Lectures: Probability Theory
Fall 2019

Copyright 2019 by James Bernhard
1 BASIC DEFINITIONS
   Tuesday, September 3
   Wednesday, September 4

2 DISCRETE RANDOM VARIABLES
   Friday, September 6
   Monday, September 9
   Tuesday, September 10

3 COMBINATORIAL TECHNIQUES
   Wednesday, September 11
   Friday, September 13
   Monday, September 16
   Tuesday, September 17
   Wednesday, September 18
   Friday, September 20
CLASS BUSINESS

Items of business:

- take roll
- go over syllabus
Some advice on how to check yourself:

- try a smaller problem
- try to solve the problem in general
- try to solve the problem in more than one way
**Definition**

An **experiment** is a procedure that satisfies:

1. It is repeatable, in theory infinitely many times, and
2. Each time it is conducted, it produces a single result from a well-defined set of possible results.

**Definitions**

Conducting an experiment a single time is called a **trial**. The result of an experiment is called an **outcome**. A set of outcomes is called an **event**. The set of all possible outcomes of an experiment is called the **sample space** of the experiment, often denoted by $S$. 
DEFINITIONS

For any set $A$, the number of elements of $A$ is called the order of $A$ and is denoted by $|A|$. If $|A|$ is finite, then $A$ is called finite.
**Example**

Flipping a particular coin in a particular way can be modeled as an experiment: it can be modeled as being repeatable and producing one of two possible results each time. The two possible outcomes of this experiment are that the coin lands *heads* and that the lands *tails*. We denote these outcomes by $H$ and $T$. The sample space $S$ is

$$S = \{H, T\}.$$

The four events associated with this experiment are:

$$\emptyset, \{H\}, \{T\}, \text{and } S.$$
**Example**

Rolling a particular 6-sided die in a particular way can be modeled as an experiment: it can be modeled as being repeatable and producing one of six possible results each time.

The six possible outcomes of this experiment are that the die lands with 1, 2, 3, 4, 5, or 6 facing up. We denote these outcomes by 1, 2, 3, 4, 5, and 6. The sample space $S$ is

$$S = \{1, 2, 3, 4, 5, 6\}.$$

There are lots of events associated with this experiment, such as:

$\emptyset$, $\{1, 3, 5\}$, $\{3, 4, 5, 6\}$, and $S$. 
EXAMPLE

Choosing a card from a shuffled 52-card deck of cards can be modeled as an experiment: it can be modeled as being repeatable and producing one of 52 possible results each time.

The 52 possible outcomes of this experiment are that each one of the 52 cards is selected.

We’ll denote these outcomes by the name of the card selected, such as 5 of diamonds. Since the sample space $S$ contains 52 outcomes, which we won’t list out here.

There are lots of events associated with this experiment. One event is:

$$\{3 \text{ of spades}, 4 \text{ of hearts}, 5 \text{ of spades}, \text{king of spades}\}.$$
Two ways to model the real world are: *deterministically* and *probabilistically*.

**Definition**

In **deterministic models**, initial conditions completely determine future states. In **probabilistic** models, future states are determined by initial conditions and inherent variability.
In probabilistic models, we can talk about and measure the inherent variability.

What is *randomness*? Variability that we cannot predict deterministically.

Even if this variability is deterministic, for whatever reason (limited ability to measure, etc.) we can’t predict it. A simple coin flip illustrates this well.

But what is *probability*?
First some preliminary definitions:

**Definitions**

Let $A, B$ be events associated with the same experiment. We say that $A$ and $B$ are **disjoint** if

$$A \cap B = \emptyset.$$

Also, let $A_1, A_2, \ldots, A_k$ be events associated with the same experiment. We say that $A_1, A_2, \ldots, A_k$ are **pairwise disjoint** if

$$A_i \cap A_j = \emptyset \text{ for all } i, j \in \{1, 2, \ldots, k\} \text{ with } i \neq j.$$
EXAMPLE

Consider rolling a standard 6-sided die, and let

\[ A = \{1, 2, 3\}, \]
\[ B = \{2, 4, 6\}, \text{and} \]
\[ C = \{5, 6\}. \]

The events \( A \) and \( C \) are disjoint. However, \( A \) and \( B \) are not disjoint, and \( B \) and \( C \) are not disjoint, so \( A, B, C \) are not pairwise disjoint.
The mathematical definition of probability is:

**DEFINITION**

For any experiment with sample space $S$, a **probability function** is a function $P$ from the experiment’s set of events to $\mathbb{R}$ satisfying:

1. $0 \leq P(A) \leq 1$ for all events $A$,
2. $P(S) = 1$,
3. For all disjoint events $A, B$,
   \[ P(A \cup B) = P(A) + P(B), \]
4. For all pairwise disjoint events $A_1, A_2, A_3, \ldots$,
   \[ P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i). \]
**DEFINITION**

Let $A$ be an event in a sample space $S$. The complement of $A$ in $S$, denoted by $S \setminus A$, is defined to be

$$S \setminus A = \{ x \in S \mid x \notin A \}.$$
PROPOSITION

Let $A$ be an event in a sample space $S$. Then

$$P(S\backslash A) = 1 - P(A).$$
To prove this proposition, note that $A$ and $S \setminus A$ are disjoint, so by third property in the definition of probability,

$$P(A) + P(S \setminus A) = P(A \cup S \setminus A).$$

By the definition of a complement, $A \cup (S \setminus A) = S$, so

$$P(A \cup (S \setminus A)) = P(S).$$

By the definition of probability,

$$P(S) = 1,$$

which proves the proposition.
THEOREM (EQUALLY PROBABLE OUTCOMES THEOREM)

Let $S$ be a sample space satisfying:

1. $S$ is finite, and
2. All the outcomes in $S$ are equally probable.

Let $A \subseteq S$ be an event. Then

$$P(A) = \frac{|A|}{|S|}.$$
To prove this, note that since $S$ is finite, then

$$S = \{s_1, s_2, \ldots, s_k\}$$

for some events $s_1, s_2, \ldots, s_k$. Since $\{s_1\}, \{s_2\}, \ldots, \{s_k\}$ are pairwise disjoint, then by the definition of probability

$$1 = P(S) = P(\bigcup_{i=1}^{k} \{s_i\}) = \sum_{i=1}^{k} P(s_i).$$

Since all outcomes are equally probable, then for all $j \in \{1, 2, \ldots, k\}$ this equals

$$\sum_{i=1}^{k} P(s_j) = kP(s_j).$$

Dividing both sides of this long equation by $k$ proves the theorem.
This theorem tells us that when the sample space is finite and the outcomes are all equally probable, all probabilities can be computed by *counting*.

Consequently, we’ll soon spend time learning various counting techniques (introductory combinatorics).
**Definition**

An experiment is called **fair** or **at random** if every outcome in its sample space is equally probable.

**Definitions**

A **Bernoulli trial** is an experiment with exactly two outcomes. These two outcomes are called **success** and **failure**, no matter what they represent.
The prototypes of finite sample spaces:

- coins flipped fairly (the prototype for a Bernoulli trial, with $H$ being success unless otherwise noted)
- dice rolled fairly
- card pulled from decks at random (usually standard decks, having 52 cards)
- colored balls selected at random from within urns (after Pólya)
EXAMPLE

If you flip a coin fairly, what are the probabilities of the two outcomes?
With our usual notation, the sample space is

\[ S = \{H, T\}, \]

so \(|S| = 2\). Also, all the outcomes are equally probable since the coin is flipped fairly, so by the equally probable outcomes theorem, the probability of any event \(A\) is

\[ P(A) = \frac{|A|}{|S|}. \]

When \(A\) consists of a single outcome, \(|A| = 1\), so its probability is \(P(A) = 1/2\).
EXAMPLE

If you pull a card from a standard deck at random, what is the probability that it will be a face card?
The sample space $S$ can be thought of as the set of all the cards in the deck, so $|S| = 52$.

Let $A$ be the event that we get a face card. Since there are 12 face cards in the deck, $|A| = 12$.

Each outcome in $S$ is equally probable since we are pulling a card at random, so by the equally probable outcomes theorem,

$$P(A) = \frac{|A|}{|S|} = \frac{12}{52} \approx 0.2308.$$
**Definition**

A random variable is a function \( X : S \to \mathbb{R} \).

**Definition**

A random variable is **continuous** if its possible values contain an interval in the real line. A random variable is **discrete** if its possible values are countable.

We will assume all random variables are discrete from now on, until further notice.

**Definition**

The Bernoulli random variable associated with a given Bernoulli trial is the random variable that is 1 when a success occurs and 0 when a failure occurs.
For a discrete random variable $X$, the events that we are most interested in are of the form $X = a$, for various values of $a$.

Discussing such events leads us to the following:

**Definition**

The distribution of a discrete random variable is a table of all the possible values and their probabilities.
To give the distribution of a random variable, we ordinarily write out a table of its possible values and their probabilities.

Another way to give the distribution of a discrete random variable is by specifying its **probability mass function**, which is a function $f : \mathbb{R} \to [0, 1]$ satisfying

$$f(a) = P(X = a)$$

for all $a \in \mathbb{R}$.

Notice that $f(a) = 0$ whenever $a$ is not a possible value of $X$, so by giving the nonzero values of $f$, you are in essence writing out a table of the possible values of $X$ and their probabilities.
**EXAMPLE**

Suppose a coin is flipped, and $P(H) = p$. Let $X$ be the random variable that is 1 when the coin lands $H$ and 0 when it lands $T$. What is the distribution of $X$?
The distribution of $X$ is

<table>
<thead>
<tr>
<th>$a$</th>
<th>$P(X = a)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1 - $p$</td>
</tr>
<tr>
<td>1</td>
<td>$p$</td>
</tr>
</tbody>
</table>
**Definition**

A random variable $X$ is said to have a **Bernoulli distribution with success probability** $p$ (where $0 < p < 1$) if its distribution is:

<table>
<thead>
<tr>
<th>$a$</th>
<th>$P(X = a)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1 - $p$</td>
</tr>
<tr>
<td>1</td>
<td>$p$</td>
</tr>
</tbody>
</table>

We denote this by writing

$$X \sim \text{Bernoulli}(p).$$

This gives us another way to specify the distribution of a random variable: by giving its distribution family name (Bernoulli here) and the value of its parameter(s) within that family ($p$ here).
**Definition**

The **expected value** or **mean** of $X$ is defined by

$$E(X) = \sum_a aP(X = a)$$

If $X$ has only finitely many possible values, the expected value of $X$ is the dot product of the two columns that make up the distribution table for $X$.

Since the expected value $E(X)$ is defined by a potentially infinite sum, it may or may not exist (depending on exactly what $X$ is).
EXAMPLE

Let $X \sim \text{Bernoulli}(p)$. Compute $E(X)$.
Since $X \sim \text{Bernoulli}(p)$, then the distribution of $X$ is

<table>
<thead>
<tr>
<th>$a$</th>
<th>$P(X = a)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$1 - p$</td>
</tr>
<tr>
<td>1</td>
<td>$p$</td>
</tr>
</tbody>
</table>

By the definition of expected value, this means that

$$E(X) = (0)(1 - p) + (1)(p) = p.$$
A game is called **zero-sum** if the total payoff to all players is 0. A two-person zero-sum game is **fair** if the expected value of its payoff (to either player) is 0.
**Example**

Suppose that a certain way of flipping a coin has $P(H) = p$, and that a 2-player game is played as follows:

1. The coin is flipped in that way.
2. If the coin lands $H$, then Player 1 pays $1 to Player 2.
3. If the coin lands $T$, then Player 2 pays $1 to Player 1.

For what values of $p$ is this two-person zero-sum game fair?
Let $X$ be the random variable is the payoff in dollars to Player 1 (it being negative if Player 1 has to pay money). Then

$$E(X) = (1)(p) + (-1)(1 - p).$$

Setting this equal to zero and solving for $p$, we find that $p = 1/2$ is the only case for which this game is fair.
**PROPOSITION**

Let $X$ be a discrete random variable whose expected value exists, and let $c \in \mathbb{R}$. Then

$$E(cX) = cE(X).$$
Since $X$ is a discrete random variable, then its values that have nonzero probability can be listed as $a_1, a_2, \ldots$ for some $a_1, a_2, \ldots \in \mathbb{R}$. This means that the values that have nonzero probability for $cX$ are $ca_1, ca_2, \ldots$.

By the definition of expected value,

$$E(cX) = ca_1P(cX = ca_1) + ca_2P(cX = ca_2) + \cdots.$$  

Also, by the definition of the random variable $cX$, 

$$P(cX = ca_i) = P(X = a_i)$$

for all $i$. Substituting this into the previous equation, we have that 

$$E(cX) = ca_1P(X = a_1) + ca_2P(X = a_2) + \cdots$$

$$= c(a_1P(X = a_1) + a_2P(X = a_2) + \cdots)$$

$$= cE(X),$$

which proves the proposition.
**PROPOSITION**

Let $X$ be a discrete random variable whose expected value exists, and let $h : \mathbb{R} \to \mathbb{R}$ be a function. (Technically, $h$ should be a measurable function.) Then

$$E(h(X)) = \sum_{a} h(a)P(X = a).$$
Since $X$ is a discrete random variable, then its values that have nonzero probability can be listed as $a_1, a_2, \ldots$ for some $a_1, a_2, \ldots \in \mathbb{R}$. This means that the values that have nonzero probability for $f(X)$ are $h(a_1), h(a_2), \ldots$.

By the definition of expected value,
\[
E(h(X)) = h(a_1)P(h(X) = h(a_1)) + h(a_2)P(h(X) = h(a_2)) + \cdots.
\]

Also, by the definition of the random variable $h(X)$,
\[
P(h(X) = h(a_i)) = P(X = a_i)
\]
for all $i$. Substituting this into the previous equation proves the proposition:
\[
E(h(X)) = h(a_1)P(X = a_1) + h(a_2)P(X = a_2) + \cdots = \sum_a h(a)P(X = a).
\]
**PROPOSITION**

*Let \( X, Y \) be discrete random variables whose expected values exist. Then*

\[
E(X + Y) = E(X) + E(Y).
\]

We will prove this proposition later in the course, but you may use it now.
**DEFINITION**

The **variance** of $X$ is defined to be

\[ \text{Var}(X) = E((X - E(X))^2) \]

**DEFINITION**

The **standard deviation** of $X$ is defined to be

\[ \sigma_X = \sqrt{\text{Var}(X)} \]

These both convey the same information. They both have names because they’re both useful in different contexts.
EXAMPLE

Let $X \sim \text{Bernoulli}(p)$. What are $\text{Var}(X)$ and $\sigma_X$?
Since $X \sim \text{Bernoulli}(p)$, then the distribution of $X$ is:

<table>
<thead>
<tr>
<th>$a$</th>
<th>$P(X = a)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$1 - p$</td>
</tr>
<tr>
<td>1</td>
<td>$p$</td>
</tr>
</tbody>
</table>

Also, we have computed that $E(X) = p$. 
We now have enough information to compute the variance:

\[
\begin{array}{ccc}
  a & (a - E(X))^2 & P(X = a) \\
  0 & (0 - p)^2 & 1 - p \\
  1 & (1 - p)^2 & p \\
\end{array}
\]

By the definition of variance and the definition of expected value,

\[
\text{Var}(X) = E((X - E(X))^2)
= (1 - p)(0 - p)^2 + (p)(1 - p)^2
= p(1 - p).
\]
**Definition**

The \( k \)-th moment of \( X \) is \( E(X^k) \), and the \( k \)-th central moment of \( X \) is \( E((X - E(X))^k) \).

We won't use this concept much right away, but we will later.

Note:

- The first moment of \( X \) is \( E(X) \).
- The first central moment of \( X \) is 0.
- The second central moment of \( X \) is \( \text{Var}(X) \).
- Moments and central moments are defined in terms of possibly infinite sums, so they may or may not exist.
PROPOSITION (ALTERNATIVE DEFINITION OF VARIANCE)

Let $X$ be a random variable whose expected value and variance exist. Then

$$\text{Var}(X) = E(X^2) - E(X)^2$$
To prove this, note that by the definition of variance,

$$\text{Var}(X) = E((X - E(X))^2)$$

$$= E(X^2 - 2XE(X) + E(X)^2).$$

By the proposition that the expected value of a sum equals the sum of the expected values, this equals

$$E(X^2) - E(2XE(X)) + E(E(X)^2).$$
By the proposition that the expected value of a scalar multiple equals that scalar multiple of the expected value, this equals

$$E(X^2) - (2E(X))E(X) + E(E(X)^2).$$

(We have used that $E(X)$ is a real number, not a random variable.)

Since the expected value of a constant random variable equals that constant, this equals

$$E(X^2) - (2E(X))E(X) + E(X)^2.$$

Simplifying proves the result, since this equals

$$E(X^2) - E(X)^2.$$
If you’re trying to remember which term is positive and which negative in this proposition, recall that variances are always nonnegative, and think about why \( E(X^2) \) is always greater than or equal to \( E(X)^2 \).
**Definition**

Two events $A, B$ in the same sample space are **independent** if

\[ P(A \cap B) = P(A)P(B) \]

**Definition**

Two random variables $X, Y$ are **independent** if

\[ P(X = x, Y = y) = P(X = x)P(Y = y) \]

for all possible values $x$ of $X$ and $y$ of $Y$.

This definition extends to more than two events (or random variables). For example, events $A, B, C$ are independent if

\[ P(A \cap B \cap C) = P(A)P(B)P(C). \]
There are two main interpretations of probability functions:

- **frequentist** The probability of an event is the relative frequency with which it occurs as the number of repetitions goes to infinity.

- **subjective** The probability of an event is a measure of an individual’s belief about how likely that event is (as measured by, say, the amount of money the individual is willing to risk on the event’s occurrence)
Problems with frequentist approach:

- you need to repeat things infinitely many times to measure probabilities
- doesn’t say anything about an individual trial

Problems with the subjective approach:

- involves individuals’ beliefs and so isn’t objective

Advantages are the reverse.
**Definition**

A random variable $X$ that is the success count in a sequence of $n$ independent Bernoulli trials each having success probability $p$ is said to have a **binomial distribution** with success probability $p$ and size $n$, where $0 < p < 1$ and $n$ is a positive integer. We denote this by

$$X \sim B(n, p).$$
EXAMPLE

Find the distribution table for $X$ when $X \sim B(2, \frac{3}{4})$. 
Since $X$ is the success count for two independent Bernoulli trials, each having success probability $\frac{3}{4}$, we can list the possible outcomes of these two trials and their probabilities:

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$HH$</td>
<td>$9/16$</td>
</tr>
<tr>
<td>$HT$</td>
<td>$3/16$</td>
</tr>
<tr>
<td>$TH$</td>
<td>$3/16$</td>
</tr>
<tr>
<td>$TT$</td>
<td>$1/16$</td>
</tr>
</tbody>
</table>

We have computed all of these using that the two coin flips are independent, and that by the definition of independent events,

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

whenever $A$ and $B$ are independent events.
From this, we can compute the distribution table:

<table>
<thead>
<tr>
<th>$a$</th>
<th>$P(X = a)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1/16</td>
</tr>
<tr>
<td>1</td>
<td>3/8</td>
</tr>
<tr>
<td>2</td>
<td>9/16</td>
</tr>
</tbody>
</table>
This method would in principle work to compute $P(X = k)$ when $X \sim B(n, p)$ for other $n$ and $p$ too.

