad \Big\} \text{ Trivial, as a mathematical identity.}$$

NB: There is <u>no</u> common agreement among statisticians as to whether

$$P(B|A) \quad \text{-or-} \quad \frac{P(B|A)}{P(B)} \quad \text{is called the prior.}$$

- Because of this lack of agreement, Bayes' Law is sometimes written as follows:

$$P(A|B) \underset{\uparrow \text{ proportional to}}{\propto} P(B|A)\,P(A)$$

The proportionality factor is $\frac{1}{P(B)}$.

Why don't people like to write down the proportionality factor?
Because, you really can <u>get it</u> from <u>the fact that:</u>

$$\int_{a=-\infty}^{\infty} P(A=a|B)\,da = 1 = \int_{a=-\infty}^{\infty} \frac{P(B|A=a)\,P(A=a)}{P(B)}\,da = \frac{1}{P(B)} \int_{a=-\infty}^{\infty} P(B|A=a)\,P(A=a)\,da$$

You can get $\frac{1}{P(B)}$, the proportionality factor, from the fact that probabilities (in this case, the posterior probability) must integrate to 1.

The conditional probability $P(A|B) = P_{(B)}(A)$ over the whole sample space must be 1.

So, if you don't know the denominator in Bayes' Law, you can obtain it by a simple integration:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \longleftarrow \int_{a=-\infty}^{\infty} P(B|A=a)\,P(A=a)\,da$$

This is why you will see, in the literature or when you get a summer job or so, the symbol $\propto$ appearing when people talk about Bayes' Law. Just to warn you.

- Let's see what the continuous analogue might be.
  The continuous analogue is obtained by a limiting process like we got when conditioning densities.

$$\underbrace{P(S_n=k \mid P=t)}_{\substack{\text{this is the}\\ \text{likelihood,}\\ \text{analogous to}\\ P(B|A)}} = \lim_{\Delta t \to 0} P(S_n=k \mid t < P \le t+\Delta t)$$

$$= \binom{n}{k} t^k (1-t)^{n-k}$$

The prior, analogous to $P(A)$, is:

$$\text{dens}\,(P=t)$$

Note: this is a density, not a probability.

The denominator, analogous to $P(B)$, is:

$$P(S_n=k) = \int_{-\infty}^{\infty} P(S_n=k \mid P=t)\,\text{dens}\,(P=t)\,dt$$

from the continuous Law of Alternatives, as shown on [4/10/98.7]

The LHS, analogous to $P(A|B)$, is:

$$\text{dens}\,(P=t \mid S_n=k)$$

So the Continuous analogue of Bayes' Law is:

$$\text{dens}(P=t \mid S_n=k) = \frac{P(S_n=k \mid P=t)\ \text{dens}(P=t)}{P(S_n=k)}$$

Instead of the event $(S_n=k)$, you can write an arbitrary event, of course.
The Continuous analogue of Bayes' Law is an estimate of:

the density given that some event pertaining to that density has happened, given the evidence

(in this example, the evidence is that n tosses yields k heads)

- So let's see how this works in practice.

From the previous pages, we have:

$$P(S_n=k \mid P=t) = \binom{n}{k} t^k (1-t)^{n-k}$$

$$P(S_n=k) = \int_0^1 \binom{n}{k} t^k (1-t)^{n-k}\ \text{dens}(P=t)\, dt$$

which gives:

(*) $$\overbrace{\text{dens}(P=t \mid S_n=k)}^{\text{posterior}} = \frac{\overbrace{\cancel{\binom{n}{k}} t^k (1-t)^{n-k}}^{\text{likelihood}}\ \overbrace{\text{dens}(P=t)}^{\text{prior}}}{\int_0^1 \cancel{\binom{n}{k}} t^k (1-t)^{n-k}\ \text{dens}(P=t)\, dt}$$

$$\boxed{\text{dens}(P=t \mid S_n=k) \propto t^k (1-t)^{n-k}\ \text{dens}(P=t)}$$

This is a very important conclusion, from several points of view.

So, you toss the coin n times, and you get k heads. Given this evidence, then you make a guess of what you think the bias of the coin is.

$\leftarrow$ dens $(P=t)$

For example, you may have reason to believe that the coin is strongly biased around $P=\frac{1}{2}$.
Or strongly biased with P between 0 and $\frac{1}{4}$.
Or whatever.
So you plug dens$(P=t)$ in.
On the basis of your previous guess of the bias, you compute dens$(P=t \mid S_n=k)$, which in turn allows you to obtain the cummulative probability of the bias of the coin.

$$\int_{-\infty}^{t} \text{dens}(P=s \mid S_n=k)\, ds = P(P \le t \mid S_n=k)$$

- What if you don't have any guess?
  If you don't have any guess, then you assume that <u>all</u> possible values of bias are equally probable. So continuous random variable P has a <u>uniform</u> distribution. That's called a <u>uniform prior</u>.

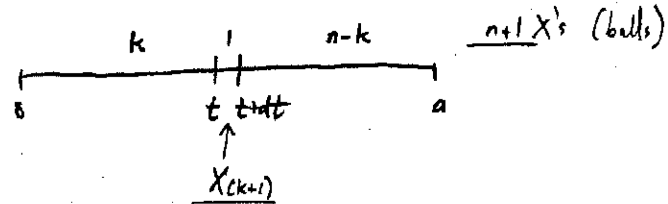
- With a <u>uniform prior</u>:

$$\text{dens}(P=t) = 1, \quad \text{for } 0 \le t \le 1$$

In this case, the Continuous analogue of Bayes' Law, equation (*), becomes:

$$\text{dens}(P=t \mid S_n=k) = \frac{t^k(1-t)^{n-k}}{\int_0^1 t^k(1-t)^{n-k}\,dt}$$

We can compute the denominator's integral in a similar fashion to that of [3/18/98.9-11], but this time we get the identity by using $n+1$ balls and focusing on $\text{dens}(X_{(k+1)}=t)$.
← the $k+1^{st}$ order statistic.



In the limit,

$$\text{dens}(X_{(k+1)}=t) = \binom{n+1}{k,1,n-k}\left(\frac{t}{a}\right)^k\left(\frac{1}{a}\right)\left(\frac{a-t}{a}\right)^{n-k}$$

and, as a consequence of the fact that integral of density over the whole sample space equals one:

$$\int_{-\infty}^{\infty} \text{dens}(X_{(k+1)}=t)\,dt = 1$$

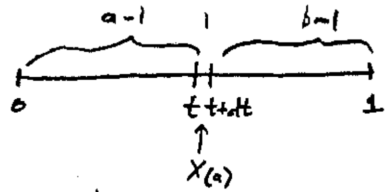
$$\int_0^1 \binom{n+1}{k,1,n-k} t^k(a-t)^{n-k}\,dt = 1$$

(**) This is another version of the <u>Beta Function Integral</u>. ⟶

$$\boxed{\int_0^1 t^k(1-t)^{n-k}\,dt = \frac{k!\,(n-k)!}{(n+1)!}}$$

- Precisely the same argument gives us the general Euler formula for the Beta Function Integral $B(a,b)$.

total $a+b-1$ X's (balls)



In the limit, we have:

$$\mathrm{dens}(X_{(a)}=t) = \binom{a+b-1}{a-1,\,1,\,b-1} t^{a-1}(1-t)^{b-1}$$

Then integrating the density to obtain probability:

$$\int_{-\infty}^{\infty} \mathrm{dens}(X_{(a)}=t)\,dt = 1$$

$$\int_0^1 \binom{a+b-1}{a-1,\,1,\,b-1} t^{a-1}(1-t)^{b-1}\,dt = 1$$

$$\boxed{B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt = \frac{(a-1)!\,(b-1)!}{(a+b-1)!}}$$

Beta Function Integral - general form

- Now that we have computed the denominator, we have:

$$\mathrm{dens}(P=t \mid S_n=k) = \frac{t^k(1-t)^{n-k}}{\left(\frac{k!\,(n-k)!}{(n+1)!}\right)}$$

Equation (**), the Beta Function Integral $\int_0^1 t^k(1-t)^{n-k}$ is exactly a proportionality constant required to ensure that:

$$\int_{-\infty}^{\infty} \mathrm{dens}(P=t \mid S_n=k)\,dt = 1$$

Since $\left(\frac{k!\,(n-k)!}{(n+1)!}\right)$ is a constant wrt t, we write:

$$\boxed{\mathrm{dens}(P=t \mid S_n=k) \propto t^k(1-t)^{n-k}}$$

- So now that we have $\text{dens}(P=t \mid S_n=k)$, what do we have philosophically? We have the answer as to what the bias of the coin is, on the basis of observing k heads in n tosses is not given by a number.
  It is given by a probability distribution, as we might expect.

$$\int_{-\infty}^{t} \text{dens}(P=s \mid S_n=k)\,ds = P(P \le t \mid S_n=k)$$

Let's see some cases of what happens if we observe $S_n=k$, for various n and k.

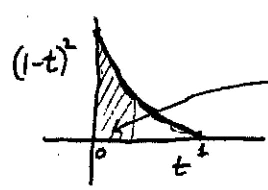
- Example – $n=2$, $k=0$ ← we observe no heads in 2 tosses
  Whats the plot of $\text{dens}(P=t \mid S_n=k)$?

  Clearly, we don't give a hoot about the proportionality constants here.
  We just plot:

$$\text{dens}(P=t \mid S_n=k) \propto t^k (1-t)^{n-k} \quad \text{-vs-} \quad t$$

  Of course, the plot needs to be normalized, so that the area under the curve integrates to 1.

$$\text{dens}(P=t \mid S_2=0) \propto t^0 (1-t)^2$$

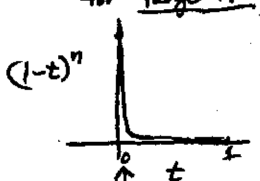


area $= \int_0^t \text{dens}(P=s \mid S_2=0)\,ds = P(P \le t \mid S_2=0)$

- Example – $n=$ large, $k=0$ ← let's say we toss it one billion times and get no heads.

$$\text{dens}(P=t \mid S_n=0) \propto (1-t)^n$$

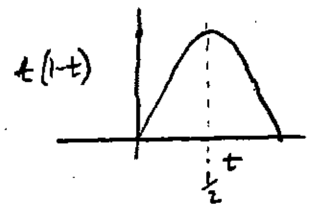
for large n



This stresses the fact that the larger n is, and if you don't get any heads, the more the area of the density is close to 0.
Which means most of the cummulative probability $P(P \le t \mid S_n=0)$ is accounted for when the bias P is close to 0.

- Example – $n=2,\ k=1$ ← 1 head in 2 tosses

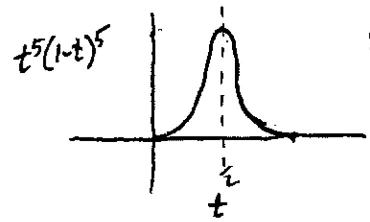
$$\text{dens}\,(P=t \mid S_2=1) \ \propto\ t\,(1-t)$$



Symmetric function around $\frac{1}{2}$

- Example – $n=10,\ k=5$ ← more tosses, but again, exactly half the tosses are heads.

$$\text{dens}\,(P=t \mid S_{10}=5) \ \propto\ t^5\,(1-t)^5$$



Symmetric function around $\frac{1}{2}$

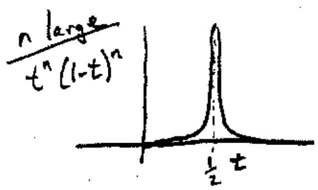
{Compared w/ above, it gets steeper and most of the area is around $\frac{1}{2}$.}

- For n even, with $k=\frac{n}{2}$ (i.e., half the tosses are heads), we have:

$$\text{dens}\,(P=t \mid S_n=\tfrac{n}{2}) \ \propto\ t^{\frac{n}{2}}\,(1-t)^{\frac{n}{2}}$$

This is a symmetric function around $\frac{1}{2}$

$$t \to 1-t \quad \Rightarrow (1-t)^{\frac{n}{2}}\,(1-(1-t))^{\frac{n}{2}}$$
$$0 \le t \le \tfrac{1}{2}$$
$$= (1-t)^{\frac{n}{2}}\, t^{\frac{n}{2}} \ \checkmark$$



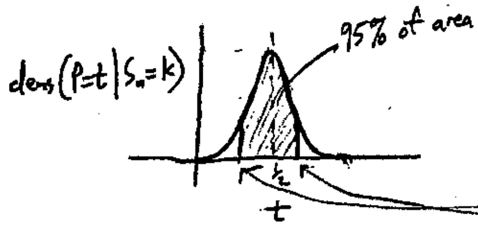
- As n gets large, you get bumps exactly where they should be.
The bumps come out in exactly the best position.
So it really works beautifully.
Except for 1 Kink.

It is my duty to inform you.
The boss will say: "I don't understand graphs. I want you to tell me whether the coin is fair, or not. Forget graphs."
What do you do?

You do confidence intervals.

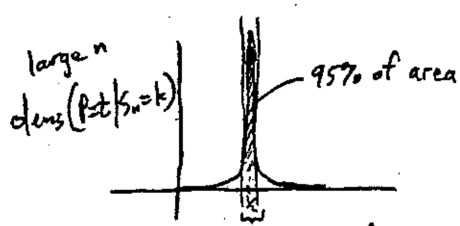
Now, let me tell you that it comes out, in the Bernoulli Process, very simple. You simply take an interval that contains 95% of the area under the graph. And that interval is called a very good confidence interval. Or else, you take an interval that contains 99% of the area, and this interval is called extremely good.

interval w/ 95% of area — <u>very good</u> confidence interval
" " 99% " " — <u>extremely good</u> confidence interval



Then you give the boss this.
With 95% confidence, the bias is between this and that.

So, for example, in this example, as n becomes large and half the tosses are heads, the curve becomes <u>very steep</u> and the <u>95% confidence interval</u> becomes <u>smaller</u>.



very good (95%) confidence interval

As n gets very large, the confidence interval gets smaller + smaller, shrinking about $\frac{1}{2}$. You get more and more evidence that the bias is $\frac{1}{2}$.

This is what statistics is all about.
People don't accept the graphs, they want a number.
So you give them this 95% (or 99%) area.
Satisfy them.

- Now, these <u>confidence intervals</u> have to be <u>computed numerically</u> w/ tables and on the computer, because you want integration of the density function such that:

$$\int_a^b \text{dens}\left(P=t \,/\, S_n=k\right) dt = .95$$, where $[a,b]$ is the very good confidence interval

$$\int_a^b \frac{t^k(1-t)^{n-k}}{\left(\frac{k!\,(n-k)!}{(n+1)!}\right)}\, dt = .95$$

Since $\int t^k(1-t)^{n-k}dt$ is <u>not</u> an elementary integral, this integral has to be computed by computer, or w/ tables.

4/10/98.16

I was going to do tables last night, but I was just too tired at 1 am.
So I didn't xerox the tables.
So I didn't give you any with this problem set. You'll get them next week.

- I'll go over this again next time, because it's too important, even though elementary.

This is what's coming next:

Why must we go through all this rigamarol if we are going to take the uniform prior?

A: We <u>don't</u> <u>always take</u> the <u>uniform prior</u>.

There is another prior, which has physical significance.
This is called the <u>Conjugate prior</u>, which is the following:

I take a coin and <u>secretly</u> toss it 100 times and keep a record.
Then I give you the coin <u>w/o</u> giving you my record.
You use the <u>uniform prior</u>.
I use the prior that corresponds to my previous tossings (from my record).
I get a <u>better confidence interval</u> than you do.

John Guidi
guidi@mit.edu

I was absent from this class.
These Lecture 25 notes are based on the following, which were kindly provided by:

Lecture tape – Jeff Lieberman
Lecture notes – Ira Gerhardt
Jeff Lieberman
Rick Monte

These Lecture 25 notes are based on my interpretation of these materials.

Bayes Estimation (cont'd)

- Last time, we derived the Beta function, which can be written in several ways:

$$\int_0^1 t^k (1-t)^{n-k}\,dt = \frac{k!\,(n-k)!}{(n+1)!}$$

And last time we studied the problem of estimating how fair a coin is on the basis of data obtained from tossing the coin a number of times.
We'll conclude the discussion today.

This discussion is, historically, one of the central chapters of statistics.
The discussion of the estimate of the fairness of the coin, using Bayes' Law – or rather, it's continuous analogue.

- As we showed last time [4/10/98.7-8], by the Continuous Law of Alternatives:

$$\underset{\text{probability that there are } k \text{ heads in } n \text{ tosses.}}{P(S_n=k)} = \int_0^1 \underbrace{P(S_n=k \mid P=t)}_{\text{probability that there are } k \text{ heads in } n \text{ tosses, given that the bias is } t.}\ \underbrace{\text{dens}(P=t)}_{\text{density of the bias}}\,dt$$

- We also have, from the Continuous analogue of Bayes' Law:

Bayes' Law for coin tossing

$$\text{dens}(P=t \mid S_n=k) = \frac{P(S_n=k \mid P=t)\ \text{dens}(P=t)}{P(S_n=k)}$$

$\longleftarrow$ $P(S_n=k)$ from above

Strictly speaking, we should find $\text{dens}(P=t \mid S_n=k)$ in two steps:

1) $P(P \le t \mid S_n=k) = \dfrac{P((P\le t) \cap (S_n=k))}{P(S_n=k)}$

then 2) $\text{dens}(P=t \mid S_n=k) = \dfrac{d}{dt} P(P\le t \mid S_n=k)$

- If we have no previous information about the bias of the coin, then the density for the random bias P is the uniform density between 0 and 1
Therefore:

$$\text{dens}(P=t) = 1 \quad , \quad 0 \le t \le 1$$

And as we showed [4/13/98.10-12]:

$$\text{dens}(P=t / S_n=k) = \frac{P(S_n=k / P=t)\,\text{dens}(P=t)}{P(S_n=k)}$$

$$= \frac{P(S_n=k / P=t)\,\text{dens}(P=t)}{\int_0^1 P(S_n=k / P=t)\,\text{dens}(P=t)\,dt}$$

(w/ $\text{dens}(P=t)=1$, this becomes the Beta Function Integral.)

Q/A: Note that the binomial coefficients for $P(S_n=k / P=t)$ in the numerator + denominator will cancel.

(*) $$\text{dens}(P=t / S_n=k) = \frac{t^k (1-t)^{n-k}}{\left(\frac{k!\,(n-k)!}{(n+1)!}\right)}$$

We're back to "Who's buried in Grant's Tomb?" We know that the integral of the probability density has to equal 1.
So:

$$1 = \int_{-\infty}^{\infty} \text{dens}(P=t / S_n=k)\,dt$$

$$= \int_0^1 \text{dens}(P=t / S_n=k)\,dt$$

$$= \int_0^1 \frac{t^k (1-t)^{n-k}}{\left(\frac{k!\,(n-k)!}{(n+1)!}\right)}\,dt \implies \int_0^1 t^k (1-t)^{n-k}\,dt = \frac{k!\,(n-k)!}{(n+1)!}$$

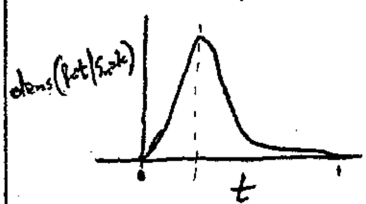
And this is precisely the Beta Function Integral, which we worked out a different way earlier [4/13/98.11] and was used in computing $\text{dens}(P=t / S_n=k)$ above.

- When people write densities, they often leave out the denominator.
For example:

$$dens(P=t \mid S_n=k) \propto t^k (1-t)^{n-k}$$

You can always recover the denominator, from the fact that the density has to integrate to 1.

- With the uniform prior, we saw last time how more + more evidence causes the graph to become steeper around the actual bias of the coin.

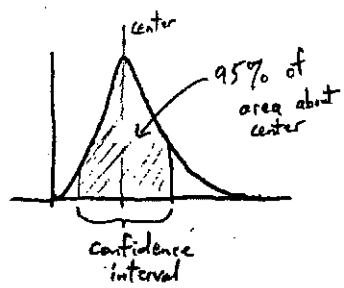


So if you toss the coin a large number of times, you are likely to get graphs that are very steep.
And this steep peak would be your guess of the bias.

Now, we have to repeat something we touched upon last time.
The integral under this curve is equal to 1, because it is a probability density.
But the woman in the street does not accept a curve as an answer.
To satisfy the man in the street, you do 2 things.
I told you one last time.

- 1) Confidence intervals

What is a confidence interval? It's simply this.
The whole area under the curve is 1.
You take some subset of the area that includes 95% of the area.



Then you look at the 2 extremes of the interval.
This is called the confidence interval.

Last night, I was going to make up some problems for you, but I was just too tired to look up the tables.
I have some tables that are horribly, poorly done.
But you want to get some excercise, as trivial as they are, on computing confidence intervals. It's of practical value – you just look it up in a table, or you get a computer program to work it out.
If $k=32$ and $n=65$, you should be able to give a 95% confidence interval.

Q: What do you do if you have a bimodal distribution?
A: The notion of confidence intervals is not well defined.
In fact, it's not well defined when it's unimodal. You may not know exactly where the middle is.
The whole thing is handwaving, really.

In general, you take a sensible center and take the 95% confidence interval about that center.
Or maybe, maybe you take 95% of the area, starting from 0.
It depends on the kind of question you are asking.
Confidence interval is not well defined.

What is well defined is the posterior.
What you do w/ the posterior depends on who you work for.

That's only the first thing.
The second thing is:

The boss says: "I don't want all these confidence intervals. I want a number."
So what do you do?
You take the expectation of the posterior.
The average.
That's called the Bayes' Estimate

2) Bayes' Estimate

$$\underbrace{E(P \mid S_n = k)}_{\text{the expected value of the posterior}}$$

The conditional expectation of P, given the event $(S_n = k)$.

(This stuff is presented in statistics books as a thick mess. It's so simple. This is clearly Mickey Mouse. But in statistics books it takes 100 pages explaining this stuff. Besides, not everyone takes 18.313.)

The conditional expectation of the random variable P — P = bias of the coin

If the boss wants a number, this is what you give her.

Let's see what it turns out to be.
This is a very interesting computation.

Q: Would you use this as the center for obtaining a confidence interval.
A: Again, often. In sensible situations, yes.
But if you have a weirdo prior, you might get a very sensitive situation, based on this weirdo prior.

By the definition of expectation:

$$E(P \mid S_n = k) = \int_{-\infty}^{\infty} t \,\mathrm{dens}(P=t \mid S_n = k)\, dt$$

$$= \int_0^1 t \,\underbrace{\mathrm{dens}(P=t \mid S_n=k)}_{\text{This density is given by equation } (*)\text{, under the uniform prior}}\, dt$$

We assume a uniform prior.

$$= \int_0^1 t \, \frac{t^k (1-t)^{n-k}}{\left(\frac{k!\,(n-k)!}{(n+1)!}\right)} \, dt$$

$$= \frac{(n+1)!}{k!\,(n-k)!} \underbrace{\int_0^1 t^{k+1} (1-t)^{n-k} \, dt}$$

This is the Beta Function Integral
See [4/10/98.12];

$$a-1 = k+1 \Rightarrow a = k+2$$
$$b-1 = n-k \Rightarrow b = n-k+1$$

$$\int_0^1 t^{k+1} (1-t)^{n-k} \, dt = \frac{(k+1)!\,(n-k)!}{(n+2)!}$$

$$= \frac{\cancel{(n+1)!}}{\cancel{k!}\,\cancel{(n-k)!}} \cdot \frac{(k+1)\cancel{!}\,\cancel{(n-k)!}}{(n+2)\cancel{!}}$$

$$\boxed{E(P \mid S_n = k) = \frac{k+1}{n+2}}$$

This is the famous Bayes' Estimate.

If you toss a coin n times and you get k heads $(S_n = k)$, the best estimate for the fairness of the coin is this number:

$$E(P \mid S_n = k) = \frac{k+1}{n+2}$$

Don't you ever forget that.
You that's what you give the boss, if she just wants a number.

- Application – the Laplace Law of Sucession

It's most impressive what it says.
Assuming that the sun has risen n times, what is the probability that it will rise tomorrow again.
We view this as tossing a coin with a bias, which we don't know.

We use Bayes' and end up w/ a Bayes' Estimate with $k=n$; ← event $(S_n = n)$
because it's always risen. Therefore:

$$E\left(\begin{smallmatrix}\text{sun will rise}\\ \text{tomorrow}\end{smallmatrix} \middle| \begin{smallmatrix}\text{sun has risen}\\ n \text{ days}\end{smallmatrix}\right) = E(P \mid S_n = n) = \frac{n+1}{n+2}$$

Now you say: Excuse me. So far we have only worked with the uniform prior.
I ran into a psychology professor the other day who said:
"All my life I have never used any other prior except uniform prior."
But that's the case.
Depending on the computations, you often use other priors.

Let's consider another prior – the conjugate prior.

↑ the term is justified from back in your previous problem sets.

Conjugate Prior

[redacted]

Before handing you the coin to test, I have privately tossed it j times and obtained i heads. I keep this information to myself.

Then you toss it, and you get k heads out of n tosses.
All you can do is use the uniform prior, because you don't know what I've done w/ the coin.

But what I can do is use the extra information, from my private tossing, which will give me a different prior.
What prior?
The posterior that comes from tossing the coin j times.
That is the conjugate prior.

Conjugate prior: $\text{dens}(P=t) \propto t^i (1-t)^{j-i}$

I take a uniform prior, then perform my private tossings and based on the observed i heads in j tosses, obtain the posterior: $\text{dens}(P=t \mid S_j = i) \propto t^i (1-t)^{j-i}$

So then, I use this private tossings posterior as my new prior, which is called a conjugate prior.

With this conjugate prior, I observe k heads in n tosses.
The posterior of this conjugate prior has the same functional form.
With this as conjugate prior, the posterior becomes:

$$\text{posterior}: \quad \text{dens}\underbrace{(P=t \mid S_n = k)} \propto t^k (1-t)^{n-k} \cdot \overbrace{t^i (1-t)^{j-i}}^{\text{dens}(P=t)}$$

this is really a lie, because I did not tell you I've already tossed the coin j times, obtaining i heads, before tossing n times & obtaining k heads.

$$\left\{ \begin{aligned} \text{dens}(P=t \mid S_n=k) &= \frac{P(S_n=k \mid P=t)\,\text{dens}(P=t)}{P(S_n=k)} \\ &\propto t^k (1-t)^{n-k}\, \text{dens}(P=t) \end{aligned} \right\}$$

↑ ignore factors that integrate to 1

posterior $\Longrightarrow$ $\text{dens}(P=t|S_n=k) \propto t^{k+i}(1-t)^{n-k+j-i}$

Therefore, the <u>posterior</u> will be <u>much steeper</u>, if I use this prior.
If you toss a coin - <u>and</u>, you take into account the information from previously tossing the coin, then your posterior is getting steeper (the powers are increased).

So, in the <u>case of coin tossing</u>, the <u>conjugate prior</u> is simply the results of the previous experiments of the same kind.

Again, why do I say <u>proportional</u>?

(**) $\text{dens}(P=t/S_n=k) \propto t^{k+i}(1-t)^{n-k+j-i}$

Because I can always integrate the density between 0 and 1 and the result must be 1, so I can determine the proportionality factors.

That's the story of the conjugate prior.
The <u>term</u> conjugate prior arose from this example.
Then it was generalized for balls into boxes, etc.
Intuitively, it means the same thing has happened before (i.e., you are using information of the same kind).

- If you take the <u>Bayes' Estimate</u> of the posterior obtained w/ the conjugate prior, of course you obtain a much sharper estimate:

Bayes' Estimate using conjugate prior:

$$E(P|S_n=k) = \int_{-\infty}^{\infty} t \underbrace{\text{dens}(P=t|S_n=k)}_{\substack{\text{posterior obtained from conjugate prior (equation (**))}}} dt$$

From [4/10/98.12], proportionality constants for density Beta Function Integral:

obtain proportionality factor by integrating density

$a-1 = k+i \rightarrow a = k+i+1$
$b-1 = n-k+j-i \rightarrow b = n-k+j-i+1$

$$\frac{(k+i)!\,(n-k+j-i)!}{(n+j+1)!} = \int_0^1 t^{k+i}(1-t)^{n-k+j-i}$$

$$= \int_0^1 \frac{(n+j+1)!}{(k+i)!\,(n-k+j-i)!} t^{k+i+1}(1-t)^{n-k+j-i}\,dt$$

$$= \frac{(n+j+1)!}{(k+i)!\,(n-k+j-i)!}\int_0^1 t^{k+i+1}(1-t)^{n-k+j-i}\,dt$$

This gives:
$a-1 = k+i+1 \Rightarrow a = k+i+2$
$b-1 = n-k+j-i \Rightarrow b = n-k+j-i+1$

$$\int_0^1 t^{k+i+1}(1-t)^{n-k+j-i}\,dt = \frac{(k+i+1)!\,(n-k+j-i)!}{(n+j+2)!}$$

$$= \frac{\cancel{(n+j)!}}{\cancel{(k+i)!}\,\cancel{(n-k+j-i)!}} \cdot \frac{(k+i+1)\cancel{!}\,\cancel{(n-k+j-i)!}}{(n+j+2)\cancel{!}}$$

$$\boxed{E(p|S_n=k) = \frac{k+i+1}{n+j+2}}$$

compare

Bayes's Estimate w/ conjugate prior

This is more accurate than Bayes' Estimate w/ uniform prior $\boxed{E(p|S_n=k) = \frac{k+1}{n+1}}$

Let's conclude our discussion about Bayes', for the moment.
We now initiate a new chapter.

- An open question which I was going to give you as a research problem:
Suppose I have the uniform process on the interval $[0,a]$, but I do not know a.
The research problem consists of 1) doing the Bayes' for a
2) finding out which priors make sense for a

It's a very interesting exercise.

You should try several priors. One of them, the prior is to know that the interval is no bigger than such + such. That's one sensible prior.

A second prior is based on the posterior, but based only on knowledge of the largest points.

4/13/98.9

The Algebra of Probability Distributions

This is one of the most counterintuitive chapters of probability.
Your intuition has to adjust to reality.
Your intuition goes wrong. Everybodies intuition goes wrong, at first, when dealing w/ this.

Suppose $X \geq 0$ is a continuous random variable that has

$$\text{dens}(X=t) = 0 \quad , \quad t<0$$

(We are taking X non-negative for simplicity of exposition. Not for any other reason.)

Now we take the random variable X^2.
And we ask, what is the density of the random variable X^2?

$$\text{dens}(X^2=t) = ?$$

The temptation to give a stupid answer is <u>wrong</u>.
For example:

$$\text{dens}(X^2=t) \neq \left[\text{den}(X=t)\right]^2$$

Why is it <u>not</u> the density squared?

Because densities integrate to 1.

We know that:

$$\int_0^\infty \text{dens}(X=t)\,dt = 1$$

Thus:

$$\int_0^\infty \left[\text{dens}(X=t)\right]^2 dt \neq 1$$

Sorry.

There is one and only one way of doing this.
Let's take the event $(X^2 \leq t)$, where $t \geq 0$

This event is the <u>same</u> as the event:

$$(X^2 \leq t) = (X \leq \sqrt{t})$$ ← if you don't see it, I can't explain it.

$X^2 \leq t \Rightarrow X \leq \sqrt{t}$.
We're just jazzing it up w/ events. It's the same fact.

Therefore:

$$\underbrace{P(X^2 \le t)}_{\substack{\text{Cummulative probability} \\ \text{distribution of } X^2}} = P(X \le \sqrt{t})$$

But the cumulative probability distribution is the integral of the density.

$$= \int_0^{\sqrt{t}} dens(X=s)\,ds$$

But, my friends, the density is the derivative of the cumulative probability distribution.

$$dens(X^2=t) = \frac{d}{dt} P(X^2 \le t)$$

$$= \frac{d}{dt} P(X \le \sqrt{t})$$

$$= \frac{d}{dt} \int_0^{\sqrt{t}} dens(X=s)\,ds$$

Now, we use our finest resources from 18.01.
This is <u>the derivative of a function of a function.</u>
I will not insult you by reminding you of the formula for the derivative of a function of a function.
Right? You all know it inside and out.

$$h(t) = f(g(t)) = \int_0^{\sqrt{t}} dens(X=s)\,ds$$

where $u = g(t) = \sqrt{t}$

By the <u>Chain Rule</u>:

$$\frac{d}{dt} h(t) = \frac{d\, f(g(t))}{d g(t)} \frac{d\, g(t)}{dt}$$

$$= \underbrace{\frac{d f(u)}{du}}_{ⓐ} \underbrace{\frac{du}{dt}}_{ⓑ}$$

ⓐ $f(u) = \int_0^{u} dens(X=s)\,ds$

$$= P(X \le u)$$

$$\frac{d f(u)}{du} = \frac{d}{du} P(X \le u)$$

$$= dens(X=u)$$

$$= dens(X=\sqrt{t})$$

ⓑ $\frac{du}{dt} = \frac{d}{dt}\sqrt{t}$

$$= \frac{1}{2\sqrt{t}}$$

$$dens(X^2=t) = dens(X=\sqrt{t})\ \frac{1}{2\sqrt{t}}$$

- If we are not sure of our results, we double <u>check</u> it by verifying that it does indeed integrate out to 1.
This is an unnecessary computation, but we are insecure.

$$\int_{-\infty}^{\infty} dens(X^2=t)\,dt = \int_0^{\infty} dens(X^2=t)\,dt$$ ← limits of integration 0 and ∞ since $dens(X^2=t)=0$ for $t<0$

$$= \int_0^{\infty} dens(X=\sqrt{t})\frac{1}{2\sqrt{t}}\,dt$$

Let's <u>substitute</u>:
$\sqrt{t} = s \implies t = s^2$
$dt = 2s\,ds$

$$= \int_0^{\infty} dens(X=s)\,\frac{1}{\cancel{2s}}\,\cancel{2s}\,ds$$

$$= \underbrace{\int_0^{\infty} dens(X=s)\,ds}_{\text{This is 1, as discussed in [4/13/98.9], so it checks.}}$$

$$= 1 \checkmark$$

And that's the way it works.

- More <u>generally</u>:
Suppose we have a random variable X.
Let $g(x)$ be "any" function, in the ordinary sense of the word function.
↑ little x

Then $g(X)$ is a <u>random variable</u>.
You want the <u>density</u> of random variable $g(X)$:

$$dens(g(X)=t) = ?$$

That's no joke.
In general, it's pretty tough.
We'll work it out only for the following special case.

function g | 4/13/98.12

- Special Case — $X \geq 0$, $\overbrace{\text{continuous}}^{\text{function } g}$, and function g has the property: $g'(x) \geq 0$ ← little x

Since the derivative is non-negative, the function g is strictly increasing.

Under these assumptions, we can repeat the reasoning we just performed for $g(X) = X^2$ word by word. But it's worth doing, because it's such an unusual computation that it takes a while to adjust to it.

$$P(g(X) \leq t) = P(X \leq g^{-1}(t))$$

← the inverse function of g

We always compute the cummulative distribution first, because that's the natural probability.

This equality is because the event $(g(X) \leq t)$ is the same as the event that $(X \leq g^{-1}(t))$. $g(x)$ is continuous, it is increasing, so there is only one inverse function.

(*) $$= \int_0^{g^{-1}(t)} \text{dens}(X = s)\, ds$$ ← {By definitions of cummulative distribution and density.}

{One of the problems you have in the Problem Set due Friday involves this type of computation.}

- Now, may I remind you that inverse function means:

$$g(g^{-1}(t)) = t$$

Therefore, differentiating gives:

$$g'(g^{-1}(t)) \cdot g^{-1\prime}(t) = \cancel{t} \; 1$$

(**) $$\Longrightarrow \boxed{g^{-1\prime}(t) = \frac{1}{g'(g^{-1}(t))}}$$

An important formula

- In view of this formula, we can take equation (*)

$$P(g(X) \le t) = \int_0^{g^{-1}(t)} \text{dens}(X=s)\,ds$$

and <u>differentiate</u> <u>both sides</u>:

$$\underbrace{\frac{d}{dt} P(g(X) \le t)}_{\text{dens}(g(X)=t)} = \frac{d}{dt}\int_0^{g^{-1}(t)} \text{dens}(X=s)\,ds$$

Identical to [4/13/98.10], we use the Chain Rule:

$$h(t) = r(s(t)) = \int_0^{g^{-1}(t)} \text{dens}(X=s)\,ds$$

$$u = s(t) = g^{-1}(t)$$

$$\frac{d}{dt}h(t) = \frac{d\,r(s(t))}{ds(t)} \cdot \frac{d\,s(t)}{dt}$$

$$= \underbrace{\frac{d\,r(u)}{du}}_{(a)} \cdot \underbrace{\frac{du}{dt}}_{(b)}$$

(a) $r(u) = \int_0^u \text{dens}(X=s)\,ds$

$= P(X \le u)$

$\frac{d\,r(u)}{du} = \text{dens}(X=u)$

$= \text{dens}(X = g^{-1}(t))$

(b) $\frac{du}{dt} = \underbrace{g^{-1\prime}(t)}_{\text{equation (**) gives:}}$

$= \frac{1}{g'g^{-1}(t)}$

$$= \text{dens}(X=g^{-1}(t))\,\frac{1}{g'g^{-1}(t)}$$

And we have:

$$\boxed{\text{dens}(g(X)=t) = \text{dens}(X=g^{-1}(t)) \cdot \frac{1}{g'g^{-1}(t)}}$$

That's the formula for the <u>density of a function</u>, under the given assumptions.

- You can <u>check</u>, just like in the previous example, that this integrates to 1. It's almost a miracle that this integrates to 1 <u>automatically</u>.

- Now, as a last word.
Suppose you have to find the density:

$$\text{dens}\,(X=s \mid Y^2=t) = \;?$$

This implies a fairly delicate computation w/ the cumulative distribution.
You can not just close your eyes and square.

$$\text{dens}\,(X=s \mid Y^2=t) \neq \left[\text{dens}\,(X=s \mid Y=t)\right]^2$$

This is the kind of problem we have for Friday.
Don't wait until the last minute.

The only way to do this is to ① rephrase everything in terms of probabilities.
For example, in terms of:

$$\left.\begin{array}{l} P(s < X \le s+\Delta s \mid t < Y^2 \le \Delta t) \\ \\ \text{or in terms of cumulative distributions.} \end{array}\right\} \text{ Something like that}$$

Then, once you have things in terms of probabilities, you are on safe ground.
Then, when the time is right, you differentiate ②

Such densities are highly non-trivial problems.

- {Next time, we'll continue w/ the Algebra of Probability Distributions.
We'll discuss the notion of convolution.}

The Algebra of Probability Densities (Cont'd)

X = continuous random variable w/ density $dens(X=t)$.
Then random variable $f(X)$ has a density that must be computed extremely delicately. ← {where f is a suitable function}

$$dens(f(X)=t) = ?$$

← In particular, one has to guard against oversimplification.

What we saw is that $dens(f(X)=t)$ has to be computed by going back to the cumulative distribution and reasoning w/ the cumulative distribution.

• We worked this out, last time, where

$X \geq 0$; X is a non-negative random variable
$f'(x) > 0$; $f(x)$ is a strictly increasing function

In more general cases, which we did not work out, working it out is kind of a mess.
It's one of the messiest computations in probability.
Fortunately, in special cases, there is always a way out, which is dictated by the example under consideration.
So, you can try to avoid such messy computations.

• So let's do one last example.
Suppose we take the random variable $3X$.
What is:

$$dens(3X=t) \text{ in terms of } dens(X=t) = ?$$

It is not $\frac{1}{3} dens(X=t)$, because it doesn't integrate to 1

$$\int_{-\infty}^{\infty} \frac{1}{3} dens(X=t) = \frac{1}{3} \neq 1$$

And the density must integrate to 1.
So that's wrong.
This is what I call an oversimplification.

Well, the following events are the same.

$$(3X \le t) \quad = \quad (X \le \tfrac{t}{3})$$

If you don't see it, I can't explain it.

Therefore:

$$P(3X \le t) = P(X \le \tfrac{t}{3})$$

cumulative distribution of the random variable $3X$.

$$= \int_{-\infty}^{\frac{t}{3}} dens(X=s)\,ds$$

Taking derivatives of both sides:

$$\frac{d}{dt} P(3X \le t) = \frac{d}{dt} \int_{-\infty}^{\frac{t}{3}} dens(X=s)\,ds$$

The derivative of this cumulative distribution is $dens(3X=t)$, by the fundamental theorem of the calculus:

$$\frac{d}{dx}\int_a^x f(t)\,dt = f(x)$$

$$\frac{d}{dt}\int_{-\infty}^{t} dens(3X=s)\,ds = dens(3X=t)$$

$$h(t) = f(g(t)) = \int_{-\infty}^{\frac{t}{3}} dens(X=s)\,ds$$

where $u = g(t) = \frac{t}{3}$

$$h'(t) = \underbrace{f'(g(t))}_{a} + \underbrace{g'(t)}_{b}$$

a $f(u) = \int_{-\infty}^{u} dens(X=s)\,ds$

$= P(X \le u)$

$\frac{d}{du} f(u) = \frac{d}{du} P(X \le u)$

$= dens(X=u)$

$= dens(X = \frac{t}{3})$

b $g(t) = \frac{t}{3}$

$g'(t) = \frac{1}{3}$

By the Chain Rule:

$$\frac{d}{dt}\int_{-\infty}^{t/3} dens(X=s)\,ds = dens(X=\tfrac{t}{3}) \cdot \frac{1}{3}$$

$$\boxed{dens(3X=t) = dens(X=\tfrac{t}{3}) \cdot \frac{1}{3}}$$

And this integrates to 1.

$$\int_{-\infty}^{\infty} dens(3X=t)\,dt = \int_{-\infty}^{\infty} dens(X=\tfrac{t}{3}) \cdot \frac{1}{3}\,dt = 1$$

This is the density of 3X.
I don't know of anyone who has an intuitive grasp of this computation w/ density.
It involves a very deep understanding of inverse functions.
There is something philosophically, extremely deep in this computation.

Don't let me get into it.
Because I'll never stop.

Let's do some other computations of this kind, so you get used to it.
Let's do a couple of computations w/ joint densities.

Joint Densities

Suppose we have continuous random variables X, Y, with given joint density: $\text{dens}(X=s,\ Y=t)$

Recall, the marginal density:

$$\text{dens}(X=s) = \int_{-\infty}^{\infty} \text{dens}(X=s,\ Y=t)\, dt$$

in other words, if you wish one of the marginal densities, you integrate out the other. You let the other random variable take any value it could, which amounts to integrating.

Given this joint density, we want:

$$\text{dens}(X^2=u,\ Y=t) = ?$$

How do we compute that one?
Don't give me an oversimplified answer.
There is no way of guessing — actually some Physicists guess and get it right,
How they do it, I don't know.

I guess the safest way to work this out is to reduce this to the case of one variable.
X^2 X

How do we do that?
Simple. We conditionalize it. That's easy.

Note that we have the following identity w/ densities, analogous to conditional probability:

$$\text{dens}(X=s \mid Y=t) = \frac{\text{dens}(X=s,\ Y=t)}{\text{dens}(Y=t)}, \quad \text{dens}(Y=t) \neq 0$$

by definition of density:

Proof

$$\text{dens}(X=s \mid Y=t) = \frac{d}{ds} P(X \le s \mid Y=t)$$

In the limiting process, the following events are equal:

$$\lim_{\Delta t \to 0} (t < Y \le t+\Delta t) = (Y=t)$$

$$= \frac{d}{ds} \lim_{\Delta t \to 0} \underbrace{P(X \le s \mid t < Y \le t+\Delta t)}_{\text{this can be expressed as an ordinary conditional probability.}}$$

$$= \frac{d}{ds} \lim_{\Delta t \to 0} \frac{P((X \le s) \cap (t < Y \le t+\Delta t))}{P(t < Y \le t+\Delta t)}$$

Only the numerator involves s.

$$= \lim_{\Delta t \to 0} \frac{\frac{d}{ds} P((X \le s) \cap (t < Y \le t+\Delta t))}{P(t < Y \le t+\Delta t)}$$

Multiply by $1 = \frac{\frac{1}{\Delta t}}{\frac{1}{\Delta t}}$

$$= \lim_{\Delta t \to 0} \frac{\frac{d}{ds} \frac{P((X \le s) \cap (t < Y \le t+\Delta t))}{\Delta t}}{\frac{P(t < Y \le t+\Delta t)}{\Delta t}}$$

$$= \lim_{\Delta t \to 0} \frac{\frac{\partial}{\partial s} \frac{P((X \le s) \cap (Y \le t+\Delta t)) - P((X \le s) \cap (Y \le t))}{\Delta t}}{\frac{P(Y \le t+\Delta t) - P(Y \le t)}{\Delta t}}$$

The numerator is precisely $\frac{\partial^2}{\partial s \partial t} P((X \le s) \cap (Y \le t))$

The denominator is precisely $\frac{d}{dt} P(Y \le t)$

$$= \frac{\frac{\partial^2}{\partial s \partial t} P((X \le s) \cap (Y \le t))}{\frac{d}{dt} P(Y \le t)}$$

{definition of the joint density dens(X=s, Y=t)}

{definition of the density dens(Y=t)}

(*) $$\text{dens}(X=s \mid Y=t) = \frac{\text{dens}(X=s, Y=t)}{\text{dens}(Y=t)}$$

QED.

- With this identity in hand, we return to finding the joint density $\text{dens}(X^2=u, Y=t)$. This can be written as follows, using the previous identity.

$$\text{dens}(X^2=u, Y=t) = \text{dens}(X^2=u \mid Y=t)\,\text{dens}(Y=t)$$

assume $X \geq 0$

Density is an infinitesimal analogue of conditional probability, as we've discussed.

We've already computed $\text{dens}(X^2=u)$ [4/13/98.10]. Computing the conditional density follows exactly the same steps:

$$\begin{aligned}\text{dens}(X^2=u \mid Y=t) &= \frac{d}{du} P(X^2 \leq u \mid Y=t)\\ &= \frac{d}{du} P(X \leq \sqrt{u} \mid Y=t)\\ &= \frac{d}{du}\int_0^{\sqrt{u}} \text{dens}(X=s \mid Y=t)\\ &= \text{dens}(X=\sqrt{u} \mid Y=t)\cdot \frac{1}{2\sqrt{u}}\end{aligned}$$

$$= \overbrace{\text{dens}(X=\sqrt{u} \mid Y=t)\cdot\frac{1}{2\sqrt{u}}}^{\text{dens}(X^2=u \mid Y=t)} \cdot \text{dens}(Y=t)$$

$$= \frac{1}{2\sqrt{u}}\,\underbrace{\text{dens}(X=\sqrt{u} \mid Y=t)\cdot\text{dens}(Y=t)}$$

Using the general identity $(*)$, we have:

$$\text{dens}(X=\sqrt{u} \mid Y=t) = \frac{\text{dens}(X=\sqrt{u}, Y=t)}{\text{dens}(Y=t)}$$

$$\boxed{\text{dens}(X^2=u, Y=t) = \frac{1}{2\sqrt{u}}\,\text{dens}(X=\sqrt{u}, Y=t)}$$

So we got it.
You have some problems like this in the problem set.
I can see no one has looked at the problem set yet.
You are waiting until Thursday night, as usual. Right?
Thursday night at 2:00AM.
That's shocking.
Even though I did it too.
It's still shocking.