Later we will use other methods find a formula to compute that, so we won’t do it now.

For now, you may type “binomial distribution calculator” into Wolfram Alpha (or use other software if you prefer) to compute it.
We’re interested in *sampling*, or selecting items from a *population*, or set of objects.

Two aspects of sampling:

1. Are we replacing each item from the population after selecting it?
2. Do we care what order we select the items in?

Both of these can be thought of in terms of choosing a hand from a deck of cards.
We’ll discuss four basic counting situations:

1. order matters, with replacement (multiplication)
2. order matters, without replacement (permutations)
3. order doesn’t matter, without replacement (combinations), and
4. order doesn’t matter, with replacement (stars and bars).
To help us discuss these, we first prove a preliminary lemma.

**Lemma (multistage counting lemma)**

If an experiment consists of $k$ stages, and if the number of outcomes for the $i$-th stages individually is $n_i$ no matter what outcomes have occurred in the previous stages, then the number of possible outcomes for the experiment is $n_1 n_2 \cdots n_k$. 
We prove this by induction on $k$.

For the base case, when $k = 1$, there are $n_1$ options in the first and only stage, and there are $n_1$ possible outcomes for the experiment, so the result holds.
To prepare to prove the inductive step, we first prove that this holds when $k = 2$. Suppose that the possible outcomes of the first stage are $a_1, a_2, \ldots, a_{n_1}$. Also suppose that the possible outcomes of the second stage when the value of the first stage is $a_i$ are $b_{i,1}, b_{i,2}, \ldots, b_{i,n_2}$.

Then the possible outcomes of the 2-stage process are

\[
\begin{align*}
(a_1, b_{1,1}) &\quad (a_1, b_{1,2}) &\quad \cdots &\quad (a_1, b_{1,n_2}) \\
(a_2, b_{2,1}) &\quad (a_2, b_{2,2}) &\quad \cdots &\quad (a_2, b_{2,n_2}) \\
&\vdots &\ddots &\vdots \\
(a_{n_1}, b_{n_1,1}) &\quad (a_{n_1}, b_{n_1,2}) &\quad \cdots &\quad (a_{n_1}, b_{n_1,n_2})
\end{align*}
\]

This shows that there are $n_1 n_2$ possible outcomes of the 2-stage process.
For the inductive step, suppose that this result holds for $k$. Any $(k+1)$-stage process is a $k$-stage process followed by a 1-stage process, making it in essence a 2-stage process (the first stage itself having $k$ sub-stages).

We have just shown that the number of outcomes for this 2-stage process is the product of the number of outcomes in the first stage (with its $k$ sub-stages) and the number of outcomes in the second stage.
By the inductive hypothesis, for the first $k$ stages, there are $n_1 n_2 \cdots n_k$ possible outcomes.

By assumption, there are $n_{k+1}$ possible outcomes for the last stage.

This implies that the total number of possible outcomes for the $(k + 1)$-stage process is $n_1 n_2 \cdots n_k n_{k+1}$, which proves the lemma.
An immediate corollary of this lemma is the following lemma:

**LEMMA (COUNTING ORDERINGS LEMMA)**

*For any integer \( n \geq 1 \), there are \( n! \) ways to order a set of \( n \) elements.*
To prove this, think of an ordering as an $n$-stage experiment: first decide what the first element will be, then decide what the second element will be, and so on until you have decided on the order for all the elements.

The first stage of this process always has $n$ possible outcomes, the next stage $n - 1$, the third stage $n - 2$, and so on.

Since the number of possible outcomes at each stage is the same no matter what the outcomes of prior stages were, then by the multistage counting lemma the number of possible outcomes for the entire process is

$$n(n - 1)(n - 2) \cdots (1) = n!,$$

which proves the lemma.
Order matters, with replacement (multiplication)

The prototype for this is flipping a coin multiple times.

Flip a coin once, and there are exactly 2 possible outcomes: $H$ and $T$.

What about twice? Three times? $n$ times?
Possible outcomes for 4 coin flips:

<table>
<thead>
<tr>
<th>Basic definitions</th>
<th>H</th>
<th>H</th>
<th>H</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete random variables</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>T</td>
</tr>
<tr>
<td>Combinatorial techniques</td>
<td>H</td>
<td>H</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>Multiple random variables</td>
<td>H</td>
<td>T</td>
<td>H</td>
<td>H</td>
</tr>
<tr>
<td>Markov chains</td>
<td>H</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>Continuous random variables</td>
<td>H</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>Sums of random variables</td>
<td>T</td>
<td>H</td>
<td>H</td>
<td>H</td>
</tr>
<tr>
<td>Statistical applications</td>
<td>T</td>
<td>H</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>Further topics</td>
<td>T</td>
<td>T</td>
<td>H</td>
<td>H</td>
</tr>
<tr>
<td>Student presentations</td>
<td>T</td>
<td>T</td>
<td>H</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>H</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>
Looking at the pattern in the above table may help to suggest the general result.
**PROPOSITION**

Suppose that order matters, and we are sampling with replacement. Then the number of distinct samples of size $n$ drawn from a population of $m$ items is $m^n$. 
Think of the sampling as an $n$-stage process: we sample (and replace), sample (and replace), etc.

Since the number of outcomes at each stage is $m$ no matter what other outcomes have occurred in the earlier stages, then by the multistage counting lemma, the number of outcomes for the entire process is the product of these, meaning $m^n$. 
Order matters, without replacement (permutations)

A good prototype for this is a “lotto ball” draw.

**Example**

Suppose that there are 50 numbered balls whirling around, and 6 of them are chosen in sequence at random. You win if you match all 6 numbers in the order in which they are drawn. How many different ordered 6-ball draws are there?
View this as a 6-stage process. The first stage has 50 possible outcomes, the next 49, and so on down to 45. By the multistage counting lemma, the number of possible outcomes is

\[(50)(49)(48)(47)(46)(45) = 11,441,304,000.\]
PROPOSITION (COUNTING PERMUATIONS)

Suppose that order matters, and we are sampling without replacement. Then the number of distinct samples of size $k$ drawn from a population of $n$ items is

$$\frac{n!}{(n - k)!}.$$

We often denote this quantity by $p_{n,k}$ and refer to it as “the number of permutations (ordered arrangements) of $k$ elements out of $n$”. 
To prove this, view this sampling process as a $k$-stage process: sample from the $n$ items, sample from the remaining $n - 1$ items, sample from the remaining $n - 2$ items, and so on until you sample from the remaining $n - k + 1$ items.

Since the number of outcomes at the $i$-th stage is $n - i + 1$ no matter what other outcomes have occurred in the earlier stages, then by the multistage counting lemma, the number of outcomes for the entire process is

$$n(n - 1) \cdots (n - k + 1) = \frac{n!}{(n - k)!}.$$
Order doesn’t matter, without replacement (combinations)

A good prototype for this is the number of card hands.

**Example**

How many (unordered) 5-card hands are there from a standard deck of cards?
If we count different orders as being different, then by the counting permutations proposition, there are

$$\frac{52!}{(52 - 5)!} = 311,875,200$$

possible samples.

But if order doesn’t matter, we are counting the same samples multiple times, once for each possible ordering. By the counting orderings lemma, there are 5! = 120 possible orderings of each of these.
To get the number of distinct samples when order doesn’t matter, we divide the number when order does matter by the number of times that we are counting each of those as distinct when we shouldn’t be, which gives

\[
\frac{52!}{(52 - 5)! \times 5!} = \frac{311,875,200}{120} = 2,598,960.
\]
**PROPOSITION (COUNTING COMBINATIONS)**

Suppose that order doesn’t matter, and we are sampling without replacement. Then the number of distinct samples of size \( k \) drawn from a population of \( n \) items is

\[
\frac{n!}{(n - k)! \cdot k!}.
\]

Since unordered samples drawn without replacement are simply subsets, this proposition also gives a formula for the number of subsets of order \( k \) in a set of order \( n \).
To prove this, note first that if we count different orders as being different, then by the **counting permutations proposition**, there are

$$\frac{n!}{(n-k)!}$$

possible samples.

But if order doesn’t matter, we are counting the same samples multiple times, once for each possible ordering. By the **counting orderings lemma**, there are $k!$ possible orderings of each of these samples of size $k$. 
To get the number of distinct samples when order doesn’t matter, we divide the number when order does matter by the number of times that we are counting each of those as distinct when we shouldn’t be, which gives

\[
\frac{n!}{(n-k)! \ k!}.
\]
The quantity in the previous proposition is important enough that we give it its own notation and name.

**Definition**

Let \( n, k \) be nonnegative integers with \( 0 \leq k \leq n \). We define the **binomial coefficient** \( n \) choose \( k \) (denoted \( \binom{n}{k} \)) by:

\[
\binom{n}{k} = \frac{n!}{(n-k)! \ k!}.
\]

We can also refer to this quantity as “the number of **combinations** (unordered arrangements) of \( k \) elements out of \( n \)”

As mentioned before, this is also the number of subsets of order \( k \) in a set of order \( n \)
EXAMPLE

In poker with a 5-card hand, compute the probability of 4 of a kind.
One possible method: Suppose order *doesn’t* matter here. Then an outcome is an unordered hand, and the sample space $S$ is the set of all unordered hands. By the unordered without replacement proposition,

$$|S| = \binom{52}{5} = 2,598,960.$$ 

Let $A$ be the event that we get 4 of a kind. To compute $|A|$, first choose which card should have 4: there are 13 ways to do this. Then there are 48 cards from which to choose the remaining card. By the multiplication principle,

$$|A| = (13)(48) = 624.$$
Since $S$ is finite and all outcomes are equally probable, then by the equally probable outcomes theorem,

$$P(A) = \frac{|A|}{|S|} = \frac{624}{2,598,960} \approx 0.000240096.$$
Another possible method: Suppose order matters. Then an outcome is an ordered hand, and the sample space $S$ is the set of all ordered hands. By the ordered without replacement theorem,

$$|S| = \frac{52!}{47!} = 311,875,200.$$
Let \( A \) be the event that we get 4 of a kind. To compute \( A \), first choose which 4 cards are to match: there are \( \binom{5}{4} = 5 \) ways to do this. Then there are 13 choices of which card will have four of a kind, and \( 4! \) ways to rearrange the four matching cards. Also, there are 48 choices for the remaining card, and \( 1! \) ways to rearrange it. By the multiplication principle,

\[
|A| = \left( \binom{5}{4} \right)(13)(4!)(48)(1!) = 74,880.
\]

Since \( S \) is finite and all outcomes are equally probable, then by the equally probable outcomes theorem,

\[
P(A) = \frac{|A|}{|S|} = \frac{74,880}{311,875,200} \approx 0.000240096.
\]
EXAMPLE
Suppose we have 8 different cards, and we’d like to distribute them among 3 different people, whose favorite colors are red, black, and blue respectively. If we care only which people the cards go to (and not what order the cards are given to people in), how many different ways are there to distribute the cards so that the numbers of cards for the three people are:

<table>
<thead>
<tr>
<th>Color</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>red</td>
<td>2</td>
</tr>
<tr>
<td>black</td>
<td>5</td>
</tr>
<tr>
<td>blue</td>
<td>1</td>
</tr>
</tbody>
</table>
One way to figure this out is to suppose that order matters, and think of this as arranging the 8 cards (numbered 1 to 8) into 8 slots in such a way that the first 2 are red, the next 5 are black, and the last 1 blue.

Some examples of doing this:

1  2  3  4  5  6  7  8
5  3  7  2  4  1  6  8
8  7  2  4  1  6  3  5
5  6  3  4  2  1  8  7
By the **counting orderings lemma**, there are $8! = 40320$ arrangements of the cards where order matters.

But we don’t care about order within the red, and again by the **counting orderings lemma**, there are $2! = 2$ arrangements of those cards where order matters.

For each rearrangement of the red, there are similarly $5! = 120$ arrangements of the black cards.

For each rearrangement of the red and black, there are similarly $1! = 1$ arrangements of the blue cards.
Viewing the arrangements of the cards within the color as a multistage process, then by the **multistage counting lemma**, the total number of multiple countings for each arrangement is the product of these, meaning:

\[ 2! \times 5! \times 1! \]

Dividing the total number of arrangements by the number of times that we are multiple counting each arrangement, we get the number of distinct ways to distribute the cards as requested:

\[ \frac{8!}{2! \times 5! \times 1!} = 168 \]
In doing this example, we have sketched a proof of the following proposition.

**Proposition (Multinomial Counting)**

Suppose that we have a set of \( n \) objects, and that we are given positive integers \( k_1, \ldots, k_r \) for which \( k_1 + \cdots + k_r = n \). The number of ways can we subdivide our set into \( r \) subsets of sizes exactly \( k_1, k_2, \ldots, k_r \) (where order within the subsets doesn’t matter) is

\[
\frac{n!}{k_1! \, k_2! \, \cdots \, k_r!}.
\]

This coincides with the formula for \( \binom{n}{2} \) when \( r = 2 \).
As was the case with the binomial coefficients, this formula from Friday is important enough to get its own name.

**Definition**

Let $n$ be a nonnegative integer, and let $k_1, \ldots, k_r$ be nonnegative integers with $k_1 + k_2 + \cdots + k_r = n$. We define the multinomial coefficient $n$ choose $k_1, \ldots, k_r$ (denoted $\binom{n}{k_1, \ldots, k_r}$) by:

$$\binom{n}{k_1, \ldots, k_r} = \frac{n!}{k_1! \, k_2! \cdots k_r!}.$$
**EXAMPLE (THE BIRTHDAY PROBLEM)**

Assuming that birthdays are uniformly distributed over 365 days, what is the probability that among $n$ people, no two people have a common birthday?
As usual, let’s be specific about how we’re setting this up and what the sample space is.

One way (of many possible ways) to solve this is suppose that order matters, and to suppose that we are sampling birthdays at random with replacement.

In this case, the sample space $S$ consists of sets of $n$ birthdays. Let $A$ be the event that no two birthdays are on the same day.

By the counting permutations proposition

$$|S| = 365^n.$$
Also, an outcome in which no two birthdays are the same is precisely an ordered subset of \( n \) days out of 365 days, so by the counting permutations proposition,

\[
|A| = \frac{365!}{(365 - n)!}.
\]

Since we have assumed that birthdays are at random, then by the equally probable outcomes theorem,

\[
P(A) = \frac{|A|}{|S|} = \frac{365!}{(365 - n)!} \cdot \frac{1}{365^n}.
\]

When \( n = 22 \), this gives approximately 0.5243; when \( n = 23 \), this gives approximately 0.4927.
**Example**

What is \((a + b)^n\), if you multiply it out?
If we write the product out, we have:

\[(a + b)(a + b) \cdots (a + b)\].

From this, we see that the product is a sum over all \(a^k b^{n-k}\), where \(k\) ranges from 0 to \(n\).

But how many times does \(a^k b^{n-k}\) occur in this product?

The product above has \(n\) terms (each in parentheses), and it gives rise to a term of the form \(a^k b^{n-k}\) each time an \(a\) contributes from exactly \(k\) of those terms (and \(b\) from \(n - k\))
So we get exactly one $a^k b^{n-k}$ term for each subset of order $k$ out of the $n$ terms in the product (that being the subset that contributes the $a$'s)

By the **counting combinations proposition**, there are $\binom{n}{k}$ of these. This means that

\[
(a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.
\]

This is where the term *binomial coefficient* comes from.
\textbf{EXAMPLE}

What is \((a + b + c)^n\), if you multiply it out?
If we write the product out, we have:

\[(a + b + c)(a + b + c) \cdots (a + b + c)\].

So we get exactly one \(a^{k_1} b^{k_2} c^{n-(k_1+k_2)}\) term for each partition of the \(n\) terms in the product into three subsets of size \(k_1, k_2, n - (k_1 + k_2)\).
Similar to the binomial case, the multinomial counting proposition implies that

$$(a + b + c)^n = \sum_{k_1, k_2 = 0}^{n} \binom{n}{k_1, k_2, n - (k_1 + k_2)} a^{k_1} b^{k_2} c^{n-(k_1+k_2)}$$

This is where the term *multinomial coefficient* comes from.
PROPOSITION (BINOMIAL DISTRIBUTION FORMULA)

Let $p \in \mathbb{R}$ with $0 < p < 1$, let $n$ be a positive integer, and let $X \sim B(n, p)$. Then

$$P(X = k) = \begin{cases} \binom{n}{k}p^k(1-p)^{n-k} & \text{if } k = 0, 1, \ldots, n \\ 0 & \text{otherwise.} \end{cases}$$
To prove this, note that since $X \sim B(n, p)$, then by the definition of a binomial distribution, $X$ is a success count for a sequence of $n$ independent Bernoulli trials, each having success probability $p$.

This means that if $k \notin \{0, 1, \ldots, n\}$, then $P(X = k) = 0$.

If $k \in \{0, 1, \ldots, n\}$, then let $S$ be the sample space for these $n$ Bernoulli trials, and let $A$ be the event that exactly $k$ successes occur.

Since the trials are independent with the same success probability, all the outcomes have equal probability. So the equally probable outcomes theorem implies that

$$P(A) = \frac{|A|}{|S|}.$$
Each outcome in $A$ can be uniquely specified by stating which subset of order $k$ out of the $n$ trials the successes formed.

This means that by the counting combinations proposition,

$$|A| = \binom{n}{k}.$$ 

Also, since the Bernoulli trials are independent, each outcome in $A$ has probability

$$p^k(1 - p)^{n-k}$$

(since the outcome has exactly $k$ successes and $n - k$ failures).
Since the outcomes in $A$ are pairwise disjoint and have union equal to $A$, then by the definition of a probability function

$$P(A) = \sum_{x \in A} P(x)$$

$$= |A| p^k (1 - p)^{n-k}$$

$$= \binom{n}{k} p^k (1 - p)^{n-k}.$$

Since $X = k$ if and only if $A$ happens, this gives us that

$$P(X = k) = P(A) = \binom{n}{k} p^k (1 - p)^{n-k},$$

which completes the proof of the proposition.
**EXAMPLE**

With 12 orange balls and 8 purple balls in an urn, what is the probability of getting exactly 2 orange and 2 purple if you pull 4 balls from the urn at random?
One way to solve this is to suppose that order doesn’t matter, so the sample space \( S \) is the (unordered) set of selections of 4 balls from the urn.

By the **counting combinations proposition**, the total number of ways of selecting 4 balls is \( \binom{20}{4} \), so \( |S| = \binom{20}{4} \).
Let $A$ be the event of getting exactly 2 orange and 2 purple.

By the **same proposition**, the total number of ways of selecting 2 orange balls is $\binom{12}{2}$.

By the **same proposition**, the total number of ways of selecting 2 purple balls is $\binom{8}{2}$.

Since exactly 2 orange and 2 purple can be thought of as a 2-stage process (first select exactly 2 orange and then select 2 purple), then the number of ways to get both is the product of these:

$$|A| = \binom{12}{2} \binom{8}{2}.$$
Since all outcomes in this sample space are equally probable, then by the **equally probable outcomes theorem**, So the probability is

\[
P(A) = \frac{|A|}{|S|}
\]

\[
= \frac{\binom{12}{2} \binom{8}{2}}{\binom{20}{4}}
\]

\[
= \frac{66 \cdot 28}{4845}
\]

\[
= \frac{616}{1615}
\]

\[
\approx 0.3814241.
\]
We have studied the first three of these four counting situations:

1. order matters, without replacement: **permutations**
   \( P_{n,k} \) being the number of ordered samples of size \( k \) out of \( n \) objects, sampled without replacement.

2. order doesn’t matter, without replacement: **combinations**
   \( C_{n,k} \) being the number of unordered samples of size \( k \) out of \( n \) objects, sampled without replacement.

3. order matters, with replacement: **multiplication**
   \( n^k \) being the number of ordered samples of size \( k \) out of \( n \) objects, sampled with replacement.

4. order doesn’t matter, with replacement: **stars and bars**.
What if we are trying to sample with replacement, but order doesn’t matter?

**Example**

How many different sets of 5 letters are possible if we draw 5 times with repetition from the set of letters \{a, b, c, d\}? 

Some examples of such sets of letters:

\{a, a, b, b, c\}  
\{a, a, b, c, d\}  
\{b, b, c, c, d\}  
\{b, b, c, d, d\}  
\{a, b, b, c, c\}  
\{a, a, b, b, c\}
If we let \( n_a, n_b, n_c, n_d \) be the numbers of each of the letter types in our sample, then a sample can be specified by giving these 4 numbers.

We know that these numbers are nonnegative integers, and that they sum to 5.

From this point of view, we are trying to figure out how many ways to express 5 as a sum of 4 nonnegative integers.
From another point of view, we have 5 stars (or slots to put letters in) and 3 bars (or dividers) to place among them.

For example, \textit{aaabd} corresponds to this stars and bars diagram:

\[
\begin{array}{c|c|c|c}
* & * & * & |
\end{array}
\]

And \textit{bcccc}:

\[
\begin{array}{c|c|c|c}
| & * & * & * & *
\end{array}
\]
Looking at it this way, we just need to count how many ways there are to choose the 3 places to put the bars, out of the $5 + 3 = 8$ possible slots for a star or a bar.

By the counting combinations proposition, this is

$$\binom{5+3}{3} = \binom{8}{3}.$$ 

To generalize this, we need only realize that the term 3 in the formulas above was the order of the set from which we were drawing the sample, minus 1.
**Proposition (counting stars and bars)**

Suppose that order doesn’t matter, and we are sampling with replacement. Then the number of distinct samples of size $k$ drawn from a population of $n$ items is

\[
\binom{n + k - 1}{n - 1},
\]

which is the same as

\[
\binom{n + k - 1}{k}.
\]

The first formula comes from choosing where to place the *bars*; the second from where to place the *stars*. 
**Example**

How many ways are there to place $k$ indistinguishable balls into $n$ urns?
Doing this can be viewed as partitioning $k$ into a sum of $n$ nonnegative integers (the first number in the sum is the number that goes into Urn 1, the second number goes into the second, etc.), which is another way of viewing sampling with replacement where order doesn’t matter.

Or, another way: we are drawing a sample of size $k$ with replacement from the list of $n$ urns, where order doesn’t matter.

By the counting stars and bars proposition, there are

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

ways to do this.
**Example**

Suppose we’re going to put 7 cookies to into 4 piles, one for each kid in the neighborhood. (And we know which pile is for which kid.) The cookies are indistinguishable, so we don’t know which cookies are in which piles; we just know how many cookies are in each kid’s pile. How many different sets of 4 piles for the 4 kids is it possible to make in this way?

Note: if the piles are rearranged so that a different pile goes to a different kid, that counts as a different set of piles.
Since the cookies are indistinguishable, we’re just looking for a way to express 7 as a sum of 4 non-negative integers (the \( i \)-th integer giving the number of cookies in the \( i \)-th kid’s pile).

Or, put another way, we draw a sample of size 7 with replacement where order doesn’t matter from a population of size 4 (namely, the list of kids).

By the **counting stars and bars proposition**, there are

\[
\binom{4 + 7 - 1}{4 - 1} = 120
\]

ways to do this.
EXAMPLE

Draw 5 cards at random from an ordinary deck. What is the probability that exactly 2 are kings?
One way to approach this problem is to suppose that order doesn’t matter. In this case, the sample space $S$ consists of unordered sets of 5 cards, and all outcomes are equally probable.

By the counting combinations proposition,

$$|S| = \binom{52}{5} = 2,598,960.$$ 

Let $A$ be the event that exactly 2 are kings.

Since all outcomes are equally probable, then by the equally probable outcomes theorem,

$$P(A) = \frac{|A|}{|S|}.$$
Events in $A$ are those with exactly 2 kings and 3 non-kings.

We can view selection of 2 kings and 3 non-kings as a 2-stage process. By the multistage counting lemma, $|A|$ is the product of the number of possibilities at each stage.

By the counting combinations proposition, the number of possibilities at the first and second stages are

$$\left(\binom{4}{2}\right) = 6 \text{ and } \left(\binom{48}{3}\right) = 17,296,$$

so

$$|A| = \left(\binom{4}{2}\right)\left(\binom{48}{3}\right) = 103,776.$$
Putting this all together, we have that

\[ P(A) = \frac{|A|}{|S|} = \frac{\binom{4}{2}\binom{48}{3}}{\binom{52}{5}} = \frac{103,776}{2,598,960} \approx 0.03992982. \]
Suppose we have two sets, $A$ and $B$, that are not necessarily disjoint:

How can we compute $|A \cup B|$?
From the picture, we see that if we add $|A|$ and $|B|$, we’re double-counting $A \cap B$.

If we subtract off the part that we’re double-counting, then we have it:

$$|A \cup B| = |A| + |B| - |A \cap B|.$$
Suppose we have three sets, $A$, $B$, and $C$, that are not necessarily disjoint:

How can we compute $|A \cup B \cup C|$?
From the picture, we see that if we add $|A|$, $|B|$, and $|C|$, we’re double-counting $A \cap B$, $A \cap C$, and $B \cap C$.

But if we just subtract the sizes of those off, we have subtracted off too much: we were triple-counting $A \cap B \cap C$, but we subtracted it off three times, so we’re not counting at all now! We need to add it back in, and then we have it:

\[
|A \cup B \cup C| = |A| + |B| + |C| - (|A \cap B| + |A \cap C| + |B \cap C|) + |A \cap B \cap C|.
\]

Continuing this line of reasoning leads to...
**Proposition (Inclusion-Exclusion for Sets)**

Let $A_1, \ldots, A_n$ be sets. Then

\[
| \bigcup_{i=1}^{n} A_i | = \sum_{i} |A_i| - \sum_{\text{distinct } i_1, i_2} |A_{i_1} \cap A_{i_2}| + \sum_{\text{distinct } i_1, i_2, i_3} |A_{i_1} \cap A_{i_2} \cap A_{i_3}| - \ldots + (-1)^{n+1} |A_1 \cap A_2 \cap \cdots \cap A_n|.
\]

Also, through the first line is an upper bound, through the second line a lower bound, through the third line an upper bound, and so on.
**Proposition (Inclusion-Exclusion for Probabilities)**

Let $A_1, \ldots, A_n$ be events for an experiment. Then

\[
P(\bigcup_{i=1}^{n} A_i) = \sum_{i} P(A_i) - \sum_{\text{distinct } i_1, i_2} P(A_{i_1} \cap A_{i_2}) + \sum_{\text{distinct } i_1, i_2, i_3} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots + (-1)^{n+1} P(A_1 \cap A_2 \cap \cdots \cap A_n).
\]

Also, through the first line is an upper bound, through the second line a lower bound, through the third line an upper bound, and so on.
If you roll three 6-sided dice fairly and independently what is the probability of at least one 6?
We will assume that order matters, so the sample space $S$ is the set of all ordered rolls of the three dice.

Let $A$ be the event that we get at least one 6.
The key to this problem is in recognizing that we should define events $A_i$ as “you get a 6 on $i$-th roll”. Then $A_1 \cup A_2 \cup A_3 = A$.

The events $A_1, A_2, A_3$ are not disjoint, but by the principle of inclusion-exclusion,

$$P(A) = P(A_1) + P(A_2) + P(A_3)$$
$$- (P(A_1 \cap A_2) + P(A_1 \cap A_3) + P(A_2 \cap A_3))$$
$$+ P(A_1 \cap A_2 \cap A_3).$$
Since the dice are rolled fairly, then by the equally probable outcomes theorem,

\[ P(A_i) = \frac{1}{6} \]

for \( i = 1, 2, 3 \).

Also, since the dice are rolled independently, then

\[ P(A_i \cap A_j) = P(A_i)P(A_j) = \frac{1}{36} \]

and

\[ P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k) = \frac{1}{216}. \]
Putting this all together, we get that

$$P(A) = 3\left(\frac{1}{6}\right) - 3\left(\frac{1}{36}\right) + \frac{1}{216} = \frac{91}{216} \approx 0.4212963.$$ 

As an alternative approach to this particular problem, you can compute that the complement of $A$ has probability

$$\left(\frac{5}{6}\right)^3 = \frac{125}{216}.$$
**EXAMPLE**

If you choose 7 cards at random from a standard deck, what is the probability of three of a kind?

Note: Four of a kind does not count as three of a kind. However, having three of a kind with two different types of card does count as three of a kind.
Let’s assume that order doesn’t matter, so the sample space $S$ is the set of unordered hands of 7 cards. All the outcomes are equally probable.

Let $A$ be the event that you get three of a kind.

By the equally probable outcomes theorem, $P(A) = \frac{|A|}{|S|}$.

By the counting combinations proposition, $|S| = \binom{52}{7} = 133,784,560$. 
To compute, $|A|$, let $A_i$ be the event that you get three of a kind for the $i$-th kind of card. Then $\bigcup_{i=1}^{13} A_i = A$.

Since the intersection of three or more of the $A_i$ is empty, the principle of inclusion-exclusion tells us that

$$|A| = \sum_i |A_i| - \sum_{i \neq j} |A_i \cap A_j|.$$
Since order doesn’t matter, getting an outcome in $A_i$ can be viewed as a 2-step process: first choose 3 of the 4 cards of the $i$-th kind, and then choose 4 of the remaining 48 cards not of that kind.

By the multistage counting lemma, $|A|$ is the product of the number of possibilities at each stage.

By the counting combinations theorem, there are $\binom{4}{3}$ possibilities in the first stage and $\binom{48}{4}$ in the second, so

$$|A_i| = \binom{4}{3} \binom{48}{4} = 778,320.$$
Similarly getting an outcome in $A_i \cap A_j$ can be viewed as a 3-step process: first choose 3 of the 4 cards of the $i$-th kind, then choose 3 of the 4 cards of the $j$-th type, and then choose 1 of the remaining 44 cards not of those kinds.

By the **multistage counting lemma**, $|A|$ is the product of the number of possibilities at each stage.

By the **counting combinations theorem**, there are $\binom{4}{3}$ possibilities in the first stage, $\binom{4}{3}$ in the second, and $\binom{44}{1}$ in the third, so

$$|A_i \cap A_j| = \binom{4}{3} \binom{4}{3} \binom{44}{1} = 704.$$
There are 13 one-fold intersections, and by the **counting combinations theorem**, there are \( \binom{13}{2} \) two-fold intersections.

Putting this all together, we have that

\[
|A| = (13) \left( \binom{4}{3} \binom{48}{4} \right) - \binom{13}{2} \left( \binom{4}{3} \binom{4}{3} \binom{44}{1} \right)
\]

\[
= 10,063,248
\]
This gives us that

$$P(A) = \frac{|A|}{|S|} = \frac{10,063,248}{133,784,560} \approx 0.0752198$$
EXAMPLE

Pick 13 cards at random out of a deck of 52. What is the probability that you will not get a card from every suit?

Something to ponder: it might be helpful to compare this to the case where there are only 2 suits with 4 cards per suit, and you are drawing 4 cards.
Suppose that order doesn’t matter, so that the sample space $S$ consists of 13-card hands. All the outcomes in $S$ are equally probable.

By the **counting combinations proposition**,

$$|S| = \binom{52}{13}.$$

Let $A$ be the event that you don’t get a card from every suit.

Since all the outcomes are equally probable, then by the **equally probable outcomes theorem**,

$$P(A) = \frac{|A|}{|S|}.$$
To find $|A|$, let $A_i$ be the event that you don’t get a card from the $i$-th suit. Then $A = A_1 \cup A_2 \cup A_3 \cup A_4$.

Since the intersection of four or more of the $A_i$ is empty, the principle of inclusion-exclusion tells us that

$$|A| = \sum_i |A_i| - \sum_{i \neq j} |A_i \cap A_j| + \sum_{\text{distinct } i,j,k} |A_i \cap A_j \cap A_k|.$$
By the **counting combinations theorem**,

\[ |A_i| = \binom{39}{13}, \quad |A_i \cap A_j| = \binom{26}{13}, \quad \text{and} \quad |A_i \cap A_j \cap A_k| = \binom{13}{13}, \]

for all distinct \( i, j, k \in \{1, 2, 3, 4\} \).

There are \( 4 = \binom{4}{1} \) of the \( A_i \).

By the **counting combinations theorem**, there are \( \binom{4}{2} \) two-fold intersections, and \( \binom{4}{3} \) 3-fold intersections.
Putting all this together, we have that

\[
|A| = \binom{4}{1} \binom{39}{13} - \binom{4}{2} \binom{26}{13} + \binom{4}{3} \binom{13}{13} = 32,427,298,180,
\]

which implies that

\[
P(A) = \frac{|A|}{|S|} = \frac{32,427,298,180}{635,013,559,600} \approx 0.0510655.
\]
As a follow-up to this example, what if we had only 8 cards and 2 suits, and we drew 4 cards?
Adapting the previous approach, we find that:

\[ |A| = \binom{2}{1} \binom{4}{4} - \binom{2}{2} \binom{0}{4} = 2 \]

and

\[ |S| = \binom{8}{4} = 70. \]

By the equally probable outcomes theorem, this implies that

\[ P(A) = \frac{\binom{2}{1} \binom{4}{4} - \binom{2}{2} \binom{0}{4}}{\binom{8}{4}} = \frac{2}{70} \approx 0.0285714. \]
If you solved the larger version well, solving the smaller version should be no problem.

And if you solved the smaller version well, solving the larger version should also be no problem!

So even if you’re assigned the larger version, it might make sense to try the smaller version first.
EXAMPLE

Find a formula for the area of a triangle on the unit sphere in terms of its angles.
To find this area, use the principle of inclusion-exclusion on the surfaces of orange slices subtended by the angles of the triangle.

The result is:

\[ \text{Area} = a + b + c - \pi, \]

where \( a, b, c \) are the angles of the triangle (measured in radians).
DEFINITION

Let $A, B$ be events with $P(B) \neq 0$. The conditional probability of event $A$ given event $B$, denoted by $P(A|B)$, is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$
EXAMPLE

Suppose you roll a 6-sided die fairly. Let $A$ be the event that the roll is even, and $B$ be the event that it is not 1. What are $P(A|B)$ and $P(B|A)$?
We have that

\[ A = \{2, 4, 6\} \quad \text{and} \quad B = \{2, 3, 4, 5, 6\}. \]

This implies that \( A \cap B = \{2, 4, 6\} \).

Since the die is rolled fairly, all 6 outcomes in the sample space are equally probable, so by the **equally probable outcomes theorem**:

\[ P(A) = 3/6 \quad \text{and} \quad P(B) = 5/6 \quad \text{and} \quad P(A \cap B) = 3/6. \]
We compute directly from the definition of conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{3/6}{5/6} = \frac{3}{5}.$$ 

By the same definition,

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{3/6}{3/6} = 1.$$ 

These values should make sense intuitively for this particular example.
**Example**

Flip a coin twice fairly and independently. Let $A$ be the event that the first flip lands *heads*, and let $B$ be the event that the second flip lands *tails*. What are $P(A|B)$ and $P(B|A)$?
The sample space for this random process is:

$$S = \{HH, HT, TH, TT\}.$$  

Also,

$$A = \{HH, HT\} \quad \text{and} \quad B = \{HT, TT\}. $$

This implies that $$A \cap B = \{HT\}.$$ 

Since the rolls are done fairly and independently, the outcomes in $$S$$ are all equally probable.
By the equally probable outcomes theorem:

\[ P(A) = \frac{2}{4} \quad \text{and} \quad P(B) = \frac{2}{4} \quad \text{and} \quad P(A \cap B) = \frac{1}{4}. \]
We compute directly from the definition of conditional probability:

\[ P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/4}{2/4} = \frac{1}{2}. \]

By the same definition,

\[ P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{1/4}{2/4} = \frac{1}{2}. \]
Notice that in this example

\[ P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B). \]

In other words, knowing that one of the events happened gave us no further information about whether the other event happened.

This leads us to the following proposition...
**PROPOSITION**

Let $A, B$ be events with $P(B) \neq 0$. Then $A$ and $B$ are independent if and only if

$$P(A|B) = P(A).$$
To prove the forwards direction, assume that \( A \) and \( B \) are independent.

Then we compute directly from the definition of conditional probability:

\[
P(A|B) = \frac{P(A \cap B)}{P(B)}
\]

\[
= \frac{P(A)P(B)}{P(B)} \quad \text{(by the definition of independent)}
\]

\[
= P(A).
\]
To prove the backwards direction, assume that \( P(A|B) = P(A) \). Then by the definition of conditional probability:

\[
P(A \cap B) = P(A|B)P(B) = P(A)P(B) \quad \text{(by assumption)}.
\]
This proposition tells us that the term *independent* is appropriately defined: events $A$ and $B$ with nonzero probability are independent if knowing that one has happened doesn’t give any information about whether the other will happen.

Note that this applies only to events with nonzero probability.

Our definition of independence applies to *all* events, not just those of nonzero probability, so it is a better definition to work with.

This proposition is conceptually useful, but our definition of independence is often more useful when doing computations.
**PROPOSITION (CONDITIONAL TREES)**

Let $A_1, \ldots, A_n$ be events with nonzero probability. Then

\[
P(A_n \cap A_{n-1} \cap \ldots \cap A_1)
= P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap \cdots \cap A_{n-1}).
\]
We prove this proposition by induction on \( n \).

For the base case, when \( n = 1 \), the statement is

\[
P(A_1) = P(A_1).
\]

To show us how to do the inductive step, we look at the case when \( n = 2 \). The statement to prove is:

\[
P(A_1 \cap A_2) = P(A_1)P(A_2|A_1).
\]

This follows immediately from the definition of conditional probability.
For the inductive step, we assume that for some positive integer $n$:

$$P(A_n \cap A_{n-1} \cap \ldots \cap A_1) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap \cdots \cap A_{n-1}).$$

By the definition of conditional probability,

$$P(A_{n+1} \cap A_n \cap \ldots \cap A_1) = P(A_n \cap \ldots \cap A_1)P(A_{n+1}|A_n \cap \ldots \cap A_1).$$

By the inductive assumption, this equals

$$(P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap \cdots \cap A_{n-1})) P(A_{n+1}|A_n \cap \ldots \cap A_1),$$

which proves the inductive step and the proposition.
This shows how trees work.

In a tree representing a multi-stage process, the edges are labeled with the conditional probabilities, and we want to compute

$$P(A_1 \cap A_2 \cap \cdots \cap A_n),$$

where each $A_i$ is an event in the $i$-th stage.

According to the proposition, this probability is the product of all the conditional probabilities along the path leading to the corresponding outcome on the tree.
**Example**

Consider flipping a coin fairly and independently three times, where the probability of $H$ each time is $p$. What is $P(HHT)$?
According to the theorem,

\[ P(H_1H_2T_3) = P(H_1)P(H_2|H_1)P(T_3|H_1 \cap H_2). \]

Even if they weren’t independent, we could compute this from the associated tree.