- Let's do another one.
  Again, X and Y being as in the previous problem.
  X and Y are not independent.
  Let's compute:

$$\text{dens}(X+Y=u) = ?$$

- First – an intuitive, physicist type argument.

  The event $(X+Y=u)$ is meaningless, because it is an event of probability 0.
  $P(X+Y=u)=0$.

$$(X+Y=u) \quad \text{iff} \quad (X=s,\ Y=u-s) \quad \text{for some } s$$

$$\text{dens}(X+Y=u) = \int_{-\infty}^{\infty} \text{dens}(X=s,\ Y=u-s)\,ds \longleftarrow \{\text{integrate out the } s\}$$

This is what we guess.
Now, let's check it by working out probabilities and then differentiating.

- Second – formally

by definition:

$$\underbrace{P(X+Y\le u)}_{\substack{\text{Now we are talking sets. The event}\\ (X+Y\le u) \text{ is meaningful}}} = \iint \text{dens}(X=s,\ Y=t)\,ds\,dt$$

Now, the painful part is to choose the domain of integration.

The domain of integration can be expressed set theoretically as:

$$\{s,t : s+t\le u\}$$

Which doesn't help that much.
But at least you get some idea.

So we do a change of variables.
$s+t\le u \Rightarrow t\le u-s$

$$P(X+Y\le u) = \int_{t=-\infty}^{u}\int_{s=-\infty}^{\infty} \text{dens}(X=s,\ Y=t-s)\,ds\,dt$$

Therefore, we can take the derivative of both sides:

$$\underbrace{\frac{d}{du} P(X+Y \le u)}_{\text{this is: } dens(X+Y=u)} = \frac{d}{du} \int_{t=-\infty}^{u} \int_{s=-\infty}^{\infty} dens(X=s,\ Y=t-s)\, ds\, dt$$

From the 2^{nd} Fundamental Theorem of the Calculus:

$$\frac{d}{du} \int_a^u f(t)\, dt = f(u)$$

(**) $$\boxed{dens(X+Y=u) = \int_{-\infty}^{\infty} dens(X=s,\ Y=u-s)\, ds}$$

Which is the same as our intuitive argument, as desired.

- As you suspect, this stuff can get pretty rough.
  You can take one whole term of <u>measure theory</u>.
  That's what this is all about.

- <u>Special Case</u> - suppose X and Y are <u>independent</u>

  By definition, if <u>X and Y</u> are <u>independent</u>:

$$P((X \le s) \cap (Y \le t)) = P(X \le s) \cdot P(Y \le t)$$

$$\underbrace{\frac{\partial^2}{\partial s \partial t} P((X \le s) \cap (Y \le t))}_{dens(X=s,\ Y=t)} = \frac{\partial}{\partial t}\left[\frac{\partial}{\partial s} P(X \le s) P(Y \le t)\right]$$

$$= \frac{\partial}{\partial t}\left[P(Y \le t) \underbrace{\frac{\partial}{\partial s} P(X \le s)}_{dens(X=s)}\right]$$

$$= dens(X=s) \underbrace{\frac{\partial}{\partial t} P(Y \le t)}_{dens(Y=t)}$$

$$\boxed{dens(X=s, Y=t) = dens(X=s)\, dens(Y=t)}$$

- Under this special case, where <u>X and Y</u> are <u>independent</u>, the formula for the density of the sum, equation (**), becomes something very interesting.

$$\text{dens}(X+Y=u) = \int_{-\infty}^{\infty} \underbrace{\text{dens}(X=s,\ Y=u-s)}\,ds$$

$$= \int_{-\infty}^{\infty} \overbrace{\text{dens}(X=s)\,\text{dens}(Y=u-s)}\,ds$$

We have seen this integral in 18.03.
This is the <u>convolution</u> of two functions.
That's were convolution comes from.
We have therefore found a meaning to convolution.

Let $h(u) = \text{dens}(X+Y=u)$
$f(s) = \text{dens}(X=s)$
$g(u-s) = \text{dens}(Y=u-s)$

Then:
$$h(u) = \int_{-\infty}^{\infty} f(s)\,g(u-s)\,ds$$

- In particular, if $\overbrace{X \text{ and } Y \geq 0}^{\text{non-negative}}$, then:

(*) $$\text{dens}(X+Y=t) = \int_{0}^{t} \text{dens}(X=s)\,\text{dens}(Y=t-s)\,ds$$

{ X and Y ≥ 0 cuts down limits of integration. If $X<0$, $\text{dens}(X=s)=0$ }

This is the <u>convolution of two functions</u>, as per 18.03.
The whole of the <u>Laplace transform</u> can be <u>interpreted probabilistically</u>.
And we will do that.
It's too bad we can't do that in 18.03 when we just had to dish out: This is convolution, take it home.

- Now you see that:

<u>Convolution</u> corresponds to taking the <u>sum</u> of <u>two independent random variable's</u>.

$$\underbrace{\text{Convolution of their densities}}_{\int_0^t \text{dens}(X=s)\,\text{dens}(Y=t-s)\,ds} \text{ equals } \underbrace{\text{density of their sums}}_{\text{dens}(X+Y=t)}$$

This is very interesting, if you ask me.

- Tough Example

Let X, Y be independent random variables, uniformly distributed in $[0,a]$.

$$\text{dens}(X=s) = \underbrace{\begin{cases} \frac{1}{a} & \text{if } s \in [0,a] \\ 0 & \text{otherwise} \end{cases}}$$

We will do this w/ indicator functions, which I discussed in SuperClass 1 [SC 2/27/98.1-4].

A short review.
- (Open parentheses:

Indicator random variables

This is a pretty useful set theoretic device for computing probabilities.

A = an event

$$\left.\begin{aligned} I_A &= \text{indicator random variable} \\ &= \begin{cases} 1 & \text{if } A \text{ happens} \\ 0 & \text{if } A \text{ does not happen} \end{cases} \end{aligned}\right\} \text{Probabilistic}$$

How do you change this into set theoretic language?
This is just a probabilistic way of saying that:

$$\left.\underbrace{I_A(\omega)}_{\text{read: "I sub A of sample point } \omega\text{"}} = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}\right\} \text{Set Theoretic}$$

The probabilistic version is superior.
Because it gives you a better understanding of set theory of events.

As John von Neumann, who contributed greatly in developing probability of quantum mechanics, said:

"Probability must be pointless."

Whenever you use sample points, your argument is not perfect.

There is a mathematician who developed all of probability pointlessly. Very interesting reading.
Some things are a little tougher to define — like convolution, w/o using points.

I_A = indicator random variable of the event A

You will find in the book the algebra of indicator random variables and the relationship bt. Boolean algebra and the algebra of indicator random variables. I leave it to you to look up.

For the present purpose, the most important thing to observe is this [sc 2/27/98.2]:

$$E(I_A) = \sum_n n\, P(I_A = n)$$

$$= 1 \cdot P(I_A = 1) + 0 \cdot P(I_A = 0)$$

$$= P(A)$$

since $P(I_A=1) = P(A)$

$$\boxed{E(I_A) = P(A)}$$

Namely, the probability of event A equals the expectation of the indicator random variable.

The use of indicator random variables is useful, because it reduces the computation of probability to the computation of expectation. And expectations are easier to deal w/ than probabilities.

So much for indicator random variables.
Close parentheses)

- Now back to our problem.
We want to write the density of X and of Y using indicator random variables.

$$\text{dens}(X=s) = \frac{1}{a}\, \underbrace{I_{[0,a]}(s)}$$

Indicator random variable on the interval $[0,a]$ of s.

$$\text{dens}(Y=t) = \frac{1}{a}\, I_{[0,a]}(t)$$

Therefore, from equation (*):

$$\text{dens}(X+Y=t) = \int_0^t \text{dens}(X=s)\, \text{dens}(Y=t-s)\, ds$$

Now we have to work in the above formulas for the densities.
Careful!

Watch me carefully.

$$= \int_0^t \frac{1}{a} I_{[0,a]}(s) \, \frac{1}{a} I_{[0,a]}(t-s)\, ds$$

$$= \int_0^t \frac{1}{a^2} I_{[0,a]}(s)\, I_{[0,a]}(t-s)\, ds$$

Now, there are <u>2 cases.</u>
Another von Neumann story:
Everytime he went to a lecture and the lecturer said "Now there are 3 cases," von Neumann got up + left. He couldn't stand it.
We are lucky here. We have only 2.

<u>Case 1</u> : $t \le a$

$$I_{[0,a]}(s) = 1 \longleftarrow \left\{ \begin{array}{l} s \text{ ranges from } 0 \text{ to } t. \\ \text{Since } t \le a, \text{ we are assured} \\ \text{that the event } s \in [0,a] \text{ happens.} \end{array} \right\}$$

$$dens(X+Y=t) = \int_0^t \frac{1}{a^2} \cancelto{1}{I_{[0,a]}(s)}\, I_{[0,a]}(t-s)\, ds$$

$$dens(X+Y=t) = \int_0^t \frac{1}{a^2} \underbrace{\cancelto{1}{I_{[0,a]}(t-s)}}\, ds$$

$$I_{[0,a]}(t-s) = 1 \longleftarrow \left\{ \begin{array}{l} s \in [0,t] \\ \text{So } t-s \in [0,t] \text{ and} \\ \text{since } t \le a, \text{ we are} \\ \text{assured that the event} \\ t-s \in [0,a] \text{ happens.} \end{array} \right\}$$

$$= \int_0^t \frac{1}{a^2}\, ds$$

$$\boxed{dens(X+Y=t) = \frac{t}{a^2} \quad , \quad t \le a}$$

- Case 2 : $a < t \le 2a$

$I_{[0,a]}(s) = 1$, as long as $s \in [0,a]$, 0 otherwise

$$\text{dens}(X+Y=t) = \int_0^t \frac{1}{a^2} I_{[0,a]}(s)\, I_{[0,a]}(t-s)\,ds$$

$$= \int_0^a \frac{1}{a^2} I_{[0,a]}(t-s)\,ds$$

restrict limits of integration to domain where $I_{[0,a]}(s) = 1$. It is 0 everywhere else.

Indicator function $I_{[0,a]}(t-s) = 1$

when: $0 \le t-s \le a$

$\Longrightarrow$ Lower bound on s:

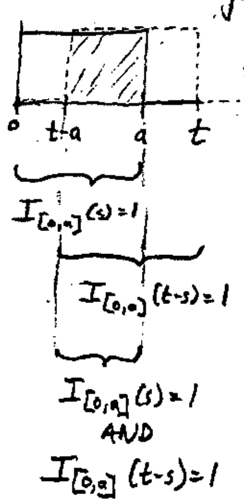
$t-a \le s$

Upper bound on s:

$s \le t$

$$I_{[0,a]}(t-s) = \begin{cases} 1 & \text{if } s \in [t-a, t] \\ 0 & \text{otherwise} \end{cases}$$

Note that the limits of integration are currently $\int_0^a$. Due to the behavior of $I_{[0,a]}(t-s)$, we can change the limits:



$\Longrightarrow \int_0^a \longrightarrow \int_{t-a}^a$

$$= \int_{t-a}^a \frac{1}{a^2}\,ds$$

$$= \frac{1}{a^2}\left(a - (t-a)\right)$$

$$= \frac{2a - t}{a^2}$$

$$\boxed{\text{dens}(X+Y=t) = \frac{2}{a} - \frac{t}{a^2}, \qquad a < t \le 2a}$$

So we have:

$$\text{dens}(X+Y=t) = \begin{cases} \frac{t}{a^2} & , \ t \le a \\ \frac{2}{a} - \frac{t}{a^2} & , \ a < t \le 2a \\ 0 & , \ \text{o.w.} \end{cases}$$

That's the density.
Now where is the proof of the pudding?
Integrating the density has to equal 1.
If it doesn't equal 1, I've made a mistake.

Check:

$$\int_{-\infty}^{\infty} \text{dens}(X+Y=t)\,dt \overset{?}{=} 1$$

$$\int_{-\infty}^{\infty} \text{dens}(X+Y=t)\,dt = \int_{0}^{2a} \text{dens}(X+Y=t)\,dt$$

$$= \int_0^a \underbrace{\text{dens}(X+Y=t)}_{\frac{t}{a^2}}\,dt + \int_a^{2a} \underbrace{\text{dens}(X+Y=t)}_{\frac{2}{a} - \frac{t}{a^2}}\,dt$$

$$= \frac{t^2}{2a^2}\Big|_{t=0}^{a} + \left[\frac{2t}{a} - \frac{t^2}{2a^2}\right]_{t=a}^{2a}$$

$$= \frac{1}{2} + \cancel{4} - \cancel{2} - \cancel{2} + \frac{1}{2}$$

$$= 1 \checkmark$$

So it *is* a density.

Now you see what you get into in this business.

- Research Problem:
  Add n independent, uniformly distributed random variables.
  Find a nice formula for:

$$\text{dens}\,(X_1 + X_2 + \dots + X_n = t) = ?$$

I want to do a seemingly disparate topic.

- Cauchy's functional equation

Suppose I have a function f that is continuous and satisfies the following functional equation:

$$f(x+y) = f(x)f(y)$$
$$\text{for } x, y \geq 0$$

What can you say about f?
What you can say is this.

$$\text{Then } f(x) = c^x$$

some constant c, raised to the x.

Why?
For the following reason.
First of all:

$$\begin{aligned} f(x+x) &= f(x)f(x) \\ &= c^x c^x \\ &= c^{2x} \end{aligned}$$

$$f(nx) = [f(x)]^n$$

simply by iteratively applying $f(x+y) = f(x)f(y)$ to:

$$f(\underbrace{x+x+\dots+x}_{n}) = \underbrace{f(x)f(x)\dots f(x)}_{n}$$

If we set $x = \frac{1}{n}$,

$$f(n\tfrac{1}{n}) = [f(\tfrac{1}{n})]^n$$

$$f(1) = [f(\tfrac{1}{n})]^n$$

And since $f(x) = c^x$,

$$f(1) = c$$

Therefore: $f(1) = c = [f(\frac{1}{n})]^n \implies [f(1)]^{\frac{1}{n}} = c^{\frac{1}{n}} = f(\frac{1}{n})$

Now, take:

$$f\left(\frac{k}{n}\right) = f\Big(\underbrace{\tfrac{1}{n} + \tfrac{1}{n} + \dots + \tfrac{1}{n}}_{k}\Big)$$

Therefore, you can apply the functional equation k times.

$$= \left[f\left(\tfrac{1}{n}\right)\right]^k$$

Using the previously proved identity;
$\left[f(1)\right]^{\frac{1}{n}} = c^{\frac{1}{n}} = f\left(\frac{1}{n}\right)$

$$= \left[f(1)\right]^{k/n} = c^{\frac{k}{n}}$$

$$\boxed{f\left(\frac{k}{n}\right) = c^{\frac{k}{n}}}$$

Therefore, for every rational number $\frac{k}{n}$, we have the conclusion that this f is assumed to be continuous.
Every number is a limit of rational numbers, so the above holds for all numbers.

$$\left.\begin{matrix} f(x+y) = f(x)f(y) \\ x, y \geq 0 \end{matrix}\right\} \Rightarrow \quad f(t) = c^t \text{ for all numbers } t$$

This is a very important fact.
The only function that satisfies the functional equation $f(x+y) = f(x)f(y)$ is $f(x+y) = c^{x+y}$

- An alternative proof ⟵ a cheapo proof, as done by physicists

Assume, in addition to being continuous, that f is differentiable.
Take the functional equation and take 1) the derivative wrt x on both sides, and then 2) " " " y on both sides.

$$\left.\begin{matrix} 1)\ f'(x+y) = f'(x)f(y) \\ 2)\ f'(x+y) = f(x)f'(y) \end{matrix}\right\} \frac{d}{dx}f(x+y) = \frac{d}{dy}f(x+y)$$

Combining gives:

$$f'(x)f(y) = f(x)f'(y)$$

Blindly assume $f(x) \neq 0$.

$$\frac{f'(x)}{f(x)} = \frac{f'(y)}{f(y)}$$

Now we apply the Principal of Ignorance.
$x + y$ do <u>not</u> know of each other's existence.

Since variables don't think, the only way to get the LHS = the RHS is if LHS = RHS = a constant:

$$\underbrace{\frac{f'(x)}{f(x)}}_{\text{only } x} = \underbrace{\frac{f'(y)}{f(y)}}_{\text{only } y} = k$$

Solve the differential equation:

$$f'(x) - kf(x) = 0$$

to obtain the exponential:

$$f(x) = c^x$$

$$\int_0^x \frac{f'(t)}{f(t)}\,dt = \int_0^x k\,dt$$

Let $u = f(t)$

$$\int_{t=0}^x \frac{1}{u}\,du = kt\Big|_0^x$$

$$\ln u\Big|_{t=0}^x = kx$$

$$\ln f(x) - \ln f(0) = kx$$

$$\ln \frac{f(x)}{f(0)} = kx$$

$$f(x) = \underbrace{f(0)}\, e^{kx}$$

John Guidi
guidi@mit.edu

Today: Some examples

Quiz: April 27 (Monday)

The examples are motivated by the fact that the marathon is coming up on Monday.

Example 1 - "Record Values"

If you watched the Winter Olympics, it was kind of exciting.
In speed skating, they have these new kind of skates, the clap skates.
Every time they ran a race, they'd get a new world's record.
Then the next race, they'd get another new world's record.
And so on.

Let's try to analyze how often this happens, where you have a lot of turnover in world record's.

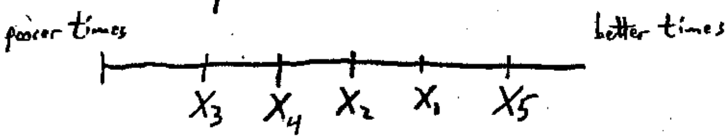
Or, another way of saying this in the opposite direction, how long will the world record you just set last.

Let $X_1, X_2, \ldots$ be an infinite number of independent and identically distributed (i.i.d.) continuous random variables (we don't care about their distribution) that represent (skater's time).

Let's assume that the X_i happen sequentially. X_1 happens, then X_2, etc.

Let $N = n$, which is the first time $X_n > X_1$ (i.e., that X_n breaks X_1's world record.)

For example:



Here, $N=5$, since X_5 is the first $> X_1$

The event $(N=n)$ is:

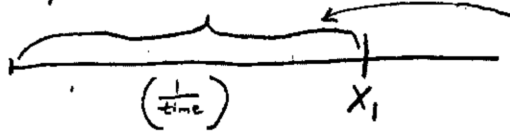
$$(N=n) = (X_2 < X_1) \cap (X_3 < X_1) \cap \ldots \cap (X_{n-1} < X_1) \cap (X_n > X_1)$$

Rather than compute this probability directly, lets get at it indirectly, by considering an easier probability. Namely the probability of the event $(N > n)$.

The event $(N > n)$ = X_1, the first racer (the first label), is larger than points at labels $X_2, \ldots, X_n$.
You just need to consider the permutation of labels.

Event $(N>n)$:

points $\overbrace{X_2, \ldots, X_n}^{n-1 \text{ points}}$ lie to the left of X_1



$$P(N>n) = \frac{(n-1)!}{n!}$$

{Total number of ways to permute labels X_2 through X_n, while keeping X_1 fixed.}

{Total number of ways of permuting labels $X_1, \ldots, X_n$.}

$$= \frac{1}{n}$$

Consider the event $(N > n-1)$. This is equivalent to:

$$(N > n-1) = \underbrace{(N=n) \cup (N>n)}_{\text{union of disjoint events}}$$

Taking probabilities:

$$P(N>n-1) = P(N=n) + P(N>n)$$

Which we can rearrange to give the probability of the event we first considered $(N=n)$:

$$P(N=n) = P(N>n-1) - P(N>n)$$

$$= \frac{1}{n-1} - \frac{1}{n}$$

$$\boxed{P(N=n) = \frac{1}{n(n-1)}}$$

Once we have this, we can then calculate the expected value:

$$E(N) = \sum_{n=2}^{\infty} n P(N=n)$$

{lower limit is 2, since you can't break a record unless a first skater sets it.}

$$= \sum_{n=2}^{\infty} \cancel{n} \frac{1}{\cancel{n}(n-1)}$$

$$\simeq \infty$$

$$\boxed{E(N) \simeq \infty}$$

X_1 holds the record and the expected amount of time that X_1 continues to hold the record is infinite.

$$(N=\infty) \Rightarrow X_\infty > X_1$$

{there are an infinite number of points until we have the first occurrence of a point $> X_1$}

So the fact that in the Olympics, these speed skaters kept skating new world records was actually quite a freak. There were people who had records all of 2 minutes. They should have expected their records to last forever.

{Note that the differing athletic prowess, abilities, and performance capacities were not taken into consideration in arguing $P(N>n) = \frac{1}{n}$. We've blithely assumed all athletes are equal, which is a gross oversimplification.}

Example 2 – "$E(X+C) \neq E(X)+C$" ← C is a constant

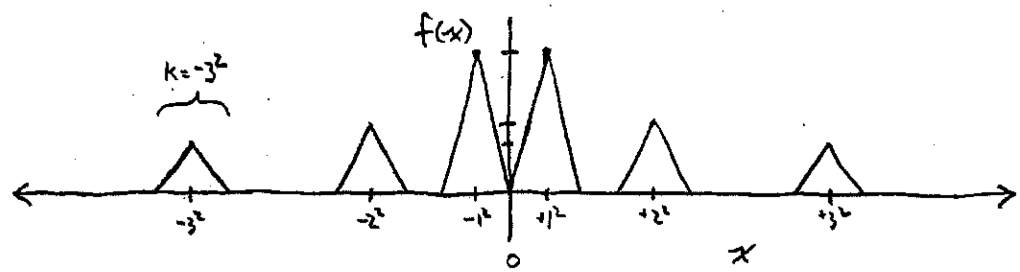
this should shock you.

Take $a = \frac{3}{\pi^2}$

Let X = continuous random variable with density:

k all squared integers

$$\text{dens}(X=x) = f(x) = \begin{cases} a\frac{1}{|k|}\left(1-|k-x|\right) & \text{for all } k = \pm 1^2, \pm 2^2, \ldots \text{ and } |k-x| \le 1 \\ 0 & \text{otherwise} \end{cases}$$



Just looking at this, you can guess the expected value of X, because of the symmetry about $x=0$.

$$E(X) = \int_{-\infty}^{\infty} x \overbrace{\text{dens}(X=x)}^{f(x)}\, dx$$

$$= 0$$

This gives us that:

$$E(X) + C = 0 + C$$

$$\boxed{E(X) + C = C}$$

Next, I have to show you that the $\overbrace{\text{expected value of } X+C}^{E(X+C)}$ is something different.

First, we note that $f(x)$ is a valid density function.
Check that integration equals 1.

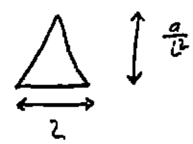
$$\int_{-\infty}^{\infty} f(u)\,du \stackrel{?}{=} 1$$

$$\int_{-\infty}^{\infty} f(u)\,du \;=\; \text{area under the curve above}$$

Because $f(x)$ is symmetric about $x=0$, we can focus on the curve to the right of the origin.

$$= 2\int_0^{\infty} f(u)\,du$$

Each triangle for $k=i^2$, $i=1,\ldots,\infty$, has area:



$$\text{Area} = \tfrac{1}{2}\,2\,\frac{a}{i^2} = \frac{a}{i^2}$$

$$= 2\sum_{i=1}^{\infty}\frac{a}{i^2}$$

$$= 2a\underbrace{\sum_{i=1}^{\infty}\frac{1}{i^2}}$$

This is one of those beautiful identities you can prove compliments of the Fourier Series expansion of $x+x^2$.

$$\sum_{i=1}^{\infty}\frac{1}{i^2} = \frac{\pi^2}{6}$$

$$= 2a\,\frac{\pi^2}{6}$$

And note that we had conveniently set

$$a = \frac{3}{\pi^2}$$

which ensures that the above equals 1.

$$\int_{-\infty}^{\infty} f(u)\,du = 1 \quad \checkmark$$

Which confirms that $f(x)$ is a valid density function.

$$E(X+C) = \int_{-\infty}^{\infty} (x+C)\,f(x+C)\,d(x+C)$$

$$= \lim_{b\to\infty} \int_{-b-C}^{b-C} (u+C)\,f(u)\,du$$

We break this integral up into <u>3 parts</u>:

$$= \lim_{b\to\infty}\left[\underbrace{\int_{-b-C}^{b+C} u\,f(u)\,du}_{\substack{\text{Since } f(u) \text{ is symmetric about } x=0, \text{ and the limits of integration are}\\ \text{negatives of each other, this equals } 0.\\ 0 = E(X)}} - \underbrace{\int_{b-C}^{b+C} u\,f(u)\,du}_{} + \underbrace{\int_{-b-C}^{b-C} C\,f(u)\,du}_{\lim_{b\to\infty} C \underbrace{\int_{-b-C}^{b-C} f(u)\,du}_{\text{This integral evaluates to 1 in the limit, as } f(u) \text{ is a valid density}}} \right]$$

In order that $E(X+C) \neq E(X)+C$, this integral must be $\neq 0$.

$$= C$$

The trick is to note the following:

$$\int_{k-1}^{k+1} u\,f(u)\,du \;\geq\; \underbrace{(k-1)}_{\text{lowest limit}} \int_{k-1}^{k+1} f(u)\,du \;\Leftarrow\; \left\{ \int_{k-1}^{k+1} u\,f(u)\,du \geq \int_{k-1}^{k+1} (k-1)\,f(u)\,du \right.$$

for any $k = 2^2, 3^2, 4^2, \ldots$, the above holds true

$$\int_{k-1}^{k+1} f(u)\,du = \text{the area of the triangle}$$
$$= \frac{1}{2}\, 2\, \frac{a}{k}$$
$$= \frac{a}{k}$$

$$\boxed{\int_{k-1}^{k+1} u\,f(u)\,du \;\geq\; (k-1)\frac{a}{k} \;>\; 0 \quad \text{for any } k = 2^2, 3^2, 4^2, \ldots}$$

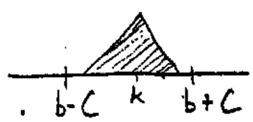
With the knowledge that $\int_{k-1}^{k+1} u f(u)\,du > 0$ (i.e., the limits of integration are those of the triangle centered at $x = k$) for $k = 2^2, 3^2, \ldots$,
Let's consider, again, the segment:

$$\lim_{b\to\infty} \int_{b-C}^{b+C} u f(u)\,du$$

Let's impose that $C > 1$.

1) If, for some large value of b, there is a squared integer k, such that $[k-1, k+1] \in [b-C, b+C]$,

then $$\int_{b-C}^{b+C} u f(u)\,du > 0$$



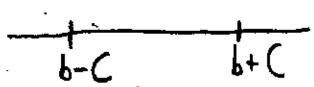
In other words, the area of the triangle for k is included in the above integral.
Since this area is non-zero, this lends credence to the claim that:

$$E(X+C) \neq E(X) + C$$

2) BUT

If $[k-1, k+1] \not\subseteq [b-C, b+C]$, in other words $[b-C, b+C]$ does not cover any triangle, then:

$$\int_{b-C}^{b+C} u f(u)\,du = 0$$



So, in fact, the limit:

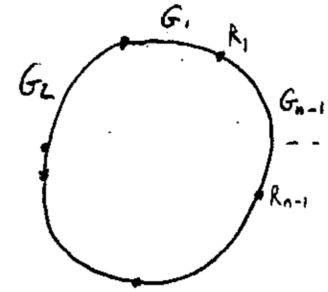
$$\lim_{b\to\infty} \int_{b-C}^{b+C} u f(u)\,du$$

does not exist. The above two cases are disparate and, as a result, the above limit is not defined.

Example 3 – "Runners on a track"

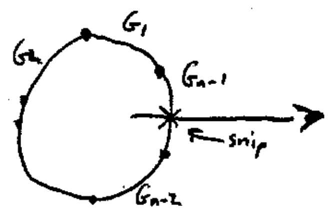
You have $(n-1)$ runners on a circular track of circumference a.

Let $G_1, G_2, \ldots, G_{n-1}$ be gaps between runners



← Set up a camera near the track
What is the distribution of runners near the camera?
(i.e., what is the distribution of gaps nearest the camera?)

Let G_{n-1} be the gap nearest the camera



We can think of this as a uniform process, where we snip the circle at the point closest to the camera.

$\Downarrow$ unrolling

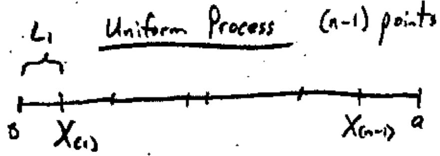
← $(n-1)$ points on interval $[0, a]$

$G_1, G_2, \ldots, G_{n-2}$ are equidistributed according to the same distribution as L_1. $P(G_1 > t) = \ldots = P(G_{n-2} > t) = P(L_1 > t)$.

↖ first gap in the uniform process

See [3/18/98.12-13, 3/30/98.11-12]

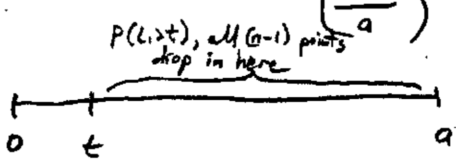
Uniform Process $(n-1)$ points dropped on interval $[0, a]$



event $(L_1 > t) = (X_{(1)} > t)$

$$P(G_1 > t) = P(L_1 > t) = P(X_{(1)} > t)$$

$$= \left(\frac{a-t}{a}\right)^{n-1}$$

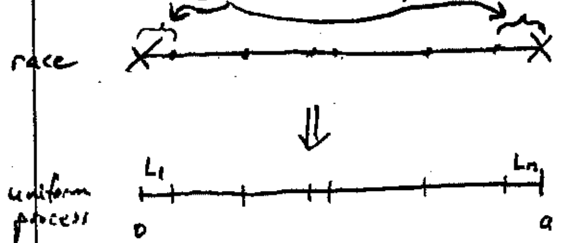


$$\underbrace{P(G_1 > t) = P(G_2 > t) = \dots = P(G_{n-2} > t)}_{\text{gaps } G_1 \text{ through + including } G_{n-2} \text{ are equidistributed, even though not independent}} = P(L_1 > t) = \left(\frac{a-t}{a}\right)^{n-1}$$

gaps G_1 through + including G_{n-2} are <u>equidistributed</u>, even though not independent

- <u>However</u> – gap G_{n-1} has a <u>different</u> distribution

  <u>G_{n-1}</u> is the gap that was closest to the camera and was subsequently snipped.



$$G_{n-1} = L_1 + L_n$$

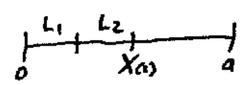
event $(G_{n-1} > t) = (L_1 + L_n > t)$

Recall that gaps in the uniform process are <u>exchangable</u>

$$= (L_1 + L_2 > t)$$

Taking the probabilities of these events:

$$P(G_{n-1} > t) = P(\underbrace{L_1 + L_2}_{X_{(2)} = L_1 + L_2} > t)$$



The <u>second order statistic</u> is precisely the <u>sum $L_1 + L_2$</u>.

$= P(X_{(2)} > t)$ ← ⓐ we can compute this de nuovo

– or –

$= 1 - P(X_{(2)} \le t)$ ← ⓑ since we already have computed $\text{dens}(X_{(2)} = t)$ [3/18/98.8-10], we can compute this readily, since:

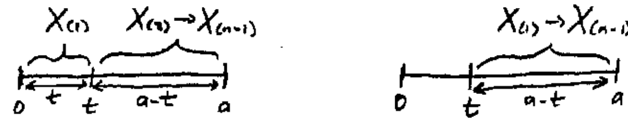
$$P(X_{(2)} \le t) = \int_{-\infty}^{\infty} \text{dens}(X_{(2)} = s)\,ds$$

ⓐ Let's compute $P(X_{(2)} > t)$ de nuovo:

$$\text{event } (X_{(2)} > t) = \big((X_{(1)} \le t) \cap (X_{(2)} > t)\big) \cup \big((X_{(1)} > t) \cap (X_{(2)} > t)\big)$$

union of disjoint events

$$P(X_{(2)} > t) = P\big((X_{(1)} \le t) \cap (X_{(2)} > t)\big) + P\big((X_{(1)} > t) \cap (X_{(2)} > t)\big)$$



$$= \binom{n-1}{1,\, n-2} \frac{t}{a} \left(\frac{a-t}{a}\right)^{n-2} + \left(\frac{a-t}{a}\right)^{n-1}$$

$$P(X_{(2)} > t) = (n-1)\,\frac{t\,(a-t)^{n-2}}{a^{n-1}} + \left(\frac{a-t}{a}\right)^{n-1}$$

And we have:

$$\boxed{P(G_{n-1} > t) = (n-1)\frac{t\,(a-t)^{n-2}}{a^{n-1}} + \left(\frac{a-t}{a}\right)^{n-1}}$$

{Explaination of apparent contradiction.}

And we observe that:

uniform process w/ n-2 points

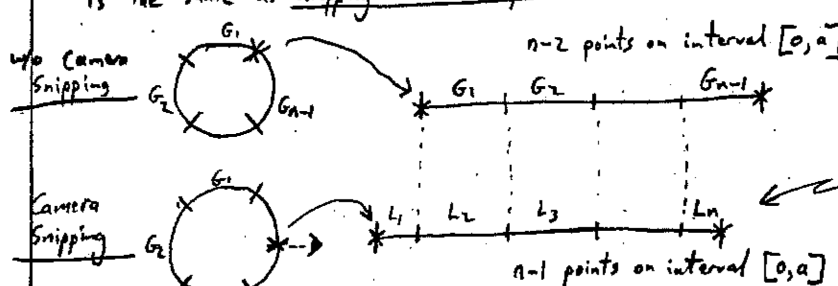
But probability calculated on basis of uniform process w/ n-1 points.

$$P(G_1 > t) = \dots = P(G_{n-2} > t) = \left(\frac{a-t}{a}\right)^{n-1} \neq P(G_{n-1} > t) = \left(\frac{a-t}{a}\right)^{n-1} + (n-1)\frac{t(a-t)^{n-2}}{a^{n-1}}$$

On the surface, this seems kind of strange.
Simply by observing this process at a certain point, I changed the probability distribution of the observed gap. Now the gap nearest to the camera is always bigger:

$$P(G_{n-1} > t) > \big(P(G_1 > t) = \dots = P(G_{n-2} > t)\big)$$

This is an apparent paradox.
How did the runners (i.e., gap G_{n-1}) know that the camera was there?
Well, they don't. There is no paradox.
Snipping the circle at the point in the gap G_{n-1} that the camera falls on is the same as dropping an extra point in the circle that breaks gap G_{n-1}.



There is an extra point on this interval.

Example 4 – "Inspector's Paradox"

An inspector is sent to find out something about the process of milk being sold in a certain store.
Milk turns out to be sour every once in a while.

Any particular bottle of milk has probability p of being sour.
The inspector wants to know, when she arrives, what the expected amount of time between the last sour bottle of milk and the next sour bottle of milk is. We are assuming there is at least one sour bottle of milk.

Let integer random variable $T = I + J$

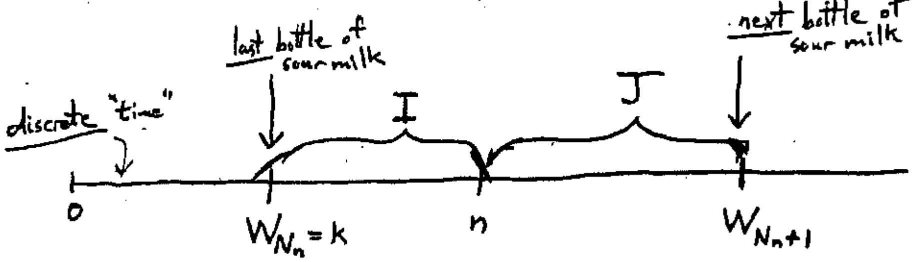
$$I = n - \underbrace{W_{N_n}}_{\text{waiting time for } N_n^{th} \text{ sour bottle of milk.}}$$

N_n = number of sour milk bottles in first n bottles
At the $W_{N_n}^{th}$ time step, there was a bottle of sour milk

I = "time" (measured in bottles) since the last bottle of sour milk arrived. Includes both point W_{N_n} and point n.

$$J = W_{N_n+1} - n$$

J = "time" from n^{th} bottle until the next bottle of sour milk arrives.
Does not include point n, but includes point W_{N_n+1}.



So the integer random variable $T = I + J$ is just a gap in the Bernoulli Process.
We want to figure out the distribution of the gaps in this process, given that the inspector has arrived.

The inspector takes data from various stores, noting the length of the gap from the point when she arrives for inspection.

The inspector claims that $P(T=k) \neq P(\underbrace{W_1 = k}_{})$

Implying that, somehow, gap T is special.

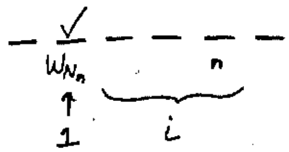
event that the 1^{st} waiting time for a sour bottle of milk.

First, let's determine the distributions of I and J.

$$P(I=i) = p\,q^{i}$$

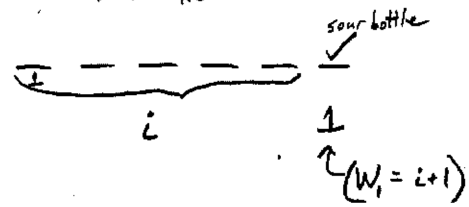
there was a sour bottle of milk at W_{N_n}

then the following i bottle's of milk were <u>not</u> sour.



$$P(W_1 = i+1) = q^{i}p$$

W_1 = 1st waiting time for a bottle of sour milk

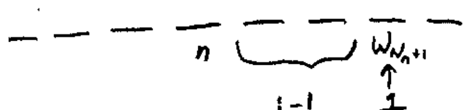


$$P(I=i) = P(W_1 = i+1) = pq^{i}$$

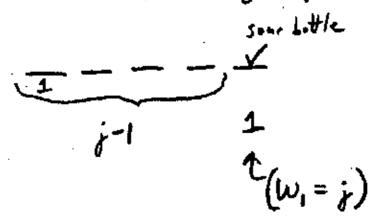
$$P(J=j) = q^{j-1}\,p$$

<u>not</u> including point n

there is a bottle of sour milk at W_{N_n+1}



$$P(W_1 = j) = q^{j-1}p$$

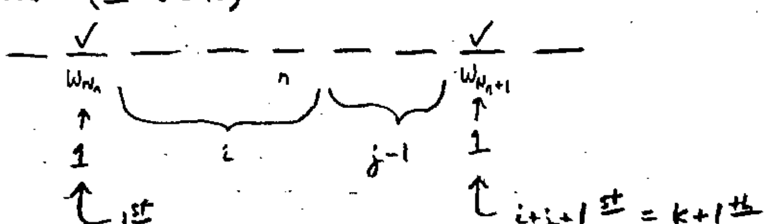


$$P(J=j) = P(W_1 = j) = q^{j-1}p$$

So we have:

$$P(T=k) = P(I+J=k)$$

Event $(I+J=k)$



$$\text{Event } (I+J=k) = \overbrace{(W_2 = k+1)}^{\Uparrow}$$

{Event $(I+J=k)$ equivalent to waiting time for the 2nd sour bottle of milk at $k+1$.}

Hence:

$$P(I+J=k) = P(W_2 = k+1) \neq P(W_1 = k)$$

The apparent paradox is that someone comes and observes the gap T and, because of this, changes the probability distribution of $P(T=k)$ so that it is no longer:

$$P(T=k) \neq P(W_1=k)$$

The explaination of the apparent contradiction is to note that the manner in which we defined random variable $T = I + J$ requires the occurrence of 2 sour bottles of milk, one at each end of the gap T.

{Thermodynamics is a sore thumb of mathematics.
It is a major failure of Mathematicians.}

The Poisson Process (beginning)

So far, we've studied basically 2 stochastic processes - the Bernoulli Process and the Uniform Process.

Both of these processes correspond to very intuitive experiments:

Bernoulli Process - corresponds to tossing the same coin repeatedly

Uniform Process - corresponds to picking n points on an interval

Now we have to bite the bullet and face up to the fact that there are, in nature, stochastic processes that we didn't know about.
And we have to adjust our intuition accordingly.
Not everything is intuitive at first, otherwise you wouldn't have anything to learn.

These two new processes we will cover from today through the end of the course.
They are:

Poisson Process
Theory of the Normal Distribution
} They are the central core of both probability and statistics.

You could say that everything we've done so far was preparatory to studying these stochastic processes.

We begin w/ the Poisson Process.
The Poisson Process exists in nature. It is extremely common.
But it is something you have not heard about, probably.

It's good to develope it by developing an analogy with the Bernoulli Process,
So we have a table of analogies - what happens w/ Bernoulli
what happens w/ Poisson

Philosophically, Poisson is continuous Bernoulli.

So let's go back to Bernoulli and do Bernoulli from a different point of view.

Back to Bernoulli

You remember that when we did Bernoulli, we introduced the waiting time W_1 [2/20/98.10-13].

W_1 = waiting time for first heads

As it's easy enough to show:

$$P(W_1 > n) = q^n$$

Because the event $(W_1 > n)$ means that the <u>first n</u> tosses are <u>tails</u>.

$$(W_1 > n) = \Omega - (W_1 = 1) - (W_1 = 2) - \dots - (W_1 = n)$$

$$P(W_1 > n) = P[\Omega - (W_1 = 1) - (W_1 = 2) - \dots - (W_1 = n)]$$

all disjoint events, and since probabilites subtract:

$$= P(\Omega) - P(W_1 = 1) - P(W_1 = 2) - \dots - P(W_1 = n)$$

$$= 1 - p - qp - \dots - q^{n-1}p$$

$$= 1 - p\underbrace{(1 + q + q^2 + \dots + q^{n-1})}_{\frac{1-q^n}{1-q}}$$

$$= \cancel{1} - \cancel{p}\frac{(\cancel{1} - q^n)}{\cancel{(1-q)}} \leftarrow 1 - q = p$$

$$P(W_1 > n) = q^n \quad \checkmark$$

Lets compute this probability now:

$$P(W_1 > n + k \mid W_1 > n) = ?$$

We could call this the <u>fools paradise probability</u>.
Everybody who plays <u>roulette</u> believes, if you waited <u>n times for red $(W_1 > k)$</u> ← {event that first n spins are <u>not</u> red}

$$\underbrace{P(W_1 > n + k \mid W_1 > n)}_{\text{whats the probability that we wait k more times, given that you've already waited n times.}}$$ becomes <u>more</u> probable.

By definition of conditional probability:

$$P(W_1 > n+k \mid W_1 > n) = \frac{P((W_1 > n+k) \cap (W_1 > n))}{P(W_1 > n)}$$

The event $(W_1 > n+k)$ contains $(W_1 > n)$

$$(W_1 > n+k) \supseteq (W_1 > n)$$

So the intersection is:

$$(W_1 > n+k) \cap (W_1 > n) = (W_1 > n+k)$$

$$= \frac{P(W_1 > n+k)}{P(W_1 > n)}$$

$$= \frac{q^{n+k}}{q^n}$$

$$P(W_1 > n+k \mid W_1 > n) = q^k$$

$P(W_1 > k)$

And we have that:

$$\boxed{P(W_1 > n+k \mid W_1 > n) = P(W_1 > k)}$$

In other words, the fact that you already waited n times does not change the probability that you have to wait k more times.

Facts are facts.

- We could have developed the entire theory of the Bernoulli Process using the random variable W_1.
Because we can define all other events pertaining to the Bernoulli Process in terms of the random variable W_1.
For example, sums:
$W_2 = W_1 + W_1$
{waiting time for second}

so forth and so on.

It's an interesting exercise.

That's what we are forced to do when we do the continuous analogue, for the Poisson process.

Buoyed by this example, we ask the following question:

- Is there a continuous waiting time that has the same property:

$$P(W_1 > n+k \mid W_1 > n) = P(W_1 > k)$$

Let's give this property a name.
It's called memoryless property.

W_1, the waiting time for the first head in the Bernoulli Process is memoryless.

A waiting time for something to happen is memoryless when knowledge that you have already waited so many units of time doesn't make the event more probable.

Now, we have the language to express that.
And we can ask the following question:

Is there a continuous analogue of this waiting time, where n is replaced by any positive number t, representing continuous time?

- Is there a continuous random variable $\overbrace{T \geq 0}^{\text{times are non-negative}}$ that represents a memoryless waiting time?

The remarkable answer is Yes.
There is one, and only one.

{The interesting situation is when there is only one. We have that situation here.}

This is very remarkable that there is only one.
So, we will show there is only one.

- Let's write precisely what we mean:

$$P(T > t+s \mid T > t) = P(T > s)$$

This is the memoryless property we want.

On the basis of this property alone, in a couple of lines, we derive the random variable.

By the definition of conditional probability:

$$P(T>\cancel{t+s} \mid T>t) = \frac{P((T>t+s) \cap (T>t))}{P(T>t)}$$

The event $(T>t+s) \supseteq (T>t)$.

If $(T>t+s)$ happens, then a posteriori $(T>t)$ happens.

$(T>t+s) \cap (T>t) = (T>t+s)$

$$P(T>t+s \mid T>t) = \frac{P(T>t+s)}{P(T>t)}$$

And since the memoryless property we want is (bottom of previous page):

$$P(T>t+s \mid T>t) = P(T>s)$$

Combining, we want:

$$\frac{P(T>t+s)}{P(T>t)} = P(T>s)$$

(*) $$\boxed{P(T>t+s) = P(T>t)\,P(T>s)}$$

- Set $P(T>t) = f(t)$

Then equation (*) becomes:

$$f(t+s) = f(t)f(s)$$

We saw this two lectures ago [4/15/98.14-16].

It's Cauchy's functional equation, which we elaborately solved.

Assuming continuity alone, we proved that [4/15/98.15]:

$$f(t) = c^t \longleftarrow \text{which can be written as: } e^{-\alpha t}$$

Let $c = e^{-\alpha} \Rightarrow c^t = (e^{-\alpha})^t = e^{-\alpha t}$

$$P(T>t) = \underbrace{f(t) = e^{-\alpha t}}_{\text{this has to be a probability.}}$$

From Cauchy's functional equation, we have:

$$\boxed{P(T>t) = e^{-\alpha t}}$$

- Therefore, the <u>density</u>:

$$\text{dens}(T=t) = \frac{d}{dt} P(T \leq t)$$

$$= \frac{d}{dt}\left[1 - P(T>t)\right]$$

$$= \frac{d}{dt}\left(1 - e^{-\alpha t}\right)$$

$$\boxed{\text{dens}(T=t) = \alpha e^{-\alpha t}}$$

- Now we can find the value of α.
We know that densities are required to integrate to 1.

$$1 = \int_{-\infty}^{\infty} \text{dens}(T=t)\,dt = \int_0^{\infty} \text{dens}(T=t)\,dt$$

$$= \int_0^{\infty} \alpha \underbrace{e^{-\alpha t}}\,dt$$

$e^{-\alpha t} > 0$, which implies that α is <u>positive</u>.
$\alpha > 0$

$$1 = \int_0^{\infty} \alpha e^{-\alpha t}\,dt$$

$$= \alpha \int_0^{\infty} e^{-\alpha t}\,dt$$

Let $u = -\alpha t$
$du = -\alpha\,dt$

$$= \cancel{\alpha} \int_{t=0}^{\infty} e^{u}\left(\frac{-1}{\cancel{\alpha}}\right) du = -e^{-\alpha t}\Big|_{t=0}^{\infty} = (-0) - (-1) = 1$$

{Vacuously true, <u>Regardless</u> of α.}

So: $\int_0^\infty \alpha e^{-\alpha t}\, dt = 1$, for any $\alpha > 0$.

therefore, for every α positive, there is a random variable T, whose density is:

$$\text{dens}(T=t) = \alpha e^{-\alpha t}$$

and which represents a memoryless waiting time.