But since these are independent, then of course by the definition of independent events this equals

\[ P(H_1)P(H_2)P(T_3). \]
PROPOSITION (PARTITIONED PROBABILITY)

If $B_1, B_2, \ldots, B_k$ are pairwise disjoint events with $\bigcup_{i=1}^k B_i = S$ and each $P(B_i)$ having nonzero probability, then for any event $A$:

$$P(A) = \sum_{i=1}^k P(A|B_i)P(B_i)$$
To prove this, we first use some basic properties of sets:

\[
P(A) = P(A \cap S) \\
= P(A \cap (\bigcup_{i=1}^{k} B_i)) \\
= P(\bigcup_{i=1}^{k} (A \cap B_i)).
\]
Since $B_1, \ldots, B_k$ are pairwise disjoint, this equals:

$$\sum_{i=1}^{k} P(A \cap B_i).$$

By the definition of conditional probability, this equals

$$\sum_{i=1}^{k} P(B_i)P(A|B_i),$$

which proves the proposition.
**Theorem (Bayes’ Theorem)**

If $A, B$ are events with nonzero probabilities, then

\[
P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}
\]

More generally, Bayes’ theorem states that if $S$ is the sample space, and if $B_1, \ldots, B_k$ are pairwise disjoint events with nonzero probabilities with $\bigcup_{i=1}^{k} B_i = S$, then

\[
P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + \cdots + P(A|B_k)P(B_k)}.
\]
To prove this, note that by the definition of conditional probability, we have

\[ P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)}. \]

Again by the definition of conditional probability, this equals

\[ P(A|B)P(B) \]

By the partitioned probability proposition, this equals

\[ \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}. \]

To prove the more general version, simply apply the partitioned probability proposition to \( B_1, \ldots, B_k \) instead of to \( B, B^c \).
Bayes’ theorem can also be written:

\[
P(B|A) = P(B) \frac{P(A|B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}.
\]

From a subjective (or Bayesian) point of view, this says that to update your belief about \(B\) with the new information \(A\), multiply \(P(B)\) by the fraction above.

This has major ramifications for the scientific process, and for updating one’s beliefs in general.

As an aside, note that if \(P(B) = 0\) or \(P(B) = 1\), then \(P(B)\) never changes with new information.
EXAMPLE

Suppose you’re not sure which of two types of coins someone is using. You believe a coin has a probability of $\frac{2}{3}$ of being fair and a $\frac{1}{3}$ probability of being $H$ all the time. If we flip the coin and it comes up $HHHHH$, what probabilities do you now believe about the coin (assuming that your beliefs are rational)?
Let $F$ be the event that the coin being used is fair.

Bayes’ theorem implies that

$$P(F|HHHHH) = \frac{P(HHHHH|F)P(F)}{P(HHHHH|F)P(F) + P(HHHHH|F^c)P(F^c)}$$

$$= \frac{\left(\frac{1}{32}\right)\left(\frac{2}{3}\right)}{\left(\frac{1}{32}\right)\left(\frac{2}{3}\right) + (1)(\frac{1}{3})}$$

$$= \frac{1}{17}$$

$$\approx 0.0588235.$$
The following example illustrates a common application of Bayes’ theorem.

**Example**

Suppose that a certain drug test is 99% sensitive (meaning $P(\text{positive}|\text{using}) = 0.99$) and 98% specific (meaning $P(\text{not positive}|\text{not using}) = 0.98$). Also, suppose that 0.5% of the population uses the drug. What is the probability that someone who tests positive is actually using the drug?
Let $U$ be the event that the person is using the drug.

Let $T$ be the event that the test is positive.

Bayes’ theorem implies that

$$P(U|T) = \frac{P(T|U)P(U)}{P(T|U)P(U) + P(T|U^c)P(U^c)}$$

$$= \frac{(0.99)(0.005)}{(0.99)(0.005) + (0.02)(0.995)}$$

$$= 0.1991952.$$

This is much smaller than most people would think, and it helps to explain why followup tests are usually needed even if a particular test is positive.
Begin test.
**Definition**

The **odds** of an event $A$ is defined to be

$$\text{odds}(A) = \frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)}.$$ 

Probability ranges from 0 to 1, while odds range from 0 to $\infty$.

You can obtain the probability from the odds and vice versa...
If $P(A) = p$ and odds$(A) = o$, then by the definition of odds,

$$o = \frac{p}{1 - p}.$$ 

This implies that

$$o - op = p,$$

so

$$p = \frac{o}{1 + o}.$$
Bayes’ theorem can also be expressed in terms of odds:

$$\text{odds}(A|B) = \frac{?}{?}.$$
We compute:

\[
\text{odds}(A|B) = \frac{P(A|B)}{P(A^c|B)} = \frac{P(B|A)P(A)}{(P(B|A)P(A) + P(B|A^c)P(A^c))} \frac{P(B|A^c)P(A^c)}{(P(B|A^c)P(A^c) + P(B|A)P(A))} = \frac{P(B|A)P(A)}{P(B|A^c)P(A^c)}.
\]
This gives us the odds form of Bayes’ theorem:

\[
\text{odds}(A|B) = \text{odds}(A) \frac{P(B|A)}{P(B|A^c)}.
\]

The term on the left is the \textit{(conditional) odds of } A \textit{ given } B.

The first term on the right is the \textit{unconditional odds of } A.

The second term on the right is called the \textit{likelihood ratio}.

In scientific reasoning, \(A\) is a hypothesis, and \(B\) is new data.
**Definition**

Let $X, Y$ be random variables. Then $(X, Y)$ is also random variable, and its distribution is called the **joint distribution** of $X$ and $Y$.

**Definition**

Let $X$ and $Y$ be discrete random variables. The **joint probability mass function** $f_{X,Y} : \mathbb{R}^2 \rightarrow \mathbb{R}$ of $X$ and $Y$ is defined by

$$f_{X,Y}(k, \ell) = P(X = k, Y = \ell)$$

for all $k, \ell$.

Ordinarily the joint distribution of two discrete random variables is specified by giving their joint probability mass function (by a formula or in a table).
**Example**

Roll two 4-sided dice fairly and independently, and let $X$ be the maximum of the two rolls and $Y$ be the sum of the two rolls. What is the joint distribution of $X$ and $Y$?
The joint distribution table is

<table>
<thead>
<tr>
<th></th>
<th>Y=2</th>
<th>Y=3</th>
<th>Y=4</th>
<th>Y=5</th>
<th>Y=6</th>
<th>Y=7</th>
<th>Y=8</th>
</tr>
</thead>
<tbody>
<tr>
<td>X=1</td>
<td>1/16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>X=2</td>
<td>0</td>
<td>1/8</td>
<td>1/16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>X=3</td>
<td>0</td>
<td>0</td>
<td>1/8</td>
<td>1/8</td>
<td>1/16</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>X=4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/8</td>
<td>1/8</td>
<td>1/8</td>
<td>1/16</td>
</tr>
</tbody>
</table>
**DEFINITION**

Let $X_1, X_2, \ldots, X_n, Y$ be random variables. The **marginal distribution** of $Y$ is defined to be the distribution of $Y$ ignoring $X_1, X_2, \ldots, X_n$. 
EXAMPLE

What are the marginal distributions of $X$ and $Y$ in the previous example?
Marginal for $X$ (for values of 1, 2, 3, 4):

\[
\begin{array}{cccc}
1 & 3 & 5 & 7 \\
16' & 16' & 16' & 16
\end{array}
\]

Marginal for $Y$ (for values of 1, 2, 3, 4, 5, 6, 7, 8):

\[
\begin{array}{ccccccc}
0, & 1 & 1 & 3 & 1 & 3 & 1 \\
16', & 8', & 16', & 4', & 16', & 8', & 16
\end{array}
\]

Of course, these should be displayed in tables rather than just as the numbers written here.
DEFINITION

Let $X_1, X_2, \ldots, X_n, Y$ be random variables. The **conditional distribution** of $Y$ given $X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n$ is the distribution of $Y$ when the values of $X_1, X_2, \ldots, X_n$ are known to be $k_1, k_2, \ldots, k_n$ respectively.
**Example**

What are the conditional distributions of $X|Y = 5$ and of $Y|X = 3$?
The conditional distribution of $X \mid Y = 5$ is

\[
0, 0, \frac{1}{2}, \frac{1}{2}
\]

The conditional distribution of $Y \mid X = 3$ is

\[
0, 0, 0, \frac{2}{5}, \frac{2}{5}, \frac{1}{5}, 0, 0
\]

Again, these should be formatted better.
**Example**

Suppose that $X$ is a student's score on the AP calculus exam, and $Y =$ grade in calculus class

Here is the joint distribution:

<table>
<thead>
<tr>
<th></th>
<th>4</th>
<th>3</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>.1</td>
<td>.05</td>
<td>0</td>
</tr>
<tr>
<td>$B$</td>
<td>.15</td>
<td>.15</td>
<td>0</td>
</tr>
<tr>
<td>$C$</td>
<td>.1</td>
<td>.15</td>
<td>.1</td>
</tr>
<tr>
<td>$D$</td>
<td>0</td>
<td>.05</td>
<td>.1</td>
</tr>
<tr>
<td>$F$</td>
<td>0</td>
<td>0</td>
<td>.05</td>
</tr>
</tbody>
</table>

What are the marginal and, say, two conditional distributions?
**Definition**

Two discrete random variables $X$ and $Y$ are called **independent** if

$$f_{X,Y}(k, \ell) = f_X(k)f_Y(\ell)$$

for all $k, \ell$. 
Are the previous (two examples ago, and the most recent example) $X$ and $Y$ independent?
**Proposition**

For any random variables $X$, $Y$ and real number $a$:

- $E(X + Y) = E(X) + E(Y)$
- $E(aX) = aE(X)$

We have already shown the second of these, directly from the definition of $E(X)$. 
To prove the first, let $V_X$ and $V_Y$ denote the set of possible values for $X$ and $Y$.

Then the set of possible values for $X + Y$ is

$$\{x + y \mid x \in V_x, y \in V_y\}.$$  

Also, by the definition of expected value,

$$E(X + Y) = \sum_{x \in V_X} \sum_{y \in V_Y} P(X = x, Y = y)(x + y)$$

$$= \sum_{x \in V_X} x \sum_{y \in V_Y} P(X = x, Y = y) + \sum_{y \in V_Y} y \sum_{x \in V_X} P(X = x, Y = y)$$

$$= \sum_{x \in V_X} xP(X = x) + \sum_{y \in V_Y} yP(Y = y)$$

$$= E(X) + E(Y).$$
**PROPOSITION**

*If random variables $X$ and $Y$ are independent, then*

$$E(XY) = E(X)E(Y).$$

*Note that independence is required for this.*
Use $V_X$ and $V_Y$ to denote the set of possible values for $X$ and $Y$. Then, using that these are independent random variables (to get the second line):

$$E(XY) = \sum_{x \in V_X} \sum_{y \in V_Y} xyP(X = x, Y = y)$$

$$= \sum_{x \in V_X} \sum_{y \in V_Y} xyP(X = x)P(Y = y)$$

$$= \sum_{x \in V_X} xP(X = x) \sum_{y \in V_Y} yP(Y = y)$$

$$= E(X)E(Y).$$
PROPOSITION

If random variables $X$ and $Y$ are independent, then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

This is emphatically not true if $X$ and $Y$ are not independent.
To prove this, we compute:

\[
\text{Var}(X + Y) = E((X + Y)^2) - (E(X + Y))^2
\]

\[
= E(X^2 + 2XY + Y^2) - (E(X))^2 - 2E(X)E(Y) - (E(Y))^2
\]

\[
= E(X^2) + 2E(XY) + E(Y^2) - (E(X))^2 - 2E(X)E(Y) - (E(Y))^2
\]

\[
= \text{Var}(X) + \text{Var}(Y) + 2(E(XY) - E(X)E(Y))
\]

\[
= \text{Var}(X) + \text{Var}(Y).
\]
The quantity in parentheses is called the *covariance* of $X$ and $Y$.

**Definition**

Let $X, Y$ be random variables. The **covariance** of $X$ and $Y$ is defined to be

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y).$$

From the above proof, we can see that twice the covariance is how far the variance is from being additive for the two variables.
Notice that if $X, Y$ are independent, then $E(XY) = E(X)E(Y)$, so $\text{Cov}(X, Y) = 0$.

It is \textit{not} true that if $\text{Cov}(X, Y) = 0$, then $X, Y$ are independent, however.
According to Wikipedia, the Monty Hall problem is “loosely based on the American television game show Let’s Make a Deal and named after its original host, Monty Hall”.

It “was originally posed (and solved) in a letter by Steve Selvin to the American Statistician in 1975. It became famous as a question from a reader’s letter quoted in Marilyn vos Savant’s “Ask Marilyn” column in Parade magazine in 1990.

The way she phrased it was as follows...
Example (The Monty Hall Problem)

Suppose you’re on a game show, and you’re given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what’s behind the doors, opens another door, say No. 3, which has a goat. He then says to you, “Do you want to pick door No. 2?” Is it to your advantage to switch your choice?
Number the doors so that you always choose Door 1 and the host always opens Door 3.

$C_i$ is the car being behind Door $i$, and $H_3$ is the event that the host opens Door 3. (Number so that these are correct.) Then:

$$P(C_2 | H_3) = ?$$

$$P(C_i) = \frac{1}{3}$$

Also, the host's protocol is:

$$P(H_3 | C_1) = \frac{1}{2}$$
$$P(H_3 | C_2) = 1$$
$$P(H_3 | C_3) = 0.$$
So

\[ P(C_2|H_3) = \frac{P(H_3|C_2)P(C_2)}{P(H_3|C_1)P(C_1) + P(H_3|C_2)P(C_2) + P(H_3|C_3)P(C_3)} \]

\[ = \frac{(1)(1/3)}{(1/2)(1/3) + (1)(1/3) + (0)(1/3)} \]

\[ = \frac{2/3}{1} \]

\[ = \frac{2}{3}. \]
Notice that to solve this problem, you need to know the host’s protocol, which is not given in Marilyn vos Savant’s formulation of the problem.

The problem may be unintuitive, and that contributed to the overall confusion, but as stated in Parade magazine, it couldn’t be solved without additional information.

That also contributed to the confusion over it.
Let’s make a deal:
https://en.wikipedia.org/wiki/Monty_Hall_problem


Or: http://www.montyhallproblem.com/

Or: https://math.ucsd.edu/~crypto/Monty/monty.html
Simpson’s paradox is a phenomenon in which a pattern observed in a variable reverses in every case when the variable is conditioned on another random variable.
Simpson’s paradox: Alaska Airlines

<table>
<thead>
<tr>
<th>Airport</th>
<th>No. On-time</th>
<th>No. Delayed</th>
<th>Pct Delayed</th>
</tr>
</thead>
<tbody>
<tr>
<td>los angeles</td>
<td>497</td>
<td>62</td>
<td>11.1</td>
</tr>
<tr>
<td>phoenix</td>
<td>221</td>
<td>12</td>
<td>5.2</td>
</tr>
<tr>
<td>san diego</td>
<td>212</td>
<td>20</td>
<td>8.6</td>
</tr>
<tr>
<td>san francisco</td>
<td>503</td>
<td>102</td>
<td>16.9</td>
</tr>
<tr>
<td>seattle</td>
<td>1841</td>
<td>305</td>
<td>14.1</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>3274</strong></td>
<td><strong>501</strong></td>
<td><strong>13.3</strong></td>
</tr>
</tbody>
</table>

America West

<table>
<thead>
<tr>
<th>Airport</th>
<th>No. On-time</th>
<th>No. Delayed</th>
<th>Pct Delayed</th>
</tr>
</thead>
<tbody>
<tr>
<td>los angeles</td>
<td>694</td>
<td>117</td>
<td>14.4</td>
</tr>
<tr>
<td>phoenix</td>
<td>4840</td>
<td>415</td>
<td>7.9</td>
</tr>
<tr>
<td>san diego</td>
<td>383</td>
<td>65</td>
<td>14.5</td>
</tr>
<tr>
<td>san francisco</td>
<td>320</td>
<td>129</td>
<td>28.7</td>
</tr>
<tr>
<td>seattle</td>
<td>201</td>
<td>61</td>
<td>23.3</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>6438</strong></td>
<td><strong>787</strong></td>
<td><strong>10.9</strong></td>
</tr>
</tbody>
</table>
Several other examples can be found at the Simpson's paradox wikipedia page: https://en.wikipedia.org/wiki/Simpson%27s_paradox
EXAMPLE

A frog named Andrei lives on two lily pads. At the start of each day he flips a coin and decides whether or not to hop to the other lily pad based on the result: $H$ means hop, while $T$ means stay in place.

The coin on Lily Pad 1 has $P(H) = 0.7$, while the coin on Lily Pad 2 has $P(H) = 0.4$.

If Andrei begins on Lily Pad 1, what is the probability that he’ll be on Lily Pad 1 at the end of the day? At the end of two days? At the end of $n$ days?
We can draw the situation as follows:

![Diagram](image)

In this picture, the two lily pads are **states**.

A **distribution** is a probability distribution on these states, meaning a nonnegative real number assigned to each state, in such a way that the numbers sum to 1.
The first question is easy to answer: the only way that he can still be on Lily Pad 1 at the end of the day is if he stays that morning, which happens with probability 0.3.
To answer the second question, let’s set up a transition matrix $T = [t_{i,j}]$ for which

$$t_{i,j} = P(\text{Lily Pad } i \text{ at end of day} | \text{Lily Pad } j \text{ at start of day}).$$

This means

$$T = \begin{bmatrix} 0.3 & 0.4 \\ 0.7 & 0.6 \end{bmatrix}.$$ 

Notice that $T$ is column stochastic, meaning the entries are all nonnegative, and the columns sum to 1.
To answer the second question, we use the law of total probability.

Let $A_i$ be the event that Andrei is on Lily Pad 1 after $i$ days, and let $B_i$ be the event that Andrei is on Lily Pad 2 after $i$ days. Then

$$P(A_2|A_0) = P(A_2, A_1|A_0) + P(A_2, B_1|A_0)$$

$$= P(A_2|A_1, A_0)P(A_1|A_0) + P(A_2|B_1, A_0)P(B_1|A_0)$$

$$= t_{1,1}t_{1,1} + t_{2,1}t_{1,2}.$$
This is equal to the first row of $T$ dotted with the first column of $T$.

That is equal to the $(1,1)$ entry of $T^2$. Coincidence? I think not.

This indicates that putting the transition probabilities into a matrix was probably a good idea.
This has given us a basic illustration of what a Markov chain is.

There are several ways to view Markov chains, such as...
1. Many Markov chains can be viewed as a *finite state automaton* consisting of a finite nonempty set $S$ of *states*, an initial state $s_0$, and a collection $p_{ij}$ of *transition probabilities* from $s_j$ to $s_i$.

2. A Markov chain is be defined to be a collection of random variables $X_0, X_1, X_2, \ldots$ with values in a finite nonempty set $S$ and satisfying the *Markov condition*:

$$P(X_i = s_{\ell_i} \mid X_0 = s_{\ell_0}, X_1 = s_{\ell_1}, \ldots, X_{i-1} = s_{\ell_{i-1}}) = P(X_i = s_{\ell_i} \mid X_{i-1} = s_{\ell_{i-1}})$$

for all $i > 0$ and all $s_{\ell_1}, s_{\ell_2}, \ldots, s_{\ell_{i-1}}, s_{\ell_i} \in S$.