{And we have seen that this is the only one.
From the definition, we've derived the only possible solution.}

- Let's compute the expectation:

$$E(T) = \int_{-\infty}^{\infty} t \,\text{dens}(T=t)\, dt$$

$$= \int_0^\infty t\, \alpha e^{-\alpha t}\, dt$$

$$= \alpha \int_0^\infty t e^{-\alpha t}\, dt$$

$$\int \underbrace{t}_{u}\, \underbrace{e^{-\alpha t}\, dt}_{dv}$$ Calculate this via integration by parts:

$\boxed{\begin{matrix} u = t \\ du = dt \end{matrix}}$ $\boxed{dv = e^{-\alpha t}\, dt}$

$$v = \int e^{-\alpha t}\, dt$$

let $w = -\alpha t$

$dw = -\alpha\, dt$

$$v = \int e^w \left(\frac{-1}{\alpha}\right) dw$$

$$= \frac{-1}{\alpha} e^w$$

$$\boxed{v = \frac{-1}{\alpha} e^{-\alpha t}}$$

$$\int \underbrace{t}_{u}\, \underbrace{e^{-\alpha t} dt}_{dv} = \underbrace{t}_{u} \underbrace{\left(\frac{-1}{\alpha}\right) e^{-\alpha t}}_{v} - \int \underbrace{\left(\frac{-1}{\alpha}\right) e^{-\alpha t}}_{v} \underbrace{dt}_{du}$$

$$= -\frac{t}{\alpha} e^{-\alpha t} + \frac{1}{\alpha} \underbrace{\int e^{-\alpha t} dt}$$

we've already computed this

$$= \frac{-t}{\alpha} e^{-\alpha t} + \frac{1}{\alpha}\left(\frac{-1}{\alpha}\right) e^{-\alpha t}$$

$$\int t e^{-\alpha t} = \frac{1}{\alpha}\left(-t - \frac{1}{\alpha}\right) e^{-\alpha t}$$

Which gives:

$$= \cancel{\alpha}\left[\frac{1}{\cancel{\alpha}}\left(-t - \frac{1}{\alpha}\right) e^{-\alpha t}\right]_{t=0}^{\infty}$$

$$= 0 + \left(0 + \frac{1}{\alpha}\right)$$

$$\boxed{E(T) = \frac{1}{\alpha}}$$

- A memoryless waiting time is completely determined by its expectation. For every expectation, there is a memoryless waiting time, which has density:

$$\text{dens}(T=t) = \alpha e^{-\alpha t}$$

and cumulative probability distribution:

$$P(T \le t) = 1 - e^{-\alpha t}$$

as the one, and only one, density and cumulative probability distribution.

This distribution is called the exponential distribution.

- Now you ought to ask a question.
What question must you ask?

Q: What are the analogues of the second, and so forth, waiting times?
That's a good question.
It's not the one I had in mind, but it's good.
I'll answer your question, but let me ask the question you might have asked instead.

The question you must ask is this:

Q: Where on earth is the sample space?
Random variables exist on a sample space.
We worked out this w/o a sample space!
Where is it? Where are the sample points?
Where are the events?
What's the probability?

That's the question.
See, all this makes sense, but it assumes some sample space has been taken.

Making the sample space explicit is the theory of the Poisson Process.

- There are many natural phenomena that obey the exponential distribution. For example, you are at a telephone booth and someone is using the phone. The waiting time for that person to finish is an exponential distribution. The fact that you have already waited so many minutes does not make more probable that the person will finish.
  Although, the expected waiting time is consumed some by that person.

- The confusion in people's minds, both in the case of winning at roulette and in this continuous case of the telephone booth, is the confusion between:

  the expected waiting time —and— the conditional probability of waiting more than you have so far.

  $$E(T) \qquad\qquad P(T > t+s \mid T > t)$$

  You have to make a careful distinction between these two.
  That's what you learn probability for.
  And you make money.

- So any situation where something can terminate with a random time

  You are waiting for something to happen.
  You know, on the average, but you don't know exactly when it happens.
  All the data you have is the average and it fits the exponential distribution.
  Marvellous, and it works perfectly.

- But the question is:

  Where is the sample space?

  So now we construct a sample space on which the exponential distribution exists.
  For example, if we just had the waiting time for Bernoulli $P(W_1 > n) = q^n$, then we would have known, sooner or later, that the sample space consists of sequences of 0 and 1, and that W_1 is the waiting time for the first 1.

  $\omega = 0\,0\,1\,0\,1 \ldots$

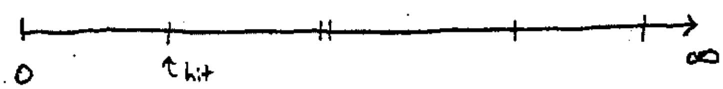
  sample point is sequence of 0's and 1's

We reason <u>backwards</u> from the waiting time W_1, and <u>derive</u> the full sample space of sequences of 0's and 1's, from just analysis of this waiting time.
Now we are in the situation where we are forced to do this <u>backwards reasoning</u> from <u>random variable</u> to <u>derivation</u> of <u>sample space</u> in spades, Because we don't have anything – it's <u>continuous</u>.

- Let me give you first the physicist's definition of the Poisson Process.

<u>Physicist's Definition of the Poisson Process</u>

You have the interval $[0, \infty)$



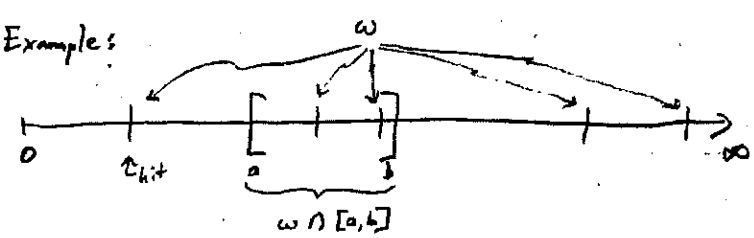
And, at any instant, there is an <u>infinitesimal little person</u>, who is tossing an infinitesimal coin.
The coin is infinitesimally expected to be 1, but mostly it is 0.

So mostly you get 0's, and every once in a while, you get a hit (1).
The idea is that the hits are <u>rare events</u>.

<u>Definition</u> – a <u>rare subset</u> ω, of hits

$\omega \subseteq [0, \infty)$ is a subset of $[0, \infty)$, with the property that for every finite interval $[a, b]$, the set $\omega \cap [a, b]$ is finite.

Example:

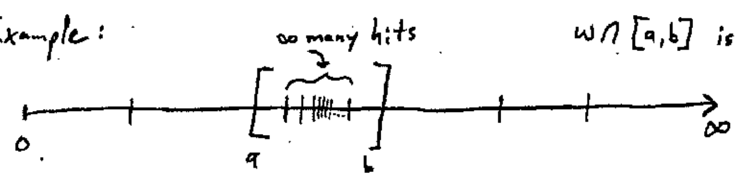


ω is a rare subset

if you take <u>any</u> finite interval $[a, b]$, then only a <u>finite number of hits</u> are in it.
Every finite interval has at most finitely many hits.

For a rare subset, you can <u>not</u> have any <u>condensation</u> (i.e., an infinite number of hits) anywhere. In <u>any</u> finite interval, there can not be an infinite number of hits.

Example:



$\omega \cap [a,b]$ is <u>infinite</u>

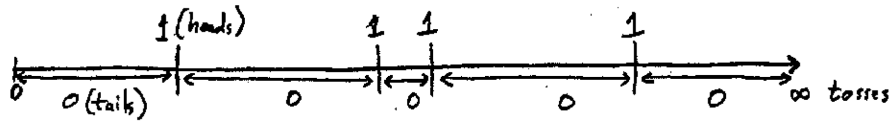
ω is <u>not</u> a rare subset

Saying a subset is rare is a way of avoiding this crisis.
Rare means that it happens every once in a while.

Our sample space Ω consists of all rare subsets of $[0,\infty)$

$$\text{Sample Space } \Omega = \underline{\overset{\text{set of}}{\text{all}}} \text{ rare subsets of } [0,\infty)$$

- A rare subset is analogous to the 1's in the Bernoulli Process.
The elements not in the rare subset are the 0's.
The idea is that we have phenomena that occur every <u>once in a while.</u>
It occurs, then you wait, it occurs, you wait, it occurs, ...
That's the idea and this is <u>carried</u> by the word <u>rare</u>.



{The sample space $\Omega_{Poisson}$ = <u>set</u> of all rare subsets of $[0,\infty)$.
Compare to sample space $\Omega_{Bernoulli}$ = all infinite sequences of 0's and 1's.}

when you consider the rare subset in analogy to the Bernoulli Process, it's a sequence of 0's and 1's, but <u>mostly</u> 0's. Intuitively. You can't say that, of course.

- You should view a <u>rare subset</u> as a <u>sequence of hits.</u>
Or a <u>time</u> when an event happened, i.e., when the phone booth becomes available, or a signal comes in from outer space.

- So this is the analogue of all sequences of 0's and 1's in Bernoulli.
Now we have to define <u>events</u> and <u>their probabilities.</u>

In the case of Bernoulli, how did we define events?
The event was $(W_i = n) \Leftarrow$ {you had the i^{th} head at the n^{th} toss.}

You <u>can't</u> do that <u>now</u>, because here things are <u>continuous.</u>
So we have to do some sort of continuous analogue of that.

Poisson Event:

$$\begin{bmatrix} (a,b] \\ n \end{bmatrix} = \left\{ \omega \in \Omega : \left| \omega \cap (a,b] \right| = n \right\}$$

This looks like a rather complicated definition.
Let's try to re-express it in <u>probabilistic language.</u>

Given ω = a rare set

Then you can always write ω as a sequence of hits (blips):

$$\omega = (\omega_1, \omega_2, \omega_3, \dots) \quad \text{blips}$$

where $\omega_1 < \omega_2 < \omega_3 < \dots$

$\omega_n \to \infty$

{Whereas in the Bernoulli Process, we had sequences of 0's and 1's, in the Poisson Process, here we just note the places where the 1's happen.}

$\begin{bmatrix} (a, b] \\ n \end{bmatrix}$ is the event that we get n blips between time interval a and time interval b.

Probabilistically, that's the event that we get n blips in the interval $(a, b]$.

These are not all of the events, of course.
We start w/ these elementary events, as we've done w/ other processes, and obtain all possible events:

Any subset of Ω obtained from the elementary events $\begin{bmatrix} (a, b] \\ n \end{bmatrix}$ by taking unions, intersections, complements, and unions of countably disjoint events shall be called events.

{In practice, if you want to show that there are subsets of Ω that are not events, you have to use the axiom of choice.
But, for practical purposes, any subset of Ω is an event.}

Any property of rare sets (or blips) that you want to talk about is computed from these basic properties.
An event is the rare set where we get n blips in the interval $(a, b]$.

- Now, I'll dish you out the probability.

Then, we'll spend an hour justifying the probability until you realize that this is what it must be.
I'll just tell you what it is. It shall be set, because I say so, to:

$$(*) \quad \text{Set } \underset{\text{(probability depends on } \alpha)}{P_\alpha} \underbrace{\left(\begin{bmatrix} (a, b] \\ n \end{bmatrix} \right)}_{\text{probability that you get n blips in interval } (a,b]} = \frac{(b-a)^n \alpha^n}{n!} e^{-(b-a)\alpha}$$

Now we have the following theorem:

Theorem - This definition makes sense

What does this theorem mean?
What does it mean to make sense?
Two things:

1) This definition is not contradictory

You can not manipulate it and get two different probabilities

2) This definition extends to all events

By the rules of boolean algebra, you can extend it to all events.

It is non-trivial to prove all the details.
But let's just check the main point.

Main Point:
The main point is this:

$$\begin{bmatrix} (a,b] \cup (b,c] \\ n \end{bmatrix} = \begin{bmatrix} (a,c] \\ n \end{bmatrix}$$

These events are equal, as the intervals $(a,b]$ and $(b,c]$ don't have any overlap and $(a,b] \cup (b,c]$ has no gaps.

This, in turn, is equal to:

$$\begin{bmatrix} (a,c] \\ n \end{bmatrix} = \bigcup_{i=0}^{n} \left(\begin{bmatrix} (a,b] \\ i \end{bmatrix} \cap \begin{bmatrix} (b,c] \\ n-i \end{bmatrix} \right)$$

Let's argue this probabilistically.
The event that you get n blips in the interval $(a,c]$ means that you get some blips in the interval $(a,b]$ and the rest of the blips in the interval $(b,c]$.
That's exactly what this says.

This is the set theoretic analogue of the binomial theorem.
It's really trivial.

This is a disjoint union.
Therefore, if our probability equation (*) is to be consistent, then we better have:

$$P_\alpha(\text{LHS}) = P_\alpha(\text{RHS})$$

Otherwise, we are in deep trouble.

Proof $P_\alpha(LHS) = P_\alpha(RHS)$ 4/22/98.14

$$P_\alpha(LHS) = P_\alpha\left(\begin{bmatrix}(a,c]\\ n\end{bmatrix}\right) = \frac{(c-a)^n \alpha^n}{n!} e^{-(c-a)\alpha}$$

(**)

$$P_\alpha(RHS) = P_\alpha\Bigg(\underbrace{\bigcup_{i=0}^{n}\left(\begin{bmatrix}(a,b]\\ i\end{bmatrix} \cap \begin{bmatrix}(b,c]\\ n-i\end{bmatrix}\right)}_{\text{disjoint union}}\Bigg) = \sum_{i=0}^{n} P_\alpha\left(\begin{bmatrix}(a,b]\\ i\end{bmatrix} \cap \begin{bmatrix}(b,c]\\ n-i\end{bmatrix}\right)$$

There is no reason why these should be equal.
Because I've left out an axiom, for the purposes of drama.
The only way these can be equal is if we have the following axiom, in addition to (*).

- **Axiom**

If $(a,b]$ and $(c,d]$ are disjoint intervals and k and n any non-negative integers, the events:

$$\begin{bmatrix}(a,b]\\ n\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}(c,d]\\ k\end{bmatrix}$$

are independent.

In other words, the number of blips occuring in a given interval does not influence the number of blips occuring in a disjoint interval.
The blips are random.

- With this assumption of the axiom of independence, equation (**) simplifies to:

$$\left\{\begin{array}{l}\text{events } A+B \text{ independent}\\ \text{has:}\\ P(A \cap B) = P(A)P(B)\end{array}\right\} \Longrightarrow \quad = \sum_{i=0}^{n} P_\alpha\left(\begin{bmatrix}(a,b]\\ i\end{bmatrix}\right) P_\alpha\left(\begin{bmatrix}(b,c]\\ n-i\end{bmatrix}\right)$$

Using the definition we have for P_α, i.e., equation (*):

$$= \sum_{i=0}^{n} \overbrace{\frac{(b-a)^i \alpha^i}{i!} e^{-(b-a)\alpha}} \; \overbrace{\frac{(c-b)^{n-i} \alpha^{n-i}}{(n-i)!} e^{-(c-b)\alpha}}$$

$e^{-(c-a)\alpha}$

$$= \sum_{i=0}^{n} \alpha^n \frac{1}{i!(n-i)!} (b-a)^i (c-b)^{n-i} e^{-(c-a)\alpha}$$

Multiply by $1 = \frac{n!}{n!}$

$$= \frac{\alpha^n}{n!} e^{-(c-a)\alpha} \underbrace{\sum_{i=0}^{n} \overbrace{\frac{n!}{i!(n-i)!}}^{\binom{n}{i}} (b-a)^i (c-b)^{n-i}}$$

What's this?
The binomial theorem!
These are the binomial coefficients.

$$\left((\cancel{b}-a)+(c-\cancel{b})\right)^n = \sum_{i=0}^{n} \binom{n}{i} (b-a)^i (c-b)^{n-i}$$

$$(c-a)^n = \sum_{i=0}^{n} \binom{n}{i} (b-a)^i (c-b)^{n-i}$$

This is just:
$(c-a)^n$

$$P_\alpha\left(\bigcup_{i=0}^{n}\left(\begin{bmatrix}(a,b]\\ i\end{bmatrix} \cap \begin{bmatrix}(b,c]\\ n-i\end{bmatrix}\right)\right) = \frac{(c-a)^n\alpha^n}{n!} e^{-(c-a)\alpha} \quad \checkmark$$

So, we have $P_\alpha(LHS) = P_\alpha(RHS)$, as desired.
It checks.
So now you are beginning to believe there is something to this, after all.
So the Poisson Process is consistent.

- Let's now justify our derivation of the exponential distribution $P(T>t) = e^{-\alpha t}$.
The easist thing we can do is this.
Let's compute the waiting time for the first blip in the Poisson Process.
You have random blips occuring. What's the waiting time for the first blip?
Give it a name.

T = waiting time for the first blip

$$P(T>t) = P(\underline{\text{no blips}} \text{ in interval } (0,t])$$

$$= P_\alpha\left(\begin{bmatrix}(0,t]\\ 0\end{bmatrix}\right)$$

The definition of P_α, i.e., equation (*), gives:

$$= \frac{(t-0)^0\alpha^0}{0!} e^{-(t-0)\alpha}$$

$$P(T>t) = e^{-\alpha t}$$

} Precisely the exponential distribution, which we also derived using Cauchy's functional equation [4/22/98.4-6].

• In conclusion, we have set up a sample space where the continuous waiting time makes sense.

And now, we're going to do all sorts of experiments on this sample space to convince ourselves that it is the only one possible.

Fact: The Poisson Process is more common than the Bernoulli Process in nature.

The Poisson Process is the process when you spread little seeds around a big field. Or you bomb someplace by dropping 18.313 students by parachute. Each 18.313 student you drop is a blip. And you are dropping over a big country. All you have is the average number of 18.313 students dropped per unit area.

That's the idea.
The idea is that all the data you have is:

number of blips per unit length

Everything else is completely random.

Under these circumstances, there is only one way of modeling this.
That's the Poisson Process.
It's a deep message.

John Guidi
guidi@mit.edu

Seven Properties of The Poisson Process

The Poisson Process is very frequent in nature.
As a matter of fact, it occurs more frequently than the Bernoulli Process, or the Uniform Process.
But, it's something that we haven't seen.
So we have to get used to it.
We need to focus our intuition.

What's the idea?
The idea, roughly, speaking as a physicist is:

Instead of a finite interval, as we had in the Uniform Process,
we have a half infinite interval.

{a half interval, because we want to see the interpretation in terms of waiting times, as one of several interpretations.}

{Then, we imagine flying over this interval, at a steady rate, w/ an indefinite supply of blips, and dropping blips at random.

The only decision is that the average number of blips per unit area is given by α.

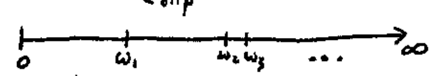
$$\underbrace{\alpha}_{\text{intensity}} = \text{average number of blips per unit area}$$

That's it.
This is the way you interpret picking infinitely many points.
Because if you pick infinitely many points, you can't divide by the random interval enclosing them.
And what turns out to be the case is that this is described by an exponential distribution.

We did the brutal, dry, mathematical definitions last time.
Now, we want to understand it from a different point of view.

Let's repeat our definition:
Our definition was based on the sample space consisting of rare subsets.

$$\Omega = \{\underbrace{\omega}_{\text{sample points}} = \text{rare subsets, where } \omega = (\underbrace{\omega_1}_{\text{blip}}, \omega_2, \dots)\}$$



Rare means you can not have any accumulation points.
This is best stated by saying – Look at any finite interval on $[0,\infty)$ and there must be only a finite number of blips.

what we are studying is the appearance of these blips, randomly.
That's the idea.

- The Poisson Process is one of 4 fundamental stochastic processes that make up all of probability theory.

The 4 fundamental stochastic processes

1) Bernoulli Process
   Tossing a coin.

2) Uniform Process
   Picking points randomly from a finite interval.

our current subject ⟹ 3) Poisson Process

4) Process pertaining to the Normal Distribution

That's it.
From a conservative viewpoint, you can define probability theory as the study of these 4 stochastic processes and their degeneration.

- Event – $\begin{bmatrix}(a,b]\\ n\end{bmatrix}$ = you have n blips in the interval $(a,b]$

We view this probabilistically as blips appear at random (someone drops them from an airplane).
And the event is that we look at the area in the interval $(a,b]$ and we see n blips (random blips).

You have to get used to the fact that whats random is:
   a) it is random when the blips appear
   b) there are infinitely many blips

It takes a little getting used to.

From these events, by the operations of boolean algebra, together with infinite unions intersections of sequences, we generate all possible events.

- Main Result

We assign a probability to this basic event.
The main result of the theory of the Poisson Process states that this probability uniquely determines the probability of such events.
The probability we assign to such events is:

$$P_\alpha\left(\begin{bmatrix}(a,b]\\ n\end{bmatrix}\right) = \frac{\alpha^n (b-a)^n}{n!} e^{-(b-a)\alpha}$$

α = given intensity

Furthermore, the number of blips appearing in disjoint intervals is independent.

If you have two events:

$$\text{event } A = \begin{bmatrix} (a,b] \\ m \end{bmatrix} \qquad \begin{bmatrix} (c,d] \\ n \end{bmatrix} = \text{event } B$$

If these intervals are disjoint, then events $A+B$ are independet. Regardless of the bottom value.

This is a way of exactly rendering the fact that the number of blips that appear in a given interval has no influence, whatsoever, on the number of blips that appear in a disjoint interval.

It is amazing that this is the only consistent way of defining a probability. And that is what we have to convince ourselves of now.

Let's introduce something analogous to what we have done for the Bernoulli and the Uniform Process. We introduce the random function $N_\alpha(t)$.

1) Bernoulli Process : random variable S_n = number of heads (1's) in the first n tosses

2) Uniform (Dirichlet) Process : [3/18/98.6] random function $U_a(t)$ = number of points choosen in the interval $[0,t)$

3) Poisson Process : random function $N_\alpha(t)$ = number of blips in the interval $[0,t)$

depends on the intensity α of the Poisson Process.

NB: α is the only data on which the Poisson Process depends

The entire theory of the Poisson Process can be reformulated in terms of the random function $N_\alpha(t)$.

What's a random function?

A random function is a function of t, which has the property that for every t, $N_\alpha(t)$ is a random variable [3/18/98.6].

So, it's a random variable, depending on the parameter t.

- What is the probability distribution of this <u>random function</u>?

$$P(N_\alpha(t) = n) = \;?$$

This is easily derived, once we realize that:

$$\text{event } \underbrace{\left(N_\alpha(t) = n\right) = \begin{bmatrix} [0,t) \\ n \end{bmatrix}}$$

So, we really have <u>2 notations</u> for the <u>same event</u>.
We could have used the notation $(N_\alpha(t)=n)$ from the start.

Therefore, we have, trivially:

$$\boxed{P(N_\alpha(t) = n) = \frac{\alpha^n t^n}{n!} e^{-\alpha t}}$$

The Poisson Distribution w/ intensity α

- Furthermore, the fact, that the <u>number of blips</u> in <u>disjoint intervals</u> are <u>independent</u>, can be expressed in terms of the random function, as follows:

    if $[a,b)$ and $[c,d)$ are <u>disjoint intervals</u>, then the <u>random variables</u>

    $N_\alpha(b) - N_\alpha(a)$ and $N_\alpha(d) - N_\alpha(c)$

    are <u>independent</u>.

This is a <u>repetition</u> of what we worked in to the definition of the Poisson Process.
Why?

Because $(N_\alpha(b) - N_\alpha(a) = m)$ the number of blips in the interval $[a,b)$ $= \begin{bmatrix} [a,b) \\ m \end{bmatrix}$

Similarily $(N_\alpha(d) - N_\alpha(c) = n)$ the number of blips in the interval $[c,d)$ $= \begin{bmatrix} [c,d) \\ n \end{bmatrix}$

since $[a,b)$ and $[c,d)$ are <u>disjoint</u>, the originating random variables are <u>independent</u>.

Again, we could have defined the Poisson Process using random functions from the start.

- Observe, already, one decisive advantage of the Poisson Process over the Uniform Process.

  The random function $U_n(t)$ of the Uniform Process is not independent. If you take the number of points of 2 disjoint intervals of the Uniform Process, is not independent, because for have a fixed number of n total points.

  But for the Poisson Process, the number of blips in disjoint intervals is independent.

  This is very good, for all sorts of purposes.

  So, you could say that the Poisson Process achieves something that we wanted for the Uniform Process, but couldn't get on a finite interval.

- Let's compute the expectation:

$$E(N_\alpha(t)) = \sum_{n=0}^{\infty} n \underbrace{P(N_\alpha(t)=n)}_{\text{by definition}}$$

this probability given by the formula we just gave for the Poisson Distribution w/ intensity α.

$$= \sum_{n=0}^{\infty} n \frac{\alpha^n t^n}{n!} e^{-\alpha t}$$

$$= \sum_{n=1}^{\infty} \frac{\alpha t\, \alpha^{n-1} t^{n-1}}{(n-1)!} e^{-\alpha t} + 0$$

$n=0$

Set $j=n-1$ to obtain:

$$= \alpha t e^{-\alpha t} \sum_{j=0}^{\infty} \frac{\alpha^j t^j}{j!}$$

The Taylor Series expansion of $e^{\alpha t}$ is:

$$e^{\alpha t} = \sum_{j=0}^{\infty} \frac{f^{(j)}(0)}{j!} t^j$$

$$= \sum_{j=0}^{\infty} \frac{\alpha^j \cancel{e^{\alpha 0}}}{j!} t^j$$

$$= \sum_{j=0}^{\infty} \frac{\alpha^j t^j}{j!}$$

$$= \alpha t \cancel{e^{-\alpha t}} \cancel{e^{\alpha t}}$$

$$\boxed{E(N_\alpha(t)) = \alpha t}$$ A major result

To repeat, the expectation of the random variable $N_\alpha(t)$, for t fixed, equals αt.

$$E(N_\alpha(t)) = \alpha t$$

- We can repeat this, as follows:

$$\underbrace{\frac{1}{t} E(N_\alpha(t)) = \alpha}$$

This is exactly what I've said already 5 times.

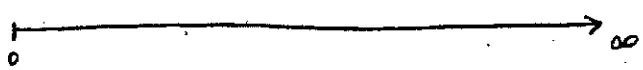
The $\underbrace{\text{average number of blips}}_{E(N_\alpha(t))}$ per unit $\underset{t}{\underline{\text{length}}}$ is α.

This tells it to you loud and clear.

We now discuss the 7 fundamental properties of the Poisson Process.

Property (1) – Law of Rare Events

I take a half infinite time interval.



At every interval $\frac{1}{j}$ of time, I flip a coin.
I'm doing exactly a Bernoulli Process, except I'm flipping exactly at the interval of time.

Flip at times $\frac{1}{j}, \frac{2}{j}, \ldots, \frac{n}{j}, \ldots$ a coin which is biased w/ bias $p_j = \frac{\alpha}{j}$.

What is the probability of getting k heads (1's) at time $\frac{n+\varepsilon}{j}$?

$$\left\{ \begin{array}{l} \text{immediately after } \frac{n}{j} \text{ and before } \frac{n+1}{j} \\ \left.\begin{array}{l}\varepsilon \to 0 \\ \varepsilon > 0\end{array}\right\} \ \frac{n}{j} < \frac{n+\varepsilon}{j} < \frac{n+1}{j} \end{array} \right\}$$

This probability is just the binomial distribution – one of the first things you learned in this class.

You have n tosses and you want k ones.
The bias is $\left(\frac{\alpha}{j}\right)$, so you have:

$$P\left(k \text{ 1's at time } \frac{n+\varepsilon}{j}\right) = \binom{n}{k}\left(\frac{\alpha}{j}\right)^k\left(1-\frac{\alpha}{j}\right)^{n-k}$$

Now, we let $j \to \infty$.
Which means, we <u>toss</u> the coin <u>faster and faster.</u>
In the <u>limit</u>, you toss the coin <u>continuously.</u>

$$\lim_{j\to\infty} \frac{1}{j}, \frac{2}{j}, \dots, \frac{n}{j}, \dots \quad = \text{continuous}$$

Set $t = \frac{n}{j} \implies j = \frac{n}{t}$

Let $n \to \infty$,
t remain fixed.

$$\lim_{n\to\infty} \frac{n+\varepsilon}{j} = \lim_{n\to\infty}\left(1+\cancel{\frac{\varepsilon}{n}}^{\,0}\right)t = t$$

$$P(N_\alpha(t) = k) = \lim_{n\to\infty} P\left(k \text{ 1's at time } \frac{n+\varepsilon}{j}\right)$$

} in the limit, $P(k \text{ 1's at time } t)$

$$= \lim_{n\to\infty} \binom{n}{k}\left(\frac{\alpha}{j}\right)^k\left(1-\frac{\alpha}{j}\right)^{n-k}$$

Substitute $j = \frac{n}{t}$ and rewriting expressions gives:

$$= \lim_{n\to\infty} \frac{(n)_k}{k!}\,\frac{\alpha^k}{\left(\frac{n}{t}\right)^k}\left(1-\frac{\alpha}{\left(\frac{n}{t}\right)}\right)^n\left(1-\frac{\alpha}{\left(\frac{n}{t}\right)}\right)^{-k}$$

$$(*) \quad = \lim_{n\to\infty} \underbrace{\frac{(n)_k}{n^k}}_{1} \cdot \frac{\alpha^k t^k}{k!}\underbrace{\left(1-\frac{\alpha t}{n}\right)^n}\underbrace{\left(1-\cancel{\frac{\alpha t}{n}}^{\,0}\right)^{-k}}_{1.}$$

The numerator and denominator each have k factors. In the limit, the ratio each term is 1.

$$\lim_{n\to\infty} \frac{\overbrace{n(n-1)(n-2)\cdots(n-k+1)}^{k}}{\underbrace{n \;\; n \;\; n \;\; \cdots \;\; n}_{k}}$$

$$= \lim_{n\to\infty}\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\cdots\left(\frac{n-k+1}{n}\right)$$

$$= 1$$

By definition,
$$\lim_{n\to\infty}\left(1+\frac{r}{n}\right)^n = e^r$$
So, $\lim_{n\to\infty}\left(1-\frac{\alpha t}{n}\right)^n = e^{-\alpha t}$

And we have:

$$P(N_\alpha(t)=k) = \frac{\alpha^k t^k}{k!} e^{-\alpha t}$$

And we've just shown that the Poisson Process can be viewed as continuous coin tossing. This limiting process is called the Law of Rare Events.

We'll talk about this more later.
But for the moment, observe that the Law of Rare Events has notable practical applications.

Example – Mother bakes cookies w/ raisins.
She wants 1 raisin per cookie.
She mixes the dough, throws in the raisins at random, and then bakes the cookies.

What's the probability that a cookie has no raisins?

$$\frac{1}{e}$$

Can't you read?
Repeat:
We have a Poisson Process with intensity $\alpha = 1$
There is, on the average, 1 raisin per cookie.
The probability of getting n raisins in a cookie is given by the Poisson Distribution:

$$\left\{ \begin{array}{l} \text{average number of blips per unit area} \\ = \frac{1 \text{ raisin}}{\text{cookie}} \end{array} \right\}$$

$$P(N_\alpha(1)=n) = \frac{\alpha^n 1^n}{n!} e^{-\alpha(1)}$$

and $\alpha = 1$, so this is:

$$= \frac{1}{n!} \cdot \frac{1}{e}$$

So the probability that a cookie has no raisins is:

$$P(N_{\alpha=1}(1)=0) = \frac{1}{e}$$

Example – You have to set up my lectures in LateX, which I would greatly appreciate ☺ for next years students.
And you make 1 mistake per page.
What's the probability that I read your page and find 3 mistakes?
$\alpha=1$, so we again have:

$$P(N_{\alpha=1}(1)=n) = \frac{1}{n!}\cdot\frac{1}{e} \implies P(N_{\alpha=1}(1)=3) = \frac{1}{3!}\cdot\frac{1}{e}$$

Why?
Because you <u>approximate</u> the <u>Bernoulli Process</u> of making a mistake for each letter by the <u>Law of Rare Events</u> of the <u>Poisson Process</u>.

Therefore, you can consider this as a Poisson Process with <u>intensity 1</u>.
And, therefore, you can tell immediately what the <u>probability</u> is of having 1000 mistakes per page.
w/ tremendous accuracy.

Now you begin to see what it's about.

Property (2) — <u>Limit of Uniform Process</u>

Let's get the <u>Poisson Process</u> as a <u>limit</u> of the <u>Uniform Process</u>
In other words, let's see that the Poisson Process is <u>really</u> <u>dropping</u> α <u>blips per unit length.</u>

We take a uniform process w/ random function $U_a(t)$.
We <u>set a</u> so that:

$$\alpha = \frac{n}{a}$$

We let a vary, together w/ the number of points n that you drop.
But the <u>ratio</u> $\frac{n}{a}$ is constant.
We drop n points in the interval $[0, a]$, where $\frac{n}{a} = \alpha$.

$$P(U_a(t)=k) = \binom{n}{k}\left(\frac{t}{a}\right)^k\left(1-\frac{t}{a}\right)^{n-k}$$

Now, let's work in $a = \frac{n}{\alpha}$:

$$= \frac{(n)_k}{k!}\,\frac{t^k}{\left(\frac{n}{\alpha}\right)^k}\left(1-\frac{t}{\left(\frac{n}{\alpha}\right)}\right)^n\left(1-\frac{t}{\left(\frac{n}{\alpha}\right)}\right)^{-k}$$

$$= \underbrace{\overbrace{\cancel{\frac{(n)_k}{n^k}}}^{1}\cdot\frac{\alpha^k t^k}{k!}\overbrace{\cancel{\left(1-\frac{\alpha t}{n}\right)^n}}^{e^{-\alpha t}}\overbrace{\cancel{\left(1-\frac{\alpha t}{n}\right)^{-k}}}^{0}}_{\text{in the limit, as } n\to\infty\text{, this is the same as equation } (\ast)\text{.}}$$

$$\lim_{n\to\infty} P(U_a(t)=k) = \frac{\alpha^k t^k}{k!}e^{-\alpha t}$$

$$\boxed{\lim_{n\to\infty} P(U_a(t)=k) = P(N_\alpha(t)=k)}$$

So, we see that the Poisson Process really is dropping α blips per unit length.

Property (3) – Memorylessness

Connection w/ memorylessness. See [4/22/98.5–8].
Recall that the waiting time for the first blip in the Poisson Process is exponentially distributed:

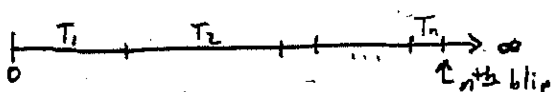
$$P(T > t) = e^{-\alpha t} \quad \Rightarrow \quad P(T \leq t) = 1 - e^{-\alpha t}$$

Remember, if you waited so long for the first blip, that wait doesn't make expected waiting time any shorter (i.e., the expected waiting time does not change).
The expected waiting time $E(T) = \frac{1}{\alpha}$.
In other words, the higher the intensity α, the shorter the expected waiting time. Period.

Recall that the density of the exponential distribution is:

$$\text{dens}(T = t) = \alpha e^{-\alpha t} \qquad \text{exponential density}$$

We reason as follows:



Here are my blips.
I wait an exponential waiting time, as they say in the business, for the first blip.

After the first blip has occurred, I wait another exponential waiting time for the second blip. And so on.
That means that the waiting time for the n^{th} blip is the sum of n independent, exponentially distributed, identically distributed random variables.
And that should give us back the Poisson Process.
Let's see if that checks.

$T_1, T_2, \ldots, T_n$ are i.i.d., exponentially distributed random variables
What is the density of their sum?

$$\text{dens}(T_1 + T_2 + \ldots + T_n = t) = \ ?$$

Let's first do the density of 2 random variables.
Then we'll guess at a more general result.

$$\text{dens}(T_1 + T_2 = t) =$$ where T_1 and T_2 are independent, we've shown that the density of their sums is the convolution of their density functions.

[4/15/98.8]

Since T_1 and T_2 are independent and $T_1 \geq 0$ and $T_2 \geq 0$, we have the convolution:

$$\text{dens}(T_1 + T_2 = t) = \int_0^t \text{dens}(T_1 = s)\,\text{dens}(T_2 = t-s)\,ds$$

Since $T_1 + T_2$ are identically distributed, they have the same density function. Namely:

$$\text{dens}(T_1 = x) = \text{dens}(T_2 = x) = \alpha e^{-\alpha x}$$

$$= \int_0^t \alpha \cancel{e^{-\alpha s}}\, \alpha e^{-\alpha(t \cancel{-s})}\, ds$$

$$= \int_0^t \alpha^2 e^{-\alpha t}\, ds$$

$$= \alpha^2 e^{-\alpha t} \int_0^t ds$$

$$\boxed{\text{dens}(T_1 + T_2 = t) = \alpha^2 t e^{-\alpha t}}$$

- Proceeding by induction, we find:

$$\text{dens}(T_1 + T_2 + \dots + T_n = t) = \underbrace{\frac{\alpha^n t^{n-1}}{(n-1)!} e^{-\alpha t}}$$

This is called the Gamma Distribution

{ strictly speaking, a misnomer. This should be called the Gamma Density, not distribution. }

Proof by induction

Basis: $n = 1$

$$\text{dens}(T_1) = \frac{\alpha^1 t^0}{0!} e^{-\alpha t} = \alpha e^{-\alpha t} \checkmark$$

Inductive Step: assume induction hypothesis holds for n.

$$\text{dens}(\underbrace{T_1 + \dots + T_n}_{U} + T_{n+1} = t) \overset{?}{=} \frac{\alpha^{n+1} t^n}{n!} e^{-\alpha t}$$

$$\text{dens}(U + T_{n+1} = t) = \int_0^t \text{dens}(U = s)\,\text{dens}(T_{n+1} = t-s)\,ds$$ ← Convolution of densities since U, T_{n+1} independent

$$= \int_0^t \text{dens}(T_1 + \dots + T_n = s)\,\text{dens}(T_{n+1} = t-s)\,ds$$

The induction hypothesis gives

$$= \int_0^t \frac{\alpha^n s^{n-1}}{(n-1)!} e^{-\alpha s}\, \alpha e^{-\alpha(t-s)}\, ds$$

$$= \int_0^t \frac{\alpha^{n+1} s^{n-1}}{(n-1)!} e^{-\alpha t}\, ds$$

$$= \frac{\alpha^{n+1}}{(n-1)!} e^{-\alpha t} \int_0^t s^{n-1}\, ds$$

$$= \frac{\alpha^{n+1}}{(n-1)!} e^{-\alpha t} \frac{1}{n} t^n$$

$$= \frac{\alpha^{n+1} t^n}{n!} e^{-\alpha t} \quad \checkmark$$

Which completes the inductive proof of the Gamma Density:

$$\text{dens}(T_1 + \dots + T_n = t) = \frac{\alpha^n t^{n-1}}{(n-1)!} e^{-\alpha t}$$

What is it the density of?
The Gamma Density, also known as the Gamma Distribution, is the waiting time for the n^{th} blip, we guess, in the Poisson Process.
It should be.

In other words, we should have that:

$$\text{event } \underbrace{(N_\alpha(t) = n)}_{\text{event that number of blips in the interval } [0,t) \text{ is } n} = \underbrace{(T_1 + T_2 + \dots + T_n < t) - (T_1 + T_2 + \dots + T_n + T_{n+1} < t)}_{\substack{\text{The event that at least } n \text{ blips occur in } [0,t) \text{ minus the event that at least } n+1 \text{ blips occur in the same interval} \\ \text{is exactly the same as saying there are exactly } n \text{ blips in the interval } [0,a).}}$$

The events are equal.
Now we want to check the probabilities:

$$P(N_\alpha(t) = n) \stackrel{?}{=} P(T_1 + T_2 + \dots + T_n < t) - P(T_1 + T_2 + \dots + T_n + T_{n+1} < t)$$

Using the fact that:

$$P(X \le x) = \int_{-\infty}^{\infty\ x} \text{dens}(X = s)\, ds$$

and the Gamma Density (Distribution), we get the probabilities of the RHS above.

$$= \int_0^t \frac{\alpha^n s^{n-1}}{(n-1)!} e^{-\alpha s}\, ds - \int_0^t \frac{\alpha^{n+1} s^n}{n!} e^{-\alpha s}\, ds$$

$$= \frac{\alpha^n}{(n-1)!} \int_0^t s^{n-1} e^{-\alpha s}\, ds - \frac{\alpha^{n+1}}{n!} \int_0^t s^n e^{-\alpha s}\, ds$$

We evaluate the 2^{nd} integral by parts, in anticipation of performing miracles of cancellation.

$$\int \underbrace{s^n}_{u} \underbrace{e^{-\alpha s}\,ds}_{dv} = \underbrace{s^n}_{u}\underbrace{\left(\frac{-1}{\alpha}e^{-\alpha s}\right)}_{v} - \int \underbrace{\left(\frac{-1}{\alpha}e^{-\alpha s}\right)}_{v} \underbrace{n s^{n-1}\,ds}_{du}$$

$$u = s^n$$
$$du = n s^{n-1} ds \qquad dv = e^{-\alpha s} ds$$
$$v = \int e^{-\alpha s} ds$$
$$= \frac{-1}{\alpha} e^{-\alpha s}$$

$$\int s^n e^{-\alpha s} ds = -\frac{s^n}{\alpha} e^{-\alpha s} + \frac{n}{\alpha}\int s^{n-1} e^{-\alpha s} ds$$

Using this indefinite integral equality above gives:

$$= \cancel{\frac{\alpha^n}{(n-1)!}\int_0^t s^{n-1} e^{-\alpha s} ds} - \frac{\alpha^{n+1}}{n!}\left[-\frac{s^n}{\alpha}e^{-\alpha s}\right]_{s=0}^{t} - \cancel{\frac{\alpha^{n+1}}{n!}\left(\frac{n}{\alpha}\int_0^t s^{n-1} e^{-\alpha s} ds\right)}$$

$$= \frac{\alpha^n}{n!}\left[s^n e^{-\alpha s}\right]_{s=0}^{t}$$

$$P(N_\alpha(t)=n) = \frac{\alpha^n t^n}{n!} e^{-\alpha t} \quad \checkmark$$

This is the Poisson Distribution

$$= P(N_\alpha(t) = n)$$

So, it checks. As desired.

- This completes the first 3 properties of the Poisson Process.
  So far, we've done only theory.
  Now we need applications – various properties of the Poisson Process.

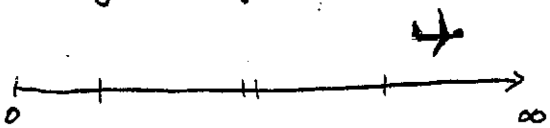
Seven Properties of the Poisson Process (cont'd)

We continue our in-depth study of the Poisson Process, which is one of the fundamental processes of probability.

Intuitively, again, what is the Poisson Process?
You go over a continuum and you sprinkle blips.
The sprinkling is completely random, subject only to one condition

the average number of blips per unit area equals α (the intensity).

Nothing else is given.



Under these conditions, there is only one possible formula that gives the formula for the probability and that is the formula for the Poisson Process.

That's extraordinary.
Don't you ever forget it.

In symbols:

Our sample space Ω consists of rare events, which are sequences of blips.

$$\Omega = \{ \omega = (\underset{\uparrow \text{blip}}{\omega_1}, \omega_2, \ldots) \}$$

And we define a probability on the set of all such rare sequences.

More generally, we define a random function:

$$N_\alpha(t) = \text{number of blips in } [0,t)$$

subject to the following condition:

$$P\Big(\underbrace{N_\alpha(b) - N_\alpha(a) = n}_{\text{event that number of blips in interval } [a,b) \text{ is } n}\Big) = \frac{\alpha^n (b-a)^n}{n!} e^{-\alpha(b-a)}$$

and furthermore

intervals disjoint

if $\underbrace{[a,b) \cap [c,d)}$ $= \phi$, then the random variables

$\underbrace{N_\alpha(b) - N_\alpha(a)}_{\text{number of blips in } [a,b)}$ and $\underbrace{N_\alpha(d) - N_\alpha(c)}_{\text{number of blips in } [c,d)}$ are independent.

This summarizes our description of the sample space, which we have given, here, in terms of events. Equivalent to the description of the event.
We're now trying to get a good feeling for this fundamental process.
As this process is not something we're used to, even though it occurs in nature more frequently than the others.

What have we seen so far?

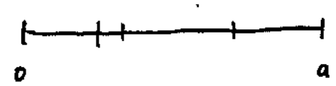
We have seen 3 ways in which we can obtain the Poisson Process as a limiting process of known processes

(1) We started tossing coins on intervals that became thicker and thicker $(\frac{1}{\delta}, \frac{2}{\delta}, \ldots, \frac{n}{\delta}, \ldots)$, with bias $p_\delta = \frac{\alpha}{\delta}$ that varied w/ the thickness of the interval.

And we saw that the limiting process that came from this coin tossing process was the Poisson Process.
That was the Law of Rare Events.
Limits of Bernoulli Coin tossing

So you can view Poisson Process as some sort of infinitesimal coin tossing.

(2) We take the uniform process on the interval $[0,a)$.



$\lim_{a \to \infty}$

We let a tend to infinity, subject to the condition that the average number of points per unit length remains equal to α.

Under these conditions, we saw that the limiting process is, again, the Poisson Process,

(3) We considered the relationship between the Poisson Process and the exponential distribution.

The exponential distribution, as we have seen, is the only memoryless continuous distribution.

Taking this as our starting point, we view a rare sequence as follows:

We wait a memoryless time of intensity α.
The expectation is $\frac{1}{\alpha}$.
As soon as there is a blip, you start waiting again, with intensity α for the next blip.
And so forth.
In this way, you create a rare sequence.

In other words, we took the sum :

$$\underbrace{T_1 + T_2 + \dots + T_n}$$

n independent random variables, which are identically distributed with an exponential distribution.

We envision the Poisson Process by using the set theoretic identity :

$$\text{event} \quad (T_1 + T_2 + \dots + T_n < t) - (T_1 + T_2 + \dots + T_n + T_{n+1} < t)$$

$$= (N_\alpha(t) = n)$$

Then by convolution, we obtain the density of the sum of n independent and identically distributed random variables as the Gamma Distribution

more correctly, Gamma Density.

By fooling around w/ the Gamma Distribution and integration by parts, we saw that the Poisson Process comes out.

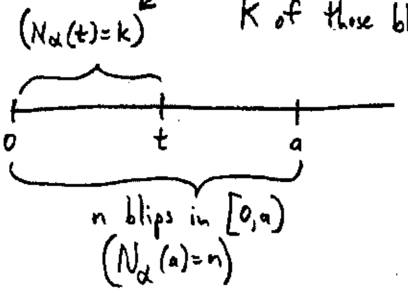
- So we have, to date, 3 heavy duty justifications for the Poisson Process.
You should be fairly convinced.
Now, let's go to number (4).

Property (4) $P(N_\alpha(t)=k \mid N_\alpha(a)=n) = P(U_a(t)=k)$ } uniform process obtained by conditioning Poisson Process

Let's compute an interesting probability.
What does this mean?

$$\underbrace{P(N_\alpha(t)=k \mid N_\alpha(a)=n)}_{} = ? \qquad t<a$$

assuming we have n blips in the interval $[0,a)$,
what then is the probability that
K of those blips are in the interval $[0,t)$?