3. A Markov chain can also be given by an initial distribution of states and an $n \times n$ column stochastic matrix $T$ (where $n = |S|$) called the *transition matrix*. The term *column stochastic* means that the sum of the entries in each column equals 1.
The first description is conceptually useful.

The second description is the most general, and it generalizes (to continuous space and time) most easily.

The third description emphasizes the connection with linear algebra.
Theorem (Chapman-Kolmogorov Theorem)

Let $n$ be a positive integer. For a Markov chain $X_1, X_2, \ldots$ with 1-step transition matrix $T$, the $n$-step transition matrix is given by $T^n$. 
We prove this by induction on $n$.

The base case (where $n = 1$) is immediate.

For the inductive step, assume that for some positive integer $n$, the $n$-step transition matrix is given by $T^n$. 
We want to compute

\[ P(X_{n+1} = i | X_0 = j). \]

For this it helps to realize how \( X_{n+1} = i \) might arise.

Either \( X_{n+1} = i \) and \( X_n = 1 \), or else \( X_{n+1} = i \) and \( X_n = 2 \), or else\ldots, or else \( X_{n+1} = i \) and \( X_n = s \) (where \( s \) is the number of states).

Since the events \( X_n = 1, X_n = 2, \ldots, X_n = s \) partition the sample space, we can apply the partitioned probability proposition\ldots
\[ P(X_{n+1} = i | X_0 = j) = \sum_{\ell = 1}^{k} P(X_{n+1} = i | X_n = \ell, X_0 = j) P(X_n = \ell | X_0 = j) \]

\[ = \sum_{\ell = 1}^{k} P(X_{n+1} = i | X_n = \ell) P(X_n = \ell | X_0 = j) \]

\[ = \sum_{\ell=1}^{k} [T_i]_{\ell, j} [T^n]_{\ell, j} \]

\[ = \sum_{\ell=1}^{k} [T^{n+1}]_{i, j}. \]
In other words, the Chapman-Kolmogorov theorem is really a specific case of the partitioned probability proposition.
DO NOT DO THIS IN YOUR ASSIGNMENTS! THIS IS ONLY TO SHOW THE MOTIVATION FOR WHAT COMES NEXT. YOU WILL BE COUNTED OFF HEAVILY IF YOU DO THIS INSTEAD OF CITING THE PROPER THEOREM.

Let’s look at some $n$-step transition matrices for the frog example:

\[
T = \begin{bmatrix} 0.3 & 0.4 \\ 0.7 & 0.6 \end{bmatrix}, \quad T^{10} = \begin{bmatrix} 0.3636363637 & 0.3636363636 \\ 0.6363636363 & 0.6363636364 \end{bmatrix}
\]

and

\[
T^{100} = \begin{bmatrix} 0.3636363636363632 & 0.3636363636363632 \\ 0.6363636363636334 & 0.6363636363636334 \end{bmatrix}
\]

The columns of this are very close to \[ \begin{bmatrix} 4/11 \\ 7/11 \end{bmatrix} \]. Why?
Suppose that \( \lim_{n \to \infty} T^n \) exists, and write it as

\[
S = \lim_{n \to \infty} T^n.
\]

Since linear transformations can be passed through limits (when they exist), then

\[
TS = T \lim_{n \to \infty} T^n = \lim_{n \to \infty} T^{n+1} = \lim_{n \to \infty} T^n = S.
\]
If the columns of $S$ are $v_1, v_2$, then the columns of $TS$ are $T(v_1), T(v_2)$.

Since $TS = S$, then

$$T(v_1) = v_1 \quad \text{and} \quad T(v_2) = v_2.$$ 

That is, $v_i$ is an eigenvector of $T$ with eigenvalue 1.

Also, each $v_i$ is a probability distribution, so its entries are nonnegative and sum to 1.
**Definition**

For an $s$-state Markov chain with transition matrix $T$, a stationary distribution is a vector $\mathbf{v} \in \mathbb{R}^s$ satisfying:

1. The entries of $\mathbf{v}$ are all non-negative,
2. The entries of $\mathbf{v}$ sum to 1, and
3. $\mathbf{v}$ is an eigenvector of $T$ with eigenvalue 1 (meaning $T(\mathbf{v}) = \mathbf{v}$).
In the frog example, both columns appear to be the same eigenvector.

Does it always happen that \( \lim_{n \to \infty} T^n \) exists, and if so, does it always converge to a matrix with identical columns?
**Theorem (Markov Chain Convergence Theorem)**

For any $s$-state Markov chain with transition matrix $T$, if there exists a positive integer $k$ such that all the entries of $T^k$ are nonzero, then:

- $1$ is the largest eigenvalue of $T$, and it occurs with (algebraic and geometric) multiplicity $1$.
- The Markov chain has a unique stationary distribution $\mathbf{v}$, and
- $\lim_{n \to \infty} T^n = S$ exists, and the columns of $S$ are all $\mathbf{v}$.

These imply that

$$\lim_{n \to \infty} T^n \mathbf{x} = \mathbf{v}$$

for all distributions $\mathbf{x} \in \mathbb{R}^s$. 
This tells us that the frog example illustrates a more general phenomenon.

The transition matrix is

\[ T = \begin{bmatrix} 0.3 & 0.4 \\ 0.7 & 0.6 \end{bmatrix}. \]

Since \( T^1 \) has all nonzero entries, then the Markov Chain Convergence Theorem applies.
In linear algebra class, you learned how to find the eigenvectors of $T$ and their eigenvalues by hand, and that’s good practice.

However, in this class, you are welcome to use a computer to obtain them.
WolframAlpha gives us an eigenvector of $T$ with eigenvalue 1 as:
\[
\begin{bmatrix}
\frac{4}{7} \\
1
\end{bmatrix}.
\]
Dividing this by the sum of the entries, we get the unique stationary state to be:
\[
\begin{bmatrix}
\frac{4}{11} \\
\frac{7}{11}
\end{bmatrix} \approx \begin{bmatrix} 0.36364 \\ 0.63636 \end{bmatrix}.
\]
By the Markov Chain Convergence Theorem, this is the distribution to which the Markov Chain will converge as the number of steps that it runs tends to $\infty$, no matter what the initial distribution is.
EXAMPLE
Consider a miniature board game consisting of 6 spaces arranged in a circle. To play the game, first place your game piece on one of the spaces. Then for a turn, flip two coins fairly and independently, and advance one space for each $H$. If you play this game for a really long time (lots of turns), with about what probability will your piece be on each of the 6 spaces?
We model this as a Markov chain in which each space on the board is a state, so there are 6 states, which we will number 1 through 6.

Since each state is a space on the board, the game rules mean that we advance two spaces with probability $1/4$, one space with probability $1/2$, and zero spaces with probability $1/4$.

With the states labeled and ordered in this way, this means that the transition matrix for the Markov chain is:

$$ T = \begin{bmatrix} 
1/4 & 0 & 0 & 0 & 1/4 & 1/2 \\
1/2 & 1/4 & 0 & 0 & 0 & 1/4 \\
1/4 & 1/2 & 1/4 & 0 & 0 & 0 \\
0 & 1/4 & 1/2 & 1/4 & 0 & 0 \\
0 & 0 & 1/4 & 1/2 & 1/4 & 0 \\
0 & 0 & 0 & 1/4 & 1/2 & 1/4 \\
\end{bmatrix} $$
Although $T$ has some zero entries,

$$T^3 \approx \begin{bmatrix} 0.031 & 0.094 & 0.234 & 0.312 & 0.234 & 0.094 \\ 0.094 & 0.031 & 0.094 & 0.234 & 0.312 & 0.234 \\ 0.234 & 0.094 & 0.031 & 0.094 & 0.234 & 0.312 \\ 0.312 & 0.234 & 0.094 & 0.031 & 0.094 & 0.234 \\ 0.234 & 0.312 & 0.234 & 0.094 & 0.031 & 0.094 \\ 0.094 & 0.234 & 0.312 & 0.234 & 0.094 & 0.031 \end{bmatrix}$$

has all nonzero entries, so the Markov Chain Convergence Theorem applies.
Using WolframAlpha, we find an eigenvector with eigenvalue 1 is

\[
\begin{bmatrix}
1 \\
1 \\
1 \\
1 \\
1 \\
1 \\
1
\end{bmatrix}
\]

Dividing this by the sum of the entries, we get that the unique stationary distribution guaranteed by the Markov Chain Convergence Theorem is:

\[
\begin{bmatrix}
1/6 \\
1/6 \\
1/6 \\
1/6 \\
1/6 \\
1/6 \\
1/6
\end{bmatrix}
\]
By the Markov Chain Convergence Theorem, this Markov chain will converge to the unique stationary distribution no matter what its initial distribution.

This means that no matter where we place our board game piece at the start of the game, after the game has been played for a long time, there is a probability of approximately 1/6 that our piece will be on each of the spaces.
(From Durrett, *Elementary Probability Theory for Applications*, page 142, Example 4.25)

**Example**

Suppose a 2-year college has four statuses for students: 1 (Year 1), 2 (Year 2), D (dropped out), and G (graduated). The probabilities of:

- repeating Year 1 is 0.25.
- continuing to Year 2 from Year 1 is 0.6.
- dropping out after Year 1 is 0.15.
- repeating Year 2 is 0.2.
- dropping out after Year 2 is 0.1.
- graduating after Year 2 is 0.7.

How long on average does a student spend at this school?
we model this as a Markov Chain whose states are: Year 1, Year 2, Dropout, Graduate.

With the states in this order, the information given translates into the following transition matrix:

\[
T = \begin{bmatrix}
0.25 & 0 & 0 & 0 \\
0.6 & 0.2 & 0 & 0 \\
0.15 & 0.1 & 1 & 0 \\
0 & 0.7 & 0 & 1
\end{bmatrix}.
\]
Since we are not being asked about long-term convergence, the Markov Chain Convergence Theorem doesn’t help here. (And in fact, it doesn’t apply here.)

Instead we take a different approach.
What makes this problem of a fundamentally different type is that this Markov chain has *absorbing states*.

**Definition**

An *absorbing state* of a Markov chain is a state for which the probability of transitioning out of it (to any other state) is 0.
As this example illustrates, one question we often ask about Markov chains with absorbing states is:

Given an initial distribution, how long on average does it take to reach an absorbing state?

We don’t ask *which* absorbing state; we just ask about how long until entering *some* absorbing state.
To address this question, the following concept is useful:

**Definition**

The reduced matrix $R$ of Markov chain is the transition matrix of the Markov chain with the rows and columns that correspond to absorbing states removed.

The row and column indices of $R$ will be the same as those of $T$ if you place the absorbing states at the end of the ordered list of states.

In what follows, we assume that the states are ordered in this way.
Assume the initial state is $j$.

Let $C_n$ be the Bernoulli random variable for which success is being in State 1 (Year 1) in the $n$-th year of the Markov chain.
Then the number of years spent in State 1 (Year 1) is \( \sum C_n \), so the average number of years in State 1 is

\[
E(\sum_{n} C_n) = \sum_{n=0}^{\infty} E(C_n) = \sum_{n=0}^{\infty} P(X_n = 1|X_0 = j)
\]

\[
= \sum_{n=0}^{\infty} [T^n]_{1,j}
\]

\[
= [\sum_{n=0}^{\infty} T^n]_{1,j}
\]

\[
= [\sum_{n=0}^{\infty} R^n]_{1,j}
\]

\[
= [(I - R)^{-1}]_{1,j}.
\]
The step in which $R$ replaces $T$ holds because the absorbing columns play no role in computing those entries.

Because $T$ is a transition matrix, then there is a theorem (which we won’t prove here) that establishes that $(I - R)^{-1}$ exists.

The last equation above is the matrix analogue of the formula for a geometric sum:

$$
\sum_{n=0}^{\infty} r^n = \frac{1}{1 - r}.
$$
This gives us a method to find average times spent in each state prior to absorption.

**Proposition (Average Absorption Times)**

For a Markov chain with absorbing states, the expected value of the number of steps spent in State $i$ given that the initial state was State $j$ is

$$[(I - R)^{-1}]_{i,j},$$

meaning the $(i,j)$ entry of $(I - R)^{-1}$.

Note that the initial state determines the column of $(I - R)^{-1}$ to look at, and the state in which you’re computing expected value of steps spent in gives you the row.
In this particular case, we compute that

\[ R = \begin{bmatrix} 0.25 & 0 \\ 0.6 & 0.2 \end{bmatrix}, \quad I - R = \begin{bmatrix} 0.75 & 0 \\ -0.6 & 0.8 \end{bmatrix} \]

and

\[ (I - R)^{-1} = \begin{bmatrix} 4/3 & 0 \\ 1 & 5/4 \end{bmatrix} \approx \begin{bmatrix} 1.33333333333333 & 0 \\ 1 & 1.25 \end{bmatrix} \]

So, given an initial state of 1 (Year 1), the expected value of the amount of time spent in State 1 (Year 1) is 4/3 and in State 2 (Year 2) is 1, for a total of \(7/3 \approx 2.3333333\).
(From Durrett, *Elementary Probability Theory for Applications*, page 143, Example 4.26)

**Example**

In tennis starting at deuce, one must win by 2. Suppose that the server wins each point with probability 0.6 (independently of other points). Starting at deuce, on average how many points will be played before the game is over?
We model this as a Markov chain with states (in order): win, ad in, deuce, ad out, lose.

Since win is an absorbing state but is not at the end of the list, we’ll need to be careful about which states are which when we form the reduced matrix.

With this order of states, the transition matrix is:

\[
T = \begin{bmatrix}
1 & 0.6 & 0 & 0 & 0 \\
0 & 0 & 0.6 & 0 & 0 \\
0 & 0.4 & 0 & 0.6 & 0 \\
0 & 0 & 0.4 & 0 & 0 \\
0 & 0 & 0 & 0.4 & 1
\end{bmatrix}.
\]
Since we’re looking at average time to absorption, we compute that

\[
R = \begin{bmatrix}
0 & 0.6 & 0 \\
0.4 & 0 & 0.6 \\
0 & 0.4 & 0
\end{bmatrix}
\quad \text{and} \quad
I - R = \begin{bmatrix}
1 & -0.6 & 0 \\
-0.4 & 1 & -0.6 \\
0 & -0.4 & 1
\end{bmatrix}.
\]

This gives us that

\[
(I - R)^{-1} = \begin{bmatrix}
1.462 & 1.154 & 0.692 \\
0.769 & 1.923 & 1.154 \\
0.308 & 0.769 & 1.462
\end{bmatrix}.
\]
Since the initial state is deuce (the middle column in both $T$ and $R$), we look at the middle column of $R$.

We want to know on average how many points are played, meaning the total number of points played in ad in, deuce, and ad out.

For this, we sum all the entries in $R$ in the column we’re interested in.

This gives us that the expected value of the total number of points played from deuce is

$$3.8461538.$$
As stated before, if a Markov chain has absorbing states, with probability 1 the chain will eventually reach an absorbing state.

This begs the question: with what probability will it reach each of the absorbing states?

Let’s look at this in the tennis example...
What is the probability of winning, starting from deuce?

To win (get to State 1), you need to transition around the non-absorbing states for \( n \) steps (for some \( n \)) to some intermediate non-absorbing state, and then transition to the \textit{win} State 1.

So the total probability of winning (getting to State 1) starting in deuce (which is Column 2 in \( R \)) is:

\[
\sum_{j} \sum_{n=0}^{\infty} [T]_{1,j} [R^n]_{j,2} = \sum_{j} T_{1,j} \sum_{n=0}^{\infty} [R^n]_{j,2} = \sum_{j} T_{1,j} [(I - R)^{-1}]_{j,2}.
\]

In other words, we’re taking the dot product of the “win” row 1 of \( T \) (only the non-absorbing columns) with the “deuce” column 2 of \( (I - R)^{-1} \).
This gives us a way to compute the probability of eventually entering a particular absorbing state, given an initial state:

**Proposition (Absorption Probabilities)**

Suppose that State $i$ is an absorbing state of a Markov chain, and State $i$ is not, and suppose that the initial state of the Markov chain is State $k$. Let $v$ be the State $i$ row of $T$ without the absorbing state columns. Let $w$ be the State $k$ column of $(I - R)^{-1}$. Then the probability that the Markov chain will eventually get absorbed into State $i$ is $v \cdot w$. 
For the tennis example, this gives the dot product of the “win” row 1 of $T$ (only the non-absorbing columns) with the “deuce” column 2 of $(I - R)^{-1}$

\[
\begin{bmatrix}
0.6 \\
0 \\
0
\end{bmatrix} \cdot \begin{bmatrix}
1.15384615384615 \\
1.92307692307692 \\
0.769230769230769
\end{bmatrix} = 0.6923077.
\]
(From Durrett, *Elementary Probability Theory for Applications*, page 143, Example 4.31)

**Example**

Suppose a coin flipping method has \( P(H) = 0.25 \). What is the average number of tosses until we see \( HH \)?
We model this as a Markov chain in which a state consists of the results of the most recent two flips of the coin. (This means the chain doesn’t begin until after the *second* flip of the coin.)

This means that there are four states, which we list in order as: *HH, HT, TH, TT*. 
With these states in this order, the transition matrix is:

\[
T = \begin{bmatrix}
0.25 & 0 & 0.25 & 0 \\
0.75 & 0 & 0.75 & 0 \\
0 & 0.25 & 0 & 0.25 \\
0 & 0.75 & 0 & 0.75
\end{bmatrix}.
\]

However, since want to know when we first arrive at \( HH \), we change the rules so that \( HH \) is an absorbing state.

Then we know how to compute how long it will take to reach that state...
With these modified rules, we have:

\[
T = \begin{bmatrix}
0 & 0 & 0.25 & 0 \\
0 & 0 & 0.75 & 0 \\
0 & 0.25 & 0 & 0.25 \\
1 & 0.75 & 0 & 0.75
\end{bmatrix}.
\]
This means that

$$R = \begin{bmatrix} 0 & 0.75 & 0 \\ 0.25 & 0 & 0.25 \\ 0.75 & 0 & 0.75 \end{bmatrix},$$

so

$$I - R = \begin{bmatrix} 1 & -0.75 & 0 \\ -0.25 & 1 & -0.25 \\ -0.75 & 0 & 0.25 \end{bmatrix},$$

which implies that

$$(I - R)^{-1} = \begin{bmatrix} 4 & 3 & 3 \\ 4 & 4 & 4 \\ 12 & 9 & 13 \end{bmatrix}.$$
To figure out the initial distribution, recall that it takes two flips to get the chain started.

- If the first two flips are **HH** (probability 0.0625), the expected value of further flips is 0.
- If the first two flips are **HT** (probability 0.1875), then we sum Column 1 to get the expected value of further flips to be 20.
- If the first two flips are **TH** (probability 0.1875), then we sum Column 2 to get 16.
- If the first two flips are **TT** (probability 0.5625), then we sum Column 3 to get 20.
The expected number of flips then is

\[2 + (0.1875)(20) + (0.1875)(16) + (0.5625)(20) = 20,\]

or approximately 20 flips.
EXAMPLE (Gambler’s ruin)

A gambler has $3 dollars to begin with. The gambler repeatedly bets $1 that a coin will land heads, and the probability of that happening is 1/2. The gambler loses upon running out of money but wins upon reaching $8. What is the probability that the gambler will win?
We model this as a Markov chain, where a state is a dollar amount that the gambler has.