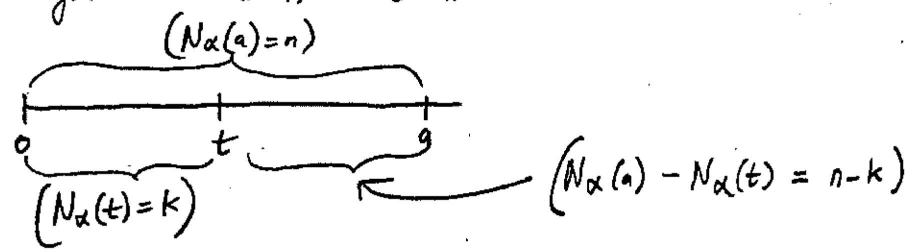
By the definition of conditional probability:

$$P(N_\alpha(t)=k \mid N_\alpha(a)=n) = \frac{P((N_\alpha(t)=k) \cap (N_\alpha(a)=n))}{P(N_\alpha(a)=n)}.$$

The events in the numerator are not independent

We can do a slight computation that is done very often, which is the following.
We want to rewrite the intersection in terms of independent events.

All you have to do is realize that:



$$(N_\alpha(t)=k) \cap (N_\alpha(a)=n) = \underbrace{(N_\alpha(t)=k) \cap (N_\alpha(a)-N_\alpha(t)=n-k)}_{\text{these are disjoint events}}$$

$$= \frac{P((N_\alpha(t)=k) \cap (N_\alpha(a)-N_\alpha(t)=n-k))}{P(N_\alpha(a)=n)}$$

And since the numerator is the probability of the intersection of two disjoint events, that probability is the product of the probabilities.

$$= \frac{P(N_\alpha(t)=k)\, P(N_\alpha(a)-N_\alpha(t)=n-k)}{P(N_\alpha(a)=n)}$$

where $P_\alpha\left(\begin{bmatrix}[c,d)\\ m\end{bmatrix}\right) = \dfrac{\alpha^m (d-c)^m}{m!} e^{-\alpha(d-c)}$

$$= \frac{\dfrac{\cancel{\alpha^k} t^k}{k!} \cancel{e^{-\alpha t}} \dfrac{\cancel{\alpha^{n-k}} (a-t)^{n-k}}{(n-k)!} \cancel{e^{-\alpha(a-t)}}}{\dfrac{\cancel{\alpha^n} a^n}{n!} \cancel{e^{-\alpha a}}}$$

$$= \underbrace{\frac{n!}{k!(n-k)!}}_{\binom{n}{k}} t^k \frac{(a-t)^{n-k}}{a^n}$$

$$= \underbrace{\binom{n}{k} \frac{t^k (a-t)^{n-k}}{a^n}}$$

What's this?
This is the Uniform Process

$$\boxed{P(N_\alpha(t)=k \mid N_\alpha(a)=n) = P(U_a(t)=k)}$$

$U_a(t)$ is the random function of the uniform process

Ha. Ha. Ha.

Conclusion: When you condition the number of blips to be n in the interval $[0,a)$, then, if you look inside the interval $[0,t)$, it's same as if it were a uniform process.
As you would expect.

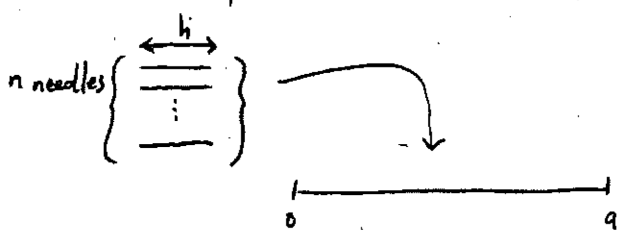
Therefore, I, in a sense, cheated you.
We can derive all the properties of the uniform process cheapo.
We didn't have to derive them the way we did.
We could have derived them from the fact that:

$$P\left(N_\alpha(t)=k \mid N_\alpha(a)=n\right) = P(U_a(t)=k)$$

But it would have been overwhelming to do it this way, from this stuff.
For example, let's work out an ambitious property of the uniform process.
{Will do an ambitious property, rather than a Mickey Mouse property, so we're going against the given Law — Always do Mickey Mouse problems first.}

Example — Needles on a stick

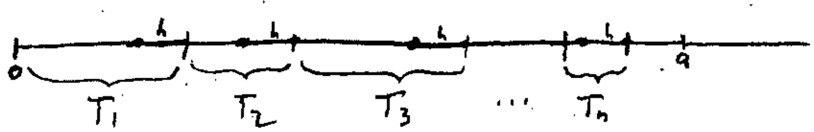
Remember how we worked this out elaborately? [4/10/98.1-6]
Now this problem comes out in a couple of lines.



We have n needles of length h. We drop them on a stick of length a such that they don't stick out.
What is the probability that no two needles overlap?

$$\frac{(a-nh)^n}{a^n}$$

Let's see if we can derive this from the newly acquired knowledge that the uniform process is obtained by conditioning the Poisson process.



We want the event:

$$\left((T_1>h) \cap (T_2>h) \cap \ldots \cap (T_n>h)\right)$$

This is the event that the first n blips are at least a distance h from each other.

What's the probability of this event? (It's not the probability we are after).

$$\underbrace{P\left((T_1 > h) \cap (T_2 > h) \cap \dots \cap (T_n > h)\right)}$$

this is the intersection of independent events, because they are independent random variables that are identically distributed w/ an exponential distribution.

$$= P(T_1 > h)\, P(T_2 > h) \dots P(T_n > h)$$

$$\left\{ \begin{array}{l} \text{And since [4/24/98.10]:} \\ \quad P(T > h) = e^{-\alpha h} \end{array} \right\}$$

$$= \left(e^{-\alpha h}\right)^n$$

$$\boxed{P\left((T_1 > h) \cap \dots \cap (T_n > h)\right) = e^{-n\alpha h}}$$

But again, that's not the probability we are after.
What we want is:

$$P\left((T_1 > h) \cap (T_2 > h) \cap \dots \cap (T_n > h) \mid N_\alpha(a) = n\right)$$

This will give us the probability that the needles will fall on the stick.
Unfortunately, when we expand this, by the definition of conditional probability, we have:

$$P\left((T_1 > h) \cap \dots \cap (T_n > h) \mid N_\alpha(a) = n\right) = \frac{P\left(\left((T_1 > h) \cap \dots \cap (T_n > h)\right) \cap \left(N_\alpha(a) = n\right)\right)}{P(N_\alpha(a) = n)}$$

I don't see an easy way to obtain this numerator

So, instead, we compute the desired probability indirectly.
Recall, from above, that:

$$P\left((T_1 > h) \cap \dots \cap (T_n > h)\right) = e^{-n\alpha h}$$

Observe that the events $(N_\alpha(a) = 0)$, $(N_\alpha(a) = 1)$, ... are a partition of the sample space.

Therefore, we can compute the above LHS by the Law of Alternatives

{which we use more than any other formula}

By Law of Alternatives:

$$\underbrace{P\big((T_1>h)\cap\ldots\cap(T_n>h)\big)}_{e^{-n\alpha h}} = \sum_{k\geq 0} P\big((T_1>h)\cap\ldots\cap(T_n>h)\,\big|\,N_\alpha(a)=k\big)\,\underbrace{P(N_\alpha(a)=k)}_{\frac{\alpha^k a^k}{k!}e^{-\alpha a}}$$

$$e^{-n\alpha h} = e^{-\alpha a}\sum_{k\geq 0} P\big((T_1>h)\cap\ldots\cap(T_n>h)\,\big|\,N_\alpha(a)=k\big)\,\frac{\alpha^k a^k}{k!}$$

$$\underbrace{e^{\alpha(a-nh)}} = \sum_{k\geq 0} P\big((T_1>h)\cap\ldots\cap(T_n>h)\,\big|\,N_\alpha(a)=k\big)\,\frac{\alpha^k a^k}{k!}$$

We expand this exponential into a power series,

$f(x) = e^x$

$$= \sum_{k\geq 0}\frac{f^{(k)}(0)}{k!}x^k \quad \longleftarrow \text{Maclaurin series}$$

$$= \sum_{k\geq 0}\frac{x^k}{k!}$$

Let $x = \alpha(a-nh)$, which gives:

$$e^{\alpha(a-nh)} = \sum_{k\geq 0}\frac{(\alpha(a-nh))^k}{k!}$$

$$\sum_{k\geq 0}\frac{\alpha^k (a-nh)^k}{k!} = \sum_{k\geq 0} P\big((T_1>h)\cap\ldots\cap(T_n>h)\,\big|\,N_\alpha(a)=k\big)\,\frac{\alpha^k a^k}{k!}$$

We now have 2 power series in α, which are equal.
Since the Taylor expansion of a function is unique, the coefficients must be individually equal.

Therefore, for $k=n$, we obtain:

$$\frac{\cancel{\alpha^n}(a-nh)^n}{\cancel{n!}} = P\big((T_1>h)\cap\ldots\cap(T_n>h)\,\big|\,N_\alpha(a)=n\big)\,\frac{\cancel{\alpha^n}a^n}{\cancel{n!}}$$

which gives us the desired probability:

$$P\big((T_1>h)\cap\ldots\cap(T_n>h)\,\big|\,N_\alpha(a)=n\big) = \frac{(a-nh)^n}{a^n} \quad \checkmark$$

{ Q/A: Where did we use the fact that $a-nh$ is positive?
No where. There is some place where it must be implicitly used.
I don't know where. }

- An interesting example for you to work out. The density of the k^{th} order statistic. Use the same method. } exercise

Property (5) Schrödinger Randomization
(an application of the preceding)

I take the interval $[0,1)$.
And I take the Poisson Process $N_\alpha(t)$

Now we divide the interval $[0,1)$ into n equal sub-intervals.

$$0 \quad \tfrac{1}{n} \quad \tfrac{2}{n} \quad \cdots \quad \tfrac{n-1}{n} \quad 1=\tfrac{n}{n}$$

Then, we take:

$$\underbrace{N_\alpha\left(\tfrac{1}{n}\right)}_{\text{This is a random variable for given } \alpha, n.} = \Theta_1$$

we'll call it Θ_1

Similarly, we have the following random variables:

$$\left.\begin{aligned} N_\alpha\left(\tfrac{1}{n}\right) &= \Theta_1 \\ N_\alpha\left(\tfrac{2}{n}\right) - N_\alpha\left(\tfrac{1}{n}\right) &= \Theta_2 \\ &\vdots \\ N_\alpha(1) - N_\alpha\left(\tfrac{n-1}{n}\right) &= \Theta_n \end{aligned}\right\}$$

These are occupation numbers when you drop a random number of blips.
(balls/blips into boxes)
except here, you are dropping a random Poisson number of balls into boxes.

Why is this good?
It's terrific.
Because dropping a random Poisson number of balls into the boxes has the property that the number of balls in different boxes are independent.

Which was not the case when we dropped k balls into n boxes (from our discussion of Maxwell-Boltzmann statistics [2/20/98.5]).

Here, $P(\Theta_1=32)$, $P(\Theta_2=17)$ and so forth are independent

$$P_{MB}(\theta_1 = j) = P\left(N_\alpha\left(\tfrac{1}{n}\right) = j \;\middle|\; N_\alpha(1) = k\right)$$

4/29/98.10

- If you want Maxwell-Boltzmann statistics, we have to condition.
  And we can get any problem of Maxwell-Boltzmann statistics, relative to occupation numbers, by conditioning. It's automatic now.
  So even that I could have skipped.
  We could have started this course w/ the Poisson Process.

  Maxwell-Boltzmann statistics is obtained by conditioning the Poisson Process.

- Example

$$P\left(\left(\theta_1 = \left\{\begin{smallmatrix}0\\1\end{smallmatrix}\right.\right) \cap \left(\theta_2 = \left\{\begin{smallmatrix}0\\1\end{smallmatrix}\right.\right) \cap \ldots \cap \left(\theta_n = \left\{\begin{smallmatrix}0\\1\end{smallmatrix}\right.\right) \;\middle|\; N_\alpha(1) = k\right)$$

So we have k balls into n boxes where each box has 0 or 1 ball.

The boxes even look like boxes now.
And it better come out the way we computed it when we did the first computation we ever did in this course.
It better come out right.

Again, as in the previous example, I don't know how to compute this probability directly.
However, I do know how to compute:

$$P\Big(\underbrace{\left(\theta_1 = \left\{\begin{smallmatrix}0\\1\end{smallmatrix}\right.\right) \cap \left(\theta_2 = \left\{\begin{smallmatrix}0\\1\end{smallmatrix}\right.\right) \cap \ldots \cap \left(\theta_n = \left\{\begin{smallmatrix}0\\1\end{smallmatrix}\right.\right)}_{\text{these are all independent events}}\Big)$$

Let's consider $P\left(\theta_1 = \left\{\begin{smallmatrix}0\\1\end{smallmatrix}\right.\right)$:

these are disjoint events

$$P\left(\theta_1 = \left\{\begin{smallmatrix}0\\1\end{smallmatrix}\right.\right) = P\big((\theta_1 = 0) \cup (\theta_1 = 1)\big)$$
$$= \underbrace{P(\theta_1 = 0)} + \underbrace{P(\theta_1 = 1)}$$

$$P(\theta_1 = 0) = P\left(N_\alpha\left(\tfrac{1}{n}\right) = 0\right) = P_\alpha\left(\left[\begin{smallmatrix}[0, \frac{1}{n})\\ 0\end{smallmatrix}\right]\right) = \frac{\alpha^0 \left(\frac{1}{n}\right)^0 e^{-\alpha\left(\frac{1}{n}\right)}}{0!} = e^{-\frac{\alpha}{n}}$$

$$P(\theta_1 = 1) = P\left(N_\alpha\left(\tfrac{1}{n}\right) = 1\right) = P_\alpha\left(\left[\begin{smallmatrix}[0, \frac{1}{n})\\ 1\end{smallmatrix}\right]\right) = \frac{\alpha^1 \left(\frac{1}{n}\right)^1 e^{-\alpha\left(\frac{1}{n}\right)}}{1!} = \frac{\alpha}{n} e^{-\frac{\alpha}{n}}$$

$$P\left(\Theta_1=\begin{Bmatrix}0\\1\end{Bmatrix}\right) = e^{-\frac{\alpha}{n}} + \frac{\alpha}{n}e^{-\frac{\alpha}{n}}$$

Similarily, for $\Theta_2, \Theta_3, \ldots, \Theta_n$.
These independent random variables have the <u>same probability</u>:

$$P\left(\Theta_i=\begin{Bmatrix}0\\1\end{Bmatrix}\right) = P\left((\Theta_i=0) \cup (\Theta_i=1)\right)$$

$$= P(\Theta_i=0) + P(\Theta_i=1)$$

$$= P_\alpha\left(\begin{bmatrix}[\frac{i-1}{n}, \frac{i}{n})\\0\end{bmatrix}\right) + P_\alpha\left(\begin{bmatrix}[\frac{i-1}{n}, \frac{i}{n})\\1\end{bmatrix}\right)$$

$$= \frac{\alpha^0\left(\frac{i}{n}-(\frac{i-1}{n})\right)^0}{0!}e^{-\alpha\left(\frac{i}{n}-(\frac{i-1}{n})\right)} + \frac{\alpha^1\left(\frac{i}{n}-(\frac{i-1}{n})\right)^1}{1!}e^{-\alpha\left(\frac{i}{n}-(\frac{i-1}{n})\right)}$$

$$= e^{-\frac{\alpha}{n}} + \frac{\alpha}{n}e^{-\frac{\alpha}{n}}$$

Therefore, we have:

because the events are <u>independent</u>:

$$P\left(\left(\Theta_1=\begin{Bmatrix}0\\1\end{Bmatrix}\right)\cap\ldots\left(\Theta_n=\begin{Bmatrix}0\\1\end{Bmatrix}\right)\right) = P\left(\Theta_1=\begin{Bmatrix}0\\1\end{Bmatrix}\right)P\left(\Theta_2=\begin{Bmatrix}0\\1\end{Bmatrix}\right)\cdots P\left(\Theta_n=\begin{Bmatrix}0\\1\end{Bmatrix}\right)$$

$$= \left(e^{-\frac{\alpha}{n}} + \frac{\alpha}{n}e^{-\frac{\alpha}{n}}\right)^n$$

$$= \left(e^{-\frac{\alpha}{n}}\left(1+\frac{\alpha}{n}\right)\right)^n$$

$$= e^{-\alpha}\left(1+\frac{\alpha}{n}\right)^n$$

Very nice.
But we want the <u>conditioning</u>.
How do we get the conditioning?
We do it this way.
We use the <u>Law of Alternatives</u>

$$\underbrace{P\left(\left(\Theta_1=\begin{Bmatrix}0\\1\end{Bmatrix}\right)\cap\ldots\cap\left(\Theta_n=\begin{Bmatrix}0\\1\end{Bmatrix}\right)\right)}_{e^{-\alpha}\left(1+\frac{\alpha}{n}\right)^n} = \sum_{k\geq 0}^{n} P\left(\left(\Theta_1=\begin{Bmatrix}0\\1\end{Bmatrix}\right)\cap\ldots\cap\left(\Theta_n=\begin{Bmatrix}0\\1\end{Bmatrix}\right)\middle| N_\alpha(1)=k\right)\underbrace{P(N_\alpha(1)=k)}_{\frac{\alpha^k 1^k}{k!}e^{-\alpha(1)}}$$

$$\cancel{e^{-\alpha}}\left(1+\frac{\alpha}{n}\right)^n = \sum_{k\geq 0}^{n} P\left(\left(\Theta_1=\begin{Bmatrix}0\\1\end{Bmatrix}\right)\cap\ldots\cap\left(\Theta_n=\begin{Bmatrix}0\\1\end{Bmatrix}\right)\middle| N_\alpha(1)=k\right)\frac{\alpha^k}{k!}\cancel{e^{-\alpha}}$$

$$\sum_{k\geq 0}^{n} P\Big(\big(\theta_1=\{{}^0_1\}\big)\cap\ldots\cap\big(\theta_n=\{{}^0_1\}\big)\Big|N_\alpha(1)=k\Big)\frac{\alpha^k}{k!} = \underbrace{\left(1+\frac{\alpha}{n}\right)^n}$$

we use the binomial theorem to expand this:

$$\left(\frac{\alpha}{n}+1\right)^n = \sum_{k=0}^{n}\binom{n}{k}\left(\frac{\alpha}{n}\right)^k(1)^{n-k}$$

$$= \sum_{k\geq 0}^{n}\frac{(n)_k}{k!}\;\frac{\alpha^k}{n^k}$$

And we have that:

$$P\Big(\big(\theta_1=\{{}^0_1\}\big)\cap\ldots\cap\big(\theta_n=\{{}^0_1\}\big)\Big|N_\alpha(1)=k\Big) = \frac{(n)_k}{n^k} \quad ✓$$

$(n)_k \leftarrow$ {total # of ways of selecting k unique boxes (equivalent to boxes w/ occupation numbers 0, 1)}

$n^k \leftarrow$ {total # of ways of dropping k balls into n boxes}

This is one of the first computations we did in this course.
The occupation numbers with 0, 1.

Q: Why is it that we can say: if the sums are equal, the terms are equal?
A: Because if two functions have the same power series for the Taylor expansions, then they are equal. Equality of polynomial functions.
We have 2 functions of the variable α, which are equal.
α doesn't know that it is a variable.
Therefore, the Taylor series expansions are equal, therefore the coefficients are equal.

You can view the Schrödinger method as a probabilistic interpretation of Taylor expansion.

- Example - Any problem related to occupation numbers can be done using the Poisson Process

Trivially. Always using the aforementioned trick.

What is the probability that all the boxes are occupied? $P((\theta_1>0)\cap\ldots\cap(\theta_n>0)\,|\,N_\alpha(1)=k)=?$

$$\underbrace{P((\theta_1>0)\cap\ldots\cap(\theta_n>0))}=?$$

again, the occupation numbers of different intervals are independent random variables.
Hence, this is the intersection of independent events.

$$P((\theta_1>0)\cap\ldots\cap(\theta_n>0))=P(\theta_1>0)\,P(\theta_2>0)\ldots P(\theta_n>0)$$

for $i=1,\ldots,n$:

$$\begin{aligned}
P(\theta_i>0) &= 1-P(\theta_i\le 0) && \leftarrow \text{by definition of probability.}\\
&= 1-P(\theta_i=0) && \leftarrow \text{negative occupation numbers are meaningless here.}\\
&= 1-P_\alpha\left(\begin{bmatrix}[\frac{i-1}{n},\frac{i}{n})\\ 0\end{bmatrix}\right)\\
&= 1-\frac{\alpha^0(\frac{1}{n})^0}{0!}e^{-\alpha(\frac{1}{n})}\\
P(\theta_i>0) &= 1-e^{-\frac{\alpha}{n}}
\end{aligned}$$

$$\begin{aligned}
&= \left(1-e^{-\frac{\alpha}{n}}\right)^n\\
&= \left(\frac{e^{-\frac{\alpha}{n}}}{e^{-\frac{\alpha}{n}}}-e^{-\frac{\alpha}{n}}\right)^n\\
&= \left[e^{-\frac{\alpha}{n}}\left(e^{\frac{\alpha}{n}}-1\right)\right]^n
\end{aligned}$$

$$\boxed{P((\theta_1>0)\cap\ldots\cap(\theta_n>0)) = e^{-\alpha}\left(e^{\frac{\alpha}{n}}-1\right)^n}$$

By the Law of Alternatives:

$$\underbrace{P((\theta_1>0)\cap\ldots\cap(\theta_n>0))}_{e^{-\alpha}(e^{\frac{\alpha}{n}}-1)^n}=\sum_{k\ge 0}^{\infty}P((\theta_1>0)\cap\ldots\cap(\theta_n>0)\,|\,N_\alpha(1)=k)\,\underbrace{P(N_\alpha(1)=k)}_{\frac{\alpha^k 1^k}{k!}e^{-\alpha(1)}}$$

For convenience, as in the text, let's denote events as follows:

$A = ((\theta_1 > 0) \cap \ldots \cap (\theta_n > 0))$ $\qquad$ $B_k = ((\theta_1 > 0) \cap \ldots \cap (\theta_n > 0) \mid N_\alpha(1) = k)$

{these are the events whose probability we want to determine.}

and we've already showed that:

$$P(A) = e^{-\alpha}\left(e^{\frac{\alpha}{n}} - 1\right)^n$$

Using this notation, we write the expression obtained from the Law of Alternatives as:

$$P(A) = \sum_{k \geq 0} P(B_k) \frac{\alpha^k}{k!} e^{-\alpha}$$

multiplying both sides by e^{α}:

$$e^{\alpha} P(A) = \sum_{k \geq 0} P(B_k) \frac{\alpha^k}{k!}$$

using $P(A)$, as obtained above, we can rewrite the LHS:

$$(*) \quad \underbrace{\left(e^{\frac{\alpha}{n}} - 1\right)^n}_{\text{expand this, using the binomial theorem}} = \sum_{k \geq 0} P(B_k) \frac{\alpha^k}{k!}$$

$$\left((-1) + e^{\frac{\alpha}{n}}\right)^n = \sum_{j=0}^{n} \binom{n}{j} (-1)^j \left(e^{\frac{\alpha}{n}}\right)^{n-j}$$

$$= \sum_{j=0}^{n} \binom{n}{j} (-1)^j \underbrace{e^{(n-j)\alpha/n}}_{\text{expand using Maclaurin Series } e^x = \sum_{k \geq 0} \frac{x^k}{k!}, \text{ with } x = \frac{(n-j)\alpha}{n}}$$

$$= \sum_{j=0}^{n} \binom{n}{j} (-1)^j \left[\sum_{k \geq 0} \frac{\left(\frac{(n-j)\alpha}{n}\right)^k}{k!}\right]$$

$$= \sum_{j=0}^{n} \binom{n}{j} (-1)^j \left[\sum_{k \geq 0} \left(1 - \frac{j}{n}\right)^k \frac{\alpha^k}{k!}\right]$$

$$= \sum_{j=0}^{n} \sum_{k\geq 0} \binom{n}{j}(-1)^j \left(1-\frac{j}{n}\right)^k \frac{\alpha^k}{k!}$$

$$\left(e^{\frac{\alpha}{n}}-1\right)^n = \sum_{k\geq 0} \underbrace{\sum_{j=0}^{n} \binom{n}{j}(-1)^j \left(1-\frac{j}{n}\right)^k}\ \frac{\alpha^k}{k!}$$

So now we have:

$$\sum_{k\geq 0} \underbrace{\sum_{j=0}^{n} \binom{n}{j}(-1)^j \left(1-\frac{j}{n}\right)^k}\ \frac{\alpha^k}{k!} = \sum_{k\geq 0} \underbrace{P(B_k)}\ \frac{\alpha^k}{k!}$$

The LHS and RHS are both polynomials in α.
Since they are equal, for each term α^k, the corresponding coefficients must be equal

$$\boxed{P(B_k) = \sum_{j=0}^{n} \binom{n}{j}(-1)^j \left(1-\frac{j}{n}\right)^k}$$

And this is precisely the probability we desire:

$$P(B_k) = P\big((\theta_1>0)\cap \dots \cap(\theta_n>0) \,/\, N_\alpha(1)=k\big) = \sum_{j=0}^{n} \binom{n}{j}(-1)^j \left(1-\frac{j}{n}\right)^k$$

This is the same as the formula we obtained by using the inclusion-exclusion principle early in the course [2/18/98.4-5].

- This Schrödinger Randomization is one of the most powerful probability techniques known.

- Another solution of equation (*):

$$\underbrace{\left(e^{\frac{\alpha}{n}}-1\right)^n}_{\substack{\text{This is a fuction of } \alpha. \\ f(\alpha)}} = \underbrace{\sum_{k\geq 0} P(B_k)\frac{\alpha^k}{k!}}$$

Note that we sum over all terms α^k.
So this is just the Taylor expansion of the RHS.

From course 18.01, we have that:

$$f(\alpha) = \sum_{k\geq 0} \frac{c_k}{k!}\alpha^k ,$$

where $c_k = \Big[D^k f(\alpha)\Big]_{\alpha=0}$

{the k^{th} derivative of $f(\alpha)$, evaluated at $\alpha = 0$.}

Combining these two equations for $f(\alpha)$, we have:

$$\sum_{k\geq 0} \underbrace{P(B_k)}\frac{\alpha^k}{k!} = \sum_{k\geq 0} \underbrace{\Big[D^k\big(e^{\frac{\alpha}{n}}-1\big)^n\Big]_{\alpha=0}}\frac{\alpha^k}{k!}$$

This means that:

$$P(B_k) = P\big((\theta_1>0)\cap\ldots\cap(\theta_n>0) \,/\, N_\alpha(1)=k\big) = \Big[D^k\big(e^{\frac{\alpha}{n}}-1\big)^n\Big]_{\alpha=0}$$

And this is a nice, closed form expression.
Not bad.
It's better than the formula that we obtained on the previous page, as well as w/ inclusion-exclusion:
It's an interesting exercise to show that this formula is the same as the formula obtained before.

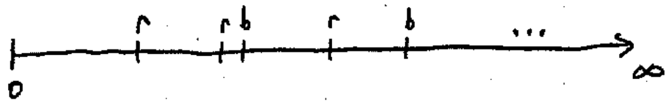
Probability that all n boxes are occupied when k balls dropped:

$$P(B_k) = \Big[D^k\big(e^{\frac{\alpha}{n}}-1\big)^n\Big]_{\alpha=0} = \sum_{j=0}^{n}\binom{n}{j}(-1)^j\Big(1-\frac{j}{n}\Big)^k$$

Property (6) <u>Poisson blips in 2 colors</u>

We have Poisson blips, but now the blips come in <u>2 colors</u> – red and black.
We have a machine that <u>only registers</u> <u>red blips</u>
it <u>misses</u> <u>black blips</u>.

How do we work this in?



Q: What remains to be said?
A: We have to tell what the probability that a given blip will be red or black.
Otherwise, you haven't specified it.

Therefore, we have a <u>Poisson process</u> <u>together</u> with the <u>probability</u> that a <u>blip will be red or black</u>.

How do we say this correctly? Once we say it correctly, of course we can compute it.
Here's how we say it correctly.

$X_1, X_2, \ldots$ = sequence of independent Bernoulli random variables where $P(X_i = 1) = p$

$\begin{cases} 1 = \text{red} \\ 0 = \text{black} \end{cases}$

Then, how do we write that the machine counts <u>only</u> the red blips?
This is an important realization.

$$X_1 + X_2 + \ldots + X_{N_\alpha(t)} = \text{number of red blips in the interval } [0,t]$$

$\{N_\alpha(t) \Rightarrow$ we randomize this interval $[0,t]\}$

Next time, we will compute probability:

$$P\left(X_1 + X_2 + \ldots + X_{N_\alpha(t)} = n\right)$$

by the <u>Law of Alternatives.</u>

Note, of course, that $N_\alpha(t)$ is <u>Independent</u> of the X_i.

This is an exact way of expressing that we have a machine that measures red blips.

John Guidi
guidi@mit.edu

The Poisson Process (Cont'd) / Random Walk and the Normal Distribution

The last lecture we did, on Schrödinger Randomization, was probably the hardest lecture in the whole course.
Basically we did the lecture in one hour.
It took me several years to understand it.

There are 2 properties of the Poisson Process left to describe.
Then we start on the last big topic of this course, which is the normal distribution and the stochastic process associated w/ the normal distribution – namely, Brownian motion.

Property (6) Poisson blips in 2 colors

Suppose we have a Poisson Process where blips are red or black

red w/ probability p
black w/ probability $q = 1-p$

We are only interested in registering the red blips.
How many red blips do we get in a given time interval?

$$R_\alpha(t) = \text{number of red blips in interval } [0,t]$$

↑ this is a random function

Our objective is to compute the probability distribution of this random function:

$$P(\underbrace{R_\alpha(t) = k}_{\text{this is a joint distribution}}) = ?$$

$P(X_i = 1) = p$
$P(X_i = 0) = q$

As we observed last time, for each X_i, you get count of $\left\{ \begin{matrix} 1 \text{ w/ probability } p \text{ (red)} \\ 0 \text{ w/ } \quad '' \quad q \text{ (black)} \end{matrix} \right\}$

$$R_\alpha(t) = X_1 + X_2 + \dots + X_{\underbrace{N_\alpha(t)}}$$

All the X_i are independent.

↑ there are, altogether, $N_\alpha(t)$ blips in the interval $[0,t]$

So, we want to compute the probability:

$$P(R_\alpha(t) = k) = P(X_1 + X_2 + \dots + X_{N_\alpha(t)} = k) = ?$$

How do you think we compute this?
The only way known to man. The Law of Alternatives.

by the Law of Alternatives:

$$P(X_1+X_2+\dots+X_{N_\alpha(t)}=k) = \sum_{n\geq 0} \underbrace{P(X_1+X_2+\dots+X_{N_\alpha(t)}=k \mid N_\alpha(t)=n)} P(N_\alpha(t)=n)$$

Since it is given that $N_\alpha(t)=n$, we have that event:

$$(X_1+X_2+\dots+X_{N_\alpha(t)}=k \mid N_\alpha(t)=n) = (X_1+X_2+\dots+X_n=k)$$

$$= \sum_{n\geq 0} P(X_1+X_2+\dots+X_n=k)\, \underbrace{P(N_\alpha(t)=n)}_{\frac{\alpha^n t^n}{n!}e^{-\alpha t}}$$

the formula for the Poisson distribution

$$= \sum_{n\geq 0} P(X_1+X_2+\dots+X_n=k)\,\frac{(\alpha t)^n}{n!}e^{-\alpha t}$$

Next, we have to remind ourselves what probability distribution of the sum of n independent random variables is. This is called the binomial distribution. Remember. The good old days. [2/20/98.7 - 9]

Let's perform an interchange of notation to remind ourselves of the formula.

$$S_n = X_1+X_2+\dots+X_n$$

and we have:

$$P(S_n=k) = \begin{cases} \binom{n}{k} p^k q^{n-k} & , \ n\geq k \\ 0 & \text{otherwise} \end{cases}$$

$$= \underbrace{\sum_{n\geq 0}} \underbrace{P(S_n=k)}\,\frac{(\alpha t)^n}{n!}e^{-\alpha t}$$

$$= \sum_{n\geq k} \overbrace{\binom{n}{k} p^k \underbrace{q^{n-k}}}\,\frac{(\alpha t)^n}{n!}e^{-\alpha t}$$

$$= \sum_{n\geq k} \frac{(n)_k}{k!} p^k q^{-k}\,\frac{(\alpha t q)^n}{n!}e^{-\alpha t}$$

$$= \frac{1}{k!} p^k q^{-k} e^{-\alpha t} \sum_{n \geq k} \frac{(n)_k}{n!} (\alpha t q)^n$$

Let $j = n-k$

$$= \frac{1}{k!} p^k q^{-k} e^{-\alpha t} \sum_{j \geq 0} \underbrace{\frac{(j+k)_k}{(j+k)!}} (\alpha t q)^{j+k}$$

$$\frac{(j+k)_k}{(j+k)!} = \frac{\cancel{(j+k)\cdots(j+1)}}{\cancel{(j+k)\cdots(j+1)}\, j!}$$

$$= \frac{1}{j!}$$

$$= \frac{1}{k!} p^k \cancel{q^{-k}} e^{-\alpha t} (\alpha t \cancel{q})^k \underbrace{\sum_{j \geq 0} \frac{(\alpha t q)^j}{j!}}$$

This is <u>nice</u>.
This is the Taylor expansion for $e^{\alpha t q}$

$$= \frac{1}{k!} (p\alpha)^k t^k \underbrace{e^{-\alpha t} e^{\alpha t q}}$$

$$e^{-\alpha t} e^{\alpha t q} = e^{-\alpha t (1-q)}$$
$$= e^{-\alpha t p}$$

$$\boxed{\underbrace{P(X_1 + X_2 + \dots + X_{N_\alpha(t)} = k)}_{P(R_\alpha(t) = k)} = \frac{(p\alpha)^k t^k}{k!} e^{-(p\alpha)t}}$$

What is this trying to tell us?
It's trying to tell us that the random function $R_\alpha(t)$, the count of the red blips, is a <u>Poisson process</u>, where the <u>intensity has changed</u> <u>from α to $p\alpha$</u>,

$$\alpha \to p\alpha \qquad \frac{(p\alpha)^k t^k}{k!} e^{-(p\alpha)t}$$

This is intuitively obvious.
Especially after ½ hour of computation.

- Conclusion

  If you sample only the red blips, you still have a Poisson process, but the intensity is changed by multiplying by the probability of a red blip.

  $$\alpha \to (p\alpha)$$

- See, the Poisson Process has all these nice properties. Last one (but we could go on forever).

Property (7) Laplace Transform

If you are given only the average number of blips in a unit interval, then the only stochastic process that fits these data is the Poisson process. } We've said this several times already. Loudly.

Complete randomness, except for the intensity α.

Now you say: Excuse me, that's not completely random.
If it were completely random, then even α would be random.

Let's see what happens if we take the Poisson process and you make the intensity α into a random variable, continuous, with a given density.
We've got the technique.

Suppose α is random with density $\text{dens}(\alpha=s) = f(s)$

{I want to give $\text{dens}(\alpha=s)$ a name.}

Next, we compute the probability that there are no blips in the interval $[0,t)$:

$$P(N_\alpha(t)=0) = \int_{-\infty}^{\infty} P(N_\alpha(t)=0 \mid \alpha=s)\, \overbrace{\text{dens}(\alpha=s)}^{f(s)}\, ds$$

{The only way known to man to do this, The continuous Law of Alternatives.}

$$= \int_0^{\infty} P(\underbrace{N_\alpha(t)=0 \mid \alpha=s}_{})\, f(s)\, ds$$

since it is given that $\alpha=s$, we have that events:

$$(N_\alpha(t)=0 \mid \alpha=s) = (N_s(t)=0)$$

$$= \int_0^{\infty} \underbrace{P(N_s(t)=0)}_{\frac{s^0 t^0}{0!}e^{-st} = e^{-st}} f(s)\, ds$$

$$P(N_\alpha(t)=0) = \underbrace{\int_0^{\infty} e^{-st} f(s)\, ds}_{\text{The Laplace Transform}}$$

This is where the Laplace transform comes from.
This is a probabilistic interpretation of the Laplace transform.

Every fact of the Laplace transform can be interpreted probabilistically from this.
We already saw convolution.
Now we see the Laplace Transform.

- This stochastic process, namely the Poisson process w/ random intensity is, in the literature, called the Cox process.

We don't have enough time in class, but there are wonderful things you can say about this.

That's it kids.
That's the end of the Poisson process.

Now, we start on the most important topic of this course.
The Theory of the Normal Distribution
You could say the rest of this course was preparation for this topic.

Let's start w/ something we have strangely neglected:

- Random Walk

It's really something we have done implicitly.
Let's work it out.

Take the Bernoulli process.
We toss a coin.
If we get heads, write +1
tails, write −1

} That's the only change.
But this change makes a tremendous psychological difference, because you can graph it.

So we take:

$$Y_1, Y_2, \ldots \quad \text{i.i.d. random variables}$$

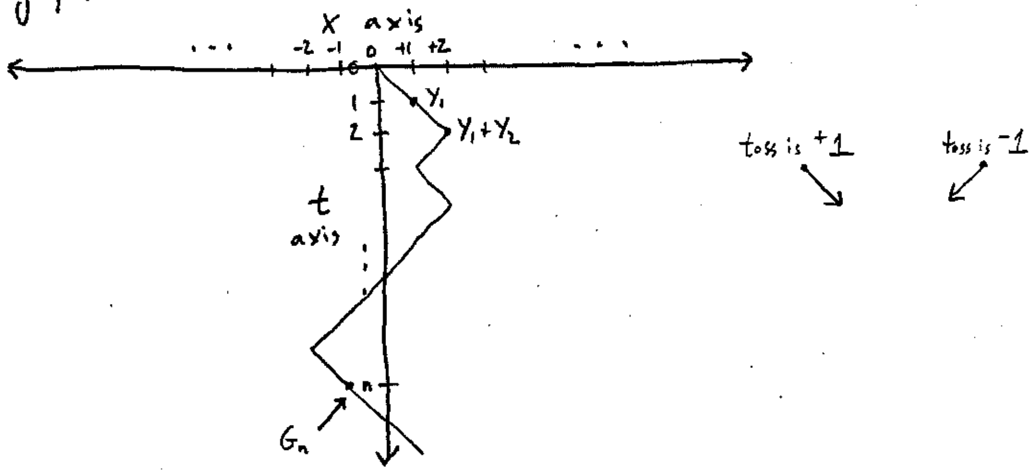
Let assume, here, that the coin is fair:

$$P(Y_i = +1) = P(Y_i = -1) = \frac{1}{2}$$

And we define the random variable:

$$\begin{aligned} G_n &= \text{position at time } n \\ &= Y_1 + Y_2 + \ldots + Y_n \end{aligned}$$

We are pretty much tossing a fair coin, but the <u>language changes.</u>
And we graph it as follows:



every sample $\omega = (\pm 1, \pm 1, \ldots)$

The sample space Ω is the set of all sequences of ± 1.
To every sample point ω, you associate a <u>continuous curve</u>, as illustrated above,
As soon as you do that, you realize there are questions to be asked.
For example:

Q: How long (i.e., number of tosses) does it take to get back to 0?

Q: How long does it take to reach some particular value of x?

All sorts of questions from this "different" viewpoint.
You'll find all sorts of these answered in the book.

Let's consider:

$$P(G_n = x) = \ ?$$

probability that at time n, you are at level x

every toss has probability $\frac{1}{2}$ of being $\begin{cases} \text{heads} \to +1 \\ \text{tails} \to -1 \end{cases}$

Suppose we have:

$$\left.\begin{aligned} u &= \#\text{ of } +1\text{'s} \\ v &= \#\text{ of } -1\text{'s} \end{aligned}\right\} \quad \begin{aligned} u + v &= n \longleftarrow \text{number of tosses} \\ u - v &= x \longleftarrow \text{level} \end{aligned}$$

$$\left.\begin{aligned} v &= u - x \\ u + (u - x) &= n \end{aligned}\right\} \Rightarrow u = \frac{n+x}{2}$$

$$\left.\begin{aligned} u &= x + v \\ (x+v) + v &= n \end{aligned}\right\} \Rightarrow v = \frac{n-x}{2}$$

Consider the event $(G_n = x)$ in terms of the standard random variable S_n in the Bernoulli process [2/20/98.7-9].
Recall S_n is the <u>number of heads</u> in <u>n tosses</u>.

$$\underbrace{(G_n = x)}_{\text{after } n \text{ tosses, you are at level } x} = \underbrace{(S_n = u)}_{\text{after } n \text{ tosses, you have } u \text{ heads}}$$

$$P(G_n = x) = P(S_n = u)$$

binomial distribution

$$= \binom{n}{u}\left(\frac{1}{2}\right)^u \left(\frac{1}{2}\right)^{n-u}$$

$n-u=v$

$$= \binom{n}{u}\frac{1}{2^n} \quad \text{and since } u = \frac{n+x}{2},$$

$$\boxed{P(G_n = x) = \binom{n}{\frac{n+x}{2}}\frac{1}{2^n}}$$

- What is :

$$P(\underbrace{G_{2n}} = 0) = ?$$

why $2n$? Because we can <u>not</u> go to 0 in an <u>odd number</u> of steps. There must be an <u>even</u> number of steps. Figure it out for yourselves.

From above, we have :

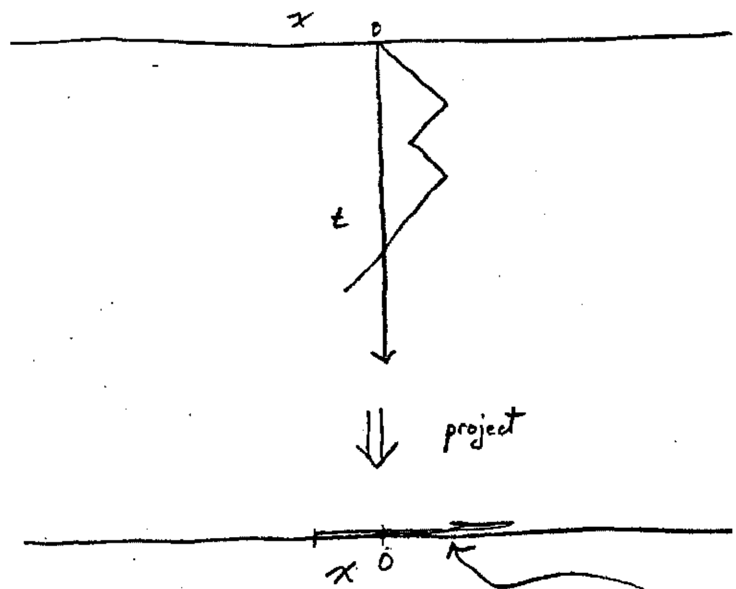
$$P(G_{2n} = 0) = \binom{2n}{\frac{2n+0}{2}}\frac{1}{2^{2n}}$$

$$(*) \qquad = \binom{2n}{n}\frac{1}{2^{2n}}$$

- Now let's define the random variable :

$$F = \text{time of first return to origin}$$

If you look just at the x axis, and you project the curve onto the x axis



Then you can visualize a random walk as a random walk. You go left, right, right, ...
Hence the name.
So it makes sense to ask when you return to the origin.

Now, let me tell you informally something that we are not going to establish, but that you ought to know.
It's part of your cultural background.

Probability that a random walk that starts at the origin ever comes back to the origin is 1, in 1 dimension and 2 dimensions.
In 3-dimensions, the probability is .397.
This is an incredible fact of life. You can make money on this.
In 1+2 dimensions, all roads go to 1.
In 3 dimensions, not.

dimension	P(random walk starting at origin returns to origin)
1	1
2	1
3	.397
4+	weirdo. Less + less weirdo than exponential

- This makes it all the more decisive to compute the probability distribution of this random variable F, the first time you come back to the origin.

Let's define:

$$f_{2n} \triangleq P(F = 2n) \quad \Bigg\} \quad \begin{array}{l} \text{what is the probability distribution?} \\ f_{2n} = P(F=2n) = \ ? \end{array}$$

You know, we've covered 3 terms of probability in this course.
Not that I want to brag.
But this is 3 terms of probability.
If you ever take another probability course later, 18.313 will be <u>vindicated</u>.
Because, you will say: why is this guy doing all this trivial stuff?
Just watch. I <u>promise</u> this is going to happen.

So we next consider the event $(G_{2n} = 0)$.
Now we use a very important trick.
If you <u>returned to 0</u> at time equal $2n$, that means there <u>must be</u> a <u>first</u> time when you return to 0.
If you don't see this, I stop.
Let's write that down.

$$\underbrace{(G_{2n} = 0)}_{\text{at time } 2n \text{ you were at origin}} = \left((F=2) \cap (G_{2n-2} = 0)\right) \cup \left((F=4) \cap (G_{2n-4} = 0)\right) \cup \ldots \cup (F = 2n)$$

$2+(2n-2)=2n$ $\qquad$ $4+(2n-4)=2n$

first return to origin was at time 2,
followed by time $2n-2$, at which time you were at origin.
Thus, at origin at time $2+(2n-2)=2n$

Now we take probabilities.
On the RHS, we have the <u>union of disjoint events.</u>
And the intersectands are <u>independent</u> random variables (F's and G's are <u>independent</u>)

$$P(G_{2n}=0) = P(F=2)P(G_{2n-2}=0) + P(F=4)P(G_{2n-4}=0) + \ldots + P(F=2n)$$

As is tradition, set $P(G_{2n}=0) = U_{2n}$. Rewriting, we obtain:

$$(**) \quad U_{2n} = f_2 U_{2n-2} + f_4 U_{2n-4} + \ldots + f_{2n}$$

But, from equation (*), we have that:

$$U_{2n} = P(G_{2n}=0) = \binom{2n}{n}\frac{1}{2^{2n}}$$

Therefore, we can solve <u>recursively</u> for the f's

$$\begin{cases} U_{2n} = \binom{2n}{n}\frac{1}{2^{2n}} \\ U_{2n} = f_2 U_{2n-2} + f_4 U_{2n-4} + \ldots + f_{2n} \end{cases}$$

giving the desired probabilities:

$$f_{2i} = P(F=2i)$$

Alternatively, we can use the z-transform.

Set: $U(z) = 1 + U_2 z^2 + U_4 z^4 + U_6 z^6 + \ldots$

$F(z) = f_2 z^2 + f_4 z^4 + f_6 z^6 + \ldots$

Then we have:

$$U(z) - 1 = U(z) F(z)$$

Proof sketch

First, multiply the two power series on the RHS.

$$\begin{aligned} U(z)F(z) = \boxed{f_2} z^2 &+ \boxed{f_4} z^4 + \boxed{f_6} z^6 + \boxed{f_8} z^8 + \ldots \\ &+ \boxed{U_2 f_2} z^4 + \boxed{U_2 f_4} z^6 + \boxed{U_2 f_6} z^8 + \ldots \\ &\qquad + \boxed{U_4 f_2} z^6 + \boxed{U_4 f_4} z^8 + \ldots \\ &\qquad\qquad + \boxed{U_6 f_2} z^8 + \ldots \end{aligned}$$

$$= \sum_{n \geq 1} \left(\left(U_{2n-2} f_2 + U_{2n-4} f_4 + \ldots + \overset{1}{\cancel{U_{2n-2n}}} f_{2n} \right) z^{2n} \right)$$

$$= \sum_{n \geq 1} \left(\left(\sum_{j=1}^{n} U_{2n-2j} f_{2j} \right) z^{2n} \right)$$

$$\underbrace{U(z) - 1}_{\text{we defined } U(z) \text{ above. It can be written as:}} = \sum_{n \geq 1} \left(\left(\sum_{j=1}^{n} U_{2n-2j} f_{2j} \right) z^{2n} \right)$$

$$U(z) = 1 + \sum_{n \geq 1} \left(U_{2n} z^{2n} \right)$$

Subtracting 1 gives the LHS above:

$$\sum_{n \geq 1} \left(U_{2n} z^{2n} \right) = \sum_{n \geq 1} \left(\left(\sum_{j=1}^{n} U_{2n-2j} f_{2j} \right) z^{2n} \right)$$

These two polynomials in z are equal only if the coefficients of each term z^{2n} are equal. Thus this requires that:

for all $n \geq 1$, $$U_{2n} = \sum_{j=1}^{n} \left(U_{2n-2j} f_{2j} \right)$$

This is a tautology.
This is precisely equation (**).
Thus we have shown that:
$U(z) - 1 = U(z) F(z)$ ✓

Since $U(z) - 1 = U(z)F(z)$, we have that:

$$F(z) = \frac{U(z)-1}{U(z)}$$

$$\boxed{F(z) = 1 - \frac{1}{U(z)}}$$

Another way of writing $U(z)$ is:

$$U(z) = \sum_{n \geq 0} \left(U_{2n} z^{2n} \right)$$

recall that we proved earlier [5/1/98.7] that:

$$U_{2n} = P(G_{2n} = 0) = \binom{2n}{n} \frac{1}{2^{2n}}$$

$$= \underbrace{\sum_{n \geq 0} \left(\binom{2n}{n} \frac{1}{2^{2n}} z^{2n} \right)}$$

"An <u>easy computation</u> shows"
Well, a straightforward computation, anyway, shows that this is the Maclaurin Series expansion of the function: $(1-z^2)^{-\frac{1}{2}}$

$$\boxed{U(z) = (1-z^2)^{-\frac{1}{2}}}$$

And we can rewrite $F(z)$ above as:

$$\underbrace{F(z)} = 1 - (1-z^2)^{\frac{1}{2}}$$

from our definition above,
$F(z) = f_2 z^2 + f_4 z^4 + \ldots$

$$\sum_{n \geq 1} \left(f_{2n} z^{2n} \right) = 1 - \underbrace{(1-z^2)^{\frac{1}{2}}}_{\text{expand by the binomial theorem w/ fractional exponents}}$$

$$= 1 - (1 + (-z^2))^{\frac{1}{2}}$$

$$= 1 - \sum_{n\geq 0} \binom{\frac{1}{2}}{n} \left(-z^2\right)^n$$

$$= \cancel{1} - \cancel{1} - \sum_{n\geq 1} \binom{\frac{1}{2}}{n} \left(-z^2\right)^n$$

$$\sum_{n\geq 1}\left(f_{2n}\, z^{2n}\right) = -\sum_{n\geq 1}\left(\binom{\frac{1}{2}}{n}(-1)^n z^{2n}\right)$$

These <u>z polynomials</u> are <u>equal</u> and, <u>thus</u>, the <u>coefficients</u> for <u>each term z^{2n}</u> are <u>equal</u>.

$$f_{2n} = \pm\binom{\frac{1}{2}}{n} \quad \text{, for all } n \geq 1$$

$$\binom{\frac{1}{2}}{n} \triangleq \frac{\overbrace{\frac{1}{2}\left(\frac{1}{2}-1\right)\left(\frac{1}{2}-2\right)\cdots\left(\frac{1}{2}-(n-1)\right)}^{n \text{ terms}}}{n!}$$

$$= \frac{\frac{1}{2}\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)\cdots\left(-\frac{(2n-3)}{2}\right)}{n!}$$

$$= (-1)^{n-1}\,\frac{\frac{1}{2}\left(\frac{1}{2}\right)\left(\frac{3}{2}\right)\cdots\left(\frac{2n-3}{2}\right)}{n!} \times \left(\frac{2^n}{2^n}\right)$$

$$= \frac{(-1)^{n-1}}{2^n}\,\frac{1\times\overbrace{1\times 3\times\cdots\times 2n-3}^{n-1 \text{ terms}}}{n!} \times \left(\frac{\overbrace{2\times 4\times\cdots\times 2n-2}^{n-1 \text{ terms}}}{2\times 4\times\cdots\times 2n-2}\right)$$

$$\frac{2^{n-1}(n-1)!}{2^{n-1}(n-1)!}$$

$$= \frac{(-1)^{n-1}}{2^n}\,\frac{(2n-2)!}{n!\,(2\times1)(2\times2)\cdots(2\times n-1)}$$

$$= \frac{(-1)^{n-1}}{2^n}\,\frac{(2n-2)!}{n!\cdot 2^{n-1}\,(n-1)!}$$

$$= (-1)^{n-1}\,\frac{(2n-2)!}{(n-1)!\,n!}\,\frac{1}{2^{2n-1}}$$

$$= (-1)^{n-1}\,\frac{1}{2n-1}\,\frac{(2n-1)!}{(n-1)!\,n!}\,\frac{1}{2^{2n-1}}$$

$$= (-1)^{n-1}\,\frac{1}{2n-1}\binom{2n-1}{n-1}\frac{1}{2^{2n-1}}$$

$$\left|\binom{\frac{1}{2}}{n}\right| = \left|\frac{1}{2n-1}\binom{2n-1}{n-1}\frac{1}{2^{2n-1}}\right| \quad \text{, for all } n\geq 1$$

And now we have the desired probability distribution we have been after - namely, that of the random variable F, the first time you come back to the origin on a random walk.

$$\boxed{f_{2n} \triangleq P(F=2n) = \left| \frac{1}{2n-1} \binom{2n-1}{n-1} \frac{1}{2^{2n-1}} \right| \quad , \forall n \geq 1}$$

This alternative computation, where we use the z-transform to solve for the probability distribution f_{2n}, is far from a trivial computation.

Q/A: This is related to the Catalan numbers.
Prof. Stanley has just finished the 2nd volume of his book on combinatorics, which includes a chapter called "62 permutations of the z-transform."

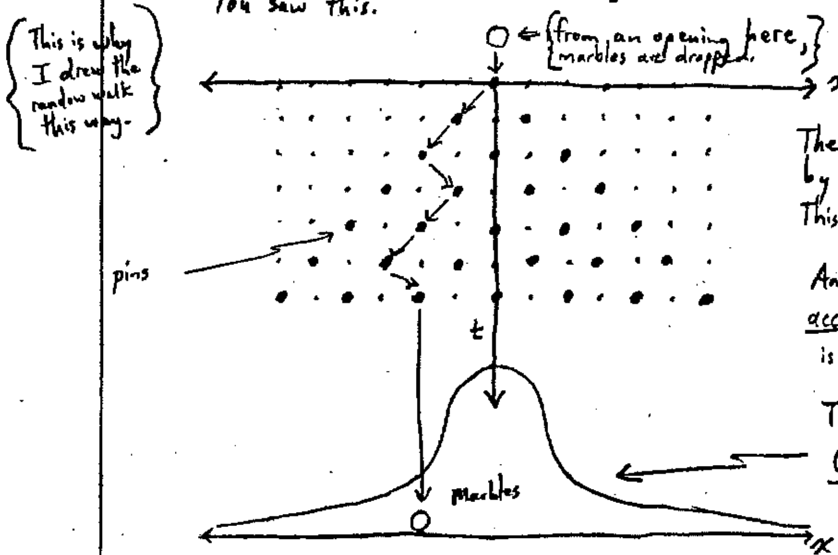
- Now we come to the real point of this lecture.
The real point is this.

Is there a continuous analogue of random walk?

In other words, from the Bernoulli process, we obtained the Poisson process by a limiting process.
Is their another limiting process where, from the random walk interpretation of the Bernoulli process the steps become shorter and shorter, and (if you don't get too much oscillation, so that the limit is well behaved), you can obtain this other limiting process?

The amazing thing is, one of the best chapters of probability, that there is such a limiting process.

How do you get it?
Didn't you all go to the Museum of Science when you were kids?
When you went there, what did you see?
You saw this.



The marbles are dropped and displaced by the pins.
This is a random walk.

And the fact that the marbles pile up according to a certain curve, means that there is a limiting process.

This limiting process gives you the Gaussian curve as the limiting curve.

Next: We will study this limiting process.

John Guidi
guidi@mit.edu

The Bell Shaped Curve

Let's continue the discourse we initiated last time, when we briefly surveyed the notion of random walk.

Recall that a symmetric random walk is just: fair coin

sequence $Y_1, Y_2, \ldots$ i.i.d. random variables with the property that:

$$P(Y_i = +1) = \frac{1}{2} \quad , \quad P(Y_i = -1) = \frac{1}{2}$$

just like coin tossing, where heads $= +1$
tails $= -1$

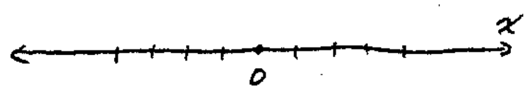
Under these circumstances, we define:

$$G_n = Y_1 + Y_2 + \ldots + Y_n$$

position at time n

- The intuitive interpretation, as we saw last time, is a random walk.

1-D

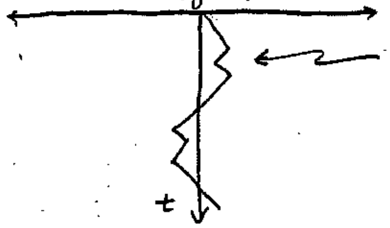


Namely, you look at the x axis w/ integer points.
Then you have a drunkard starting at the origin.
If the coin comes up heads, the drunkard moves to the right $+1$
tails, " " " " " left -1

This process has infinitely many properties that are still being studied.

- We call the random walk symmetric because we use a fair coin.
Actually, if the coin is not fair, the properties of random walk are the same as they would be for a fair coin, except for a drift.
The drift corresponds to the unfairness of the coin.
So the symmetric random walk gives you all the fundamental properties.

- We can visualize this by graphing a random walk as a polygonal curve.



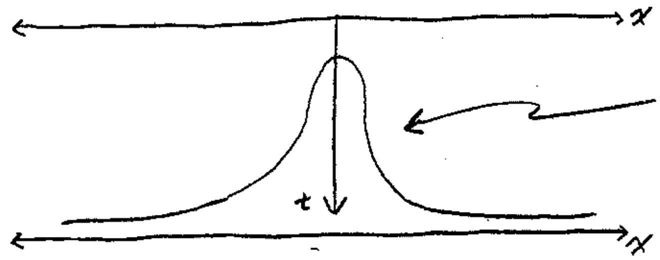
this way you get a visualization of a particular path of a random walk.

You can view a stochastic process, i.e., random walk, whose sample space Ω consists of all these paths.
The sample space is the set of all sequences of ± 1. A sample point $\omega = (\pm 1, \pm 1, \ldots)$.
[infinite]
But you can visualize it by a path.

So you are putting a probability on the set of all paths (i.e., this is equivalent to putting a probability on the set of all sequences).

The paths are infinite. For finite paths, you just condition.

- As we saw last time, from our Museum of Science contraption:



{When the balls are dropped, they form a curve, which is a bell shaped curve.}

The equation of this curve is:

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Standard Normal Distribution ⟵ the explanation of this bell shaped curve.

This is one of the important chapters in probability — in fact, of mathematics.

- Let's first verify that the Standard Normal Distribution is indeed a probability distribution.

$$\int_{-\infty}^{\infty} P(X = x)\, dx = 1$$

So, if $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ is a probability distribution, integral must be 1. $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \geq 0$.
Check:

{We did this in 18.03}

$$\underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx}_{\text{call this integral } A} \stackrel{?}{=} 1$$

$$A^2 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy}_{}$$

use a different variable of integration for the 2nd term

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{(x^2+y^2)}{2}} dx\, dy$$

Switch to polar coordinates

$$\int_D f(x,y)\, dA = \int_{r_0}^{r_n} \int_{\theta_0}^{\theta_n} f(r\cos\theta, r\sin\theta)\, r\, d\theta\, dr$$

$$= \frac{1}{2\pi} \int_{r=0}^{\infty} \underbrace{\int_{\theta=0}^{2\pi} e^{-\frac{r^2}{2}} r\, d\theta}\, dr$$

$$\int_0^{2\pi} e^{-r^2/2} r\, d\theta = 2\pi e^{-r^2/2} r$$

$$= \frac{1}{\cancel{2\pi}} \int_0^{\infty} \cancel{2\pi} e^{-\frac{r^2}{2}} r\, dr$$

$$= \int_0^{\infty} e^{-\frac{r^2}{2}} r\, dr$$

Let $u = \frac{r^2}{2}$

$du = r\, dr$

$$= \int_{r=0}^{\infty} e^{-u} du$$

$$= -e^{-u} \Big|_{r=0}^{\infty}$$

$$= -e^{-\frac{r^2}{2}} \Big|_{r=0}^{\infty}$$

$$= 0 - (-1)$$

$$A^2 = 1 \checkmark$$

Thus: $A = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 1 \checkmark$

The bell shaped curve does indeed define a probability distribution.

On which sample space?
Does there exist a sample space and a random variable on that sample space which has this probability distribution?

These questions are easily justified using a "Grant's Tomb" argument, As follows:

Sample space $\Omega = (-\infty, \infty)$
the line

sample point ω = point on the line

this is the only case we've covered where the sample points are actual points.

events are the events generated by the intervals
events like we did for the uniform process $((a,b])$, etc.

Define probability by setting:

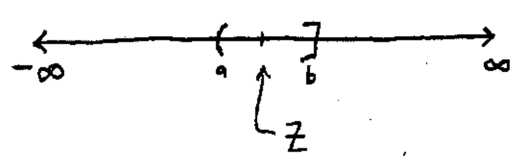
$$P\big((a,b]\big) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}}\,dx$$

that's a probability.
We just verified that the integral over the whole line is 1:

$$P\big((-\infty,\infty)\big) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx = 1$$

So this is a sample space.
It may not be what you expect for a sample space,
but you have to swallow it.

Now let's find a random variable.
The random variable is that we pick a point on the line at random.
Choose a point at random, what's the probability distribution? "Who's buried in Grant's Tomb?"



$$P(a < Z \le b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}}\,dx$$

Because we built it that way. It's circular.

random variable Z = choose a point at random

Recall the definition of density:

$$P(a < Z \le b) = \int_a^b dens(Z=t)\,dt$$

So, fact that:

$$P(a < Z \le b) = \int_a^b \underbrace{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}} dx$$

means that:

$$dens(Z=t) = \overbrace{\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}}$$

- This trick works – whenever you have a distribution and you want to construct a sample space.

  A random variable, having the given probability distribution, can be constructed this way.

  This shows that it is always possible to realize a probability distribution by a random variable.

- Z is called a standard normal distributed random variable

  this notation is universal

- Now, we want a justification for this phenomenon.
  Here, we get to the frontiers of science.
  The problem is not that there is no justification.
  The problem is there are too many.

  When there are several completely different justifications for the same phenomenon, something is funny.

  You are stunned if there is just one.
  But, when there are several, from completely different points of view, something may be missing.

I will point out the gaps in everybody's knowledge.
Which people don't like to talk about.
We'll discuss 3 justifications:

1) Gauss
2) Maxwell – Einstein
3) Central Limit Theorem of Probability

3 different justifications

- ~~First~~, let's see what is it we want to justify.
  Let's be clear as to what is it we want to justify.
  To do that, we have to introduce a new notion:

  Variance

[Character Assassination, aka Course Evaluations, done.
1) letter for his eyes only.
2) MIT evaluation.]

- Variance of a random variable

  First, we'll give the definition of variance, then we'll make comments on what it is really about.

  Let X = any random variable with expectation
  $E(X) = m$

  The variance, by definition, of X is:

$$\mathrm{Var}(X) = E\left((X-m)^2\right)$$

rewriting this using $E(X)=m$:

$$= E\Big(\underbrace{(X - E(X))^2}_{\text{expanding this}}\Big)$$

$$= E\left(X^2 - \cancel{2}E(X)^2 + E\cancel{(X)}^2\right)$$

$$\mathrm{Var}(X) = E\left(X^2 - E(X)^2\right)$$

or, if you prefer:

$$\mathrm{Var}(X) = E\left(X^2 - m^2\right)$$

Before we ask what variance is about,
let's back up and look again at random variables.

- What is a real random variable, really?

It can be one of many things, depending on what you are interested in.

View (1) A random variable describes a random phenomenon

By giving the phenomenon a probability distribution, instead of a specific value.
It's a natural, logical thing to describe such phenomena, not by a single value, but by a range of values together with their probabilities.

View (2) Random variable is the result of a search

we touched briefly on this in SuperClass 4 [4/24/98]

Every integer random variable defines a partition of the sample space.
When you do a search, the random variable is the measurement that tells you which block in this partition the unknown sample point is in.

That's the beginning of search theory, information theory, and all that.

But, now, we are seeing a 3rd concept of a random variable.

View (3) Random Variable as measurement of an imperfectly given quantity

For example, you want to measure people's height.
You measure 1000 people and get their average.
You want to know how close is this average to the average you would get by measuring everyone on the planet's height.

You can view the process of measuring someone's height as:
The process of measuring an average and the deviation from that average

Or: You have a measurement process, with a measuring device that is imperfect. So, it varies a lot.

- The variance of a random variable is related to the interpretation of the random variable as an imperfect measurement in view of the expectation, i.e., View (3) above.

The variance tells you how accurate the random variable is.
You are squaring the deviation of the random variable from it's average value.

$$Var(X) = E\left(\left(X - E(X)\right)^2\right)$$

$E(X) = m$

Var(X) will be big if the random variable is spread out
small " " " " is concentrated around it's average

Variance is a measure of spread

It is the best measure of spread

- In the limit, suppose that X is a random variable that takes the value 3 with probability 1.

$$\left.\begin{array}{l} P(X=3)=1 \\ E(X)=3 \end{array}\right\} \Rightarrow Var(X) = E\Big(\big(\underset{\substack{\uparrow \\ X=3\text{, in} \\ \text{the limit}}}{X} - \underbrace{E(X)}_{3}\big)^2\Big)$$

$$= 0$$

So, if the random variable X takes only one value, its variance $Var(X)$ is 0.

If the variables are all completely spread out, the variance will be big.
We'll see examples w/ the normal distribution.

- Properties of Variance

Whereas $E(X+Y) = E(X) + E(Y)$ always,
it is not necessarily true that

$$Var(X+Y) \stackrel{?}{=} Var(X) + Var(Y)$$

However, if X and Y are independent random variables:

$$Var(X+Y) = Var(X) + Var(Y)$$

X, Y independent, then we have equality

Let's verify this fact that variance is additive for independent random variables.

Proof

X, Y are <u>independent</u> random variables

Since $Var(Z) = E((Z-E(Z))^2)$, the variance of the sum $X+Y$ is, by definition:

$$Var(X+Y) = E(\underbrace{((X+Y)-E(X+Y))^2}_{\text{expand this out}})$$

$$= E\Big[(X+Y)^2 - 2(X+Y)\underbrace{E(X+Y)}_{(E(X)+E(Y))} + \underbrace{E(X+Y)^2}_{(E(X)+E(Y))^2}\Big]$$

$$= E\Big[X^2+2XY+Y^2$$
$$-2XE(X)-2YE(X)-2XE(Y)-2YE(Y)$$
$$+E(X)^2+2E(X)E(Y)+E(Y)^2\Big]$$

since expecation of a sum is the sum of the expectations:

$$= E(X^2)+E(2XY)+E(Y^2)$$
$$-E(2XE(X))-E(2YE(X))-E(2XE(Y))-E(2YE(Y))$$
$$+E(E(X)^2)+E(2E(X)E(Y))+E(E(Y)^2)$$

remember that, for c some constant:

$$E(cZ) = E(\underbrace{Z+Z+\dots+Z}_{c})$$
$$= cE(Z)$$

also, the <u>expectation</u> is a <u>constant</u>,

$$E(\underbrace{E(Z)}) = E(Z)$$

$E(Z)$ is a constant, so it goes <u>out of the expectation</u>

For example:

$$E(2YE(X)) = E((\underbrace{2E(X)})Y) = \underbrace{2E(X)}E(Y)$$

$$= E(X^2) + 2E(XY) + E(Y^2)$$
$$-\cancel{2}E(X)^2 - 2E\cancel{(X)E}(Y) - 2E(Y)E(X) - \cancel{2}E(Y)^2$$
$$+ E\cancel{(X)^2} + 2E\cancel{(X)E}(Y) + E\cancel{(Y)^2}$$

Since X + Y are independent, the product rule for expectation states that:

$$E(XY) = E(X)E(Y)$$

For proof, see text p. 145

$$= E(X^2) + 2E\cancel{(X)E(Y)} + E(Y^2)$$
$$-E(X)^2 - 2E\cancel{(Y)E}(X) - E(Y)^2$$

$$= \underbrace{E(X^2) - E(X)^2}_{Var(X)} + \underbrace{E(Y^2) - E(Y)^2}_{Var(Y)}$$

$$= Var(X) + Var(Y) \quad \checkmark$$

Which proves our contention that, if X and Y are independent random variables:

$$Var(X+Y) = Var(X) + Var(Y)$$

John Guidi
guidi@mit.edu

18.313

Lecture 33
5/6/98.1

Variance: The Normal Distribution

X = a random variable

$E(X) = m$ ← expectation or mean. We'll assume finite expectation

(Note that it is possible, by the way, for a random variable to have infinite expectation.
You can construct examples of this.)

The variance we define as the measure of the deviation from the expectation.
We square so that deviations on either side of the mean count the same.

$$\mathrm{Var}(X) = E((X-m)^2)$$

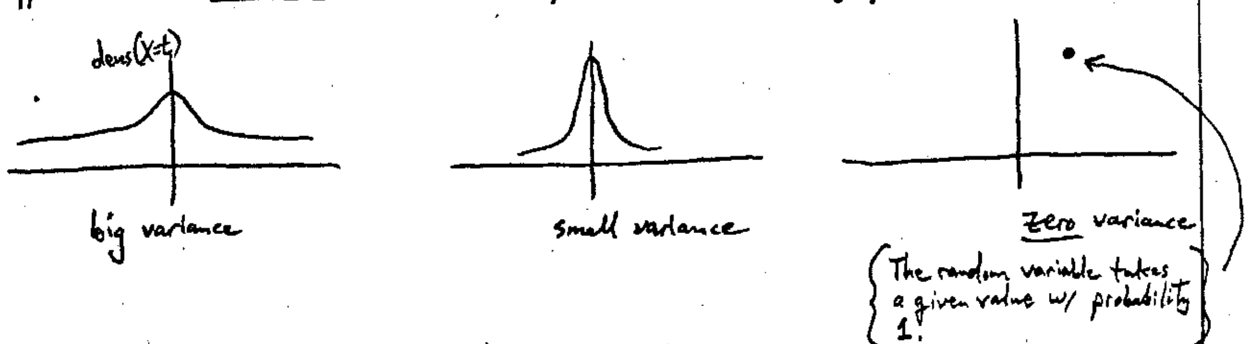
$$= E(X^2 - 2mX + m^2) = E(X^2) - E(2mX) + E(m^2)$$

$$= E(X^2) - \underbrace{2mE(X)}_{2m^2} + \underbrace{E(m^2)}_{m^2}$$

$$= E(X^2) - m^2$$

$$= E(X^2) - E(X)^2$$

Suppose we have a continuous random variable, and we have density plots as follows:



Variance is important when you study random variables from the point of view of measurement.

From this perspective:

A random variable is an inaccurate measure of the mean

90% of statistics relates to this situation – inaccurate measure of the mean.

- As we saw last time:

  If X and Y are <u>independent</u> random variables, then:
  $$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$
  This is not necessarily true if X and Y are <u>not</u> <u>independent</u>

  (dependent variables could have equality by <u>accident</u>)

  (<u>Unlike</u> where w/ <u>expectations (means)</u>, where mean of the sum equals sum of the means — no questions asked.)

- $$\begin{aligned}\underset{\uparrow\, c \text{ a constant}}{\mathrm{Var}(cX)} &= E\left((cX)^2\right) - E(cX)^2 \\ &= c^2 E(X^2) - \left(cE(X)\right)^2 \\ &= c^2 \underbrace{\left(E(X^2) - E(X)^2\right)}_{\mathrm{Var}(X)}\end{aligned}$$

  $$\boxed{\mathrm{Var}(cX) = c^2\,\mathrm{Var}(X)}$$

- That means that the variance is not dimensional the same as the random variable. So we are lead to introduce another quantity that is <u>dimensionaly</u> the <u>same as</u> the <u>random variable.</u>

  $$\sigma(X) = \sqrt{\mathrm{Var}(X)}$$
  <u>Standard Deviation</u>

  $$\begin{aligned}\sigma(cX) &= \sqrt{\mathrm{Var}(cX)} \\ &= \sqrt{c^2\,\mathrm{Var}(X)} \\ &= |c|\sqrt{\mathrm{Var}(X)}\end{aligned}$$

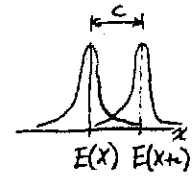
  $$\boxed{\sigma(cX) = |c|\,\sigma(X)}$$

- $\mathrm{Var}(cX) = \overset{\swarrow\text{quadratic}}{c^2}\,\mathrm{Var}(X)$ —vs— $\sigma(cX) = |c|\,\sigma(X)$

  <u>different</u> dimensionality — <u>same</u> dimensionality

- $$\mathrm{Var}(X+c) = E\left((X+c)^2 - E(X+c)^2\right)$$
$$= E\Big(X^2 + 2cX + c^2 - \underbrace{\left(E(X)+E(c)\right)^2}_{\left(E(X)+E(c)\right)^2 = \left(E(X)+c\right)^2 = E(X)^2 + 2cE(X) + c^2}\Big)$$
$$= E\left(X^2 + 2cX + \cancel{c^2} - E(X)^2 - 2cE(X) - \cancel{c^2}\right)$$
$$= E(X^2) + 2c\cancel{E(X)} - E(X)^2 - 2c\cancel{E(X)}$$
$$= E(X^2) - E(X)$$

$$\boxed{\mathrm{Var}(X+c) = \mathrm{Var}(X)}$$



Intuitively, this should be clear.
The variance is the deviation from the mean.
When you add a constant to X, you shift the mean, ← $E(X+c) = E(X)+c$
the deviation around the shifted mean does not change. ← $\mathrm{Var}(X+c) = \mathrm{Var}(X)$

This is a very important fact.
The variance of a random variable does not depend on a position.
It is an intrinsic number associated with the random variable that measures something about the density distribution, something about it's shape.

Now let's compute a couple of variances, just to get a feeling.

- Example – X is Bernoulli w/ probability distribution:

$$P(X=1) = p$$
$$P(X=0) = q = 1-p$$

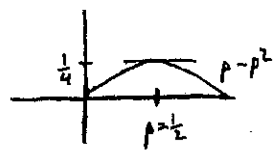
$$\mathrm{Var}(X) = \underbrace{E(X^2)}_{} - \underbrace{E(X)^2}_{}$$

$$\sum_n n\,\underbrace{P(X^2=n)}$$
$$(X^2=n) = (X=\sqrt{n})$$
$$\sum_n n\,\underbrace{P(X=\sqrt{n})}$$
$P(X=0)$ when $n = 0^2$, $(\sqrt{n}=0)$.
$P(X=1)$ when $n = 1^2$, $(\sqrt{n}=1)$.
$$0^2\cdot q + 1^2\cdot p$$
$$p$$

$$\left[\sum_n n\,P(X=n)\right]^2$$
$$(0\cdot q + 1\cdot p)^2$$
$$p^2$$

$$= p - p^2$$
$$= p(1-p)$$
$$\boxed{Var(X) = pq}$$

- Observe that $pq = p - p^2$ is a quadratic equation w/ maximum at $p = \frac{1}{2}$, $1-p = \frac{1}{2} = q$:



$$\left.\begin{array}{l} \frac{d}{dp}(p-p^2) = 1-2p = 0 \\ \frac{d^2}{dp}(p-p^2) = 2 < 0 \end{array}\right\} \ p-p^2 \text{ maximum at } p = \frac{1}{2}$$

Therefore $pq = p - p^2$ has a <u>maximum</u> value of $\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{4}$

$$Var(X) = pq \leq \frac{1}{4}$$

So the <u>maximum variance</u> of the Bernoulli random variable occurs when the bias of the coin $p = \frac{1}{2}$.

This <u>upper bound</u> is important when simplifying computations, as you will see shortly.

- The <u>standard deviation</u> of the <u>Bernoulli random variable</u> is:

$$\sigma(X) = \sqrt{Var(X)}$$
$$= \sqrt{pq}$$
$$\boxed{\sigma(X) = \sqrt{pq} \leq \frac{1}{2}}$$

- Example - Poisson Process

We have the random function $N_\alpha(t)$ of the Poisson process w/ intensity α.
Fix $\alpha + t$ so we have the random variable $N_\alpha(t)$.

We computed the expectation as [4/24/98.5]:

$$E(N_\alpha(t)) = \alpha t$$

Rewrite this as:

$$\underbrace{\frac{\overset{\text{average \# blips}}{E(N_\alpha(t))}}{\underset{\text{per unit length}}{t}}}_{} = \alpha \leftarrow \text{intensity}$$

This is how we build the Poisson process.
We want the average number of blips per unit length to be α.
That's how it comes out

A remarkable fact about the Poisson process is that the variance has the same value as the mean.

$$\boxed{Var(N_\alpha(t)) = \alpha t}$$

It's a gruesome computation,
I want you to suffer.

Fact:

The Poisson Process is the only process where the mean equals the variance.

This is very remarkable.
It characterizes the Poisson Process in a very deep way.

$Var(N_\alpha(t)) = \alpha t$ 5/6/98.5.1

Proof Following outline by Feller, An Introduction to Probability Theory + Applications, Vol 1, 3rd Ed., Wiley, 1967 p. 228.

For convenience, set $X = N_\alpha(t)$ — $\alpha + t$ fixed.

$$Var(X) = \underbrace{E(X^2)}_{} - \underbrace{E(X)^2}_{\text{we already know } E(N_\alpha(t)) = \alpha t. \;\; E(X)^2 = (\alpha t)^2}$$

We now focus on this:

$$E(X^2) = \sum_{n \geq 0} n P(X^2 = n)$$

$$= 0 + \sum_{n \geq 1} n P(X^2 = n)$$

$$= \sum_{n \geq 1} n P(X = \sqrt{n})$$

Let $k = \sqrt{n} \Rightarrow k^2 = n$

$$= \sum_{k \geq 1} k^2 P(X = k) \qquad P(X = k) = P(N_\alpha(t) = k) = \frac{\alpha^k t^k}{k!} e^{-\alpha t}$$

$$= \sum_{k \geq 1} k^2 \frac{\alpha^k t^k}{k!} e^{-\alpha t}$$

$$= e^{-\alpha t} \sum_{k \geq 1} k \frac{\alpha^k t^k}{(k-1)!}$$

$$= e^{-\alpha t} \sum_{k \geq 1} ((k-1) + 1) \frac{\alpha^k t^k}{(k-1)!}$$

$$= e^{-\alpha t} \sum_{k \geq 1} (k-1) \frac{\alpha^k t^k}{(k-1)!} + e^{-\alpha t} \sum_{k \geq 1} \frac{\alpha^k t^k}{(k-1)!}$$

$$= \cancel{e^{-\alpha t}} (\alpha t)^2 \underbrace{\sum_{k \geq 1} \frac{(\alpha t)^{k-2}}{(k-2)!}}_{\text{Maclaurin Series for } e^{\alpha t}} + \cancel{e^{-\alpha t}} (\alpha t) \underbrace{\sum_{k \geq 1} \frac{(\alpha t)^{k-1}}{(k-1)!}}_{\text{Maclaurin series for } e^{\alpha t}}$$

So we have that:

$$Var(N_\alpha(t)) = \cancel{(\alpha t)^2} + (\alpha t) - \cancel{(\alpha t)^2}$$

$$Var(N_\alpha(t)) = \alpha t \quad ✓$$

Q.E.D. Again, it is a remarkable fact that $E(N_\alpha(t)) = Var(N_\alpha(t)) = \alpha t$

- Example – Exponential Random Variable

T exponential, i.e.,

$$P(T > t) = e^{-\alpha t}$$

T is the memoryless waiting time for the first blip in the Poisson Process.

We already computed the expectation [4/22/98.7-8]:

$$E(T) = \frac{1}{\alpha}$$ } The greater the intensity of the Poisson Process, the smaller the waiting time for the first blip.

Now let's compute the variance:

$$Var(T) = \underbrace{E(T^2)}_{} - \underbrace{E(T)^2}_{\frac{1}{\alpha^2}}$$

$$\left(E(T^2) = \int_{-\infty}^{\infty} s \, dens(T^2 = s)\, ds \right.$$

$$(T^2 = s) = (T = \sqrt{s})$$

Let $t = \sqrt{s} \Rightarrow t^2 = s$

$$\left. = \int_0^\infty t^2 \, dens(T = t)\, dt \right)$$

$$= \int_0^\infty t^2 \underbrace{dens(T = t)}\, dt - \frac{1}{\alpha^2}$$

{ We've already shown [4/22/98.6]:
$dens(T = t) = \alpha e^{-\alpha t}$ }

$$= \underbrace{\int_0^\infty t^2 \alpha e^{-\alpha t}\,dt}_{} - \frac{1}{\alpha^2}$$

We need to evaluate this integral. We can obtain this in a clever way. Recall the integral for the expectation of T:

$$E(T) = \int_{-\infty}^{\infty} t \, \underbrace{dens(T=t)}_{\Downarrow} \, dt$$

$$= \int_0^\infty t \, \overbrace{\alpha e^{-\alpha t}} \, dt = \frac{1}{\alpha}$$

From this, multiplying both sides by $\frac{1}{\alpha}$, we obtain:

$$\int_0^\infty t e^{-\alpha t}\,dt = \frac{1}{\alpha^2}$$

Differentiate both sides with respect to α:

$$\frac{\partial}{\partial\alpha}\int_0^\infty t e^{-\alpha t}\,dt = \frac{\partial}{\partial\alpha}\left(\frac{1}{\alpha^2}\right)$$

$$\int_0^\infty t \frac{\partial}{\partial\alpha} e^{-\alpha t}\,dt = -2\left(\frac{1}{\alpha^3}\right)$$

$$-\int_0^\infty t^2 e^{-\alpha t}\,dt = -\frac{2}{\alpha^3}$$

Multiply both sides by $(-\alpha)$:

$$\int_0^\infty t^2 \alpha e^{-\alpha t}\,dt = \frac{2}{\alpha^2}$$

This is precisely the integral we were after. Pretty Neat.

$$= \frac{2}{\alpha^2} - \frac{1}{\alpha^2}$$

$$\boxed{Var(T) = \frac{1}{\alpha^2}}$$

- Suppose we have random variable X with a given expectation:

$$E(X) = m$$

Let's define a new random variable:

$$Y = aX + b$$

$a + b$ are constants

What are the mean and variance of Y?

- $$E(Y) = E(aX + b)$$
$$= E(aX) + E(b)$$
$$\boxed{E(Y) = aE(X) + b}$$

- $$Var(Y) = Var(aX + b)$$
$$= Var(aX) \quad \text{since } Var(Z+c) = Var(Z)$$
$$= a^2 Var(X) \quad \text{since } Var(cZ) = c^2 Var(Z)$$
$$\boxed{Var(Y) = a^2 Var(X)}$$

- Therefore, if $\begin{Bmatrix} E(X) = m \\ Var(X) = \sigma^2 \end{Bmatrix}$ then the random variable

$\frac{X-m}{\sigma}$ has mean 0 and variance 1.

Such a random variable is said to be <u>normalized</u>. {the term <u>standardized</u> is also used.}

<u>Proof</u>
$$E\left(\frac{X-m}{\sigma}\right) = E\left(\frac{1}{\sigma}X - \frac{m}{\sigma}\right)$$
$$= E\left(\frac{1}{\sigma}X\right) - E\left(\frac{m}{\sigma}\right)$$
$$= \frac{1}{\sigma}\overset{m}{\cancel{E(X)}} - \frac{m}{\sigma}$$
$$= 0 \checkmark$$

$$
\begin{aligned}
Var\left(\frac{X-m}{\sigma}\right) &= Var\left(\frac{1}{\sigma}X - \frac{m}{\sigma}\right) \\
&= Var\left(\frac{1}{\sigma}X\right) \\
&= \frac{1}{\sigma^2}\cancelto{\sigma^2}{Var(X)} \\
&= 1 \quad \checkmark
\end{aligned}
$$

- Another application of the same Mickey Mouse principle is the following:

(*) If Y is a normalized random variable, i.e., $E(Y)=0$ and $Var(Y)=1$, then the random variable $X=\sigma Y+m$ has:

$$E(X)=m \quad \text{and}$$
$$Var(X)=\sigma^2$$

<u>Proof</u>

$$
\begin{aligned}
E(X) &= E(\sigma Y+m) \\
&= E(\sigma Y)+E(m) \\
&= \sigma \cancelto{0}{E(Y)} + m \\
&= m \quad \checkmark
\end{aligned}
$$

$$
\begin{aligned}
Var(X) &= Var(\sigma Y+m) \\
&= Var(\sigma Y) \\
&= \sigma^2 \cancelto{1}{Var(Y)} \\
&= \sigma^2 \quad \checkmark
\end{aligned}
$$

i.e., Z has a standard normal density distribution

- Recall a random variable Z is <u>standard normal</u> whenever the <u>density</u> is:

$$dens(Z=t)=\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}} \qquad \left\{\text{See } [5/4/98.4\text{-}5]\right\}$$

- Let's perform the preceeding computations with standard normal random variables.

By the immediately discussion (*), it follows that:

Let Z = standard normal variable

$$\left.\begin{array}{l} E(Z) = 0 \\ Var(Z) = 1 \end{array}\right\}$$ Properties of normalized random variables.

Let's check that Z satisfies these properties:

- $$E(Z) = \int_{-\infty}^{\infty} z \; dens(Z = z) \, dz$$

$$= \int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \, dz$$

Let $u = -\frac{z^2}{2}$

$du = -z \, dz$

$$= \frac{-1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{u} \, du$$

$$= \frac{-1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \Big|_{-\infty}^{\infty}$$

$$= 0 - 0$$

$E(Z) = 0$ ✓

Standard normal variable has mean 0.

- $$Var(Z) = E(Z^2) - \cancel{E(Z)^2}^{\,0}$$

$$= \int_{-\infty}^{\infty} s \; dens(Z^2 = s) \, ds$$

Let $z = \sqrt{s} \Rightarrow z^2 = s$

$$= \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \, dz$$

$$= \frac{1}{\sqrt{2\pi}} \underbrace{\int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} \, dz}$$

{we obtain this integral by differentiating a known integral.}

To start with, we have shown that:

$$E(Z) = \int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0$$

Thus, we have the following equality:

$$\int_{-\infty}^{\infty} z e^{-\frac{z^2}{2}} dz = 0$$

Differentiating both sides, we get:

$$\frac{d}{dz} \int_{-\infty}^{\infty} z e^{-\frac{z^2}{2}} dz = \frac{d}{dz} 0$$

$$\int_{-\infty}^{\infty} \left(\frac{d}{dz} z e^{-\frac{z^2}{2}}\right) dz = 0$$

$$\int_{-\infty}^{\infty} \left[z\left(-z e^{-\frac{z^2}{2}}\right) + e^{-\frac{z^2}{2}}\right] dz = 0$$

$$\underbrace{\int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} dz} = \underbrace{\int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz}$$

this is the integral we are trying to obtain,

Recall that $\text{dens}(Z=z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

And since we know that, by the definition of density:

$$\int_{-\infty}^{\infty} \text{dens}(Z=z)\, dz = 1$$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 1$$

Which readily provides that:

$$\int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz = \sqrt{2\pi}$$

And thus, we have that:

$$\int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} dz = \sqrt{2\pi}$$

Pretty <u>slick</u>!

So,

$$Var(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz$$

$$= \frac{1}{\sqrt{2\pi}} \sqrt{2\pi}$$

$$Var(Z) = 1 \checkmark$$

So it is indeed the fact that:

$$Z = \text{a standard normal random variable} \Rightarrow \begin{cases} E(Z) = 0 \\ Var(Z) = 1 \end{cases}$$

properties of normalized random variables.

Further, from (*), we also know that:

- random variable $X = \sigma Z + m$ has

$$E(X) = m$$
$$Var(X) = \sigma^2$$

- We want to go a bit farther than that.
We want the density distribution of $\sigma Z + m$.

$$dens(X = t) = ?$$
$$X = \sigma Z + m$$

We spent a whole week on the Algebra of Probability Distributions.
Let's apply it now.
Our chickens are coming to roost.
You say: Well, we know how to do that.
Let's do it.

$$P(X \le t) = P(\sigma Z + m \le t)$$

event $(\sigma Z + m \le t) = \left(Z \le \frac{t-m}{\sigma}\right)$

observe that σ is always positive
$\sigma > 0$

$$= P\left(Z \le \frac{t-m}{\sigma}\right)$$

$$= \int_{-\infty}^{\frac{t-m}{\sigma}} \underbrace{dens(Z = s)}_{\frac{1}{\sqrt{2\pi}} e^{-s^2/2}} ds$$

The cumulative probability distribution is the integral of the density distribution.

$$P(X \le t) = \int_{-\infty}^{\frac{t-m}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-s^2/2}\, ds$$

Hence:

$$\begin{aligned} dens(X=t) &= \frac{d}{dt} P(X \le t) \\ &= \frac{d}{dt} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{t-m}{\sigma}} e^{-s^2/2}\, ds \\ &= \frac{1}{\sqrt{2\pi}} \underbrace{\frac{d}{dt} \int_{-\infty}^{\frac{t-m}{\sigma}} e^{-s^2/2}\, ds} \end{aligned}$$

Recall the 2nd Fundamental Theorem of Calculus:

$$\frac{d}{du}\left(\int_a^u f(s)\, ds\right) = f(u)$$

We evaluate the above by the <u>chain rule</u>.

Let $u = \frac{t-m}{\sigma}$

$$g(u) = \int_{-\infty}^{u} e^{-s^2/2}\, ds$$

Then:

$$\frac{d}{dt} g\left(\frac{t-m}{\sigma}\right) = \frac{d}{dt} g(u) = \frac{dg(u)}{du} \frac{du}{dt}$$

$$\frac{d}{dt} \int_{-\infty}^{\frac{t-m}{\sigma}} e^{-s^2/2}\, ds = \underbrace{\left(\frac{d}{du} \int_{-\infty}^{u} e^{-s^2/2} ds\right)}_{e^{-u^2/2}} \cdot \frac{1}{\sigma}$$

$$= e^{-\left(\frac{t-m}{\sigma}\right)^2/2} \frac{1}{\sigma}$$

$$\boxed{dens(X=t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-m)^2/2\sigma^2}}$$

As this is a density, automatically it integrates to 1. We don't have to check that.

$$\int_{-\infty}^{\infty} dens(X=t)\, dt = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-m)^2/2\sigma^2}\, dt = 1$$

Because its the <u>density</u> of a random variable, the integral has to equal 1. It's stupid to check that.

$$\text{dens}(X=t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-m)^2/2\sigma^2}$$ is called the <u>Normal Distribution</u> of <u>mean m</u> and <u>variance σ^2</u>

- <u>Conclusion</u>

If random variable Z is standard normal, then $\overbrace{\sigma Z + m}^{\text{random variable}}$ has the density:

$$\text{dens}(\sigma Z + m = t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-m)^2/2\sigma^2}$$

In statistics books, this is abbreviated as saying:

$$\sigma Z + m \text{ is } \underbrace{\mathcal{N}(m, \sigma^2)}$$

normally distributed w/ mean m and variance σ^2

- And, of course, if you have a random variable X, where:

$$X \text{ is } \mathcal{N}(m, \sigma^2)$$

then:

$$\frac{X-m}{\sigma} \text{ is standard normal}$$

- $$X = \sigma Z + m$$

$$\begin{array}{cc} X & Z \\ \updownarrow & \updownarrow \\ E(X) = m & E(Z) = 0 \\ Var(X) = \sigma^2 & Var(Z) = 1 \\ \mathcal{N}(m, \sigma^2) & \mathcal{N}(0,1) \end{array}$$

- If $X_1, X_2, \ldots, X_n$ are i.i.d. of <u>finite variance</u> with

$$E(X_1) = m$$
$$Var(X_1) = \sigma^2 < \infty$$

since X's are <u>independent</u>:

then $$Var\Big(\underbrace{\frac{X_1 + X_2 + \cdots + X_n}{n}}_{\text{average}}\Big) = Var\left(\tfrac{1}{n}X_1\right) + Var\left(\tfrac{1}{n}X_2\right) + \cdots + Var\left(\tfrac{1}{n}X_n\right)$$

$$= \frac{1}{n^2}Var(X_1) + \frac{1}{n^2}Var(X_2) + \cdots + \frac{1}{n^2}Var(X_n)$$

$$Var\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{\sigma^2}{n}$$

$$Var(cX) = c^2 Var(X)$$

- What does that mean?

It's a very important result.

It means that if you take n measurements of the same quantity and average them, then the variance of the average:

$$Var\left(\frac{X_1+X_2+\dots+X_n}{n}\right) = \frac{\sigma^2}{n}$$

goes down as n increases.

and the standard deviation of the average:

$$\sigma\left(\frac{X_1+X_2+\dots+X_n}{n}\right) = \frac{\sigma}{\sqrt{n}}$$

goes down as n increases.

In other words, it confirms our intuitive feeling that taking several measurements of the same quantity makes the measurement of the expectation more accurate. Namely, the variance of the expectation goes down.

$$Var\left(\frac{X_1+X_2+\dots+X_n}{n}\right) = \frac{\sigma^2}{n}$$

Provided that the variance is finite.

- Statistics in One Easy Lesson

90% of statistics is studying independent random variables, with finite variance.

The main assumption of statistics is this:

When you make repeated measurements, you might as well assume that the random variables you are measuring are normally distributed.

When you make several measurements of the same quantity, close your eyes and assume that the random variables are normally distributed.

In other words, suppose you have random variables:

$$X_1, X_2, \dots, X_n \quad \leftarrow \text{independent random variables}$$

You take the mean m, and the variance σ^2 and you forget everything else.

Then, you replace the X's by:

$$X_1 = \sigma Z + m$$

standard normal random variable

That's the main assumption of statistics.

In making measurements of a quantity X with $E(X) = m$
$Var(X) = \sigma^2 < \infty$

<u>assume</u> that $\underbrace{X \text{ is } \mathcal{N}(m, \sigma^2)}$

<u>Assume</u> that X is <u>normally distributed</u>
w/ mean m and variance σ^2.

- NB: This <u>assumption</u> has <u>never</u> been <u>fully justified.</u>

Engineers, etc. say who cares? It works.
There are <u>partial</u> justifications, as we will see, by the:

1) Central Limit Theorem
2) Various derivations of the Normal Law, axiomatic principles.