We order the states from 0 to 8 dollars.

With this, we compute that the transition matrix is:

\[
T = \begin{bmatrix}
1 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 1 \\
\end{bmatrix}
\]
This means that the reduced matrix $R$ is

$$R = \begin{bmatrix}
0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.5 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 \\
\end{bmatrix},$$

so

$$\left(I - R\right)^{-1} = \begin{bmatrix}
7/4 & 3/2 & 5/4 & 1 & 3/4 & 1/2 & 1/4 \\
3/2 & 3 & 5/2 & 2 & 3/2 & 1 & 1/2 \\
5/4 & 5/2 & 15/4 & 3 & 9/4 & 3/2 & 3/4 \\
1 & 2 & 3 & 4 & 3 & 2 & 1 \\
3/4 & 3/2 & 9/4 & 3 & 15/4 & 5/2 & 5/4 \\
1/2 & 1 & 3/2 & 2 & 5/2 & 3 & 3/2 \\
1/4 & 1/2 & 3/4 & 1 & 5/4 & 3/2 & 7/4 \\
\end{bmatrix}.$$
For the probability of winning, we use the absorption probabilities proposition.

That is, we take the dot product of the “win” row 9 of $T$ (only the non-absorbing columns) with the $3$ column (which is Column 3) of $(I - R)^{-1}$

This gives:

$$
\begin{bmatrix}
0 \\
0 \\
0 \\
0 \\
1/2
\end{bmatrix}
\cdot
\begin{bmatrix}
5/4 \\
5/2 \\
15/4 \\
3 \\
3/2 \\
3/4
\end{bmatrix}
= \frac{3}{8}
$$

Why does this formula make sense?
With a different value of starting money, we find that the general formula is $k/n$, where $k$ is the starting amount and $n$ is the winning amount.
**Example**

In the Gambler’s ruin problem, on average how long is spent in the $3$ state? How long is the game played in total on average?
These types of problems are familiar by now.

Summing the $3$ column of $(I - R)^{-1}$, we get that the expected length of time spent in the $3$ state is $15$ plays.

We can do a similar calculation by summing all the entries of each column to yield the expected length of time spent in each state.

Summing all of those expected lengths yields the total expected length of a gambling run to be $84$ plays.
Another way to look at Markov chains is geometrically.

I have written a paper on this subject, so if you find it interesting, just let me know and I can get it to you.

In my view, there remain plenty of interesting research questions in this area that undergraduates could pursue.
The standard \textit{n-simplex} $\Delta_n$ is the set of points in $\mathbb{R}^{n+1}$ whose entries are nonnegative and whose entries sum to 1.

To understand how to look at Markov chains geometrically, note that a distribution for an $s$-state Markov chain is really a point on $\Delta_{s-1} \subset \mathbb{R}^s$. 
Where there’s a matrix, there’s a linear transformation.

A transition matrix has all nonnegative entries and columns that sum to 1.

This represents (relative to the standard basis for $\mathbb{R}^s$) a linear transformation $T : \mathbb{R}^s \rightarrow \mathbb{R}^s$ that is **stochastic**, meaning that $T(\Delta_{s-1}) \subseteq \Delta_{s-1}$.

There is a one-to-one correspondence between finite Markov chains and stochastic linear transformations.
We can ask the question: for which \( x \in \Delta_n \) does 
\[ \lim_{n \to \infty} T^n(x) \] exist?

The condition in the Markov Chain convergence theorem translates to the condition that some power of \( T \) sends \( \Delta_{n-1} \) to its interior.

But there are plenty of interesting things to say geometrically (and about the Markov chain) when the Markov Chain convergence theorem condition doesn’t hold.

(See my paper for further discussion on these matters.)
Where are we so far? We have seen:

1. the basics of discrete probability theory (definitions, counting techniques, conditional probability)
2. some applications, in particular an introduction to Markov chains

We have used coins, dice, cards, and urns as prototypical examples for discrete probability theory.
We have also met the following distribution families:

1. Bernoulli
2. Binomial

And Poisson in passing (it is often used to model counts of infrequent events that have a relatively constant probability of occurring)
Now we’ll study continuous probability theory, working towards the central limit theorem and some of its applications.
Why study continuous probability theory separately? Consider the following example.

**Example**

Suppose we have a board game spinner that is equally probable to land in any direction along the circle. We define a random variable $X$ whose value is the angle the spinner makes with the horizontal (measured in radians) divided by $2\pi$, so $X$ takes on values in $[0, 1)$. What is $P(X = 1/4)$? $P(X = 3/8)$?
If we were to try to apply the equally probable outcomes theorem (which doesn’t apply because the sample space is finite), we’d see that

\[ P(X = x) = \frac{1}{\infty} = 0, \]

no matter what \( x \) is.

Although this theorem doesn’t apply directly, we could approximate the game by finite games with \( n \) possible outcomes, and the limit as \( n \rightarrow \infty \) would still be 0.

With any “reasonable” definition of probability, all of these probabilities should equal 0. Then how are they to sum to 1?
The problem is that we’re asking the wrong question. Instead we should ask questions like: what is \( P(1/2 < X \leq 3/4) \)?

By intuition, we suspect that that answer to this should be 1/4. How can we define continuous probabilities in such a way that this works?

We want a way to encode the distribution of \( X \) so that we can answer all of the types of questions above, not just \( P(X = x) \).
**Definition**

A **probability density function (pdf)** is a nonnegative function \( f : \mathbb{R} \rightarrow \mathbb{R} \) satisfying

\[
\int_{-\infty}^{\infty} f(x) \, dx = 1.
\]

We interpret probability density functions as follows: for any \( a, b \in \mathbb{R} \),

\[
P(a < X \leq b) = \int_{a}^{b} f(x) \, dx.
\]
What is a probability density function for $X$ in the previous example? What if the interval had been $[a, b)$ rather than $[0, 1)$?
**Definition**

A random variable $X$ is said to have a **uniform distribution on** $[a, b]$ (with $b > a$) if the probability density function of $X$ is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases}$$

In this case, we write $X \sim \text{Unif}(a, b)$. 
DEFINITION

A cumulative distribution function (cdf) is a nonnegative, everywhere nondecreasing function \( F : \mathbb{R} \to \mathbb{R} \) such that

\[
\lim_{x \to -\infty} F(x) = 0 \text{ and } \lim_{x \to \infty} F(x) = 1.
\]

For any given random variable:

\[
f(x) = F'(x) \text{ and } F(x) = \int_{-\infty}^{x} f(t) \, dt
\]

We interpret the cumulative distribution function as giving, for any \( x \in \mathbb{R} \):

\[
F(x) = P(X \leq x).
\]
**Example**

Let $X$ be a random variable with $X \sim \text{Unif}(a, b)$. What is the cumulative distribution function of $X$?
By the definition of a uniform distribution, the probability density function of $X$ is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases}$$

By the definition of a cumulative distribution function, the cumulative distribution function of $X$ is

$$F(x) = \int_{-\infty}^{x} f(t) \, dt.$$
This means that

\[ F(x) = \begin{cases} 
0 & \text{if } x < a \\
\frac{x-a}{b-a} & \text{if } a \leq x \leq b \\
1 & \text{if } x > b.
\end{cases} \]
**Example**

For what value of \( c \in \mathbb{R} \) is the following function a probability density function:

\[
f(x) = \begin{cases} 
  ce^{-\lambda x} & \text{for } x \geq 0 \\
  0 & \text{otherwise.}
\end{cases}
\]
Since a probability density function must integrate to 1, we have

\[ 1 = \int_{-\infty}^{\infty} f(x) \, dx \]
\[ = \int_{0}^{\infty} ce^{-\lambda x} \, dx \]
\[ = c \left( -\frac{1}{\lambda} e^{-\lambda x} \right) \bigg|_{0}^{\infty} \]
\[ = \frac{c}{\lambda}, \]

so \( c = \lambda \).
**Example**

What is the cumulative distribution function for a random variable whose probability density function is:

\[ f(x) = \begin{cases} 
\lambda e^{-\lambda x} & \text{for } x \geq 0 \\
0 & \text{otherwise}
\end{cases} \]
By definition, the cumulative distribution function is

\[ \int_{-\infty}^{x} f(t) \, dt = \begin{cases} 
0 & \text{if } x < 0 \\
\int_{0}^{x} \lambda e^{-\lambda t} \, dt & \text{if } x \geq 0 
\end{cases} \]

\[ = \begin{cases} 
0 & \text{if } x < 0 \\
1 - e^{-\lambda x} & \text{if } x \geq 0.
\end{cases} \]
**Definition**

A random variable $X$ is *exponentially distributed with parameter* $\lambda$, written $X \sim \text{Exp}(\lambda)$ if its probability density function is

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

EXAMPLE

For what value of $c \in \mathbb{R}$ is the following function a probability density function:

$$f(x) = \begin{cases} 
  cx^{-\rho} & \text{for } x \geq 1 \\
  0 & \text{otherwise}.
\end{cases}$$
Since a probability density function must integrate to 1, we have

\[
1 = \int_{-\infty}^{\infty} f(x) \, dx
\]

\[
= \int_{1}^{\infty} cx^{-\rho} \, dx
\]

\[
= c\left(\frac{1}{-\rho + 1}\right)x^{-\rho + 1}\bigg|_{1}^{\infty}
\]

\[
= c \frac{1}{\rho - 1},
\]

so \( c = \rho - 1 \).
**Example**

What is the cumulative distribution function for a random variable whose probability density function is:

\[
f(x) = \begin{cases} 
(\rho - 1)x^{-\rho} & \text{for } x \geq 1 \\
0 & \text{otherwise.}
\end{cases}
\]
By definition, the cumulative distribution function is

\[
\int_{-\infty}^{x} f(t) \, dt = \begin{cases} 
0 & \text{if } x < 1 \\
\int_{1}^{x} (\rho - 1)t^{-\rho} \, dt & \text{if } x \geq 1 
\end{cases}
\]

\[
= \begin{cases} 
0 & \text{if } x < 1 \\
1 - x^{-\rho + 1} & \text{if } x \geq 1.
\end{cases}
\]
**DEFINITION**

A random variable $X$ is *power law distributed with parameter $\rho > 1$*, written $X \sim \text{Power}(\rho)$ if its probability density function is

$$f(x) = \begin{cases} 
(\rho - 1)x^{-\rho} & \text{for } x \geq 1 \\
0 & \text{otherwise.}
\end{cases}$$
Example

For what value of $c \in \mathbb{R}$ is the following function a probability density function:

$$f(x) = ce^{-x^2/2}$$
This is based on a slightly trickier integral:

\[
\left( \int_{-\infty}^{\infty} e^{-x^2/2} \, dx \right)^2 = \left( \int_{-\infty}^{\infty} e^{-x^2/2} \, dx \right) \left( \int_{-\infty}^{\infty} e^{-y^2/2} \, dy \right)
\]

\[
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)/2} \, dx \, dy
\]

\[
= \int_{0}^{2\pi} \int_{0}^{\infty} e^{-r^2/2} \, r \, dr \, d\theta
\]

\[
= 2\pi \left( -e^{-r^2/2} \right) \bigg|_{0}^{\infty}
\]

\[
= 2\pi.
\]

This tells us that \( c = 1/\sqrt{2\pi} \).
With one additional change of variables \((z = (x - \mu) / \sigma)\), the same argument shows that

\[
\frac{1}{\sigma \sqrt{2\pi}} e^{-((x-\mu)/\sigma)^2/2}
\]

is a probability density function for any \(\mu \in \mathbb{R}\) and any \(\sigma > 0\).
A random variable $X$ is **normally distributed** with mean $\mu$ and variance $\sigma^2$ (with $\sigma > 0$), written $X \sim N(\mu, \sigma^2)$ if its probability density function is

$$
\frac{1}{\sigma \sqrt{2\pi}} e^{-((x-\mu)/\sigma)^2 / 2}.
$$

If $\mu = 0$ and $\sigma^2 = 1$, then $X$ is said to be **standard normally distributed**.
It is straightforward to verify that if $X \sim N(\mu, \sigma^2)$, then

1. $E(X) = \mu$,
2. $\text{Var}(X) = \sigma^2$,
3. $\frac{X - \mu}{\sigma} \sim N(0, 1)$. 
To sketch a normal distribution, note that:

- Normal distributions are bell-shaped.
- The top of the bell occurs where $x = \mu$.
- There are points of inflection at $x = \mu - \sigma$ and $x = \mu + \sigma$.

The graph of the probability density function of a standard normal distribution looks like this...
Basic definitions
Discrete random variables
Combinatorial techniques
Multiple random variables
Markov chains
Continuous random variables
Sums of random variables
Statistical applications
Further topics
Student presentations

Friday, October 25
To get a sense for distance in normal distributions, note the \textbf{68-95-99.7 rule}: if $X \sim N(\mu, \sigma^2)$, then

\begin{enumerate}
\item $P(\mu - \sigma < X < \mu + \sigma) \approx 0.68,$
\item $P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.95,$
\item $P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 0.997.$
\end{enumerate}
Normal distributions don’t have cumulative distribution functions that can be expressed in terms of elementary functions.

However, for a standard normal distribution, the cumulative distribution function is called the complementary error function:

$$F(x) = \int_{-\infty}^{x} \frac{e^{-t^2/2}}{\sqrt{2\pi}} \, dt.$$ 

Since there is no straightforward formula for this function, it is ordinarily approximated numerically by computer.
**Definition**

Let $X$ be a random variable. The **expected value** $E(X)$ of $X$ is defined by

$$E(X) = \int_{-\infty}^{\infty} xf(x) \, dx.$$ 

More generally, if $r : \mathbb{R} \to \mathbb{R}$ is a function, then the **expected value** $E(r(X))$ of the random variable $r(X)$ is defined by

$$E(r(X)) = \int_{-\infty}^{\infty} r(x)f(x) \, dx.$$ 

Compare this to the definition for a discrete random variable:

- $P(X = x)$ has been replaced with $f(x) \, dx$,
- sums have been replaced by integrals.
**Example**

Let $X \sim \text{Unif}(a, b)$. What is $E(X)$?
The probability density function of $X$ is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases}$$

By definition, the expected value of $X$ is

$$E(X) = \int_{-\infty}^{\infty} xf(x) \, dx = \int_{a}^{b} x \cdot \frac{1}{b-a} \, dx = \frac{a + b}{2}. $$
**Example**

Let $X \sim \text{Unif}(a, b)$. What is $\text{Var}(X)$?
It’s perhaps easier to use the alternative formulation of variance:

\[ \text{Var}(X) = E(X^2) - E(X)^2. \]

We have already computed \( E(X) \). We also compute:

\[
E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) \, dx \\
= \int_a^b x^2 \frac{1}{b - a} \, dx \\
= \frac{b^3 - a^3}{3(b - a)} \\
= \frac{a^2 + ab + b^2}{3}
\]
This gives us that

\[
\text{Var}(X) = E(X^2) - E(X)^2
\]

\[
= \frac{a^2 + ab + b^2}{3} - \left(\frac{a + b}{2}\right)^2
\]

\[
= \frac{4a^2 + 4ab + 4b^2 - 3a^2 - 6ab - 3b^2}{12}
\]

\[
= \frac{(b - a)^2}{12}.
\]
EXAMPLE

Let $X \sim \text{Exp}(\lambda)$. What is $E(X)$?
The probability density function of $X$ is

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

By definition, the expected value of $X$ is

$$E(X) = \int_{-\infty}^{\infty} x f(x) \, dx$$

$$= \int_{0}^{\infty} x \lambda e^{-\lambda x} \, dx$$

$$= \frac{1}{\lambda}.$$
**EXAMPLE**

Let $X \sim \text{Exp}(\lambda)$. What is $\text{Var}(X)$?
It’s perhaps easier to use the alternative formulation of variance:

$$\text{Var}(X) = E(X^2) - E(X)^2.$$ 

We have already computed $E(X)$. We also compute:

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) \, dx$$

$$= \int_{0}^{\infty} x^2 \lambda e^{-\lambda x} \, dx$$

$$= \frac{2}{\lambda^2}.$$
This gives us that

\[ \text{Var}(X) = E(X^2) - E(X)^2 \]

\[ = \frac{2}{\lambda^2} - \left( \frac{1}{\lambda} \right)^2 \]

\[ = \frac{1}{\lambda^2}. \]
EXAMPLE

Let \( X \sim \text{Power}(\rho) \). What is \( E(X) \)? What is \( \text{Var}(X) \)?
Since the computations proceed exactly as in the previous examples, we won’t write them here.

The results are that

\[ E(X) = \frac{\rho - 1}{\rho - 2} \]

\[ E(X^2) = \frac{\rho - 1}{\rho - 3} \]

\[ \text{Var}(X) = \frac{\rho - 1}{(\rho - 2)^2(\rho - 3)}. \]
**Example**

Let $X \sim \mathcal{N}(\mu, \sigma^2)$. What is $E(X)$? What is $\text{Var}(X)$?
Since the computations proceed exactly as in the previous examples, we won’t write them here.

The results are that

\[ E(X) = \mu \]
\[ \text{Var}(X) = \sigma^2. \]

Probably the easiest way to get at these are to compute them for standard normal distributions, and then note that if \( X \sim N(\mu, \sigma^2) \), then

\[ \frac{X - \mu}{\sigma} \sim N(0,1). \]
**Definition**

Let $X$ be a random variable, and let $p \in \mathbb{R}$ with $0 < p < 1$. The $p$-quantile is the number $x_p$ satisfying

$$P(X \leq x_p) = p.$$ 

The **quantile function** of $X$ is the function $q : (0, 1) \rightarrow \mathbb{R}$ sending $p$ to $x_p$.

If the cumulative distribution function of $X$ is constant on an interval, then a $p$-quantile in that range won’t be well-defined without further tweaking of the definition.

However, we won’t resolve that here. We simply mention the problem.
Notice that if the cumulative distribution function $F$ of $X$ is strictly increasing wherever it is nonzero, then it is invertible (except possibly at 0), and the quantile function is the inverse of the cumulative distribution function:

$$F(q(p)) = p \text{ and } q(F(x)) = x.$$
DEFINITIONS

Let $X$ be a random variable. Then

1. The **median** of $X$ is the $0.5$-quantile of $X$.
2. The **first quartile** of $X$ is the $0.25$-quantile of $X$.
3. The **third quartile** of $X$ is the $0.75$-quantile of $X$. 
**Example**

Let $X \sim \text{Exp}(\lambda)$. Then the cumulative distribution function of $X$ is

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

which is invertible in the region where it is nonzero. What is the quantile function of $X$?
For quantile function, we solve $F(x) = p$ for $x$ to invert the cumulative distribution function:

$$1 - e^{-\lambda x} = p, \quad \text{so} \quad x = \frac{1}{\lambda} \log \frac{1}{1 - p}.$$
**Example**

Let $X \sim \text{Exp}(\lambda)$. What is the distribution of $X^2$?
This is easier to prove if we focus on the cumulative distribution functions $F_1$ for $X$ and $F_2$ for $X^2$.

It is immediate that $F_2(x) = 0$ if $x < 0$.