None of them provides a complete justification.
To this day, there is a question how to completely, rigorously justify this assumption.

<u>We will assume it.</u>

Why do we make this assumption?
How does it work?

- Suppose we make n measurements of a quantity

$X_1, X_2, \ldots, X_n \longleftarrow$ {We assume $X_1, X_2, \ldots, X_n$ are <u>independent</u> random variables. And that each is $\mathcal{N}(m, \sigma^2)$}

Suppose the variance is <u>known</u> and is <u>finite</u>:

$Var(X_1) = \sigma^2 < \infty \longleftarrow$ {if it's <u>not known</u>, we can limit it by an <u>upper bound</u>, like we did w/ the Bernoulli random variable [5/6/98.4]}

You want to know what the mean of these random variables is.

Want to <u>estimate</u> $m = E(X_1) = E(X_2) = \ldots = E(X_n)$

You don't know the mean.

What do we do?
Use Bayes' Law.

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}$$

We have also a version of Bayes' Law for densities [4/10/98.10]:

$$\underbrace{\text{dens}(m=s \mid X_1=t_1, X_2=t_2, \ldots, X_n=t_n)}_{\text{posterior}} = \frac{\overbrace{\text{dens}(X_1=t_1, X_2=t_2, \ldots, X_n=t_n \mid m=s)}^{\text{likelihood}}\ \overbrace{\text{dens}(m=s)}^{\text{prior}}}{\text{dens}(X_1=t_1, X_2=t_2, \ldots, X_n=t_n)}$$

Remember, the denominator doesn't matter, because if you don't know it, you obtain it by integrating the numerator relative to s.

$$\int_{-\infty}^{\infty} \text{dens}(X_1=t_1, X_2=t_2, \ldots, X_n=t_n \mid m=s)\,\text{dens}(m=s)\,ds = \text{dens}(X_1=t_1, X_2=t_2, \ldots, X_n=t_n)$$

How does Bayes' Law for densities help us?

Taking n measurements of the same quantity means:

event $X_1 = t_1 \leftarrow$ measurement 1 gives t_1.

$X_2 = t_2 \leftarrow$ measurement 2 gives t_2.

$\vdots$

$X_n = t_n \leftarrow$ measurement n gives t_n.

the X's are assumed normal, with mean m and variance σ^2.

What does Bayes' Law do for you?
Bayes' Law says:

You don't know the mean m, so the mean is a random variable.
The density of the mean, given these n measurements

$$\text{dens}(m=s \mid X_1=t_1, X_2=t_2, \ldots, X_n=t_n)$$

will be equal to the density of these n measurements, given the mean, times the prior times a constant.

$$\text{dens}(m=s \mid X_1=t_1, X_2=t_2, \ldots, X_n=t_n) \propto \text{dens}(X_1=t_1, X_2=t_2, \ldots, X_n=t_n \mid m=s)\ (\text{Prior})$$

{So, if we establish a prior – namely, some preliminary guess of what the mean we are measuring is going to be like, then the likelihood is computed very simply.}

- Likelihood

$$\text{dens}(\underbrace{X_1=t_1, X_2=t_2, \ldots, X_n=t_n}_{\text{these random variables are independent. So their densities multiply.}} \mid \underbrace{m=s}_{\text{means that the mean } m \text{ is fixed to } s \text{ from the start.}})$$

$$\text{dens}(X_1=t_1, X_2=t_2, \ldots, X_n=t_n \mid m=s) = \text{dens}(X_1=t_1 \mid m=s)\,\text{dens}(X_2=t_2 \mid m=s) \ldots \text{dens}(X_n=t_n \mid m=s)$$

likelihood

Consider $\text{dens}(X_i=t_i \mid m=s)$.
It is given that mean $m=s$.
As discussed, each of the X_i is assumed to be normal with variance σ^2 and we are given mean $m=s$.
So each X_i is $\mathcal{N}(s, \sigma^2)$.
Thus:

$$\text{dens}(X_i=t_i \mid m=s) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(t_i-s)^2/2\sigma^2}$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-(t_1-s)^2/2\sigma^2}\right)\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-(t_2-s)^2/2\sigma^2}\right) \ldots \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-(t_n-s)^2/2\sigma^2}\right)$$

So the likelihood comes out real easy, because they are independent random variables.
So you plug that in to Bayes' Law for densities.

Now comes the 2nd big assumption of statistics. ← {We won't finish this today, so we'll just set it up.}
Unjustified, except that it works.
I wish one of you would work this out. Make me happy.
Justify it.
Write me a letter. Send me a reprint.

With the above as likelihood, Bayes' Law for densities becomes:

$$\text{dens}(m=s \mid X_1=t_1, X_2=t_2, \ldots, X_n=t_n) \propto \frac{1}{\sigma^n(\sqrt{2\pi})^n} e^{-((t_1-s)^2+(t_2-s)^2+\cdots+(t_n-s)^2)/2\sigma^2} \overbrace{\text{dens}(m=s)}^{\text{the prior}}$$

This looks like a mess.
But it isn't.
We've said 100 times that every density has to integrate to 1.
Now, we're going to see the real culmination of this fact.

Let's focus on the exponent of e:

$$(t_1-s)^2 + (t_2-s)^2 + \cdots + (t_n-s)^2 = ns^2 - 2s\underbrace{(t_1+t_2+\cdots+t_n)}_{\text{multiply by } 1 = \frac{n}{n}} + (t_1^2+t_2^2+\cdots+t_n^2)$$

$$\left\{ (t_i-s)^2 = t_i^2 - 2st_i + s^2 \right\}$$

$$= ns^2 - n2s\left(\frac{t_1+t_2+\cdots+t_n}{n}\right) + (t_1^2+t_2^2+\cdots+t_n^2)$$

$$\frac{t_1+t_2+\cdots+t_n}{n} = \overline{m}$$ The sample mean

We take n measurements, and we average those n measurements.

$$\frac{t_1 + t_2 + \dots + t_n}{n} = \text{the } \underline{\text{sample}} \text{ mean } \bar{m}$$

The average of the actual numbers we observed.
They are just numbers.
We averaged them.

$$= ns^2 - n2s\bar{m} + (t_1^2 + t_2^2 + \dots + t_n^2)$$

How do we get rid of these t_i?

$$= \underbrace{n(s-\bar{m})^2}_{ns^2 - n2s\bar{m} + n\bar{m}^2} + \underline{\text{crud}}$$

how do we get rid of the crud?
I'll tell you next time.

If you look carefully, you'll realize that it can be removed wonderfully, from the fact that densities must integrate to 1.
We'll see this in detail.

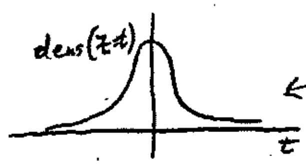
After we do this, we have the statistics we want.
And you can go to town w/ statistics.

John Guidi
guidi@mit.edu

The Normal Distribution (cont'd)

Let's continue our study of the Normal Distribution.
Last time, we noted that [5/4/98.4-5]:

if Z = a standard normal random variable,
then:
$$dens(Z=t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$



this density is the <u>bell-shaped curve</u>

Let me recite a few facts about this density.
In the old days, people carried around tables.
Today, we do these computations on computer.

$$P(-a < Z \le b) = P(Z \le b) - \underbrace{P(Z \le -a)}$$

since the density is symmetric about $x=0$,
$$P(Z \le -a) = P(Z \le \infty) - P(Z \le a)$$
$$= 1 - P(Z \le a)$$

$$= \int_{-\infty}^{b} dens(Z=t)\,dt - 1 + \int_{-\infty}^{a} dens(Z=t)\,dt$$

$$P(-a < Z \le b) = \int_{-\infty}^{b} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt + \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt - 1$$

These integrals are readily available.
(for example, text p. 177)

Example:
$$P(-1 < Z \le 1) = \underbrace{\int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt}_{0.8413} + \underbrace{\int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt}_{0.8413} - 1$$
$$= 0.6826.$$

$$P(-1 < Z \le 1) = 0.6826$$
$$P(-1.96 < Z \le 1.96) = 0.95 \longleftarrow \text{significant}$$
$$P(-2 < Z \le 2) = 0.9545$$
$$P(-2.58 < Z \le 2.58) = 0.99 \longleftarrow \text{very significant}$$

These numbers have <u>names</u> in statistics.
These are <u>technical terms</u>.

When you say standard normal random variable Z is significant, it means:

$$P(-1.96 < Z \le 1.96) = 0.95$$

By common agreement.

- We also saw last time:

The random variable $\sigma Z + m$ has the density:

$$\text{dens}(\sigma Z + m = t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-m)^2/2\sigma^2}$$

Random variable $\sigma Z + m$ has <u>mean m</u> and <u>standard deviation σ</u> (variance $= \sigma^2$)

- <u>Basic Rule of Statistics</u>

70% of all statistical computations use the following:

Given a "tough" random variable X with $E(X) = m$ and $Var(X) = \sigma^2$, <u>replace</u> it by $\underbrace{\sigma Z + m}$

In other words, replace X by the <u>simplest possible</u> random variable that has the <u>same expectation</u> and the <u>same variance</u>.
Close your eyes and hope for the best.
And it works.

I told you last time, <u>nobody</u> really knows, ultimately, why this works.
Rivers of words have been written — by philosophers, physicists, mathematicians, crackpots.
The great French mathematician Henri Poincaré once wrote that:

"The explanation of the normal distribution is something that mathematicians consider to be the physicists' task. And physicists consider to be the mathematicians' task."

So the ball bounces back and forth.

- Last time, we began to apply this basic rule, in a problem of estimation of measurement.

We started w/ Bayes' Law for densities:

$$\text{dens}(m=s \mid X_1=t_1, X_2=t_2, \dots, X_n=t_n) = \frac{\text{dens}(X_1=t_1, X_2=t_2, \dots, X_n=t_n \mid m=s)\ \text{dens}(m=s)}{\text{dens}(X_1=t_1, X_2=t_2, \dots, X_n=t_n)}$$

prior

likelihood

posterior

dens (m=s given n <u>independent</u> measurements)

This is what a lot of people make their living on.
In very sophisticated ways, to be sure.

Last time, we said "Let's take Bayes' Law and apply the Basic Rule of Statistics"
Let's take n measurements and apply the Basic Rule of Statistics.

When we apply the Basic Rule of Statistics, we will assume that:

$X_1, X_2, \ldots, X_n$ are independent normally distributed random variables, whose mean and variance we do not know.

We assume the variance ($\sigma^2 < \infty$) is finite and, therefore, we take an upper bound for the variance that is safe.
And we plug that in.

The problem is to estimate the mean of the random variable on the basis of the data given by n measurements.

$$E\left(\frac{X_1 + \cdots + X_n}{n}\right) = ?$$

- We saw last time [5/6/98.14-15] that if we take the average of n independent, identically distributed random variables, then the variance goes down by $\frac{1}{n}$.

$$\mathrm{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{\sigma^2}{n}$$

So, you get a much more accurate estimate of the average with larger n, provided that the variance is finite.

- Now, let's continue, from last time, our evaluation of Bayes' Law for densities [5/6/98.18]:

If $X_1, X_2, \ldots, X_n$ are i.i.d., normally distributed, then:

$$\mathrm{dens}(m=s \mid X_1=t_1, X_2=t_2, \ldots, X_n=t_n) \propto \text{[redacted]}$$

$$\underbrace{\frac{1}{\sigma^n(\sqrt{2\pi})^n}}_{\text{for fixed } n \text{, this is a constant } k_1.} e^{-\left((t_1-s)^2+(t_2-s)^2+\cdots+(t_n-s)^2\right)/2\sigma^2} \mathrm{dens}(m=s)$$

$$\propto e^{-\left((t_1-s)^2+(t_2-s)^2+\cdots+(t_n-s)^2\right)/2\sigma^2} \mathrm{dens}(m=s)$$

Something cute comes up when simplifying this expression.

Call $\overline{m} = \dfrac{t_1+t_2+\dots+t_n}{n}$

↑ the sample mean

In other words, $t_1, t_2, \dots, t_n$ are the actual measurements we obtained in our experiment. Numbers. No nonsense.
We take the average of these numbers.
This should give an approximation to what we want to compute.
Let's see how.

Let's focus on the exponent:

$$(t_1-s)^2+(t_2-s)^2+\dots+(t_n-s)^2 = ns^2 - \underbrace{2s\left(t_1+t_2+\dots+t_n\right)}_{\text{multiply by } 1=\frac{n}{n}} + \left(t_1^2+t_2^2+\dots+t_n^2\right)$$

$$= ns^2 - n2s\left(\frac{t_1+t_2+\dots+t_n}{n}\right) + \left(t_1^2+t_2^2+\dots+t_n^2\right)$$

$$= ns^2 - n2s\overline{m} + \left(t_1^2+t_2^2+\dots+t_n^2\right)$$

Add $0 = n\overline{m} - n\overline{m}$

$$= \underbrace{ns^2 - n2s\overline{m} + n\overline{m}}_{n(\overline{m}-s)^2} \underbrace{- n\overline{m} + \left(t_1^2+t_2^2+\dots+t_n^2\right)}_{\text{crud}}$$

crud
Contains no s.
Call this constant K_2.

$$(t_1-s)^2+(t_2-s)^2+\dots+(t_n-s)^2 = n(\overline{m}-s)^2 + K_2$$

So:

$$dens(m=s / X_1=t_1, X_2=t_2, \dots, X_n=t_n) \propto \underbrace{e^{-(n(\overline{m}-s)^2+K_2)/2\sigma^2}}_{\underbrace{e^{-K_2/2\sigma^2}}_{\text{this is just a constant } K_3} e^{-n(\overline{m}-s)^2/2\sigma^2}} dens(m=s)$$

$$\propto K_3 e^{-n(\overline{m}-s)^2/2\sigma^2} dens(m=s)$$

$$\propto e^{-n(\overline{m}-s)^2/2\sigma^2} dens(m=s)$$

Let K = constant that incorporates constants K_1, K_2, K_3.

Then Bayes' Law for densities is:

$$dens(m=s \mid X_1=t_1, X_2=t_2, \ldots, X_n=t_n) = \frac{\overbrace{k\, e^{-(\bar{m}-s)^2/2\frac{\sigma^2}{n}}}^{dens(X_1=t_1, X_2=t_2, \ldots, X_n=t_n \mid m=s)}\; dens(m=s)}{dens(X_1=t_1, X_2=t_2, \ldots, X_n=t_n)}$$

Rewrite the denominator using the continuous Law of Alternatives consequence [4/8/98.7]:

$$= \frac{k\, e^{-(\bar{m}-s)^2/2\frac{\sigma^2}{n}}\; dens(m=s)}{\int_{-\infty}^{\infty} \underbrace{dens(X_1=t_1, X_2=t_2, \ldots, X_n=t_n \mid m=s)}\; dens(m=s)\, ds}$$

(*) $$dens(m=s \mid X_1=t_1, X_2=t_2, \ldots, X_n=t_n) = \frac{k\, e^{-(\bar{m}-s)^2/2\frac{\sigma^2}{n}} \cdot \overbrace{dens(m=s)}^{prior}}{\int_{-\infty}^{\infty} k\, e^{-(\bar{m}-s)^2/2\frac{\sigma^2}{n}}\; dens(m=s)\, ds}$$

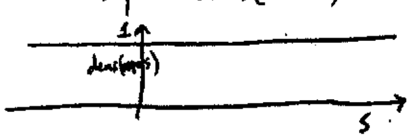
- Now you say "What shall we choose as prior?"
  There is an honest choice and a dishonest choice.

1) The honest choice

   Choose a normal distribution w/ a certain variance, which you guess.
   That can be done.
   It's a hairy computation, but it can be done.

2) The dishonest choice

   Choose as prior $dens(m=s) = 1$ for all s.



   Note: This is not a probability density.

   $$\int_{-\infty}^{\infty} dens(m=s)\, ds \neq 1$$

   Who cares? It works anyway.
   That's what they do

   This is the 2nd big assumption in statistics.

No one has ever justified this.
I tell you, because maybe you will.
It's good for you to know the truth.
Even though this irritates statisticians,
Don't ever tell them you choose as prior a <u>measure</u> that is <u>not</u> a <u>density</u>.

They choose the <u>uniform measure of the line.</u>
Which means you pick a point on the line at random, but it's <u>not</u> a <u>probability</u>.

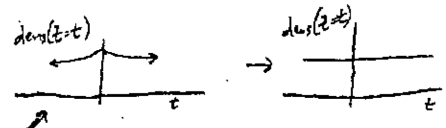
However, they say, if you choose $dens(m=s)=1$, then equation $(*)$ is still convergent, because the RHS has a negative exponential.
So it still makes sense.

Let's see what happens.
Close your eyes and make the 2^{nd} big assumption:

Choose as prior, not a probability density, but a <u>measure</u>:

$$dens(m=s) = 1, \quad \text{for all } s$$

If really pushed against the wall, they will say:



Replace this by a <u>normal distribution</u> w/ <u>very high variance</u> and that is <u>almost uniform.</u>
So why not <u>go to</u> the <u>limit</u> (even though the limit doesn't make sense)?

Then, equation $(*)$ becomes:

$$dens(m=s \mid X_1=t_1, X_2=t_2, \ldots, X_n=t_n) = K' e^{-(\bar{m}-s)^2/2\frac{\sigma^2}{n}}$$

incorporates constant K and integral in the denominator w/ $dens(m=s)=1$.

Now comes the beauty of this.
We can immediately solve for constant K', even though it comes from a mess.
Why?
Because the above is a <u>density</u> distribution, therefore the whole thing has to integrate to 1.

$$\int_{-\infty}^{\infty} dens(m=s \mid X_1=t_1, X_2=t_2, \ldots, X_n=t_n)\,ds = \int_{-\infty}^{\infty} K' e^{-(\bar{m}-s)^2/2\frac{\sigma^2}{n}}\,ds = 1$$

$$\Rightarrow K' = \frac{1}{\int_{-\infty}^{\infty} e^{-(\bar{m}-s)^2/2\frac{\sigma^2}{n}}\,ds}$$

So we know K'.

- What's this telling us?
It's telling us that, with a uniform prior (i.e., $dens(m=s) = 1$, for all s), then the posterior is <u>normally distributed</u> with mean $\overline{m}$ (the <u>sample</u> mean) and variance $\frac{\sigma^2}{n}$

$$\underset{\text{posterior}}{dens(m=s|X_1=t_1, X_2=t_2, \dots, X_n=t_n)} = k' e^{-(\hat{\overline{m}}-s)^2/2\frac{\sigma^2}{n}}$$

Posterior normally distributed with mean $\overline{m}$ and variance $\frac{\sigma^2}{n}$

And that's an interesting conclusion, even though obtained <u>not rigorously</u>.
Conclusion:
If you take n measurements,
and we take a prior, which in <u>some intuitive sense</u> means <u>no knowledge whatsoever</u>
then the <u>posterior</u> tells us:

the mean is a random variable that is normally distributed w/ variance $\frac{\sigma^2}{n}$

This really confirms of our previous computation of the variance of the mean of n independent random variables [5/6/98.14-15]:

$$Var\left(\frac{X_1+X_2+\dots+X_n}{n}\right) = \frac{\sigma^2}{n}$$

- The formula:

$$dens(m=s|X_1=t_1, X_2=t_2, \dots, X_n=t_n) = k' e^{-(\overline{m}-s)^2/2\frac{\sigma^2}{n}}$$

is the <u>basic formula</u> <u>of statistics.</u>

In other words, if you have <u>n measurements</u>, and you want to make a <u>prediction</u> <u>about their average</u>, this is the formula you use.

i.e., m is $\mathcal{N}(\overline{m}, \frac{\sigma^2}{n})$

- How do you use the formula to make a prediction about the average?
Like this:

Since m is normally distributed with $E(m) = \overline{m}$, $Var(m) = \frac{\sigma^2}{n}$

it follows that:

$$Z = \frac{m-\overline{m}}{\frac{\sigma}{\sqrt{n}}}$$ is <u>standard normal</u>.

see [5/6/98.14]

$$E(Z) = E\left(\frac{m-\bar{m}}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= E\left(\frac{m}{\frac{\sigma}{\sqrt{n}}}\right) - E\left(\frac{\bar{m}}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= \frac{1}{\frac{\sigma}{\sqrt{n}}} \underbrace{E(m)}_{\substack{\text{given that} \\ E(m)=\bar{m}}} - \frac{\bar{m}}{\frac{\sigma}{\sqrt{n}}}$$

$$= \frac{1}{\frac{\sigma}{\sqrt{n}}} \bar{m} - \frac{\bar{m}}{\frac{\sigma}{\sqrt{n}}}$$

$$E(Z) = 0 \checkmark$$

$$Var(Z) = Var\left(\frac{m-\bar{m}}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= Var\left(\frac{m}{\frac{\sigma}{\sqrt{n}}} - \frac{\bar{m}}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= Var\left(\frac{m}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= \frac{1}{\frac{\sigma^2}{n}} \underbrace{Var(m)}_{\substack{\text{given that} \\ Var(m) = \frac{\sigma^2}{n}}}$$

$$= \frac{1}{\frac{\sigma^2}{n}} \cdot \frac{\sigma^2}{n}$$

$$Var(Z) = 1 \checkmark$$

$$\therefore \quad Z = \frac{m-\bar{m}}{\frac{\sigma}{\sqrt{n}}} \text{ is } \underline{\text{standard normal.}}$$

Hence:

$$P(a < m \leq b) = \int_a^b \underbrace{dens(m=s)}\, ds$$

where m is $\mathcal{N}(\bar{m}, \frac{\sigma^2}{n})$, we have (see [5/6/98.13]):

$$= \int_a^b \frac{1}{\frac{\sigma}{\sqrt{n}}\sqrt{2\pi}}\, e^{-(s-\bar{m})^2/2\frac{\sigma^2}{n}}\, ds$$

Better is to write the probability in terms of the <u>standard normal</u> random variable. It's easier to use this, because <u>the tables</u> available provide the integral of the standard normal density distribution.

$$\overbrace{P\left(a < \frac{m-\bar{m}}{\frac{\sigma}{\sqrt{n}}} \leq b\right)}^{P(a < Z \leq b)} = \int_a^b \underbrace{dens(Z=s)}_{\text{where } Z \text{ is } \mathcal{N}(0,1)}\, ds$$

$$= \int_a^b \frac{1}{\sqrt{2\pi}}\, e^{-s^2/2}\, ds$$

$$(*) \quad \boxed{P\left(a < \frac{m-\bar{m}}{\frac{\sigma}{\sqrt{n}}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-s^2/2}\, ds}$$

← this is the integral you find in tables,
So it's always better to normalize.

Example

I toss a coin 100 times and obtain 41 heads.
Is the coin fair?

We assume $m = \mathcal{N}(\bar{m}, \frac{\sigma^2}{n})$

$n = 100$

The sample mean $\bar{m} = \frac{41}{100}$

We don't know σ.
But we do know that if X is a Bernoulli random variable,
$\mathrm{Var}(X) = \sigma^2 = pq \le \frac{1}{4}$ [5/6/98.3-4].

So we assume $\sigma = \frac{1}{2}$, to be safe

maximum possible variance.
Assume distribution as widespread as possible, to be safe.

Therefore, we have, at the significant level:

$$P\left(-1.96 < \frac{m-\bar{m}}{\frac{\sigma}{\sqrt{n}}} \le 1.96\right) = \underbrace{\frac{1}{\sqrt{2\pi}}\int_{-1.96}^{1.96} e^{-s^2/2}\,ds}_{0.95}$$

Hence:

$$P\left(-1.96 < \frac{m-\frac{41}{100}}{\frac{(\frac{1}{2})}{\sqrt{100}}} \le 1.96\right) = 0.95$$

We are assessing whether the coin is fair.
So we want to see if the above holds for $m = \frac{1}{2}$

$$P\left(-1.96 < \underbrace{\frac{\frac{1}{2}-\frac{41}{100}}{\frac{(\frac{1}{2})}{\sqrt{100}}}}_{1.8} \le 1.96\right) = 0.95$$

And, indeed, $-1.96 < 1.8 \le 1.96$.

A more informative way of setting this up is:

$$P\left(-1.96\frac{\sigma}{\sqrt{n}} + m \le \frac{h}{n} < (1.96)\frac{\sigma}{\sqrt{n}} + m\right) = 0.95$$

where h = # of heads in n tosses

$$P\left(\left(-1.96\frac{\sigma}{\sqrt{n}} + m\right)n \le h < \left(1.96\frac{\sigma}{\sqrt{n}} + m\right)n\right) = 0.95$$

all h that satisfy this inequality $\left(-1.96\frac{\sigma}{\sqrt{n}}+m\right)n \le h < \left(1.96\frac{\sigma}{\sqrt{n}}+m\right)n$ indicate the coin is fair (i.e., $m = \frac{1}{2}$) to a significant level (0.95).

In this particular example :

$$P(40.2 \le h < 59.8) = 0.95$$

The fact that we got 41 heads out of our $n = 100$ tosses means that the inequality:
$$40.2 \le 41 < 59.8$$
is satisfied.

The fact that we get between 40.2 and 59.8 heads happens 95% of the time.
Therefore, it is not surprising that I get 41 heads.
Therefore, the result 41 heads out of 100 is <u>not significant</u> at the 0.95 confidence level.

In other words, we have no right to conclude that the coin is biased, Because 95% of the time, the number of heads in 100 tosses is between $[40.2, 59.8)$.

This is how statistical reasoning works.
This <u>scheme of reasoning</u> is 75% of statistics.
Don't kid yourselves.

- The <u>scheme of reasoning</u> we followed is the scheme of reasoning for statistics.
  - We tossed a coin and observed the number of heads.
  - We conjectured that the coin was fair (i.e., $m = \frac{1}{2}$).
  - We chose a significance level (in this case, 0.95).
  - Using the standard normal random variable, we determined the <u>lower</u> and <u>upper bound</u> on the observations (i.e., # of heads) for which the integral of the standard normal density equals the significance level.
  - If the # of observations is w/in the lower + upper bounds, the conjecture is supported.
    " " " " " <u>not</u> w/in " " " " ", the conjecture is <u>not</u> supported.

- If we had tossed the coin 100 times and had obtained 39 heads, then the inequality does not hold.

$$P(40.2 \cancel{\le} (h=39) < 59.8) = 0.95$$

And this would be significant at the 0.95 level.
And we would conclude that the coin was biased and that $m \ne \frac{1}{2}$.

It's like going to the astrologer.

Is there a way of justifying all these steps?
How far have people gotten?
Let's consider a number of justifications people have put forward

Justification (1) The Central Limit Theorem

It's not really a justification.
It's a psychological justification.

This is a major result of probability.

Suppose you have:

$$\underbrace{X_1, X_2, \ldots}_{\text{an infinite sequence}} \quad \text{i.i.d. random variables with:}$$

$$E(X_1) = m$$

the main assumption

$$Var(X_1) = \sigma^2 < \infty \leftarrow \text{finite variance}$$

The random variable $X_1 + X_2 + \ldots + X_n$ has mean nm:

$$E(X_1 + X_2 + \ldots + X_n) = E(X_1) + E(X_2) + \ldots + E(X_n)$$
$$= nm$$

So we obtain a standard normal random variable by taking the sum of n random variables, subtracting nm, and normalizing the result by $\sigma\sqrt{n}$:

$$Z = \frac{X_1 + X_2 + \ldots + X_n - nm}{\sigma\sqrt{n}}$$

$$E(Z) = E\left(\frac{X_1 + X_2 + \ldots + X_n - nm}{\sigma\sqrt{n}}\right)$$
$$= \frac{1}{\sigma\sqrt{n}}\left[E(X_1 + X_2 + \ldots + X_n - nm)\right]$$
$$= \frac{1}{\sigma\sqrt{n}}\left[E(X_1) + E(X_2) + \ldots + E(X_n) - E(nm)\right]$$
$$= \frac{1}{\sigma\sqrt{n}}\left[nm - nm\right]$$

$$E(Z) = 0 \checkmark$$

$$\mathrm{Var}(Z) = \mathrm{Var}\left(\frac{X_1+X_2+\dots+X_n-nm}{\sigma\sqrt{n}}\right)$$

$$= \mathrm{Var}\left(\frac{X_1+X_2+\dots+X_n}{\sigma\sqrt{n}} - \frac{nm}{\sigma\sqrt{n}}\right) \quad \longleftarrow \mathrm{Var}(Y+c) = \mathrm{Var}(Y)$$

$$= \mathrm{Var}\left(\frac{X_1+X_2+\dots+X_n}{\sigma\sqrt{n}}\right) \quad \longleftarrow \mathrm{Var}(cY) = c^2\,\mathrm{Var}(Y)$$

$$= \frac{1}{\sigma^2 n}\mathrm{Var}(X_1+X_2+\dots+X_n)$$

Since $X_1, X_2, \dots, X_n$ are <u>independent</u> [5/4/98.8-10]:

$$= \frac{1}{\sigma^2 n}\underbrace{\left[\mathrm{Var}(X_1)+\mathrm{Var}(X_2)+\dots+\mathrm{Var}(X_n)\right]}_{n\sigma^2}$$

$$= \frac{1}{\sigma^2 n}\cdot n\sigma^2$$

$$\mathrm{Var}(Z) = 1 \ \checkmark$$

So, random variable $Z = \dfrac{X_1+X_2+\dots+X_n-nm}{\sigma\sqrt{n}}$ has $E(Z)=0$, $\mathrm{Var}(Z)=1$.

The random variable $\dfrac{X_1+X_2+\dots+X_n-nm}{\sigma\sqrt{n}}$ is <u>standard</u>

We want to consider what happens when $n\to\infty$.
If there's going to be a limit as $n\to\infty$, it's going to happen as you keep the random variables <u>normalized</u>, for each n.
And, indeed:

$$\lim_{n\to\infty} P\left(a < \frac{X_1+X_2+\dots+X_n-nm}{\sigma\sqrt{n}} \le b\right) \to \frac{1}{\sqrt{2\pi}}\int_a^b e^{-s^2/2}\,ds$$

the <u>normal distribution</u>

In other words, if you <u>add</u> <u>more and more measurements</u>, and <u>keep normalizing</u> them so they are standard normal, then in the limit, they take the <u>bell shaped curve</u> for their density.
That's the Central Limit Theorem.

This is <u>part of the justification</u> why, given <u>several independent measurements</u> of the <u>same quantity</u>, you <u>assume</u> that the <u>X_i</u> are <u>normally distributed</u>.

- What is tough is that you don't know how far in the limit you have to go until you converge.

In practice, say statisticians, this convergence is extremely fast.

For about 50 years, there was a school of Russian statisticians, whose sole purpose was to investigate the speed of convergence of the Central Limit Theorem.
There are many results on this.

Note: We have not proved The Central Limit Theorem.

There is a bad proof in my book.
There is a good proof in the Italian version of my book.

There are many proofs of this theorem.
We don't have time.
It takes a couple of hours to prove it.

Let's see how it works.

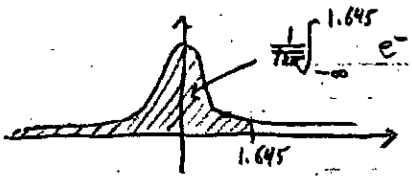
- Example

There are $\overbrace{\text{2 airlines}}^{\text{TWA, United}}$ flying between Boston and LA.

Every day, 1000 passengers fly between Boston and LA.
Each passenger chooses between TWA and United by flipping a fair coin.
How many seats must each airline have to be safe?

What do you mean by safe?
You assign confidence levels.
You can be safe 95% of the time
or very safe 99% of the time.

You look up in your computer, or p.177 of the text, and find that a standard normally distributed random variable ≤ 1.645 has area under the density distribution equal to 0.95.



$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{1.645} e^{-s^2/2}\,ds = 0.95$$

$$P(Z \le 1.645) = 0.95$$

Now, you turn the crank.
This is just a Bernoulli Process.
Let X_i = random variable that the i^{th} passenger chooses TWA

$$\left.\begin{aligned} P(X_i = 1) &= p = \tfrac{1}{2} \\ P(X_i = 0) &= 1-p = \tfrac{1}{2} \end{aligned}\right\} \text{The } X_i \text{ are i.i.d}$$

Let $S_n = X_1 + X_2 + \dots + X_n$ ← here, $n = 1000$

= total number of passengers that choose TWA

The Central Limit Theorem states:

$$P\left(a < \frac{X_1 + X_2 + \dots + X_n - nm}{\sigma\sqrt{n}} \le b\right) \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-s^2/2}\, ds \quad \text{as } n \to \infty$$

So, we have:

$$(*) \quad P\left(\frac{\overbrace{X_1 + X_2 + \dots + X_n}^{S_n} - nm}{\sigma\sqrt{n}} \le 1.645\right) \approx \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{1.645} e^{-s^2/2}\, ds}_{0.95}$$

Since this is a Bernoulli process:

a) the mean of each X_i is:

$$E(X_i) = \sum_n n\, P(X_i = n)$$

$$= 0 \cdot (1-p) + 1 \cdot p \quad \leftarrow p = \tfrac{1}{2} \text{ (coin is fair)}$$

$$\boxed{m = E(X_i) = \tfrac{1}{2}}$$

b) as we've shown earlier [5/6/98.3-4], the upper bound on the variance

$$\sigma^2 = Var(X_i) \le \tfrac{1}{4}$$

So, to be safe, we take $\boxed{\sigma = \tfrac{1}{2}}$

Now we simply need to find S_n that satisfies the inequality in equation (*). Namely, we want S_n such that:

$$\frac{S_n - nm}{\sigma\sqrt{n}} \le 1.645$$

$$S_n \le 1.645\, \sigma\sqrt{n} + nm$$

$$S_n \le 1.645\, \tfrac{1}{2}\sqrt{1000} + 1000 \cdot \tfrac{1}{2}$$

$$\boxed{S_n \le 526.01 \text{ w/ 95\% confidence}}$$

So, w/ 526 seats, you can accommodate passengers, from the 1000 that chose at random to fly TWA, 95% of the time.

That's important kids.
No nonsense.
If you provide 1000 seats, you'll go broke.

- Suppose you want to be safe 99% of the time.
Plug it all in.

1) First, we determine (from the tables) that a standard normally distributed variable Z has 99% of the area under the density distribution when:

$$P(Z \le 2.33) = 0.99$$

2) Then, we use the Central Limit Theorem:

$$P\left(\frac{S_n - nm}{\sigma\sqrt{n}} \le 2.33\right) \approx \underbrace{\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{2.33} e^{-s^2/2}\, ds}_{0.99}$$

3) Finally, we solve for S_n such that the above inequality holds:

$$\frac{S_n - nm}{\sigma\sqrt{n}} \le 2.33$$

$$S_n \le 2.33\,\sigma\sqrt{n} + nm$$

$$S_n \le 2.33\,\tfrac{1}{2}\sqrt{1000} + 1000\cdot\tfrac{1}{2}$$

$$\boxed{S_n \le 536.84 \text{ w/ 99\% confidence}}$$

- In summary:

min seats	confidence
526	95%
537	99%

With only 11 more seats, you go from 95% to being 99% sure you can handle passengers who randomly choose TWA.

All of this is non frivolous.
Because we do not know the rate of convergence of the Central Limit Theorem. We need to address how good an approximation we have for small n (i.e., $n \not\to \infty$).

- Q/A after class:
There is a nice proof of the following: $D = \frac{d}{dx}$

$$\left[D^k x^n\right]_{x=0} = \left[D^n x^k\right]_{x=0}$$

$$\left[p(D)\,g(x)\right]_{x=0} = \left[g(D)\,p(x)\right]_{x=0}$$

For example: $\left[D^4(x^5 + 2x^3)\right]_{x=0} = \left[(D^5 + 2D^3)\,x^4\right]_{x=0}$

The Normal Distribution (cont'd)

Recall that:

a random variable Z w/ $\underbrace{\text{standard normal distribution}}_{}$ means that the density is:

$$E(Z) = 0 \qquad Var(Z) = 1$$

$$dens(Z=t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

In other words:

$$P(Z \le t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-s^2} ds$$

These statements are equivalent, as we know.

The normal distribution is also known as the Gaussian Distribution

- More generally [5/6/98.9-12]:

The random variable $\sigma Z + m$ has a normal distribution w/:

$$E(\sigma Z + m) = m$$
$$Var(\sigma Z + m) = \sigma^2$$

- And we computed [5/6/98.12-14]:

$$P(\sigma Z + m \le t) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{t} e^{-(s-m)^2/2\sigma^2} ds$$

- We also observed, when we were studying variance, that [5/6/98.14]:

if $X_1, X_2, \ldots, X_n$ are i.i.d. random variables w/ $\underbrace{Var(X_1) = \sigma^2 < \infty}_{\text{i.e., } Var(X_1) \text{ finite}}$

then:

$$\underbrace{Var\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)}_{\text{variance of the average}} = Var(\tfrac{1}{n}X_1) + Var(\tfrac{1}{n}X_2) + \cdots + Var(\tfrac{1}{n}X_n)$$
$$= \frac{1}{n^2}Var(X_1) + \frac{1}{n^2}Var(X_2) + \cdots + \frac{1}{n^2}Var(X_n)$$
$$= \frac{1}{n^2}(n\,Var(X_1))$$
$$= \frac{1}{n}Var(X_1)$$

$$Var\left(\frac{X_1+X_2+\cdots+X_n}{n}\right) = \frac{\sigma^2}{n}$$

- Last time, we concluded with an exploration of this fact.
Namely, if you have n independent, identically distributed variables w/ finite variance, suppose that:

$$E(X_1) = m$$
$$Var(X_1) = \sigma^2$$

Then:

$$E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n)$$
$$= n\,E(X_1)$$
$$= nm$$

$$Var(X_1 + X_2 + \dots + X_n) = Var(X_1) + Var(X_2) + \dots + Var(X_n)$$
$$= n\,Var(X_1)$$
$$= n\sigma^2$$

How do we normalize the random variable $X_1 + X_2 + \dots + X_n$?

Normalizing means: 1) subtracting the mean
and 2) dividing by the standard deviation
so that the resulting random variable has:

mean $= 0$
variance $= 1$

$$\frac{X_1 + X_2 + \dots + X_n - nm}{\sigma\sqrt{n}} \leftarrow E(X_1 + X_2 + \dots + X_n)$$

Standard Deviation $(X_1 + X_2 + \dots + X_n)$

this random variable is standardized

- So, for every n, we keep standardizing it, so that it doesn't blow up.
This is the only hope we have of tending to a limit for the distribution.

And, as a matter of fact, under these circumstances:

finite variance
independent random variables

You have:

$$P\left(a < \frac{X_1 + X_2 + \dots + X_n - nm}{\sigma\sqrt{n}} \le b\right) \longrightarrow \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2}\, dt$$

Which means that the limit of the distribution of this standardized random variable tends to the standard normal distribution.

$$P\left(\frac{X_1+X_2+\dots+X_n-nm}{\sigma\sqrt{n}} \le t\right) \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-s^2/2}\,ds$$

This is the content of the Central Limit Theorem. ← {we don't have time to prove this}

- The Central Limit Theorem provides yet another justification for statistics. Namely, when you take several "repeated" measures" of "same quantity," you assume that the quantity in question is given by ~~the~~ a normal distribution, whose mean = sample mean
  variance = sample variance

  You plug it in and do computations.

- For example, we discussed last time the problem of passengers choosing, at random, one of two airlines [5/8/98.13-15].

  What is the tolerance level, as they say, for an airline to have full seats?

  ↑ Say 95% of the time, we want to accommodate all passengers. For the rest of the time, the hostess will come in and say "Will someone wish to give up a seat?"

  As we showed, we have:

$$P\left(\frac{S_n - nm}{\sigma\sqrt{n}} \le 1.645\right) = 0.95 \qquad \Longrightarrow \quad S_n = 526$$

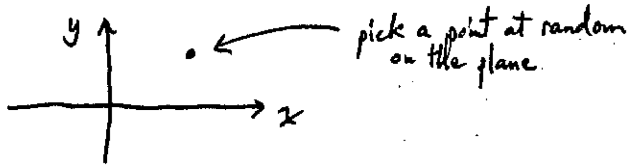
↑ we look this up in a table

- So this is another justification for statistics. We saw the original "justification" from using Bayes' Theorem and choosing dens$(m=s)=1$ [5/8/98.6]. Now let's see two additional justifications.

Justification (2) Maxwell – Einstein Derivation of the Normal Distribution

some of his best papers were on the normal distribution.

They say this:
Let's take a plane and suppose we pick a point at random on the plane.
Let's try to give meaning to picking a point at random on the plane.



Now we know we can not use the measure of the plane (area).
Because that gives you infinite area and that can not be normalized.
That's a sticky point that people have never been able to fix.

(Remember, people use measures as priors and they work, but they have no idea why it works.)

We want a real probability density in 2 variables (X, Y) which are random.
What are we to mean for random?
We mean the following things:

(1) $\underbrace{X \text{ and } Y}$ (coordinates of the point you pick) are independent random variables

you pick an independant X and an independant Y

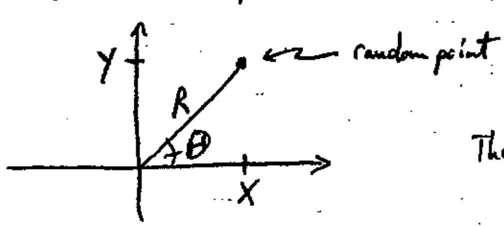
Then, the joint density equals the product of the densities [4/15/98.7]:

$$\text{dens}(X=t, Y=s) = \text{dens}(X=t)\,\text{dens}(Y=s)$$

and for reasons of symmetry, of course:

$$\text{dens}(X=t) = \text{dens}(Y=t)$$

(2) switch to polar coordinates



The radius R and angle Θ are random variables.

$$\text{dens}(X=t, Y=s)\, dt\, ds = \text{dens}(R=r, \Theta=\theta)\, r\, dr\, d\theta$$

$\Longrightarrow$ The <u>assumption</u> is that:

$$\text{dens}(R=r, \Theta=\theta) = \text{dens}(R=r)$$

↑ assumption is that the density is <u>independent</u> of Θ.

With just this assumption, that's it.
<u>Out comes the normal distribution</u>

The <u>only density</u> $\text{dens}(R=r)$ that satisfies the above is the <u>normal distribution</u> $\frac{1}{\sigma\sqrt{2\pi}} e^{-t^2/2\sigma^2}$

Lets see why.
We've actually done most of the work.

Say $\text{dens}(R=r) = \underbrace{g(r)}_{\text{some function } g \text{ of } r}$

and $\text{dens}(X=t) = f(t)$

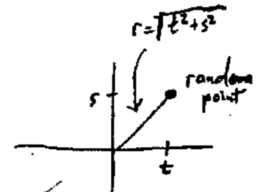
From items (1) + (2) above, we have:

$$\text{dens}(X=t, Y=s)\, dt\, ds = \text{dens}(R=r, \Theta=\theta)\, r\, dr\, d\theta$$

$$\text{dens}(X=t)\,\text{dens}(\cancel{X}=s)\, dt\, ds = \text{dens}(\cancel{R=t})\, r\, dr\, d\theta$$

$$\underbrace{\text{dens}(X=t)\,\text{dens}(X=s)}_{f(t)\,f(s)}\cancel{dt\,ds} = \underbrace{\text{dens}(R=r)}_{g(\sqrt{t^2+s^2})}\cancel{r\,dr\,d\theta}$$

$$f(t)\,f(s) = g(\sqrt{t^2+s^2})$$



This is a functional equation to solve.

(*) $$f(t)\,f(s) = g(\sqrt{t^2+s^2})$$

You have to find <u>all functions</u> $f + g$ that satisfy this functional equation.
That's not hard.

Set $s=0$:

$$\left.\begin{aligned} f(t)\,f(0) &= g(\sqrt{t^2}) \\ &= g(t), \quad t \geq 0 \end{aligned}\right\} \Longrightarrow f(t) = \frac{g(t)}{f(0)}, \quad t \geq 0$$

Similarily, set $t=0$:

$$\left.\begin{aligned} f(0)\,f(s) &= g(\sqrt{s^2}) \\ &= g(s)\,,\quad s\geq 0 \end{aligned}\right\} \Rightarrow \quad f(s) = \frac{g(s)}{f(0)}\,,\quad s \geq 0$$

So, we can rewrite functional equation (*) as:

$$\frac{g(t)}{f(0)}\,\frac{g(s)}{f(0)} = g\left(\sqrt{t^2+s^2}\right)$$

$$\frac{g(t)\,g(s)}{f(0)^2} = g\left(\sqrt{t^2+s^2}\right)$$

This is the functional equation we already had, when we were doing exponential distributions [4/15/98.14-16], [4/22/98.5].

Set $h(t) = \dfrac{g(\sqrt{t})}{f(0)}$

Then:

$$\underbrace{h(t)\,h(s) = h(t+s)}_{\text{This is the functional equation for the exponential}} \quad \Rightarrow \quad \left\{\begin{aligned} h(t^2)\,h(s^2) &= h(t^2+s^2) \\ \frac{g(\sqrt{t^2})}{f(0)}\,\frac{g(\sqrt{s^2})}{f(0)} &= \frac{g\left(\sqrt{t^2+s^2}\right)}{f(0)} \\ \frac{g(t)\,g(s)}{f(0)^2} &= \frac{g\left(\sqrt{t^2+s^2}\right)}{f(0)} \\ \frac{g(t)\,g(s)}{f(0)} &= g\left(\sqrt{t^2+s^2}\right) \\ \uparrow\ f(0) &= f(0)^2 \quad ? \end{aligned}\right.$$

$$h(t) = e^{ct}$$

Since $h(t^2) = \dfrac{g(t)}{f(0)}$,

$$g(t) = e^{c' t^2}$$

Recall that $g(t) = \text{dens}(R = t)$ is just a density distribution.
So we know that:

$$\int_{\theta=0}^{2\pi}\int_{r=0}^{\infty} e^{c' r^2}\, r\,dr\,d\theta = 1$$

So we can determine c'.

If you work this out, you end up with:

$$\boxed{g(t) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-t^2/2\sigma^2}}$$

The normal distribution.

So the <u>normal distribution</u> <u>comes out</u> just from this consideration of items (1) + (2) above.

Really wonderful stuff.

The last justification we shall consider: 5/11/98.7

Justification (3) Wiener's Characterization

This is the best one - not just because he taught this course.

Wiener says:

- When we dealt w/ tossing a coin, we didn't characterize just coin tossing. We studied the Bernoulli Process - sequences of coin tossing.
- Similarily, when we studied the waiting time for a phone call to end, namely the exponential distribution, we didn't just study the exponential distribution, we studied sums of independently distributed exponential random variables, and we were thereby lead to the Poisson Process!
- When we studied picking a point at random in the interval $[0,a]$, we immediately went over to picking n points at random in the interval $[0,a]$ and then we started doing hairy computations

In each case, when we were confronted w/ the problem of studying a probability distribution, we tried to go over to a stochastic process associated w/ that distribution.

$$\text{probability distribution} \longrightarrow \text{stochastic process}$$

That's what Wiener does w/ the Normal Distribution.

$$\text{normal distribution} \longrightarrow \underbrace{\text{Wiener Process}}_{\text{The Wiener Process or Brownian motion}}$$

It's based on the following remark.
A very important fact.

$$\text{Suppose} \quad \left.\begin{array}{l} X_1 \text{ is } \mathcal{N}(m_1, \sigma_1^2) \\ X_2 \text{ is } \mathcal{N}(m_2, \sigma_2^2) \end{array}\right\} \quad X_1 + X_2 \text{ independent}$$

X_1 is normal, with mean m_1, variance σ_1^2

So we have 2 independent random variables, which are normally distributed.

Then the basic fact is that their sum is normal, with mean the sum of their means and their variance, the sum of their variances.

$$X_1 + X_2 \text{ is } \mathcal{N}(\underbrace{m_1 + m_2}, \underbrace{\sigma_1^2 + \sigma_2^2})$$

To be expected, since expectation of sum is sum of the expectations.

But the interesting thing is that the variances behave like this.

As a matter of fact, one can show that the normal distribution is the only distribution where:

$$\left.\begin{array}{l} X_1 \text{ with } E(X_1) = m_1 \\ \qquad\quad Var(X_1) = \sigma_1^2 \\ X_2 \text{ with } E(X_2) = m_2 \\ \qquad\quad Var(X_2) = \sigma_2^2 \end{array}\right\} \quad X_1 + X_2 \text{ has } E(X_1+X_2) = m_1 + m_2, \quad Var(X_1+X_2) = \sigma_1^2 + \sigma_2^2$$

The means add and the variances add

We could check this by the convolution of 2 hairy integrals

Alternatively:

Proof

We may assume that $m_1 = 0$, $m_2 = 0$ — Since if we shift the random variable:

$$X_1 \leftarrow X_1 - m_1$$

$$E(X_1) = E(X_1 - m_1) = E(X_1) - E(m_1) = m_1 - m_1 = 0$$

So we only need to check:

X_1, X_2 independent, normal with:

$$E(X_1) = 0 \qquad E(X_2) = 0$$
$$Var(X_1) = \sigma_1^2 \quad Var(X_2) = \sigma_2^2$$

By the continuous law of alternatives consequence [4/8/98.7], the joint density is:

$$dens(X_1 + X_2 = t) = \int_{-\infty}^{\infty} \underbrace{dens(X_1 + X_2 = t / X_2 = s)} dens(X_2 = s)\, ds$$

The event $X_1 + X_2 = t$, given that $X_2 = s$ is the same as: $X_1 + s = t$
$X_1 = t - s$

$$= \int_{-\infty}^{\infty} dens(X_1 = t-s)\, dens(X_2 = s)\, ds$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sigma_1\sqrt{2\pi}} e^{-(t-s)^2/2\sigma_1^2} \frac{1}{\sigma_2\sqrt{2\pi}} e^{-s^2/2\sigma_2^2}\, ds$$

$$= \frac{1}{\sigma_1\sqrt{2\pi}} \frac{1}{\sigma_2\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(t-s)^2/2\sigma_1^2} e^{-s^2/2\sigma_2^2}\, ds$$

Let's work out the exponents:

$$\frac{-(t-s)^2}{2\sigma_1^2} - \frac{s^2}{2\sigma_2^2} = \frac{-t^2+2ts-s^2}{2\sigma_1^2} - \frac{s^2}{2\sigma_2^2}$$

$$= \underbrace{\frac{-t^2\sigma_2^2+2ts\sigma_2^2-s^2\sigma_2^2-s^2\sigma_1^2}{2\sigma_1^2\sigma_2^2}}$$

The claim is that this is a perfect square of the form:

?

$$\frac{-\left(s\sqrt{\sigma_1^2+\sigma_2^2}+\beta\right)^2}{2\sigma_1^2\sigma_2^2}$$

Note how tantalizingly close:

$$\frac{-\left(s\sqrt{\sigma_1^2+\sigma_2^2}-t\sigma_2\right)^2}{2\sigma_1^2\sigma_2^2}$$

gets.

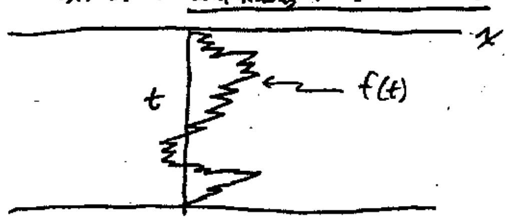
The claim is, that if you work this all out, you eventually obtain:

(*) $$\boxed{\mathrm{dens}(X_1+X_2=t) = \frac{1}{\sqrt{\sigma_1^2+\sigma_2^2}\,\sqrt{2\pi}}\, e^{-t^2/2(\sigma_1^2+\sigma_2^2)}}$$

Normal distribution w/ variance $\underbrace{\sigma_1^2+\sigma_2^2}_{\leftarrow \text{additive.}}$

So, if you take 2 random variables that are normally distributed, if you take their sum, then their sum is normally distributed with mean the sum of their means and with variance the sum of their variances.

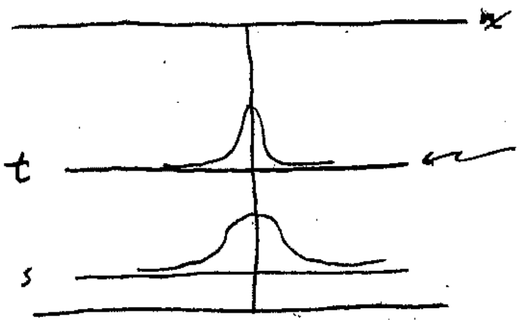
- The main idea from Wiener.
  Let's construct a continuous random walk



Can you give a consistent probability density for this situation?
The answer is very simple.

Ω = all continuous functions $f(t)$, where $f(0)=0$

events are defined as follows:



at each time t, the balls have moved around, like the Museum of Science display [5/11/98.13]. So at each time t, we have a normal distribution.

Let $W(t)$ be a random function.

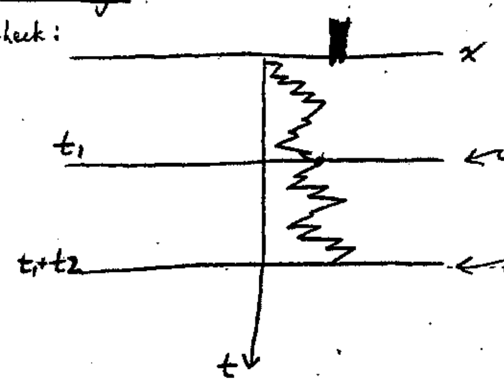
The density is:

$$\boxed{\text{dens}(W(t)=x) = \frac{1}{\sqrt{2\pi t}}\, e^{-x^2/2t}} \quad \leftarrow \quad \frac{1}{\sigma\sqrt{2\pi}}\, e^{-x^2/2\sigma^2}$$

Variance is t
$\sigma^2 = t$

There is one major thing to check.
Is this definition consistent?
Consistency

Check:



at time t_1, $W(t_1)$ will be normally distributed w/ variance t_1.

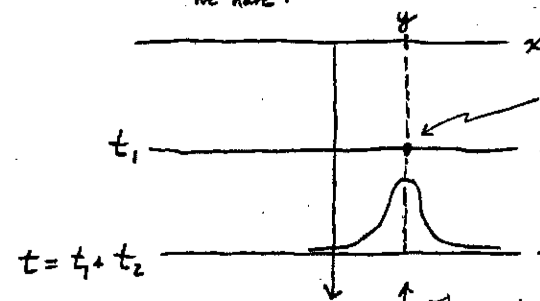
at time t_1+t_2, $W(t_1+t_2)$ will be normally distributed w/ variance t_1+t_2.

Let $t = t_1 + t_2$

By the Law of Alternatives Consequence:

$$\text{dens}\left(W(t)=x\right) = \int_{-\infty}^{\infty} \underbrace{\text{dens}\left(W(t)=x \mid W(t_1)=y\right)}_{} \underbrace{\text{dens}\left(W(t_1)=y\right)}_{\frac{1}{\sqrt{2\pi t_1}} e^{-y^2/2t_1}} dy$$

It is given that $W(t_1)=y$.
Thus, for each $y \in (-\infty, \infty)$, we have:



First, we have that $W(t_1)=y$

Second, with y as "origin", we do another random walk of time t_2.

The resulting curve will be shifted. And this will be normally distributed, with expectation y. It will be centered at y.

$$\text{dens}\left(W(t)=x \mid W(t_1)=y\right) = \frac{1}{\sqrt{t_2}\sqrt{2\pi}} e^{-(x-y)^2/2t_2}$$

$$= \int_{-\infty}^{\infty} \underbrace{\frac{1}{\sqrt{t_2}\sqrt{2\pi}} e^{-(x-y)^2/2t_2}}_{g(x-y) = \text{dens}(Z_2 = x-y)} \; \underbrace{\frac{1}{\sqrt{t_1}\sqrt{2\pi}} e^{-y^2/2t_1}}_{f(y) = \text{dens}(Z_1 = y)} dy$$

This is exactly the convolution of two normal distributions. [4/15/98.8]

Thus:

$$h(x) = \int_{-\infty}^{\infty} f(y)\, g(x-y)\, dy$$

where:

$$h(x) = \text{dens}(Z_1 + Z_2 = x)$$

$$= \text{dens}(Z_1 + Z_2 = x)$$

And this is exactly the density distribution we obtained earlier in equation (*) [5/11/98.8-9]

$$= \frac{1}{\sqrt{(\sqrt{t_1})^2 + (\sqrt{t_2})^2}\,\sqrt{2\pi}} e^{-x^2/2\left((\sqrt{t_1})^2 + (\sqrt{t_2})^2\right)}$$

And, since $t = t_1 + t_2$

$$\boxed{\text{dens}\left(W(t)=x\right) = \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t}}$$ ✓

- So it's consistent.
  Therefore, because it is consistent, it therefore exists.
  There are more technical things to verify, but this is a basic fact.

  If you stop a random walk in between, and look at it in between, and measure densities, it's consistent w/ the density measurement of the entire random walk.
  (Just the same idea as we used to show the memoryless property for the Poisson Process [4/28/98.10-12]).

- Therefore, we have a new stochastic process that obtains probabilities by a random function.
  Instead of a random walk [5/1/98.5-13],

  $$P(F = 2n)$$

  you have a random function.

  $$P(W(t) \leq x)$$

- This is how a random walk tends to this Wiener Process in the limit.
  The Central Limit Theorem tells you how random walk approaches Weiner Process.

  All computations are done in the Wiener Process, once you discover the Wiener Process exists, and you forget about random walks.
  You just do everything in the Wiener Process.

  This is the basic process of the theory of stochastic processes, which you study in 18.445.

- Law of Large Numbers ← {A strengthening of the fact that the variance goes down as n increases.}

  $X_1, X_2, \ldots$ of i.i.d. random variables
  with finite variance

  $\uparrow$ $Var(X_1) = \sigma^2 < \infty$

  [redacted]

  $$P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - m\right| > \varepsilon\right) \to 0 \quad , \text{as } n \to \infty,$$

  for any $\varepsilon > 0$.

This is the probability that the average deviates from the mean by more than ε.

Note that this law is true for any ε.

And this means, more or less, that w/ probability 1, the sample average $\frac{X_1 + X_2 + \dots + X_n}{n}$ tends to the mean m.

- A strengthening of this statement would be:

$$P\left(\lim_{n\to\infty} \frac{X_1 + X_2 + \dots + X_n}{n} = m\right) = 1$$

The Strong Law of Large Numbers

- So, now you know all the probability that you need.
I hope you take another probability course and you find it real easy.

Last Quiz on Wednesday.

I hope you all take UROPs w/ me and help me w/ my problems.

— The End —

John Guidi
guidi@mit.edu

{ Problem #9 is from a research paper. }

Q: Can you use the <u>law of successive conditioning</u> to prove the <u>inclusion-exclusion principle?</u>
A: Yes. You can prove the inclusion-exclusion principle using conditional probability. I want to give that to you as a problem.

There are dozens of proofs of the inclusion-exclusion principle.
Most proofs use induction on the number of events.

First you prove that:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

which is an axiom of probability.

Then you assume that it is true for n events and prove:

$$P(A_1 \cup A_2 \cup \ldots \cup A_n \cup \underbrace{A_{n+1}}) = \quad \ldots$$

1) condition relative to the last event
2) juggle the properties of conditional probability
3) reduce it to the case of an event with conditional probability

There's <u>another way</u> of proving the inclusion-exclusion principle, which is given in the book, which is the <u>simplest</u>.
Let's discuss it here, since I'm not going to discuss it in class.

<u>His favorite proof of the inclusion-exclusion principle</u>

This is based on the idea of an <u>indicator random variable</u>

<u>Indicator Random variable</u>

You have a sample space Ω and an event A.

Ω

You define a random variable I_A, which is called the indicator random variable of the event A.
Let me define it first probabilistically, then set theoretically.
Equivalent — two languages.

Probabilistic

$$I_A = \begin{cases} 1 & \text{if } A \text{ happens} \\ 0 & \text{if } A \text{ does not happen} \end{cases}$$

read: Indicator A

What does this mean in set theoretic language?
i.e.,

$$\underbrace{I_A(\omega)}_{I_A \text{ at a sample point } \omega.} = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$$

Set Theoretic

We see a striking contrast between using probabilistic language and set theoretic language.

I_A is a random variable.
What is it's probability distribution?

$$P(I_A = 1) = P(A)$$

$$P(I_A = 0) = 1 - P(A)$$

What is the expectation of this random variable?

$$E(I_A) = \sum_n n\, P(I_A = n)$$

$$= 1 \cdot P(I_A = 1) + 0 \cdot P(I_A = 0)$$

$$= P(A)$$

from above, since $P(I_A = 1) = P(A)$

$$\boxed{E(I_A) = P(A)}$$

This is a fundamental identity.
What does it do for you?
It <u>reduces</u> computations of <u>probabilities</u> to computations of <u>expectations</u>.

Why are computations w/ expectations nicer?
Because expectation of a sum is the sum of the expectations.
No questions asked.

So now, let's use the technique of indicator random variables to prove the inclusion-exclusion principle.

To do that, we have to do the following.
We have to set up a <u>correspondence</u> between the operations of boolean algebra of events and the operations of indicator random variables, which are the ordinary sum + times.

This is very nice. We get rid of $\cap$, $\cup$ and they get replaced by $+, *$.

Let's see how that works.
First, what is the indicator of A^c?
That's easy.

$$I_{A^c} = \underset{\substack{\uparrow \\ \text{the function } 1}}{1} - I_A \qquad \left\{ \begin{array}{l} \text{Q/A: } 1 = I_\Omega \\ \quad 1 \text{ is a function that takes the value } 1 \\ \quad \text{at every sample point.} \end{array} \right\}$$

Next. The indicator of A and B.

$$I_{A\cap B} = I_A \underset{\substack{\uparrow \\ \text{ordinary times}}}{\cdot} I_B \qquad\qquad I_A \cdot I_B = 1 \text{ iff both } A \text{ and } B \text{ happen}$$

Things get a little kinky w/ the union.

$$I_{A\cup B} = I_A + I_B - I_A \cdot I_B$$

Let's check.

If $I_{A\cup B} = 1$ then either I_A or I_B or both must be 1

$$\begin{array}{rccccc} & I_A & + & I_B & - & I_A \cdot I_B \\ 1 = & 1 & + & 0 & - & 0 \\ 1 = & 0 & + & 1 & - & 0 \\ 1 = & 1 & + & 1 & - & 1 \end{array} \quad \checkmark$$

But this can be written in a much more striking way.
And this is the key to the whole story.

$$I_{A\cup B} = 1 - (1 - I_A)(1 - I_B) \Leftarrow \left\{ \begin{array}{l} \cancel{1-}\left[\cancel{1} - I_A - I_B + I_A I_B\right] \\ I_A + I_B - I_A I_B \end{array} \right.$$

By the way, this is DeMorgan's Law:

$$\begin{aligned} I_{A\cup B} &= 1 - I_{(A\cup B)^c} \\ &= 1 - I_{A^c \cap B^c} \\ &= 1 - I_{A^c} I_{B^c} = 1 - (1 - I_A)(1 - I_B) \end{aligned}$$

This extends to several unions.

More generally:

$$I_{A_1 \cup A_2 \cup \dots \cup A_n} = 1 - (1 - I_{A_1})(1 - I_{A_2}) \cdots (1 - I_{A_n})$$

Indicator of $A_1 \cup A_2 \cup \dots \cup A_n$

This is the <u>inclusion-exclusion principle.</u>
How does the inclusion-exclusion principle come out?
By taking <u>expectations</u> of both sides!
Let's <u>expand RHS</u> by rules of algebra.

$$I_{A_1 \cup A_2 \cup \dots \cup A_n} = \sum_i I_{A_i} - \sum_{i<j} I_{A_i} I_{A_j} + \sum_{i<j<k} I_{A_i} I_{A_j} I_{A_k} - \dots + \dots$$

Taking expectations:

$$E(I_{A_1 \cup A_2 \cup \dots \cup A_n}) = E\Big(\sum_i I_{A_i} - \sum_{i<j} I_{A_i} I_{A_j} + \sum_{i<j<k} I_{A_i} I_{A_j} I_{A_k} - \dots + \dots\Big)$$

expectation is <u>additive</u>

$$= \sum_i E(I_{A_i}) - \sum_{i<j} E\big(\overbrace{I_{A_i} I_{A_j}}^{I_{A_i \cap A_j}}\big) + \sum_{i<j<k} E\big(\overbrace{I_{A_i} I_{A_j} I_{A_k}}^{I_{A_i \cap A_j \cap A_k}}\big) - \dots + \dots$$

Recall the fundamental identity [sc 2/27/98.2].
The expectation of the indicator of an event is the probability of that event.

$$E(I_A) = P(A)$$

So we have:

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \dots$$

That's it. We've proved the inclusion-exclusion principle.

So now you say: "What's the point of going through all this stuff w/ Réyni's Principle if there is a simpler way?"

- More on Réyni's Principle (from a different point of view)

Let's wax theoretical.
Suppose you are given:

$$A_1, A_2, \dots, A_n = \text{events (finite)}$$

And suppose they are allowed $\cup, \cap, {}^c$, but not infinite unions
(the 3 boolean operations, but not countable additivity)

How do we describe the most general event that is expressible in terms of $A_1, A_2, \dots, A_n$?

In other words, you're given these events.
What do they determine?
There are 2 ways of answering this question.

Cheapo answer

A: anything that can be expressed in terms of the events with $\cup, \cap, {}^c$.
That's called a boolean function.

Answer – any boolean function of $A_1, A_2, \dots, A_n$.
A boolean function is any expression that can be made up of $\cup, \cap, {}^c$.

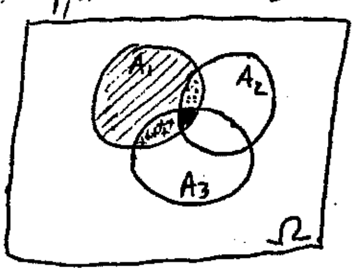
Ex:

$$\left((A_1 \cup A_2) \cap (A_3 \cup A_1^c) \cap (A_4 \cup A_2)^c\right) \cup (A_3 \cup A_4 \cup A_5^c)$$

So, in a certain sense, this answers the question.
But it's not a satisfactory answer.

The real answer

Suppose we have 3 events in our sample space Ω



$A_1 \cap A_2 \cap A_3$
$A_1 \cap A_2 \cap A_3^c$
$A_1 \cap A_2^c \cap A_3$
$A_1 \cap A_2^c \cap A_3^c$
etc

$\Bigg\}$ 2^n smallest possible sets

You see that there is a better answer to the question when you look at this picture.
Consider the smallest possible sets determined by these events, including the one outside all the events.

Any event determined by the A_i will be some union of the smallest possible events.

How do we determine precisely what it means to be the smallest possible set?
Take all the events and intersect the event or the complement.

Every event expressible in terms of $A_1, A_2, \dots, A_n$ is a <u>disjoint union</u> of events of the form:

$$A_1^{\pm} \cap A_2^{\pm} \cap \dots \cap A_n^{\pm}$$ { ± is Réyni's notation for a set or its complement. }

where $A_i^{+} = A_i$

$A_i^{-} = A_i^{c}$

These are the smallest possible sets.
These are sometimes called the <u>atoms</u>.
So any event expressible in terms of n events is the union of disjoint atoms.
This is called the <u>disjunctive normal form</u>.

Every boolean function of $A_1, A_2, \dots, A_n$ is a disjoint union of atoms.
This is called disjunctive normal form.

How many <u>distinct events</u> can be expressed in terms of n events?

$$\underbrace{2}\,^{2^n}$$ ← number of distinct atoms for n events

$\text{atom}_1 \cup \text{atom}_2 \cup \dots \cup \text{atom}_{2^n}$
For every atom, you can either take it or not take it.
Total # of disjoint unions of atoms $= 2^{\#\text{ distinct atoms}}$

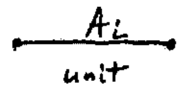
$$= 2^{2^n}$$

In logic, this result is called the construction of the <u>free boolean algebra w/ n generators</u>.

One thing I did not mention in class that is covered in the book is the <u>circuit theoretic interpretation</u> of the boolean operations.
This was worked out by Shannon when he took this course for his S.M. thesis.

Circuit-theoretic interpretation

every event is considered as a unit.

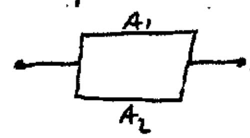


You can piece together units in series or in parallel

series

corresponds to:
$A_1 \cap A_2$

parallel



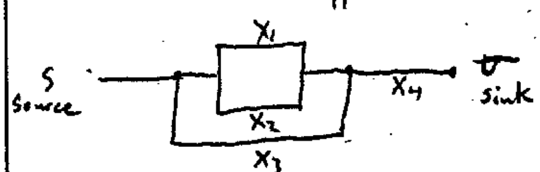
corresponds to:
$A_1 \cup A_2$

The interpretation is this.
This is a channel through which a message goes.
These are units that can break down.

If you put two units in series, the message will go through iff both of them work.
If you put two units in parallel, the message will go through iff at least one works.

Complementation - they invented something

Here is the main application of this idea:



The circuit is made up of units that may break down independently.
The probability that a unit breaks down is p.
" " " " " does not break down is $q = 1-p$.

All units have the same probability of breaking down.
Under these conditions, you want the probability that a message sent from s will arrive at t.
This computation is an important theory, which people work full time on.
It's called <u>reliability theory</u>.
There's a journal called the IEEE Transactions on Reliability that is dedicated exclusively to computing stuff like that.

Reliability of a series, parallel circuit is the probability that the circuit works.

So, in general, there are two independent problems:

1) You are given a circuit and you have to compute its reliability

2) You are given so much money and you need a given reliability (e.g., probability = .99), and you have to design the circuit.
Of course, if you put everything in parallel, it works. But that's too expensive.
That's a very complicated chore. There's a whole art to it. There are an incredible number of techniques to simplify things.

What's the basic computation of reliability?
Inclusion - exclusion.

Let's see, using the above circuit,

random variable

$$X_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } q \end{cases}$$

This is the most important application of the inclusion - exclusion principle there is.
Reliability of the circuit is the probability that at least one path goes through.

Define events:

$$A_1 = ((X_1 = 1) \cap (X_4 = 1))$$
$$A_2 = ((X_2 = 1) \cap (X_4 = 1))$$
$$A_3 = ((X_3 = 1) \cap (X_4 = 1))$$

$$\text{reliability} = P(A_1 \cup A_2 \cup A_3)$$

The inclusion-exclusion principle gives:

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) = \; & P(A_1) + P(A_2) + P(A_3) \\ & - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) \\ & + P(A_1 \cap A_2 \cap A_3) \end{aligned}$$

$$= P((X_1=1)\cap(X_4=1)) + P((X_2=1)\cap(X_4=1)) + P((X_3=1)\cap(X_4=1))$$
$$- P((X_1=1)\cap(X_2=1)\cap(X_4=1))$$
$$- P((X_1=1)\cap(X_3=1)\cap(X_4=1))$$
$$- P((X_2=1)\cap(X_3=1)\cap(X_4=1))$$
$$+ P((X_1=1)\cap(X_2=1)\cap(X_3=1)\cap(X_4=1))$$

And since all the X_i are independent:

$$= p^2 + p^2 + p^2 - p^3 - p^3 - p^3 + p^4$$
$$\underbrace{P(A_1 \cup A_2 \cup A_3) = 3p^2 - 3p^3 + p^4}_{\text{that's the reliability of that circuit}}$$

Q: Can you describe the circuit as the sum or product of their <u>indicators</u> and then take the expectation?
A: No, you can't.
You can't expand this. There is no way.
These are like the atoms.
You really have to use inclusion-exclusion.

In general, we can have a complicated circuit and computing the reliability by inclusion-exclusion can be extremely laborious.
It's a big business to simplify inclusion-exclusion computations.
Even to get <u>approximate</u> expressions by breaking off the inclusion-exclusion sums after the first terms. You have to see what the error is when you break it off.

How do you represent the <u>distributive law</u>?
You can, but it's a little contrived.

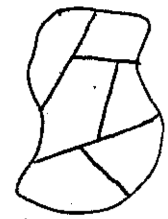
Algebra of Partitions

a partition $\pi = \underbrace{\{B\}}_{\text{family of events } B.\ B \text{ is called a block.}}$ with the properties:

(1) if $B \in \pi$ then $P(B) > 0$

(2) if $B, B' \in \pi$ and $B \neq B'$ then

$$P(B \cap B') = 0$$



(3) $P\left(\bigcup_{B \in \pi} B\right) = \underline{1}$

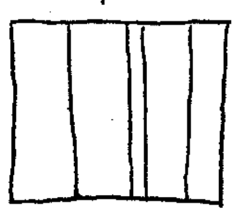
We want to define operations on partitions that are analogues of boolean $\cup$ and $\cap$ of sets.

If π and σ are two partitions of Ω, P, the <u>meet</u> $(\pi \wedge \sigma)$ is the partition whose blocks are all intersections

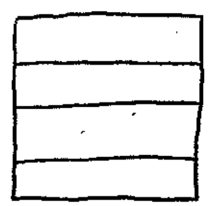
$$B \cap C\ , \text{ for } B \in \pi,\ C \in \sigma, \text{ whenever } P(B \cap C) > 0$$

We will use this later when we do entropy.

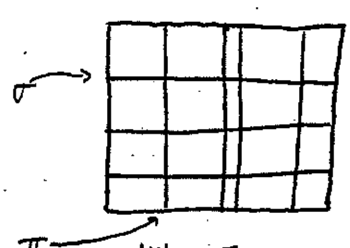
Let's see a picture



partition π



partition σ



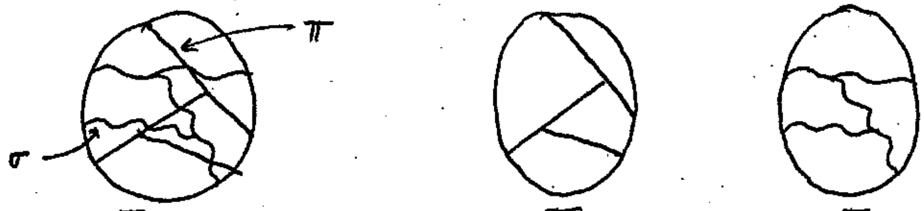
partition $\pi \wedge \sigma$

the meet is all the rectangles

Let's take another example:

π = ——— partition

σ = ~~~ partition



$\pi \wedge \sigma$ $\qquad$ π $\qquad$ σ

Meet ($\wedge$) for partitions is the analogue of intersection ($\cap$) for sets.

There is also an analogue of set union for partitions, but it's harder to visualize.

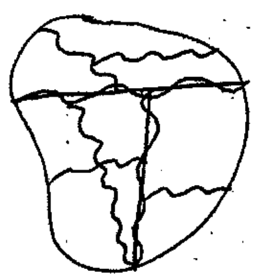
For sets, we have containment of events. $A \subseteq B$
What's the analogue for containment for partitions?

If π and σ are partitions of Ω, P, then we say π is a <u>finer</u> partition than σ:

$$\pi \leq \sigma$$

when every block of π is contained in some block of σ.

<u>Example</u> $\quad$ ~~~ = π $\quad$ ——— = σ



$\pi \leq \sigma$

π is finer than σ

How do I get a smaller partition? I just cut up the blocks of σ and then I get π.

Start w/ σ.
Cut into smaller blocks to get π.

Next time we'll discuss the analogue of set union, the join ($\vee$) of partitions. There is a beautiful algebra of partitions, which parallels the algebra of sets. But partitions do <u>not</u> satisfy the <u>distributive law.</u>

$\pi \vee \sigma$ = smallest partition of which both π and σ are finer.
$\uparrow$ join

- Many different priors occur in practice.
  Sometimes you really have information that says that:
  number of good guys is even
  or information that there is a bump in the middle
  or convex
  or monotonic, whatever.

There's a whole array of techniques for computing posterior probabilities.

- You may notice that the notation we use for Bayesian statistics is not totally rigorous.
  Because you use the random variable U for the prior and the random variable U for the posterior.

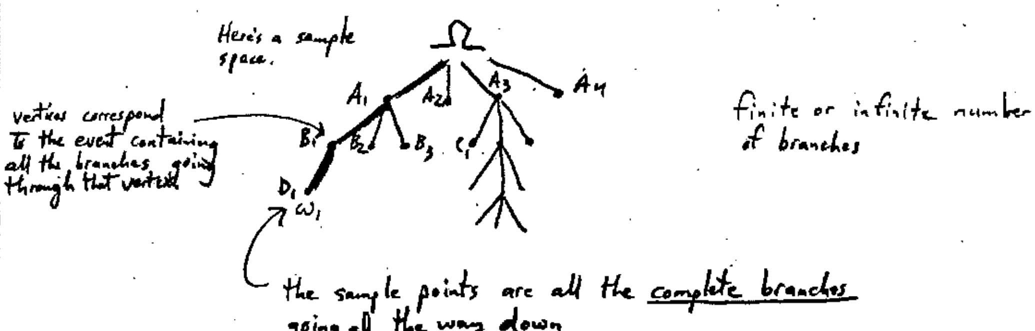
$$P(U=j\,|\,A=i) = \frac{P(A=i\,|\,U=j)\,P(U=j)}{P(A=i)}$$

These two random variables are different, because they have different distributions.
I followed standard statistical treatment, but its not completely rigorous.
To be rigorous, you would have to give them different names, like U_0 and U

prior posterior

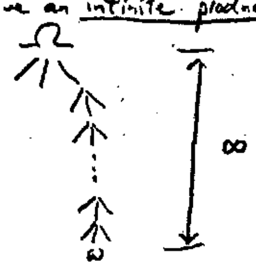
$$P(U=j\,|\,A=i) = \frac{P(A=i\,|\,U_0=j)\,P(U_0=j)}{P(A=i)}$$

- Trees
  In practice, a lot of sample spaces are defined in terms of trees.



$$P(\omega_1) = P(A_1)\,P(B_1|A_1)\,P(D_1|B_1)$$

Note that you could have an infinite product for the probability of a sample point, $P(\omega)$.
For example:



If you sum up the finite number of branches, then you have the probability of the event at a given vertex.
These probabilities are consistent with the Law of Successive Conditioning [3/2/98.2-3]
(ex: $D_1 \supseteq B_1 \supseteq A_1$, etc.).

The most notable example, which we will cover partly in class, but not completely. So we'll cover it here.

- Urn Models

This is a model that is used to cover all sorts of situations.
Traffic accidents, spread of a disease, etc.

You start w/ an urn w/ r red balls and b black balls.

r red
b black

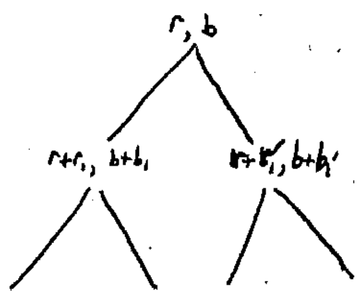
Then you extract a ball:

$$\text{Let } X_1 = \begin{cases} 1 & \text{if } 1^{st} \text{ ball is red} \\ 0 & \text{if " " " black} \end{cases}$$

You set up a replacement discipline whereby:

$\left\{ \begin{array}{l} \text{if you extract a red ball,} \\ \quad \text{you put in } r_1+1 \text{ red and } b_1 \text{ black balls} \\ \text{if you extract a black ball,} \\ \quad \text{you put in } r_1' \text{ red and } b_1'+1 \text{ black balls} \end{array} \right.$

Note: r_1, b_1, r_1', b_1' may be negative.



Then you perform successive extractions, following the above discipline.
For each extraction, you add so many red balls and so many black balls.

So, for example, if the first 3 extractions are red, the number of red balls is:

$$r \underbrace{\cancel{-1} + r_1 \cancel{+1}}_{1^{st}} \underbrace{\cancel{-1} + r_1 \cancel{+1}}_{2^{nd}} \underbrace{\cancel{-1} + r_1 \cancel{+1}}_{3^{rd}} \text{ extraction and discipline}$$

$$= r + 3r_1 \text{ red balls after } 3^{rd} \text{ extraction and discipline}$$

You can vary the discipline.
You can indicate the <u>spread of a disease</u> by saying that if the disease occurs (i.e., a black ball is extracted), you can make the disease more probable by adding more black balls.
And you can study how the disease progesses.

The case we will study in class (I don't want to spoil it now) is:

You extract one ball
You replace $c+1$ balls of the color extracted

That's the <u>Pólya Urn Model</u>

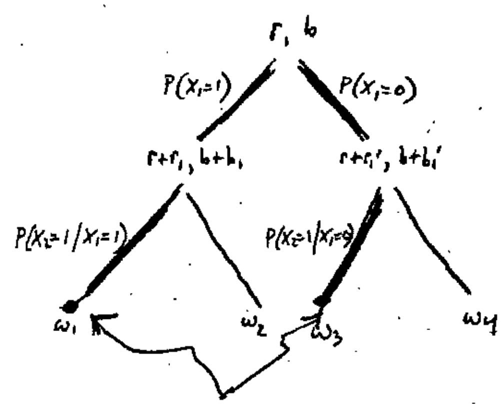
So this is the sample space.
A <u>succession of extractions</u>, with a very complex way of computing the probabilities.

Let $X_2 = \begin{cases} 1 & \text{if the } 2^{\text{nd}} \text{ ball extracted is red} \\ 0 & \text{if " " " " black} \end{cases}$

By the Law of Alternatives:

$$P(X_2=1) = P(X_2=1/X_1=0)\,P(X_1=0) + P(X_2=1/X_1=1)\,P(X_1=1)$$

or from the <u>probability tree</u>:



$P(X_2=1)$ = sum of these disjoint probabilities.

$$P(X_1=0) = \frac{b}{r+b} \qquad P(X_2=1/X_1=0) = \frac{r+r_1'}{r+r_1'+b+b_1'}$$

$$P(X_1=1) = \frac{r}{r+b} \qquad P(X_2=1/X_1=1) = \frac{r+r_1}{r+r_1+b+b_1}$$

In general computations of probabilities of these types do not simplify.
But we will see that in the Pólya Urn case, where we replace the extracted ball w/ $c+1$ of the same color, there is a simplification.

$$P(X_2=1) = \left(\frac{r+r_1'}{r+r_1'+b+b_1'}\right)\left(\frac{b}{r+b}\right) + \left(\frac{r+r_1}{r+r_1+b+b_1}\right)\left(\frac{r}{r+b}\right)$$

This can be rewritten as:

$$P(X_2=1) = \frac{br\,(r+r_1)\,(r+r_1')}{(r+b)^2\,(r+r_1'+b+b_1')(r+r_1+b+b_1)}$$

In the case of urn models, where you extract and replace balls w/ various disciplines, the conditional probabilities allow you to answer basic questions.
An example where you study the sample space by conditional probability.

- A few words about balls into boxes.

$\circ \ \circ \ \circ \ \ldots \ \circ$, k balls
— — — ... — n boxes

We've covered:

	Maxwell-Boltzmann	Bose-Einstein	Fermi-Dirac
balls	distinguishable	indistinguishable	indistinguishable, occupation number 0 or 1
boxes	distinguishable	distinguishable	distinguishable

Let's consider the case where:
- Balls – distinguishable
- Boxes – indistinguishable


k
n

A bunch of balls goes into a box.
Another bunch of balls goes into a box.

Clearly unoccupied boxes make no difference – the boxes are indistinguishable.
It makes no difference.

The count is which balls are in the same box.
You might as well assume that all the boxes are occupied.

You don't count the number of configurations.
You count the number of configurations in which all the boxes are occupied.
What is this?

a partition

A partition is a set of non-empty sets of balls belonging to the same class.
Placing distinguishable balls into indistinguishable boxes is the same as partioning the set of balls.
Because all you have for the data is which balls go together.

balls distinguishable
boxes indistinguishable (all boxes occupied) } partitions of the set of balls

Let's consider another case:
Balls – indistinguishable
Boxes – indistinguishable } both indistiguishable

k

n

You might as well assume all the boxes are occupied.
That means, to every box, there is an integer associated with it (i.e., the number of indistinguishable balls that occupy the box.

But the boxes are indistinguishable.
That means you have the set of integers, which is unordered and which add up to k.
This corresponds to:

partition of the integer k into $\leq n$ summands

Ex: $k=5 \Rightarrow$ partitions (the order of the summands doesn't matter)

5
4+1
3+2
3+1+1
2+2+1
2+1+1+1
1+1+1+1+1

} $\Sigma = 5$

indistinguishable balls
indistinguishable boxes, all boxes covered } partition of an integer

What about the situation where the <u>order of the summands</u> <u>matters</u>?

Let's take partitions of k (say $k=5$) that are ordered into (say 3) parts.

unordered partitions of size 3	if order mattered then:
3 + 1 + 1	3+1+1
2 + 2 + 1	1+3+1
	1+1+3
	2+2+1
	2+1+2
	1+2+2

$\left.\right\} 6$

Q/A: remember, all the boxes are occupied! You are not allowed 0.

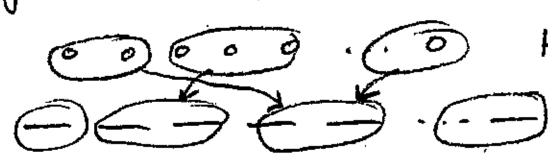
$\Uparrow$

what does this correspond to "balls into box"-wise?

Bose-Einstein statistics requiring at least one ball in each box.
All occupation numbers > 0.

$\left\{\begin{array}{l}\text{balls indistinguishable} \\ \text{boxes indistinguishable,} \\ \text{all boxes occupied}\end{array}\right\} \xrightarrow{\text{followed by}} \left\{\begin{array}{l}\text{balls indistinguishable} \\ \text{boxes distinguishable,} \\ \text{all boxes occupied}\end{array}\right\} =$ partition of integer k into <u>ordered</u> summands.

- <u>Most general case</u> – Balls into Boxes



k $\quad B$ set of balls

n $\quad \mathcal{U}$ set of boxes

given: partition π of the set B $\longleftarrow$ π = blocks of B, which are disjoint, whose union is B.
partition σ of the set $\mathcal{U}$. $\longleftarrow$ σ = likewise.

Informally:

Place the balls into the boxes.
Two balls belonging to the same block of the π partition are considered identical.
Two boxes belonging to the same block of the σ partition are considered identical.

How many ways are there of putting the balls into the boxes?

This is the most general case.
I know how to solve it, but I'm just too lazy to write it up.
There is a very simple formula for generating this.
If you want to work on this, you are welcome.
It's not hard.
It should be written up somewhere, but it isn't.

Balls in the same block are indistinguishable.
Balls in different blocks are partially distinguishable.

This is the <u>most general</u> balls into boxes problem, <u>if</u> placing the balls into boxes is a <u>function</u>.

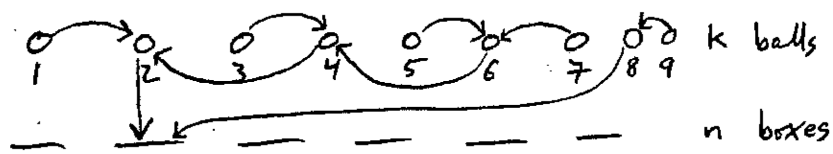
We've considered one case where placing the balls into boxes is <u>not</u> a function.
Disposition is <u>not</u> a function

↳ where we place balls into boxes and the balls in the box are <u>configured a certain way.</u>

Then all hell breaks loose.
You have a tremendous variety of placements that have to adhere to certain configurations.
You want to see an example?
One of the most famous examples is this:

• <u>Reluctant Functions</u> (This is pretty sophisticated)

<u>Occupancy Model</u>



Informally:

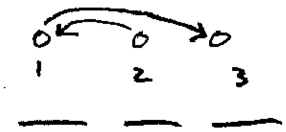
Take the 1st ball. It has a choice. Either it goes into a box, or it goes to another ball.
If it goes to a ball, wherever that ball goes, the 1st ball goes too.
This 2nd ball carries on its back the 1st ball.
Then this ball now decides – either it goes to another ball, or it goes into a box.
And so on.

Once a ball goes to another ball or into a box, it is no longer available.

Eventually, the balls have to drop into a box.
In other words, if you follow the history of any ball, it eventually drops (together w/ any balls it has gone to) into a box.

Q: if ball 1 goes to ball 3 and then ball 2 goes to ball 3, does ball 2 have an option where it goes when it piggybacks on ball 3?
A: No. The <u>balls are not ordered</u> when they go to other balls.

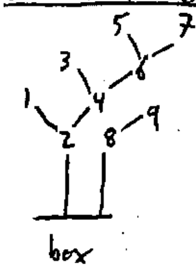
is the <u>same configuration</u> as



The above illustrates the occupancy model.
What would be the distribution model?

You view the balls in a box.
This corresponds to the balls going into a box and being configured in a certain way in the box.

The balls in the box form a forest of rooted trees



forest from our example

together, you can have several trees going into the same box.

There are a number of different ways of doing this that are very complicated.
But it's an extremely simple problem (if you can number it).
This is in a paper of mine.

The number of trees, we know.
I want the number of consequences of a reluctant function. There is a very simple formula.
I leave it to you as an exercise.
The total number is not that trivial, if you restrict occupied boxes to a single tree, rather than a rooted forest, then it becomes tough.

Remarkably enough, this does occur.
Put the balls in the boxes and configure them in a forest of rooted trees.

Disposition is a special case – the forest is restricted to one tree, one stump

Another special case that occurs very frequently is where the balls in each box are configured into several linearly ordered trees.

disjoint linear orders

Laguerre polynomials count the number of configurations for this special case of several linearly ordered trees.

We can go on forever with special cases of reluctant functions (i.e., restricting the configurations of forests to have certain properties).

The theory of Abel polynomials gives the formula for counting the number of configurations in the general case.

There's more to this than meets the eye.
You can make your living on this balls into boxes business.

Partitions

$\{\bullet \ \bullet \ \bullet \ \cdots \ \bullet\}$ a set of n elements

There are 2^n subsets

$\binom{n}{k}$ subsets with k elements

We can ask the same questions about partitions:

What is the total number of partitions of a set with n elements?

– and –

What is the total number of partitions with k blocks of a set with n elements?

The formula for the total number of partitions is very remarkable.

total number of partions:

B_{n+1} = Bell number
= number of partions of a set w/ n elements

$$B_{n+1} = \underbrace{\frac{1}{e}\left(\frac{1^n}{0!} + \frac{2^n}{1!} + \frac{3^n}{2!} + \cdots\right)}_{\text{Dobinsky's formula}}$$

(This is in a brown book, which is a collection of his papers.)

this is an integer!

How do you see that this is an integer intuitively?

This is a very remarkable formula. It's not clear that this is an integer.
And yet it gives the number of partitions of a set.
(Graham, Knuth, Patashnik have a closed form (asymptotically) for this – Ex 9.46)

When I was young, one of my first papers was to give a very simple proof of this.

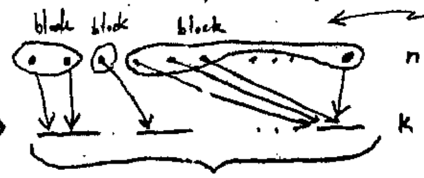
There are many other formulas for Bell numbers, but this is the only closed form formula.
We will see this later when we do the theory and the reasoning of the Poisson Process.

(It's very important to be almost right. It's more important to be almost right than right.
Why?
Because, if you are almost right, then you know it's stable under perturbation.
But if a small deviation makes a big difference, then your statement is unstable.)

- total number of partitions with k blocks ← (partition n things into k non-empty subsets.)

This can be computed by inclusion - exclusion

← partition n balls into <u>k blocks</u>



k boxes, <u>all occupied</u>

Balls into boxes
Balls are balls, boxes are boxes

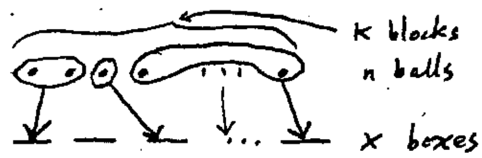
You take all functions of Maxwell - Boltzmann where <u>every box is occupied.</u>
Then you make the boxes <u>indistinguishable</u> by dividing by $k!$.
So the number of ways of placing these balls into boxes is:

1) all the boxes are occupied
2) divide by $k!$ to make the boxes indistinguishable

This is equal to:

$S(n,k)$ - Stirling Numbers of the Second kind

- $S(n,k)$ – Stirling Numbers of the second kind



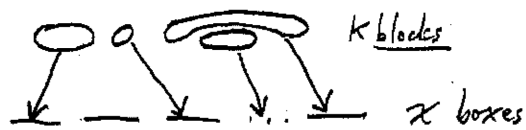
k boxes occupied

1) There are $S(n,k)$ ways to partition n balls into k blocks

$\}\ S(n,k)$

$\}$ balls into blocks

2) Now the blocks become balls.
Now we place balls into boxes.



$$(x)_k = x(x-1)(x-2)\cdots(x-k+1)$$

Lower factorial gives number of ways of placing blocks into boxes

$\}$ blocks into boxes

So we have:

$$\boxed{x^n = \sum_{k=0}^{n} S(n,k)\,(x)_k}$$

- Recall the difference of a function Δ:

$$\Delta f(x) = f(x+1) - f(x)$$

$$\Delta (x)_k = (x+1)_k - (x)_k$$

$$= \underbrace{(x+1)(x)\cdots(x-k+2)}_{(x)_{k-1}} - \underbrace{(x)(x-1)\cdots(x-k+2)}_{(x)_{k-1}}(x-k+1)$$

$$= \left[(x+1) - (x-k+1)\right](x)_{k-1}$$

$$\boxed{\Delta (x)_k = k\,(x)_{k-1}}$$

From $x^n = \sum_{k=0}^{n} S(n,k)\,(x)_k$, we have:

$$\Delta x^n = \sum_{k=0}^{n} S(n,k)\, k\, (x)_{k-1}$$

Multiple applications of Δ gives:

$$\Delta^j x^n = \sum_{k=0}^{n} S(n,k)\,(k)_j\,(x)_{k-j}$$

From this, we can show that:

$$\left[\Delta^k x^n\right]_{x=0} = k!\, S(n,k) \implies S(n,k) = \frac{\left[\Delta^k x^n\right]_{x=0}}{k!}$$

Where do Stirling numbers of the second kind come up?
All over the place. For example:
Recall the <u>differential operator</u> D , which is $\frac{d}{dx}$. Then we can show that:

$$(xD)^n = \sum_{j=0}^{n} \underbrace{S(n,j)}_{\text{Stirling number of the second kind}} x^j D^j$$

{ Combinatorial Identities
John Riordan
Wiley + Sons 1968 p. 218 }

<u>Proof by induction.</u>
Basis: $(xD)^0 = 1$

{ Note that:
$$(xD)(xDf(x)) = x\left(x'Df(x) + xD^2f(x)\right)$$
$$= xDf(x) + x^2D^2f(x)$$
$$(xD)^2 = xD + x^2D^2$$
}

Inductive step:

Assume $(xD)^n = \sum_{j=0}^{n} S(n,j)\, x^j D^j$

Then:
$$(xD)^{n+1} = (xD)(xD)^n$$
$$= (xD)\sum_{j=0}^{n} S(n,j)\, x^j D^j$$
$$= \sum_{j=0}^{n} S(n,j)\,(xD)\,x^j D^j$$

$$= \sum_{j=0}^{n} S(n,j)\, x\left[j\, x^{j-1} D^j + x^j D^{j+1}\right]$$

$$= \sum_{j=0}^{n} S(n,j)\left[j\, x^j D^j + x^{j+1} D^{j+1}\right]$$

$$= \sum_{j=0}^{n} \Big(j\, S(n,j)\, x^j D^j + \underbrace{S(n,j)\, x^{j+1} D^{j+1}}_{S(n,j-1)\, x^j D^j}\Big)$$

$\longleftarrow$ change of variable $j \to j-1$, $j+1 \to j$

$$= \sum_{j=0}^{n} \underbrace{\left[j\, S(n,j) + S(n,j-1)\right]}_{S(n+1,j)} x^j D^j$$

$\longleftarrow$ identity for Stirling numbers of the second kind.

$$(xD)^{n+1} = \sum_{j=0}^{n} S(n+1,j)\, x^j D^j \quad \checkmark$$

- Note that:

$$(xD)\,x^k = x\,k\,x^{k-1} = k\,x^k$$

$$(xD)^2 x^k = (xD)\big((xD)\,x^k\big) = (xD)\,k\,x^k = x\,k\,k\,x^{k-1} = k^2 x^k$$

And we have:

(*) $$(xD)^n x^k = k^n x^k$$

as well as:

(**) $$x^j D^j x^k = (k)_j\, x^k$$

$$\left\{ x^j \left(D^j x^k\right) = x^j\,(k)_j\, x^{k-j} = (k)_j\, x^k \right.$$

Taking the original identity:

$$(xD)^n = \sum_{j=0}^{n} S(n,j)\, x^j D^j$$

multiply both sides by x^k:

$$\underbrace{(xD)^n x^k}_{\substack{\text{from } (*), \\ k^n x^k}} = \sum_{j=0}^{n} S(n,j)\, \underbrace{x^j D^j x^k}_{\substack{\text{from } (**), \\ (k)_j\, x^k}}$$

Which gives the combinatorial identity involving the Stirling numbers of the second kind:

$$k^n x^k = \sum_{j=0}^{n} S(n,j)\,(k)_j\, x^k$$

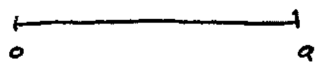
This is an <u>analytic</u> derivation.