Also, since $X$ is always nonnegative, then for any nonnegative real number $c$,

$$X^2 \leq c$$

if and only if

$$X \leq \sqrt{c}.$$
This means that for $x > 0$,

$$F_2(x) = P(X^2 \leq x) = P(X \leq \sqrt{x}) = F_1(\sqrt{x}).$$

Differentiating both sides and denoting the probability functions by $f_1$ for $X$ and $f_2$ for $X^2$, we get that for any $x \geq 0$

$$f_2(x) = \frac{1}{2\sqrt{x}}f_1(x).$$

And as we have already shown, $f_2(x) = 0$ for all $x < 0$. 
**THEOREM (TRANSFORMATION THEOREM)**

Let $X$ be a random variable with probability density function $f_X : \mathbb{R} \to \mathbb{R}$, and let $g : [a, b] \to [c, d]$ be a monotone strictly increasing function with inverse $h : [c, d] \to [a, b]$, and let $f_{g(X)}$ denote the probability density function of $g(X)$. Then

$$f_{g(X)}(y) = f_X(h(y))h'(y) \text{ for all } y \in [c, d].$$
The proof can mimics the $X^2$ example.

We focus on the cumulative distribution functions, and those change in a simple way involving $h$, the inverse of $g$.

Differentiating and applying the chain rule proves the result.

The strictly increasing assumption ensures invertibility of $g$ in the relevant region.

If $g$ is instead strictly decreasing, the result holds except that the right side of the equation needs a minus sign (because the inequality switches directions during the inverse transformation).
Continuous distributions of more than one variable

Rigorous coverage of the upcoming concepts requires real analysis, which isn’t a prerequisite for this class.

For this class, the way to understand these concepts is in terms of their parallels with discrete probability theory.
**DEFINITION**

Let $X, Y : S \rightarrow \mathbb{R}$ be random variables. The **joint cumulative distribution function** of $X$ and $Y$ is the function $F_{X,Y} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ defined by

$$F_{X,Y}(x,y) = P(X \leq x \text{ and } Y \leq y)$$

for all $x,y \in \mathbb{R}$.

**DEFINITION**

If $F_{X,Y}(x,y)$ exists and is measurable, then $X$ and $Y$ are called **jointly continuous**.

However, this is a rather technical condition and we will not dwell on it this class. We will generally just assume it to be satisfied.
**Definition**

Let $X, Y : S \rightarrow \mathbb{R}$ be random variables. The joint probability density function of $X$ and $Y$ is the function $f_{X,Y} : \mathbb{R}^2 \times \mathbb{R} \rightarrow \mathbb{R}$ defined by

$$f_{X,Y}(x,y) = \frac{\partial F_{X,Y}}{\partial x \partial y}(x,y)$$

for all $x, y \in \mathbb{R}$.

Although we won’t prove it in this class, the joint probability density function has the property that for all (reasonably nice) sets $A \in \mathbb{R}^2$,

$$P((x,y) \in A) = \int \int_A f_{X,Y}(x,y) \, dx \, dy.$$
The equation

\[ P((x, y) \in A) = \int \int_A f_{X,Y}(x, y) \, dx \, dy \]

is a good way to think about joint probability density functions.

Joint probability density functions are always nonnegative, and they integrate (over the plane) to 1.

Intuitively speaking, where the joint probability function is large, there is a high probability that \( X \) and \( Y \) will (jointly) take on values nearby.
DEFINITION

Let $X, Y : S \rightarrow \mathbb{R}$ be jointly continuous random variables. The **marginal distribution** of $X$ is the distribution of $X$ ignoring $Y$. If $f_{X,Y} : \mathbb{R}^2 \rightarrow \mathbb{R}$ is the joint probability density function of $X$ and $Y$, then the marginal probability density function $f_X : \mathbb{R} \rightarrow \mathbb{R}$ of $X$ is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dy.$$ 

The marginal distribution of $Y$ is defined similarly.
DEFINITION

Let $X, Y : S \to \mathbb{R}$ be jointly continuous random variables. The conditional distribution of $Y$ given $X = x$ is defined to be the distribution of $Y$ when the value of $X$ is $x$. The conditional distribution of $Y$ given $X = x$ is defined by

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

where $f_X(x)$ is the marginal probability density function of $X$ and $f_{X,Y}(x,y)$ is the joint probability density function of $X$ and $Y$. If $f_X, f_Y : \mathbb{R} \to \mathbb{R}$ are the marginal probability density functions of $X$ and $Y$, and if $f_{X,Y} : \mathbb{R} \to \mathbb{R}$ be their joint probability density function, then the conditional probability density function $f_{Y|X=x} : \mathbb{R} \to \mathbb{R}$ of $Y$ given $X = x$ is given by

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$
The conditional probability density function is still a probability density function, so it can still be interpreted in terms of probabilities.

More specifically,

$$P(a \leq Y \leq b \mid X = x) = \int_a^b f_{Y \mid X=x}(y) \, dy.$$ 

Also, the conditional cumulative distribution function $F_{Y \mid X=x}$ is defined as usual:

$$F_{Y \mid X=x}(b) = P(-\infty \leq Y \leq b \mid X = x) = \int_{-\infty}^b f_{Y \mid X=x}(y) \, dy.$$
**Definition**

Let $X_1, X_2, \ldots, X_n : S \rightarrow \mathbb{R}$ be jointly continuous random variables, and let $f_1, f_2, \ldots, f_n : \mathbb{R} \rightarrow \mathbb{R}$ be the individual (i.e., marginal) probability density functions of these random variables, and let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be their joint probability density function. $X_1, X_2, \ldots, X_n$ are called **independent** if

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1)f_2(x_2) \cdots f_n(x_n)$$

for all $x_1, x_2 \ldots, x_n \in \mathbb{R}$. 
**Proposition**

Let $X, Y$ be jointly continuous random variables. Then $X$ and $Y$ are independent if and only if

$$f_{Y|X=x}(y) = f_Y(y)$$

for all $x, y \in \mathbb{R}$. A similar statement could be made switching the roles of $X$ and $Y$.

We give this proposition here without proof (again simply referring to the parallels with discrete probability theory).
This example is Chapter 5, Exercise 23 from Durrett’s *Elementary Probability for Applications*

**Example**

Suppose $X \sim \text{Exp}(\lambda)$ and $Y = \log(X)$. Find the distribution function of $X$. 
Since \( g(x) = \log(x) \) is strictly increasing on the range of \( X \) (which is all positive numbers), it is invertible on the range of \( X \).

Its inverse is on this range is

\[
h(x) = e^x.
\]

Since \( X \sim \text{Exp}(\lambda) \), then the probability density function of \( X \) is

\[
f(x) = \lambda e^{-\lambda x}
\]

for \( x \geq 0 \), and 0 otherwise.
Applying the transformation theorem to this $X$ and $g$, we find that the probability density function of $Y = g(X)$ is

$$f(h(x))h'(x) = \lambda e^{-\lambda e^x} e^x.$$ 

for $x \geq 0$, and 0 otherwise.

This distribution is called the *double exponential distribution*. 
**Example**

Suppose you are waiting for an event to happen, and that your waiting time to the event $X$ has $X \sim \text{Exp}(\lambda)$. If you have already waited a time $t$, what is the probability that you will need to wait at least a time $s$ longer for the event?
Since $X \sim \text{Exp}(\lambda)$, then the cumulative distribution function of $X$ is

$$F(x) = 1 - e^{-\lambda x}$$

for $x \geq 0$, and 0 otherwise.

By the definition of conditional probability,

$$P(X > t + s | X > t) = \frac{P(X > t + s)}{P(X > t)} = \frac{1 - F(t + s)}{1 - F(t)} = e^{-\lambda(t+s) + \lambda t} = e^{-\lambda s} = P(X > s).$$
In other words, the probability that you’ll have to wait at least another $s$ minutes is the same as it was when you started!

This is called the *lack of memory* property of the exponential distribution.
Begin test.
**Definition**

A collection $X_1, X_2, X_3, \ldots, X_n$ of random variables is called **independent and identically distributed** (or IID) if the collection of random variables is independent and all the random variables have the same distribution.

This captures the idea of a *sample* in statistics.

In order to be able to work with such samples mathematically, we consider each individual observation as coming from a random variable with its own distribution.
**Definition**

Let $X_1, X_2, X_3, \ldots, X_n$ be a collection of random variables. The **sample mean** $\overline{X}_n$ of this collection of random variables is defined to be

$$\overline{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}.$$

We may omit the subscript $n$ at times, where it isn’t needed for clarity.

Note carefully that the sample mean is a *random variable*. 
Sneak preview of where we’re headed: the *central limit theorem* asserts, in rough terms, that if \( X_1, X_2, \ldots \) are IID with mean \( \mu \) and variance \( \sigma^2 \), then:

\[
\lim_{n \to \infty} \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}}
\]

has a standard normal distribution.
Means and variances of sums of random variables

We have already shown that for random variables $X, Y$: 

$$E(X + Y) = E(X) + E(Y).$$

What about variances?
Recall the following definition.

**Definition**

Let $X, Y$ be random variables. The **covariance** of $X$ and $Y$ is defined by

$$\text{Cov}(X, Y) = E((X - E(X))(Y - E(Y))).$$

Note that by its definition,

$$\text{Var}(X) = \text{Cov}(X, X).$$
PROPOSITION

Let $X, Y, Z$ be random variables, and let $c \in \mathbb{R}$. Then

1. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.
2. $\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$.
3. $\text{Cov}(X + Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z)$.
4. $\text{Cov}(X, cY) = c\text{Cov}(X, Y)$.
5. $\text{Cov}(cX, Y) = c\text{Cov}(X, Y)$. 
For the first, we have by the definition of covariance:

\[
\text{Cov}(Y, X) = E((Y - E(Y))(X - E(X))) \\
= E((X - E(X))(Y - E(Y))) \\
= \text{Cov}(X, Y).
\]
For the second, by the definition of covariance and the linearity of expected value:

\[
\text{Cov}(X, Y + Z) = E((X - E(X))(Y + Z - E(Y + Z)))
\]
\[
= E((X - E(X))(Y - E(Y) + Z - E(Z)))
\]
\[
= E((X - E(X))(Y - E(Y)) + (X - E(X))(Z - E(Z)))
\]
\[
= E((X - E(X))(Y - E(Y))) + E((X - E(X))(Z - E(Z)))
\]
\[
= \text{Cov}(X, Y) + \text{Cov}(Y, Z).
\]
The proof of the fourth is similar, and the others follow by combining the later proven results with \( \text{Cov}(Y, X) = \text{Cov}(X, Y) \).
**Proposition**

Let $X, Y$ be random variables. Then

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$
We compute that

\[ \text{Cov}(X, Y) = E((X - E(X))(Y - E(Y))) \]
\[ = E(XY) - E(X)E(Y) - E(X)Y + E(X)E(Y) \]
\[ = E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \]
\[ = E(XY) - E(X)E(Y). \]
This proposition tells us that the covariance is how far expected value is from being multiplicative.

But we also have:

**PROPOSITION**

Let $X, Y$ be random variables. Then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y).$$
We have

\[ \text{Var}(X + Y) = \text{Cov}(X + Y, X + Y) \]
\[ = \text{Cov}(X + Y, X) + \text{Cov}(X + Y, Y) \]
\[ = \text{Cov}(X, X) + \text{Cov}(Y, X) + \text{Cov}(X, Y) + \text{Cov}(Y, Y) \]
\[ = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y). \]
This proposition tells us that covariance is (half of) how far variance is from being additive (up to a factor of 2).

We have already proved the following proposition:

**Proposition**

Let $X, Y$ be random variables. If $X, Y$ are independent, the $\text{Cov}(X, Y) = 0$.

Note that this does not hold in the other direction: if $\text{Cov}(X, Y) = 0$, it does not necessarily follow that $X, Y$ are independent.

Sometimes people use the term *statistically independent* to refer to variables having covariance zero.
**THEOREM**

*Chebyshev’s inequality* Let $Y$ be a random variable whose mean and variance exist. Then

$$P(|Y - E(Y)| > \varepsilon) \leq \frac{\text{Var}(Y)}{\varepsilon^2}.$$
To prove this, denote $E(Y)$ by $\mu$. Then

$$\text{Var}(Y) = E((Y - \mu)^2)$$

$$= \int_{-\infty}^{\infty} (y - \mu)^2 f(y) \, dy$$

$$\geq \int_{-\infty}^{\mu-\varepsilon} (y - \mu)^2 f(y) \, dy + \int_{\mu+\varepsilon}^{\infty} (y - \mu)^2 f(y) \, dy$$

$$\geq \int_{-\infty}^{\mu-\varepsilon} \varepsilon^2 f(y) \, dy + \int_{\mu+\varepsilon}^{\infty} \varepsilon^2 f(y) \, dy$$

$$= \varepsilon^2 P(|Y - \mu| \geq \varepsilon).$$
Chebyshev’s inequality gives us a new way to look at the variance: the variance gives a bound on the probability that the random variable will be far from its mean.
**EXAMPLE**

Let $X_1, X_2, \ldots, X_n$ be IID random variables with mean $\mu$ and variance $\sigma^2$. What is $E(\bar{X}_n)$? What is $\text{Var}(\bar{X}_n)$?
By linearity of expected value,

$$E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \mu.$$ 

Also, since these are independent random variables, their variance add, so

$$\text{Var}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \sum_{i=1}^{n} \frac{1}{n^2} \text{Var}(X_i) = \frac{\sigma^2}{n}.$$ 

This proves the following proposition...
**PROPOSITION**

Let $X_1, X_2, \ldots, X_n$ be IID random variables with mean $\mu$ and variance $\sigma^2$. Then

$$E(\overline{X}_n) = \mu \text{ and } \text{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$ 

This is one step toward the central limit theorem: at least we now know the mean and variance of $\overline{X}_n$. 
**Theorem (Weak Law of Large Numbers)**

Let $X_1, X_2, \ldots$, be IID random variables with mean $\mu$ and variance $\sigma^2$. Then for any $\epsilon > 0$,

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \epsilon) = 0.$$
To prove this, recall that $E(\bar{X}_n) = \mu$ and $\text{Var}(\bar{X}_n) = \sigma^2/n$.

Applying Chebyshev’s inequality to $Y = \bar{X}_n$ gives

$$0 \leq P(|\bar{X}_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$ 

Taking limits and applying the sandwich theorem proves the theorem.
The strong law of large numbers asserts that the convergence is almost everywhere (meaning \textit{except at a set of measure 0}).

Here's a demonstration of the law of large numbers:

https://www.stat.berkeley.edu/~stark/Java/Html/lln.htm
The following example is from Durrett’s *Elementary Probability for Applications* (Example 6.1 on page 190).

**Example**

Suppose we roll 2 4-sided dice independently, and that each has the following distribution: $P(X = i) = i/10$. Let $X$ be the roll of the first and $Y$ the second. What is the distribution of $X + Y$?
Since $X, Y$ are discrete, then because of independence,

$$P(X + Y = k) = \sum_{j=0}^{k} P(X = j, Y = k - j)$$

$$= \sum_{j=0}^{k} P(X = j)P(Y = k - j)$$

For continuous random variables, an analogous result holds. We state it without proof:

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x) \, dx.$$
**Definition**

Let $f_X, f_Y : \mathbb{R} \rightarrow \mathbb{R}$. The convolution $f_X * f_Y : \mathbb{R} \rightarrow \mathbb{R}$ of $f_X$ with $f_Y$ is defined by

$$f_X * f_Y(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z - x) \, dx.$$ 

We won’t pursue this concept any further, other than to point out that it is a useful one that has been studied extensively.
The following example is from Durrett’s *Elementary Probability for Applications* (Example 6.5 on page 193).

**EXAMPLE**

Let $X, Y \sim \text{Unif}(0, 1)$ be independent. What is the distribution of $X + Y$?
If $0 \leq z \leq 1$, then

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x)f_Z(z-x) \, dx = \int_{0}^{z} 1 \, dx = z.$$ 

Similarly, if $1 \leq z \leq 2$, then

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x)f_Z(z-x) \, dx = \int_{z-1}^{1} 1 \, dx = 2 - z.$$ 

It is 0 otherwise.
Let $X_1, X_2$ be independent with $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$. Then $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. 
To prove this, let \( a = \sigma_1^2 \) and \( b = \sigma_2^2 \). Since expected value is linear, then \( E(X_1 + X_2) = \mu_1 + \mu_2 \), and since these are independent, then \( \text{Var}(X_1 + X_2) = a + b \). So only normality needs to be shown.

Let \( Y_1 = X_1 - \mu_1 \) and \( Y_2 = X_2 - \mu_2 \). To prove the proposition, we show that \( Y_1 + Y_2 \sim N(0, a + b) \).
We know that

\[
f_{Y_1+Y_2}(x) = \int_{-\infty}^{\infty} f_1(z) f_2(x-z) \, dz
= \frac{1}{\sqrt{2\pi a} \sqrt{2\pi b}} \int_{-\infty}^{\infty} \exp\left(-\frac{z^2}{2a} - \frac{z^2}{2b} + \frac{2xz}{2b} - \frac{x^2}{2b}\right)
\]

The quantity in the exponent is

\[
\frac{a+b}{2ab} \left(-z^2 + \frac{2a}{a+b} xz - \frac{a}{a+b} x^2\right) = \frac{a+b}{2ab} \left(-(z - \frac{ax}{a+b})^2 - \frac{ab}{(a+b)^2} x^2\right).
\]
Putting this back in, the convolution above equals

\[
\frac{1}{\sqrt{2\pi a}\sqrt{2\pi b}} e^{-x^2/(a+b)} \int_{-\infty}^{\infty} \exp \left( -\frac{a + b}{2ab} (z - \frac{ax}{a + b})^2 \right) \, dz.
\]

The integral is, by the usual method, \( \sqrt{\frac{2\pi ab}{a + b}} \), which means the convolution is

\[
\frac{1}{2\pi(a + b)} e^{-x^2/(a+b)}.
\]
Theorem (Miniature Central Limit Theorem)

If $X_1, X_2, \ldots$ are IID normally distributed random variables with mean $\mu$ and variance $\sigma^2$, then $\overline{X}_n \sim N(\mu, \sigma^2 / n)$.

This is immediate from the previous proposition, and the properties of expected value and variance.

An equivalent theorem asserts that the sum of these random variables is distributed as $N(n\mu, n\sigma^2)$.

This is only the miniature version, because the really remarkable part of the central limit theorem is the wide class of random variables that it applies to, not just normally distributed ones.

But this version does give the distribution of $\overline{X}_n$ exactly, not just in the limit as $n \to \infty$. 
(In the first part of class, we went over project topics.)

We now state the central limit theorem and explore some of its uses, deferring its proof until later...
Theorem (Central Limit Theorem for Sums)

Let $X_1, X_2, \ldots$ be (reasonably nice) IID random variables with mean $\mu$ and standard deviation $\sigma$, and let $S_n$ be the partial sum up to $n$. Also, let $\Phi(z)$ be the cdf of a standard normal distribution:

\[
\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx
\]

Then

\[
\lim_{n \to \infty} P\left( a \leq \frac{S_n - \mu}{\sigma \sqrt{n}} \leq b \right) = \Phi(b) - \Phi(a)
\]

for all $a, b \in \mathbb{R}$.