<u>All</u> the formulas of $\left\{\begin{array}{l}\text{random variables,} \\ \text{expectation,} \\ \text{probability}\end{array}\right\}$ can be obtained as <u>purely probabilistic</u> formulas, in addition to the analytic derivations.

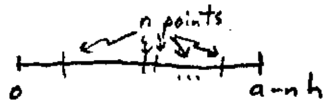
John Guidi
guidi@mit.edu

Needles on a stick (again)

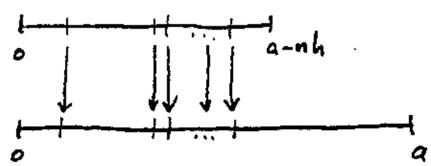
A one liner.



We have the interval $[0,a]$.
I take another interval of length $a-nh$.



On this other interval, I pick n points at random.
Then, for each point, I take the interval $[0,a]$ and select the same point.



What is the number of ways this can be done?

$$\frac{(a-nh)^n}{a^n}$$

$(a-nh)^n$ ← {number of ways of dropping n points on the interval $[0, a-nh]$}

a^n ← {number of ways of dropping n points on the interval $[0,a]$}

Note: I said number of ways when, in fact, it's a measure.

I wish we had time to work this in 2-dimensions.

Geometric Probability

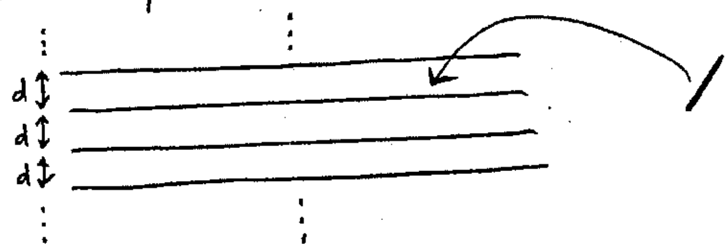
From the text:

"The theory of geometric probability is concerned w/ random points on the plane, spheres, and other geometric objects, as well as with random lines, circles, and other figures."

Let's start w/ the most famous problem of Geometric Probability, which is the Buffon Needle Problem.

Buffon Needle Problem

We draw an infinite number of straight lines on the plane, parallel and distance d apart.



You have a needle of length l.
You take the needle and drop it on the plane at random.

what is the probability that the needle will intersect at least one of the lines?

This can be solved many ways.
It can be solved by the continuous Law of Alternatives trigonometrically:

- You drop the center of the needle, then you rotate the needle. And you compute the probability of intersection by the continuous Law of Alternatives.

But now, let's give a completely probabilistic solution, using no computations.
This is a test of your understanding of probability.
It's almost pure handwaving.

Let A_l = event that needle meets at least one of the straight lines.

X_l = number of straight lines the needle meets when dropped at random.
↑ random variable

Now, we consider the expectation of X_l:

$$E(X_l) = \sum_n n P(X_l = n)$$

$$= 0 \cdot P(X_l = 0) + 1 \cdot P(X_l = 1) + 2 \cdot P(X_l = 2) + \dots$$

if the needle is short $(l < d)$:
In this case, the needle can meet at most 1 line.

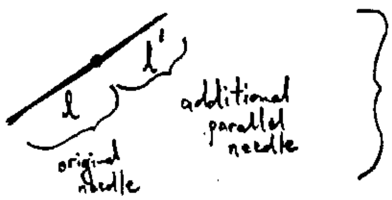
$$E(X_l) = 0 \cdot P(X_l = 0) + 1 \cdot P(X_l = 1)$$
$$= P(X_l = 1) \leftarrow (X_l = 1) \text{ is just the event } A_l$$
$$= P(A_l)$$

Therefore, by this reasoning, we have reduced the computation of the probability of an intersection of a straight line to the computation of the expectation.
This is a very frequently used technique in probability.
Expectations are easier to compute than probabilities.

Why?
Because the expectation of the sum of two random variables is the sum of the expectations! No questions asked. [2/25/98.8-9]

How do we exploit this fact?
Like this.

Take the original needle of length ℓ.
Then tack on an additional parallel needle of length ℓ'



The resulting needle has length $\ell+\ell'$

Now, I claim that the number of times that the needle of length $\ell+\ell'$ meets a straight line is the number of times the first segment meets a straight line, plus the number of times the second segment meets a straight line.
If you don't see that, I can not explain it.

$$X_{\ell+\ell'} = X_\ell + X_{\ell'}$$

Therefore:

$$E(X_{\ell+\ell'}) = E(X_\ell) + E(X_{\ell'})$$

Now, just for psychological reasons, let's set:

$$f(\ell) = E(X_\ell)$$

which gives:

$$f(\ell+\ell') = f(\ell) + f(\ell')$$

Now we ask, which kinds of functions satisfy this condition?
Constant times ℓ.
Therefore,

$$f(\ell) = c\,\ell$$

$$f(\ell+\ell') = f(\ell) + f(\ell')$$
$$c(\ell+\ell') = c\ell + c\ell'$$

universal constant c,
independent of ℓ,
but depending on d (the distance between straight lines).

$$E(X_\ell) = c\,\ell$$

- Now we use the Charles Lamb principle.
He wrote a short story - The Discovery of Roast Pork.
How was roast pork discovered?
Very simply.
One day, in a hamlet in England, a house accidentally burned down.
There were pigs in it. They were roasted.
Someone accidentally tasted the meat and liked it.
It was fantastic. It was terrific.
So what did they do?

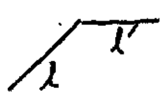
They started stuffing pigs into houses and burning the houses down.
They liked the meat.

Until one day, a great genius came along and said:
You don't have to put them into the houses.
This is how roast pork was invented.

This is the method of scientific discovery

- We're going to go through this crisis right now.
Now I say, who says the needles being combined need to be parallel to each other.
They could be crooked and the same conclusion holds

that's the roast pork part.

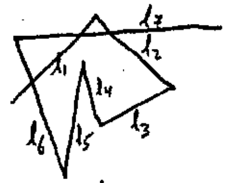


(provided the needles are rigid, of course)
The above argument works just the same for this situation.
Exactly the same conclusion.

Because if you took a crooked needle of length $\ell+\ell'$, the number of times this crooked needle meets any of the straight lines on the plane is the same as the number of times it meets on the stretch of length ℓ + the number of times it meets on the stretch of length ℓ'.
We still have:

$$X_{\ell+\ell'} = X_\ell + X_{\ell'}$$

- Proceeding along this line, we can do something like this:



needle

$$X_{l_1+l_2+\dots+l_n} = X_{l_1} + X_{l_2} + \dots + X_{l_n}$$

And it's <u>still true</u>.

$$E(X_l) = cl$$

- You can pass to the <u>limit</u>:

length l

Any rigid thing like this, you drop it.
The expected number of straight lines it meets is <u>proportional to its length</u>.
For <u>any curve</u>, whatsoever

$$E(X_l) \propto \overbrace{\text{Length of curve}}^{l} = c\,l$$

The universal constant c, in all the above, is the <u>same</u>.
It's <u>not</u> the probability, but the <u>expectation</u>.

$$E(X_l) = c\,l \quad \text{for } \underline{\text{any}} \text{ curve, whatsoever.}$$

- What we are <u>missing</u> is the <u>value</u> of the <u>constant c</u>.

We need to drop the most favorable shape that will reveal the constant c.
So we drop the needle (i.e., curve) of that shape.
So I take this needle.

circle w/ circumference πd

No matter how you drop each rigid needle (circle), it <u>always</u> intersects the straight lines <u>twice</u>.

So $E(X_{\pi d}) = 2$

$$E(X_{\pi d}) = 2$$
$$= c(\pi d) \quad \longleftarrow \text{ since } E(X_\ell) = c\,\ell$$

$$\Rightarrow \boxed{c = \frac{2}{\pi d}}$$

And we have the universal constant.

So, for any curve whatsoever,

(*) $$\boxed{E(X_\ell) = \frac{2}{\pi d}\,\ell}$$

• Now, we go back to the straight curve, where we have a short needle ($\ell < d$). We had shown [SC 4/10/98.2]:

short, straight needle ℓ

$$E(X_\ell) = P(X_\ell = 1) = P(A_\ell)$$

And we have formula (*) above, compliments of the universal constant:

$$E(X_\ell) = \frac{2}{\pi d}\,\ell$$

Combining the above, we have that:

$$\underbrace{P(A_\ell) = \frac{2}{\pi d}\,\ell}$$

This is an experimental method for computing π.
You drop needles.

[Q: What happens w/ longer straight needles that can intersect more than one straight line in the plane?
A: I've never worked it out.]

- This argument can be adapted to give a much more general result.

A <u>convex set</u> in the plane is a set that has <u>no bumps</u>.

(points on this piece of segment $\overline{xy} \notin B$)

Convex　　　　　　~~NOT~~ CONVEX

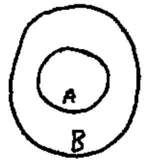
Convex is an elegant way of saying no bumps.
A nice way of saying:

given <u>any</u> two points x and y in the set,
if you join these two points by a straight line,
the <u>entire segment</u> is included in the set.
} Convex set

This is just an elegant way of excluding bumps.

- <u>Sylvester's Theorem</u>

Example 1: You take 2 convex sets A and B in the plane, one inside the other. Big set B, small set A.



$A \subseteq B$

{Notes for this example, sets A+B need not be convex}

Suppose I pick a point arbitrarily in B.
Then I ask, <u>given</u> the condition that I've picked a point in B, what is the probability that the point also belongs to A?

The answer is obvious. It's the ratio of the area of A divided by the area of B.

$$P(x \in A \mid x \in B) = \frac{P((x \in A) \cap (x \in B))}{P(x \in B)}$$

$$= \frac{P(x \in A)}{P(x \in B)}$$

$$P(x \in A \mid x \in B) = \frac{\text{area}(A)}{\text{area}(B)}$$ } See commentary on [SC 4/10/98.9-10]

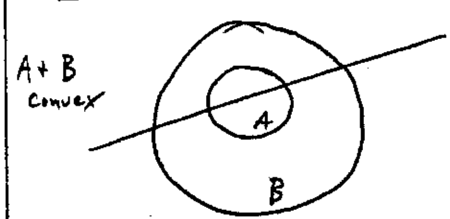
We implictly use the concept of a <u>measure</u> here. <u>Measures</u> can provide <u>conditional probabilities</u> but <u>not</u> probability of basic events. e.g., not $P(x \in A)$

Example 2:

- Now, what's the next best thing in the world after a point?
  A line.

  Now, instead of picking a point at random, I'm picking a line at random.
  I ask the same question.
  Note - for this example, the sets $A + B$ must be convex.

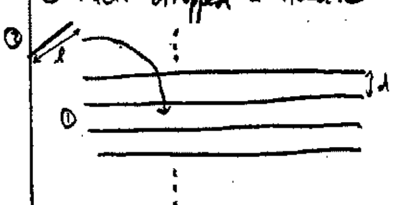
  For 2 convex sets A, B, where $A \subseteq B$,
  I pick a line at random.
  Assuming that the line intersects B, what is the probability that it also intersects A?



  This can also be done by pure handwaving.

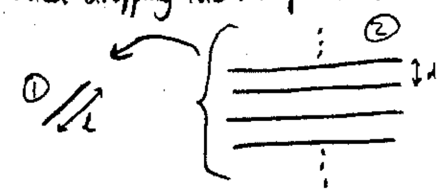
- Let's go back and analyze what we really did in the Buffon Needle Problem.
  What is really going on.

① We drew straight lines
② Then dropped a needle

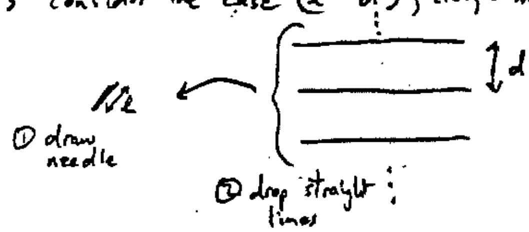


But this is statistically the same as:

① drawing the needle
② Then dropping the straight lines



From this analysis,
let's consider the case $(\ell < d)$, i.e., where the needle is of sufficiently small size.



event $\underbrace{A_\ell = \text{needle meets at least one straight line}}$

this is the same as saying:
event that a straight line intersects the needle

So,

$$P(\text{a straight line intersects the needle}) \propto l \quad \Leftarrow \text{length of the needle}$$

$$P(\text{a straight line intersects the needle}) = P(A_l) = \frac{2}{\pi d}\, l \;\propto l$$

- How do we get back to Sylvester's problem?
  Let's ask the following question:

  What does it mean to pick a random line?

  Let's go back and see what we did in Example 1, where we picked a random point in the plane.
  How did we interpret the process of picking a point at random, in terms of picking points?
  We said that we pick a point in the plane.
  That's not a uniform probability.
  It's a measure, called the area, which is infinite
  So we implicitly use the concept of a measure when picking the points.

  Then we took the ratio of two measures, and that gave us the probability.
  So we need the notion of a measure.

- Measure

  What's a measure?
  A measure is something that has all the properties of probability, except that unlike probabilities which lie between $[0, 1]$, measures lie between $[0, \infty)$.

  Measure μ – Properties

  - $\mu(A) \geq 0$
  - $\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B)$
  - $\mu(\phi) = 0$

  But $\mu(\Omega) \not\equiv 1$ $\Longleftarrow$ unlike probability, where $P(\Omega) = 1$

  - Countable additivity $\quad A_1, A_2, \ldots$ is a finite or infinite sequence of pairwise disjoint e

    $\mu(\bigcup_n A_n) = \sum_n \mu(A_n)$

example: area(A) is a <u>measure</u> in the plane

example: length(A) is a <u>measure</u> on the line

- You get <u>probabilities from measures</u> by taking <u>conditional probabilities</u>.
  If you have a <u>measure</u>, instead of probability, you can <u>only</u> talk about <u>conditional probability</u>.
  You can <u>not</u> talk about the probability of a basic event (i.e., $\cancel{P(A)}$).
  This is a little tricky, but that's the way it is.

- Let me tell you a little secret.
  One of the great <u>unsolved problems of statistics</u> is the fact that in Bayes' Law, you can use, as the <u>prior</u>, <u>measures</u> instead of probabilities.

$$P(A|B) = \frac{P(B|A)\,\cancel{P(A)}}{P(B)} \xleftarrow{\text{prior}} \mu(A) \qquad \triangleq \frac{\mu(A)}{\mu(B)}$$

conditional probability in terms of measures

For example, instead of tossing a coin (i.e., $P(A) = p$), we have a random variable on the whole straight line and we estimate Bayes by taking measurements.

If you can measure the whole straight line, you have to take, as <u>prior</u>, the <u>measure on the whole line</u>, because it can be anywhere.
<u>Without</u> taking probabilities.

People have realized this since 1939, when a book by Sir Harold Jeffries appeared, trying to justify why this is.
Why can you take as prior a measure instead of a probability?
<u>No one</u> has come up w/ a <u>suitable justification.</u>
Do you want to become famous? Justify this.

So measures occur as priors in Bayes' Law.

Q: What utility is there in taking the measure as a prior?
A: You want a uniform prior.
You are estimating a random variable (e.g., $P(X=n)$). Suppose you don't know the whole distribution.
You can use <u>measures</u>.
Note that the <u>posterior</u> is a <u>probability</u>: $P(A|B) = \dfrac{\mu(A)}{\mu(B)}$
even if the prior is a measure.

- Back to our reexamination of Example 1.
  In answering the question about picking points in A and B, we implicitly used the fact that the right probability to use is the ratio of two measures.

$$P(x \in A \mid x \in B) = \frac{\text{area}(A)}{\text{area}(B)} \longleftarrow \frac{\mu(A)}{\mu(B)}$$

• Now, we have the same problem w/ lines.
Now, we have to visualize the sample space Ω as the set of all straight lines in the plane.

Ω = all straight lines in the plane

To answer the question posed in Example 2, we need a measure.
But this measure is not any old measure.
This measure has to be independent of position, or invariant of position, as they say.

And it's remarkable that there is one, and only one, measure on a set of straight lines, which is just like the analogue of area.
It's easier than you think.

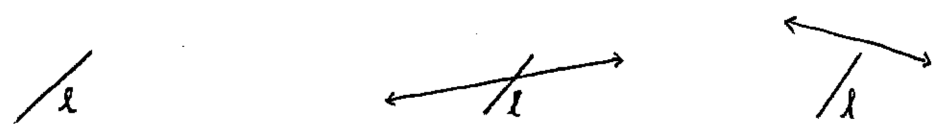
Is there an invariant measure on the set of all straight lines in the plane?

Yes. There is.
And it is unique, except for normalization factors.

{it depends on what you call measure = 1
With area, for example, it's up to you
what you call area = 1.}

Why?

• We use exactly the same reasoning as we did in the case of the Buffon Needle Problem.
Now our sample space Ω, remember, is the set of all straight lines.
A sample point $\omega \in \Omega$ is a straight line.



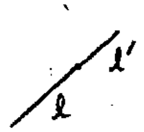
We've already picked our needle, so we picked a segment of length ℓ.

event straight line meets ℓ

event straight line does not meet ℓ

Now we consider the event that a straight line line meets ℓ.
What is the measure of that event?

First, we note that if we tack on ℓ', $\ell' \parallel \ell$, then the measure is the sum.



Let B_ℓ = event that a straight line meets a segment of length ℓ

Then we take as the <u>measure</u>:

$$\mu(B_\ell) = \underset{\uparrow}{c}\,\ell \leftarrow \text{length}$$

$\left\{\begin{array}{l}\text{the universal constant } \frac{2}{\pi d} \\ [4/10/98.5\text{-}6]\end{array}\right\}$

Every straight line either meets ℓ, or it doesn't meet.
And note that this measure is <u>independent</u> of the <u>position of ℓ</u>.

So this gives you the measure of several events, namely, the event that a straight line intersects ℓ.
That's enough. (There is another unsolved problem here, which I'll tell you about later).(*)

{ All you need is that. You <u>don't want</u> the <u>formula</u> for the <u>measure</u> of a <u>set of straight lines</u>.
All you need is the <u>measure</u> of a <u>set of straight lines</u> that $\underbrace{\text{meet a given segment } \ell}_{\text{this is the important part}}$. }

So we have the measure $\mu(B_\ell) = c\,\ell$

- Now, we use <u>exactly</u> the same argument as before, w/ the Buffon Needle Problem.


S

If I take <u>any</u> shape whatsoever, with length (s) = ℓ,
then define random variable:

$$X_\ell(\omega) = \begin{cases} 1 & \text{if straight line } \omega \text{ meets shape } S \\ 0 & \text{otherwise} \end{cases}$$

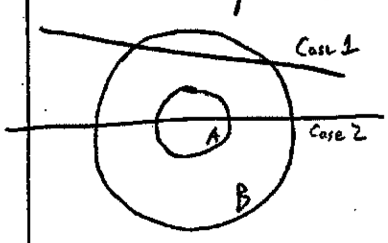
Then $X_\ell(\omega)$ is the <u>indicator</u> random variable where a straight line meets the shape.

And using the same argument presented in [SC 4/10/98.2-6], we have the expectation:

$$E(X_\ell) = c \cdot \ell$$

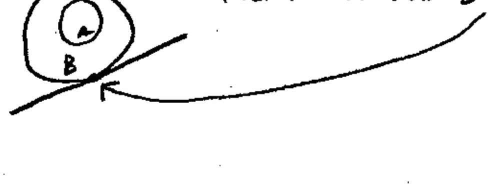
average of times the shape S, of length ℓ, is intersected.

- Back to example 2 [SC 4/10/98.8], where I have two convex sets A and B, with $A \subseteq B$.



Since I have a convex set, a line that intersects B intersects A either twice (case 2) or not at all (case 1).

We <u>ignore</u> the <u>extreme case</u> of <u>Probability = 0</u> that the line intersects B at a single point

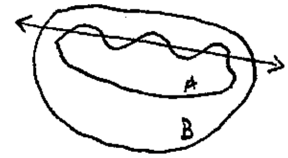


Therefore, the probability that the straight line that meets B also meets A equals:

$$P\left(\text{line meets } A \,\middle|\, \text{line meets } B\right) = \frac{\mu(B_A)}{\mu(B_B)}$$

$$= \frac{\cancel{2}\cdot \text{perimeter}(A)}{\cancel{2}\cdot \text{perimeter}(B)}$$

We take, as measure of a convex set, twice its perimeter. Note that we require convex sets to ensure that any line intersects twice or not at all (ignoring, again, the extreme case w/ probability = 0, of a single intersection).
For example, we want to avoid the pathological case:



etc.

$$\boxed{P\left(\text{line meets } A \,\middle|\, \text{line meets } B\right) = \frac{\text{perimeter}(A)}{\text{perimeter}(B)}}$$

This gives you a beautiful way of computing perimeters.

This gives you a practical way of computing perimeters of weird convex sets.
You take the average number of lines that meet it, within a known convex set B.

$$\text{perimeter}(A) = \overbrace{P\left(\text{line meets } A \,\middle|\, \text{line meets } B\right)}^{\Downarrow}\,\text{perimeter}(B)$$

This is the beginning of tomography (tomography is done in 3-dimensions).
An object has a smaller object inside, the probability that a straight line that meets the outer object will also meet the inner object is the ratio of their surface areas. By the same argument really.

You can read my recently published book on the subject, published in 1950.

- The famous <u>unsolved problem</u> I refered to (*):
It's been around quite a while.

$$\Omega = \text{all straight lines in the plane}$$

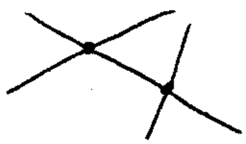
Let's define:

$$I_{A(\omega)} = \begin{cases} 1 \text{ if line } \omega \text{ meets } A \\ 0 \text{ otherwise} \end{cases}$$

For indicator random variables of <u>events</u> A and B, we've shown that (Text p. 90-91):

$$\underbrace{I_A + I_B}_{\text{events}} = I_{A \cup B} + I_{A \cap B} \quad \Big\} \text{ due to Boolean algebra}$$

For the <u>indicator random variables associated with lines</u>, you have <u>other identities</u>, besides this one.
When you are dealing with lines, you have <u>additional identities</u> that come in <u>because</u> of <u>geometry</u>.

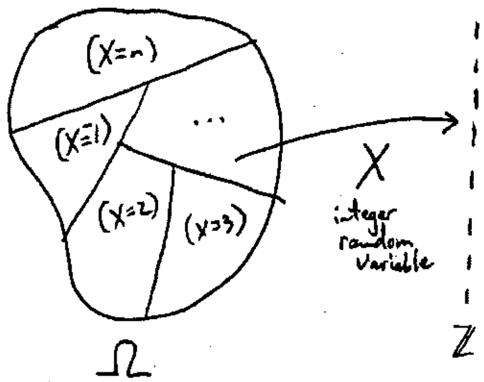


No one has been able to describe all these fundamental geometric identities.
This problem is formally described in my book.

The algebra of indicator random variables for lines is something <u>more</u> than Boolean algebra, due to the geometry.

Entropy and Information

One of the most beautiful chapters in probability.
It's too bad I can't cover it in the lectures. There's only so much one can do.

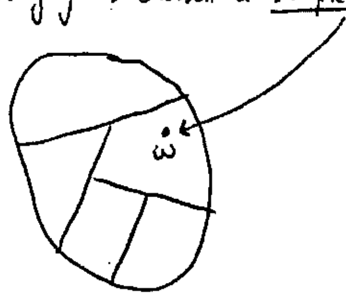


Suppose we have sample space Ω and an integer random variable X.

As we briefly mentioned [SC 2/27/98.10-11], you can have the situation where the sample space is partitioned into blocks.

The events $(X=n)$, n an integer with $P(X=n) > 0$, are a partition of the sample space Ω if they disjointly cover Ω ← strictly

This gives us another way of looking at random variables.
We view a random variable as a question.
A bad guy has choosen a sample point ω.



We try to determine the unknown sample point by asking questions.
The answers to our our questions will tell us which block the chosen sample point belongs in.
The more questions we ask, the more we narrow down where the element, chosen by the bad guys, is.

Thus, a random variable is a question that delivers a certain amount of information on the search problems of finding particular sample points.

Entropy is a numerical measurement on the amount of information that a random variable has.

Let's see what the properties of such information ought to be.

Entropy is a numerical measure of the amount of information in a partition π of Ω

{ At this point, we will drop reference to random variable and replace it w/ partition. Because all that matters is the partition. }

First, I give you the formula:

For $\pi = \{\underbrace{B}_{\text{Block}} : B \in \pi\}$, ← { any block has positive probability $(P(B) > 0,\ B \in \pi)$; intersection of any 2 blocks has probability 0 $P(B \cap B') = 0$, $B, B' \in \pi$, $B \neq B'$; probability of union of all the blocks is 1 $P\left(\bigcup_{B \in \pi} B\right) = 1$ }

$$\underbrace{H[\pi]}_{\text{entropy of a partition}} \overset{\triangle}{=} \sum_{B \in \pi} P(B) \log_2 \frac{1}{P(B)}$$

↑ by definition

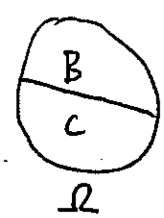
Q: Why $H[\pi]$?

A: Because Boltzmann called it H.
This formula is sculpted on Boltzmann's grave.
I'm not kidding.

Now you say: This is weirdissimo. Where did you get this? The purpose of this lecture is to show you this is not weirdissimo. It's absolutely natural.

The units of entropy are called bits ← after Prof. Shannon
bits = binary digits

- Example



Take the partition of two blocks with:

$$\pi_0 = \{B, C\}, \quad P(B) = \tfrac{1}{2} \implies P(C) = \tfrac{1}{2}$$

The entropy of this partion is:

$$H[\pi_0] = P(B)\log_2 \frac{1}{P(B)} + P(C)\log_2 \frac{1}{P(C)}$$

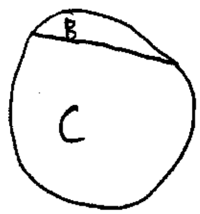
$$= \frac{1}{2}\underbrace{\log_2 \frac{1}{\left(\frac{1}{2}\right)}}_{1} + \frac{1}{2}\underbrace{\log_2 \frac{1}{\left(\frac{1}{2}\right)}}_{1}$$

$H[\pi_0] = 1$ bit } Therefore, a partition of 2 blocks w/ equal probability has 1 bit of information.

The simplest question is answered Yes or No w/ probability $\frac{1}{2}$.

Now you say, what if we had two blocks w/ different probabilities?

• Example



$$\pi_1 = \{B, C\} \quad \text{with } P(B) = \varepsilon \implies P(C) = 1 - \varepsilon$$

$$H[\pi_1] = \varepsilon \log_2 \frac{1}{\varepsilon} + (1-\varepsilon) \log_2 \frac{1}{(1-\varepsilon)}$$

What happens as $\varepsilon \to 0$?

$$\lim_{\varepsilon \to 0} H[\pi_1] = \cancel{\varepsilon}^{0} \log_2 \frac{1}{\varepsilon} + (1-\varepsilon) \cancel{\log_2 \frac{1}{(1-\varepsilon)}}^{0}$$

$$\underbrace{\lim_{\varepsilon \to 0} H[\pi_1] = 0}$$

In other words, a partition w/ one teeney weeney block and one big block has very small entropy.

In particular, if you take the partition w/ a single block of probability 1,

$$\pi = B \; , \; P(B) = 1$$

$$H[\pi] = P(B) \log_2 \frac{1}{P(B)}$$

$$= 1 \log_2 1$$

$$= 0 \text{ bits}$$

the entropy is 0.

It doesn't give you any information as to where the unknown sample point is.

• Now we come to the fundamental properties of entropy:

Property (1) if π and π_2 are independent partitions then

{independent = every block of π independent of π_2}

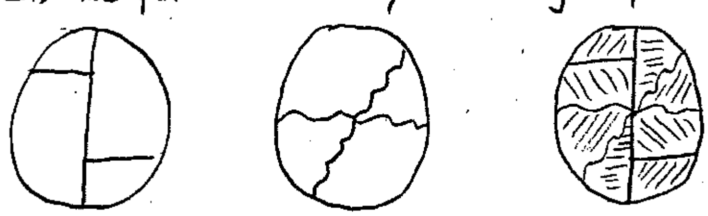
$$H[\pi \wedge \pi_2] = H[\pi] + H[\pi_2]$$

meet

Meet for partition is the analogue of intersection for sets.

What is $\pi \wedge \pi_2$?

Its the partition obtained by intersecting every block of π w/ every block of π_2.



π $\qquad$ π_2 $\qquad$ $\pi \wedge \pi_2$

Meet gives you a partition that is finer.

- Let's verify this.
This is the crucial computation of this theorem.

Suppose $\pi = \{B_i\}$, $\pi_2 = \{C_j\}$

Then

$$H[\pi \wedge \pi_2] = \sum_{B,C} P(B \cap C) \log_2 \frac{1}{P(B \cap C)} \quad \Bigg\} \text{ This is for the meet of } \underline{\text{any}} \text{ two partitions.}$$

Sum over all B's and C's

- The blocks of the meet $\pi \wedge \pi_2$ are the intersections of the blocks.
When considering all pairs of blocks in the cartesian product $B \times C$, blocks that do not intersect have $B \cap C = \phi$. In which case:

$$P(B \cap C) \log_2 \frac{1}{P(B \cap C)} = \underbrace{0 \log_2 \frac{1}{0}}_{=0 \text{, by convention.}}$$

- Since the partitions are <u>independent</u>, then blocks B+C are independent.

$$P(B \cap C) = P(B)P(C)$$

$$= \sum_{B,C} P(B)P(C) \log_2 \frac{1}{P(B)P(C)}$$

$$= \sum_{B,C} P(B)P(C)\left(\log_2 \frac{1}{P(B)} + \log_2 \frac{1}{P(C)}\right)$$

split into 2 sums:

$$= \sum_{C \in \pi_2} P(C) \underbrace{\sum_{B \in \pi} P(B) \log_2 \frac{1}{P(B)}}_{H[\pi]} \quad \Bigg\} \Leftarrow \sum_{B,C} P(B)P(C) \log_2 \frac{1}{P(B)}$$

$$+ \sum_{B \in \pi} P(B) \underbrace{\sum_{C \in \pi_2} P(C) \log_2 \frac{1}{P(C)}}_{H[\pi_2]} \quad \Bigg\} \Leftarrow \sum_{B,C} P(B)P(C) \log_2 \frac{1}{P(C)}$$

$$= \overbrace{\sum_{C \in \pi_2} P(C)}^{1} H[\pi] + \overbrace{\sum_{B \in \pi} P(B)}^{1} H[\pi_2]$$

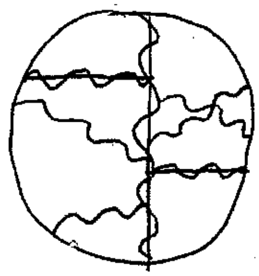
$$H[\pi \wedge \pi_2] = H[\pi] + H[\pi_2] \checkmark$$

Q/A: This has to be a partition w/ a countable number of blocks.
It can not be a continuous partition.
This is one case where the integral is not the limit of sums.
Very weird.

- Recall their is a partial order on partitions.
The partial order of finer:

$\pi \le \pi_3$ means π is finer than π_3

(i.e., every block of π is contained in some block of π_3)
[SC 2/27/98.11] (Take the blocks of π_3 and cut them up to get π.)



$\sim\sim = \pi$
$— = \pi_3$

Property (2) if $\pi \le \pi_3$

$$H[\pi] \ge H[\pi_3]$$

The finer the partition is, the bigger the entropy.

If you take a finite sample space, the maximum entropy is for the partition where each sample point is a block (i.e., the finest partition).

Property (3) $H[\pi \wedge \pi_2] \le H[\pi] + H[\pi_2]$

This is true in general, even if π and π_2 are not independent.
You have equality only if π and π_2 are independent.

- These 3 properties characterize entropy completely.

Q: Is there a property like inclusion-exclusion, something like that for entropy?
A: I wish we did.
People have tried. Very hard.
That's a very good question.

- Maximum Entropy Principle

There is a discrete version, which we are about to see.
There is a continuous version, which is much deeper.

The discrete version is this: (i.e., fix the number of blocks)

Among all partitions with n blocks, the ones with maximum entropy are the one's for which all blocks have probability $\frac{1}{n}$.

This can be proved and if we have time I'll prove it later.

$P(B) = \frac{1}{n}, \forall B \in \Pi$
all blocks have equal probability

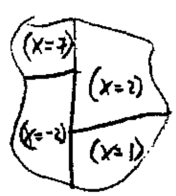
There is a deep interpretation of the maximum entropy principle.
If you don't know anything about the distribution of the random variable that takes n values, then you assume that all values are equally likely.
You assume maximum entropy.

$$\underset{\text{block}}{B_i} = \underset{\text{integer random variable}}{(X = i)}, \text{ for } n \text{ values of } i \Bigg\} \; \Pi = \{B_i\} = \{(X=i)\}$$

for all non-null events $(X=i)$

$$P(B_i) = \frac{1}{n} \implies \underbrace{P(X=i) = \frac{1}{n}}_{\text{maximum entropy}}$$

Let X = integer random variable
Π is the partition defined by X
(i.e., $\Pi = \{(X=i), \forall i \text{ s.t. } (X=i) \neq \phi\}$)

Π 

Define $H(X) = H[\Pi]$

People like to talk about entropy of random variables, even though entropy is a property of partitions.

Similary, we have the joint entropy of 2 integer random variables X and Y:

$$H(X+Y) = H[\underset{\text{partition of } X}{\Pi} \wedge \underset{\text{partition of } Y}{\Pi_4}]$$

So the random variable maximum entropy is the one of the uniform distribution. Provided the random variable takes n values.

- Now we play 20 questions.
  I think of a number between 0 and $2^n - 1$.
  And you have to guess it.
  The question discipline is that you can only ask Yes or No questions.
  (Usually the questions are restricted).

  What is the minimum number of questions?

- First, we'll do it abstractly, even though we know what the answer is.

  We can write the number in binary digits.

$$\underbrace{0110\ \ldots\ 110}_{n \text{ binary digits}}$$

The number has at most n binary digits (number $\in [0, 2^n - 1]$).
So you can guess the number in n questions.

1) Is the 1^{st} digit 1? A: No $\to 0$
   Yes $\to 1$

2) Is the 2^{nd} digit 1?

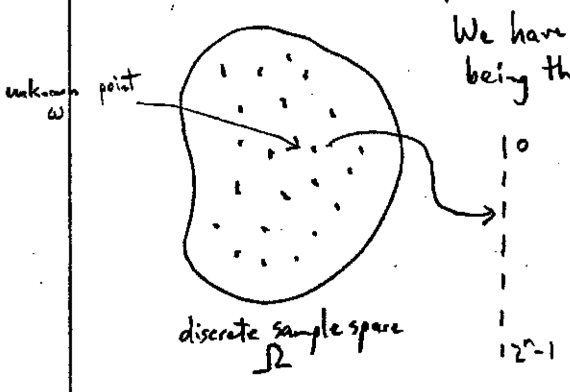
⋮

n) Is the n^{th} digit 1?

$$\Longrightarrow \overbrace{0110\ldots110}^{n \text{ digits}}$$

A_n=No ... A_2=Yes A_1=No

- Now let's do it over, jazzed up.
  We repeat this in purely probabilistic terms.



We have a sample space of size 2^n, with sample points being the numbers 0 through $2^n - 1$.

(i.e., no information)

Since we have no prior, each sample point has probability:

$$P(\omega) = \frac{1}{2^n}$$

A question is an arbitrary random variable over the number of entries.
The answer, with the unkown point ω, tells you which block the unknown point is in.

What does it mean to find the point w/ questions?
Well, you have questions $\pi_1, \pi_2, \ldots, \pi_k$ (i.e., partitions)

To find a point means that the meet of these partitions is a trivial partition with one block with one point.
That's what it means to find a point.

So the problem of search is to find the minimum number of partitions whose meet is the trivial partition.
Let's call the trivial partition $\hat{0}$.

find min K such that:

$$\pi_1 \wedge \pi_2 \wedge \ldots \wedge \pi_K = \hat{0}$$ read "zero hat"

But, we have that:

$$H[\hat{0}] = H[\pi_1 \wedge \pi_2 \wedge \ldots \wedge \pi_K]$$

and from the inequality from property 3 above:

$$\leq H[\pi_1] + H[\pi_2] + \ldots + H[\pi_K]$$

Let's calculate the entropy of the trivial partition.

$$H[\hat{0}] = \sum_{B \in \hat{0}} P(B) \log_2 \frac{1}{P(B)}$$

$$= \sum_{2^n} P(\omega) \log_2 \frac{1}{P(\omega)}$$

$$= \cancel{2^n} \left(\frac{1}{\cancel{2^n}}\right) \log_2 \frac{1}{\left(\frac{1}{2^n}\right)}$$

$$\boxed{H[\hat{0}] = n}$$

And, therefore:

$$n \leq H[\pi_1] + H[\pi_2] + \ldots + H[\pi_K]$$

So whatever question search scheme you devise, it has to consist of partitions, or random variables, whose entropies add up to at least n.

This is a constraint.
You can not guess the point unless the entropy adds up to n.

Preferably, you want to make the questions (partitions) independent, because you save entropy:

$$n = H[\hat{0}] = H[\pi_1 \wedge \pi_2 \wedge \dots \wedge \pi_k] \underset{\substack{\text{equality only}\\ \text{when } \pi_1,\dots,\pi_k \\ \text{all independent} \\ \text{(by property 1)}}}{=} \underbrace{H[\pi_1] + H[\pi_2] + \dots + H[\pi_k]}_{\text{RHS is min when } \pi_1,\dots,\pi_k \text{ all independent}}$$

- Viewed in our first abstraction, each digit of the number is an independent random variable.

$$\begin{array}{ccccc} 0 & 1 & 1 & \dots & 1 & 1 & 0 \\ \uparrow & & & & & \uparrow & \uparrow \\ A_n & & & \dots & & A_2 & A_1 \end{array}$$

Q/A: There is a concept of conditional entropy, which we will not cover.

- Philosophically, entropy : partitions :: probability : sets

  Partitions are much tougher than sets.

This stuff takes a whole term in Course 6.
I had some notes written, but I've lost them ☹

- Famous Problem of the Scales

You have 3^n coins, all of them gold, but one of them is fake ← it's lighter


scale

You are allowed the following experiment:

That is a question: { You take 2 subsets of k coins and you put them on the scale. The scale either doesn't tilt, or it tilts to the right, or it tilts to the left.

What is the minimum number of weighings needed to determine the fake coin?

{we could change this to any number and modify things accordingly} SC 4/24/98.10

The sample space Ω has 3^n sample points with:

$$P(\omega) = \frac{1}{3^n}$$

The entropy of the trival partition $\hat{0}$ is:

$$H[\hat{0}] = \sum_{B \in \hat{0}} P(B) \log_2 \frac{1}{P(B)}$$

$$= \cancel{\sum_{3^n}} \cancel{\left(\frac{1}{3^n}\right)} \log_2 \frac{1}{\left(\frac{1}{3^n}\right)}$$

$$= \log_2 3^n$$

$$= \underbrace{(\log_2 3)}_{\text{universal constant } c_u} n$$

$$\boxed{H[\hat{0}] = c_u \cdot n}$$

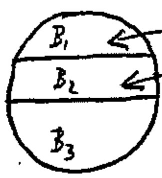
That means the number of weighings you need, where each weighing gives you some entropy, must have a total entropy adding up to $c_u \cdot n$, otherwise you can not determine the fake coin.

(Identifying the fake coin is equivalent to realizing the trivial partition)

So what's a weighing?

Weighing

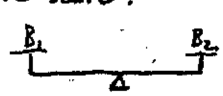
A weighing is a partition w/ 3 blocks:



Ω

2 blocks have the same number of elements

The coins in these 2 blocks are placed on the plates of the scale:



In each weighing, you know in which of the 3 blocks the fake coin is. $\Rightarrow$ $\begin{cases} \text{If the scale tilts, then you know which block of } B_1, B_2 \text{ the fake coin is in.} \\ \text{If the scale doesn't tilt, the fake coin is in block } B_3 \end{cases}$

Every weighing gives you some information as to where the fake coin is.

- By the Maximum Entropy Principle [SC 4/24/98.6], we are best off if we split the coins into 3 equal piles.

$$P(B_i) = \frac{1}{3} \quad , \quad i = 1, 2, 3$$

{abuse of notation: B_i = event that fake coin is in block B_i}

and the entropy of this partition $\pi = \{B_1, B_2, B_3\}$ is:

$$H[\pi] = \sum_{B \in \pi} P(B) \log_2 \frac{1}{P(B)}$$

$$= \cancel{\sum_3} \cancel{\left(\tfrac{1}{3}\right)} \log_2 \frac{1}{\left(\frac{1}{3}\right)}$$

$$= \underbrace{(\log_2 3)}_{\text{universal constant } c_u} 1$$

$$\boxed{H[\pi] = c_u \cdot 1}$$

This is the entropy of a single weighing.

- Since we have that the entropy of the trivial partition is:

$$H[\hat{0}] = c_u \cdot n$$

Then we need at least n weighings, each w/ entropy $H[\pi] = c_u \cdot 1$, to determine which is the fake coin.

$$H[\hat{0}] = c_u \cdot n \iff n\, H[\pi] = n\,(c_u \cdot 1)$$

You can not determine the fake coin in less than n weighings.
That's what information theory tells you.

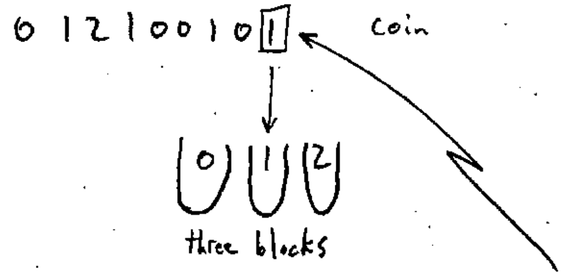
- Now that we know the min number of weighings, let's devise a way of doing it.
How do you do this in practice?
Easy,

Label each coin in the ternary system: ⟸ Coding (i.e., labelling) sample points

Coin	Label
0	0
1	1
2	2
3	10
4	11
5	12
6	20
7	21
8	22
9 etc.	100 etc.

} there are 3^n coins

Thus each coin will have a label consisting of a sequence of 0's, 1's, and 2's.
For example:



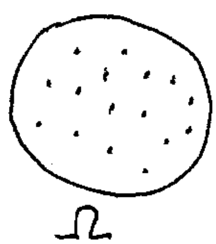
For the 1st weighing, take the first digit and split the 3^n coins into one of 3 piles, depending on this digit.

For each successive weighing, proceed similarly by splitting the 3^n coins based on the successive digit.

- Shannon Coding Theorem

In general, the Shannon Coding Theorem, which is the main theorem in information theory, goes something like this.

You have a discrete sample space Ω, with entropy of the trivial partition some number, say n.



$$H[\hat{o}] = n$$

You want to code the sample space, by assigning to each sample point, a sequence of 0, 1 digits, in such a way that the sequence is the shortest possible.

The Shannon Coding Theorem tells you that, if you are allowed shorter and longer digits and you count only the average number of digits, you can always code the sample space (regardless of its probabilities) with an average of n digits.

- That's just the beginning of information theory.
The fun comes when you have noise.
In other words: Back to the "20 questions" example of guessing a number [SC 4/24/98.7]
But, now I'm allowed 10 lies.
What's the minimum number of questions? } This is a channel w/ noise

You have to work out the entropy you need and the main theorem of Shannon's Theorem. How you work out the entropy, taking into account the noise.

Even w/ 1 lie, it is tough.

This gives you an idea about information theory.
Next week, a specialist on information theory will give the SuperClass. Professor Viola of Course 6.
He works full-time using entropy to solve interesting problems

Image Reconstruction } all done
Pattern Recognition } by entropy

There is a fundamental question. People don't really know why it works.
Why does the Maximum Entropy Principle [SC 4/24/98.6] work?

Q: Is there a relationship between the entropy of the conjugate prior and the entropy of the posterior in Bayes' Law?
A: Lot's of words.
Rivers of words.

The relation between Bayes' and entropy is unsolved.
Work is still going on. Research is going on.
People have not understood the relationship.
It's one of the major unsolved problems of science.
Perhaps you will solve it, while I'm still alive.
Please.

Within the text, pagination is of the form *lecture_date.page*. Note that a different convention is used in both the table of contents and the index, where pagination is of the form *lecture_number.page*. The table of contents provides the mapping between *lecture_number* and *lecture_date*. For example, page 1.1 in the table of contents and index corresponds to page 2/4/98.1 in the text, where the topic anomaly detection is discussed.

Index