The miniature version of this assumes that the original random variables are independent and identical normally distributed random variables, and in this case it is exact!
Theorem (Central Limit Theorem for Means)

Let $X_1, X_2, \ldots$ be (reasonably nice) IID random variables with mean $\mu$ and standard deviation $\sigma$, and let $\bar{X}_n$ be the sample mean of the first $n$. Also, let $\Phi(z)$ be the cdf of a standard normal distribution:

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Then

$$\lim_{n \to \infty} P\left( a \leq \frac{\bar{X}_n - n\mu}{\sigma / \sqrt{n}} \leq b \right) = \Phi(b) - \Phi(a)$$

for all $a, b \in \mathbb{R}$.
In practice we use the central limit theorem to approximate the distribution of $S_n$ as $N(n \mu, n \sigma^2)$, or the distribution of $\bar{X}_n$ as $N(\mu, \sigma^2/n)$.

Later this semester, we will prove the central limit theorem for functions that have all their moments.
Let’s look at another type of application of the expected values of sums.

This shows how we can sometimes use the symmetry of a situation to get at expected values as sums...
The following example is from Durrett’s *Elementary Probability for Applications* (Example 6.8 on page 197).

**Example**

Place \( n \) balls randomly into \( m \) boxes. What is the expected number of empty boxes?
Let $X_i = 1$ if the $i$-th box is empty, and 0 otherwise.

Let $M$ be the number of empty boxes.

Then

$$M = \sum_{i=1}^{m} X_i.$$
By the linearity of expected value,

\[ E(M) = \sum_{i=1}^{m} E(X_i). \]

By symmetry, all of the \( X_i \) have the same expected value, so

\[ \sum_{i=1}^{m} E(X_i) = mE(X_1). \]

The probability that \( X_1 = 1 \) is \((1 - 1/m)^n\), since this is the probability that all the balls would go into the last \( m - 1 \) boxes.

This means that

\[ E(M) = m(1 - 1/m)^n. \]
In practice, one way that the central limit theorem is used is in *hypothesis testing*.

To conduct a hypothesis test, we need:

1. A **null hypothesis** $H_0$, which is often a statement about a parameter related to the distribution of a random variable.
2. An **alternative hypothesis** $H_a$, which is usually the negation of the null hypothesis.
3. A **test statistic**, which is a random variable whose distribution under the null hypothesis (meaning “if the null hypothesis is true”) is known.
4. A **significance level** $\alpha$, typically taken to be 0.05, which is the threshold for statistical significance.
The result of a hypothesis test is a $p$-value.

**Definition**

A $p$-value is the probability that under the null hypothesis the test statistic would be at least as extreme as its observed value (derived from the data collected).

This means that in order to be able to conduct a hypothesis test, the distribution of the test statistic under the null hypothesis must be known.

One thing that is done in a mathematical statistics course is to prove that certain commonly used test statistics have particular distributions under the null hypothesis.
**Definition**

A $p$-value is called **statistically significant** if it is less than or equal to the chosen significance level.

The phrase I would like you to use when reporting the statistical significance of a $p$ value is:

- If $p \leq \alpha$: We found statistically significant evidence against the null hypothesis.
- If $p > \alpha$: We did not find statistically significant evidence against the null hypothesis.

In practice, people use this term more loosely, but I’d like you to understand the concept that it encapsulates.
EXAMPLE
Suppose we are wondering whether a certain method of flipping a certain coin is fair. Suppose we flip the coin 100 times and get 58 H. Does this coin appear unfair? (Use the usual significance level of $\alpha = 0.05$.)
We model the coin flip as a Bernoulli trial, with success being a $H$.

Let $p$ denote the probability of a $H$ in a flip.

Let $X$ be the random variable whose value is the count of $H$ in 100 flips.

Since $X$ is the success count of 100 independent Bernoulli trials with success probability $p$, then $X \sim B(100, p)$. 
This means that the expected value of $X$ is $100p$, and the variance of $X$ is

$$\sigma^2 = 100p(1 - p).$$

By the central limit theorem, the $B(100, p)$ distribution can be approximated by an $N(100p, \sigma^2)$ distribution.
We can now test

\[ H_0 : p = \frac{1}{2} \text{ and } H_a : p \neq \frac{1}{2}. \]

Under \( H_0 \),

\[ \sigma = \sqrt{100(1/2)(1 - 1/2)} = 5. \]

How extreme is this?
We compute that under $H_0$ the standardized version of the observed value of $X$ is

$$\frac{58 - (100)(1/2)}{5} = 1.6.$$ 

Using a computer, we compute the $p$-value by computing the area under a standard normal probability density function for which $|z| > 1.6$:

$$p\text{-value} = 0.1095985834.$$ 

Comparing this to the usual significance level of 0.05, we did not find statistically significant evidence that this method of flipping the coin is unfair.
**Example**

In the game of *Pass the Pigs*, the pigs that you roll can land in 7 different ways, one of which is called a *down-sider*. Suppose we roll a pig 200 times and get 52 *down-siders*. Does this pig appear unfair?
We model the pig roll as a Bernoulli trial, with success being a *down-sider*.

Let $p$ denote the probability of a *down-sider* in a roll.

Let $X$ be the random variable whose value is the count of *down-siders* in 200 rolls.

Since $X$ is the success count of 200 independent Bernoulli trials with success probability $p$, then $X \sim B(200, p)$.
This means that the expected value of $X$ is $200p$, and the variance of $X$ is

$$\sigma^2 = 200p(1 - p).$$

By the central limit theorem, the $B(200, p)$ distribution can be approximated by an $N(200p, \sigma^2)$ distribution.
We can now test

\[ H_0 : p = \frac{1}{7} \text{ and } H_a : p \neq \frac{1}{7}. \]

Under \( H_0 \),

\[ \sigma = \sqrt{200 \left( \frac{1}{7} \right) \left( 1 - \frac{1}{7} \right)} = 4.9487166. \]

How extreme is this?
We compute that under $H_0$ the standardized version of the observed value of $X$ is
\[
\frac{52 - (200)(1/7)}{4.9487166} = 4.7342722.
\]

Using a computer, we compute the $p$-value by computing the area under a standard normal probability density function for which $|z| > 4.7342722$:
\[
p\text{-value} = 0.0000021984.
\]

Comparing this to the usual significance level of 0.05, we did find statistically significant evidence against the null hypothesis.
In general when hypothesis testing with a test statistic $T$, if the null hypothesis is true, then by the definition of the significance level $\alpha$, the probability of finding statistically significant evidence against the null hypothesis is $\alpha$.

In other words, $\alpha$ is the probability that if the null hypothesis is true, you will reject it.

Because the null hypothesis is often of the form “there is no effect” or “there is no difference”, etc., this is called a false positive or Type I error.

A false negative or Type II error means not rejecting the null hypothesis when it is in fact false.
Setting $\alpha$ allows us to control the Type I error rate, when the null hypothesis is true.

Why not just set $\alpha$ really low?
Because there is a trade-off between the Type I error rate and the Type II error rate.

By making the Type I error rate lower, you will increase the Type II error rate, and vice versa.

The significance level of 0.05 is by far the standard in academic publications, but there are many situations where that choice would be inappropriate (for example, in a court case where the null hypothesis is that the defendant is not guilty, etc.).
Moment Generating Functions

**Definition**

Let $X$ be a random variable, and let $k \geq 1$. The $k$-th moment of $X$ is defined to be $E(X^k)$.

Next class, we’ll prove the Central Limit Theorem for distributions whose moments all exist.
**Definition**

Let $X$ be a random variable. The moment generating function $M_X(t)$ is defined as $E(e^{tX})$.

This will certainly exist for $t = 0$, but it doesn’t necessarily converge for other values of $t$. 
**EXAMPLE**

Let \( X \sim B(n, p) \). What is the moment generating function for \( X \)?
The probability mass function for $X$ is

$$f(x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

if $x = 0, 1, \ldots, n$, and 0 otherwise.

By the definition of a moment generating function,

$$M_X(t) = E(e^{tX})$$

$$= \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1 - p)^{n-k}$$

$$= (pe^t + q)^n.$$

This converges for all values of $t$. 
**EXAMPLE**

Let \( X \sim \text{Exp}(\lambda) \). What is the moment generating function of \( X \)?
The probability density function of $X$ is

$$f(x) = \lambda e^{-\lambda x}$$

for $x \geq 0$, and 0 otherwise.

By the definition of a moment generating function,

$$M_X(t) = E(e^{tX})$$

$$= \int_0^\infty e^{tx} (\lambda e^{-\lambda x}) \, dx$$

$$= \lambda \int_0^\infty e^{tx - \lambda x} \, dx.$$

This integral is finite if and only if $t < \lambda$.

If $t < \lambda$, we can compute further...
\[ M_X(t) = \lambda \int_0^\infty e^{tx - \lambda x} \, dx \]
\[ = \lambda \left. \frac{1}{t - \lambda} e^{tx - \lambda x} \right|_0^\infty \]
\[ = \frac{\lambda}{\lambda - t}. \]
Why is it called the *moment generating function*?

Because:

\[
M_X^{(k)}(0) = \frac{d^k}{dt^k} E(e^{tX}) \bigg|_{t=0}
\]

\[
= E\left(\frac{d^k e^{tX}}{dt^k}\right) \bigg|_{t=0}
\]

\[
= E(X^k e^{tX}) \bigg|_{t=0}
\]

\[
= E(X^k).
\]
Recall the Central Limit Theorem:

**Theorem (Central Limit Theorem for Means)**

Let $X_1, X_2, \ldots$ be (reasonably nice) IID random variables with mean $\mu$ and standard deviation $\sigma$, and let $\overline{X}_n$ be the sample mean of the first $n$. Also, let $\Phi(z)$ be the cdf of a standard normal distribution:

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Then

$$\lim_{n \to \infty} P(a \leq \frac{\overline{X}_n - n\mu}{\sigma / \sqrt{n}} \leq b) = \Phi(b) - \Phi(a)$$

for all $a, b \in \mathbb{R}$.
Before we prove this, we state and prove some properties of moment generating functions that we’ll use:

1. If $Y = aX + b$, then
   
   $$M_Y(t) = e^{bt} M_X(at).$$

2. If $X_1, \ldots, X_n$ are ind’t and $Y = X_1 + \cdots + X_n$, then
   
   $$M_Y(t) = M_{X_1}(t) \cdots M_{X_n}(t).$$
To prove these, we compute directly from the definition of a moment generating function:

\[ M_{aX}(t) = E(e^{aXt}) = E(e^{X(at)}) = M_X(at). \]
Also:

\[ M_{X+b}(t) = E(e^{(X+b)t}) \]
\[ = e^{bt}E(e^{Xt}) \]
\[ = e^{bt}M_X(t). \]

And for independent \(X_1, \ldots, X_n:\)

\[ M_{X_1+\ldots+X_n}(t) = E(e^{(X_1+\ldots+X_n)t}) \]
\[ = E(e^{X_1t} \cdots e^{X_nt}) \]
\[ = E(e^{X_1t}) \cdots E(e^{X_nt}) \]
\[ = M_{X_1}(t) \cdots M_{X_n}(t), \]

using independence in the next to last step.
What is the moment generating function of a standard normally distributed random variable $X$?
We compute it directly from the definition of a moment generating function:

\[
E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \, dx
\]

\[
= \int_{-\infty}^{\infty} \frac{e^{-(x^2-2tx+t^2)/2}e^{t^2/2}}{\sqrt{2\pi}} \, dx
\]

\[
= e^{t^2/2} \int_{-\infty}^{\infty} \frac{e^{-(x^2-2tx+t^2)/2}}{\sqrt{2\pi}} \, dx
\]

\[
= e^{t^2/2}.
\]

The last step follows because the integrand is the probability density function of a \( \mathcal{N}(t, 1) \) random variable.
Proof of the Central Limit Theorem

We now prove the Central Limit Theorem for probability density functions that have all their moments.

The Central Limit Theorem holds more generally, but more general versions are harder to prove.

To begin, let $\mathcal{F}$ denote the space of probability density functions for which all moments exist.
Two facts that we won’t prove, but which we’ll use are that in $\mathcal{F}$:

- There is a one-to-one correspondence between moment generating functions and cumulative distribution functions.
- A sequence of moment generating functions converges to a limiting moment generating function if and only if its associated cumulative distribution functions converge to the limiting moment generating function’s associated cumulative distribution function.
To prove the Central Limit Theorem, we will show that the moment generating function of the standardized sample mean converges to the moment generating function of a standard normal distribution.

Together with the above facts, this will imply that the cumulative distribution functions, and therefore the probability integrals, will converge as stated in the Central Limit Theorem.
With these, let

\[ Z_n = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}. \]

\begin{align*}
M_{Z_n}(t) &= e^{-\sqrt{n} \mu t / \sigma} M_{\bar{X}}\left(\frac{nt}{\sigma}\right) \\
&= e^{-\sqrt{n} \mu t / \sigma} M_{n\bar{X}}\left(\frac{t}{\sigma \sqrt{n}}\right) \\
&= e^{-\sqrt{n} \mu t / \sigma} \left( M_X\left(\frac{t}{\sigma \sqrt{n}}\right) \right)^n,
\end{align*}
Taking the logarithm of both sides, we get that \( \log(M_{(X-\mu)/(\sigma/\sqrt{n})}(t)) \) equals:

\[
\log(M_{Z_n}(t)) = -\frac{\sqrt{n}\mu t}{\sigma} + n \log M_X\left(\frac{t}{\sigma\sqrt{n}}\right)
\]

\[
= -\frac{\sqrt{n}\mu t}{\sigma} + n \log\left(1 + \mu_1 \frac{t}{\sigma n^{1/2}} + \mu_2 \frac{t^2}{2\sigma^2 n} + \mu_3 \frac{t^3}{\sigma^3 n^{3/2}}\right) + \ldots
\]
Expanding the logarithm as \( \log(1 + x) = x - x^2/2 + x^3/3 + \ldots \), we get

\[
= -\frac{\sqrt{n}t}{\sigma} + n\left( \mu_1 \frac{t}{\sigma n^{1/2}} + \mu_2 \frac{t^2}{2\sigma^2 n} + \mu_3 \frac{t^3}{\sigma^3 n^{3/2}} \right) + \ldots
\]

\[
- \frac{1}{2} \left( \mu_1 \frac{t}{\sigma n^{1/2}} + \mu_2 \frac{t^2}{2\sigma^2 n} + \mu_3 \frac{t^3}{\sigma^3 n^{3/2}} \right) + \ldots)^2
\]

\[
+ \frac{1}{3} \left( \mu_1 \frac{t}{\sigma n^{1/2}} + \mu_2 \frac{t^2}{2\sigma^2 n} + \mu_3 \frac{t^3}{\sigma^3 n^{3/2}} \right) + \ldots)^3 - \ldots
\]
Collecting like terms, this equals

$$(-\frac{\sqrt{n}\mu}{\sigma} + \frac{\sqrt{n}\mu_1}{\sigma})t + (\frac{\mu_2}{2\sigma^2} - \frac{\mu_1^2}{2\sigma^2})t^2 + (\frac{\mu_3}{6} - \frac{\mu_1\mu_2}{2} + \frac{\mu_1^3}{6})\frac{t^3}{\sigma^3\sqrt{n}} + \ldots$$

Using that $\mu_1 = E(X) = \mu$ and

$$\mu_2 - \mu_1^2 = E(X^2) - E(X)^2 = \sigma^2,$$

the first term vanishes and the second is $t^2/2$. 
Since all the higher order terms have powers of $\sqrt{n}$ in their denominator, this means that

$$\lim_{n \to \infty} \log M_{Z_n}(t) = \frac{1}{2} t^2$$

Since \( \log \) is continuous, we can take the log outside and exponentiate to get

$$\lim_{n \to \infty} M_{Z_n}(t) = e^{t^2/2},$$

which is the moment generating function of a standard normal distribution.

This proves the Central Limit Theorem.
**Poisson distributions**

**DEFINITION**

Let $\lambda > 0$. A discrete random variable $X$ is said to have a **Poisson distribution with parameter $\lambda$** if

$$P(X = k) = \begin{cases} e^{-\lambda} \left( \frac{\lambda^k}{k!} \right) & \text{if } k = 0, 1, 2, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

In this case, we write $X \sim \text{Poisson}(\lambda)$.

Note that the constant in front is needed for normalization.
We can view and calculate with Poisson distributions at:

http://www.distributome.org/V3/calc/PoissonCalculator.html

Notice the shape that a Poisson distribution takes on as $\lambda$ gets larger.
Begin Test.
**Example**

Let $X \sim \text{Poisson}(\lambda)$. What is $E(X)$? What is $\text{Var}(X)$?
We compute $E(X)$ directly from the definition:

$$E(X) = \sum_{k=0}^{\infty} ke^{-\lambda} \left( \frac{\lambda^k}{k!} \right)$$

$$= 0 + 1e^{-\lambda} \frac{\lambda^1}{1!} + 2e^{-\lambda} \frac{\lambda^2}{2!} + \ldots$$

$$= \lambda(e^{-\lambda} \frac{\lambda^0}{0!} + e^{-\lambda} \frac{\lambda^1}{1!} + \ldots)$$

$$= \lambda \sum_{j=0}^{\infty} e^{-\lambda} \left( \frac{\lambda^j}{j!} \right)$$

$$= \lambda.$$
Also,

\[
E(X^2) - \lambda = E(X^2) - E(X) \\
= E(X(X - 1)) \\
= \sum_{k=0}^{\infty} k(k - 1)e^{-\lambda}\frac{\lambda^k}{k!} \\
= 0 + 0 + (1)(2)e^{-\lambda}\frac{\lambda^2}{2!} + (2)(3)e^{-\lambda}\frac{\lambda^3}{3!} + \cdots \\
= \lambda^2(e^{-\lambda}\frac{\lambda}{0!} + e^{-\lambda}\frac{\lambda^1}{1!} + \cdots) \\
= \lambda^2 \sum_{j=0}^{\infty} e^{-\lambda}\frac{\lambda^j}{j!} \\
= \lambda^2, \\
\]

so \(E(X^2) = \lambda^2 + \lambda\).
From this, we get that

\[
\text{Var}(X) = E(X^2) - E(X)^2 \\
= \lambda^2 + \lambda - \lambda^2 \\
= \lambda.
\]
**PROPOSITION**

If \( \lim_{n \to \infty} p_n = 0 \) and \( \lim_{n \to \infty} np_n = \lambda \), then

\[
\lim_{n \to \infty} (1 - p_n)^n = e^{-\lambda}.
\]
To prove this, note that for $|x| < 1$,

$$\log(1 - x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \ldots.$$ 

This means that

$$\log(1 - p_n)^n = n \log(1 - p_n) = n(-p_n - p_n^2/2 - \ldots).$$
Exponentiating, we get

\[(1 - p_n)^n = e^{-np_n} e^{-np_n(p_n/2 + p_n^2/3 + ...)}\].

This means that

\[
\lim_{n \to \infty} (1 - p_n)^n = \lim_{n \to \infty} e^{-np_n} e^{-np_n(p_n/2 + p_n^2/3 + \ldots)} = e^{-\lambda} \cdot 1
\]
**PROPOSITION**

If $X_n \sim B(n, p_n)$ are random variables, and if $\lim_{n \to \infty} p_n = 0$ and $\lim_{n \to \infty} np_n = \lambda$, then

$$\lim_{n \to \infty} P(X_n = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$
In words: given a large number of independent events with small probability, then the count of events that occurs is approximately a Poisson distribution.

In practice this is used by noting that if $X_n \sim B(n, p)$, and if $p$ is small, then $X_n$ is approximately Poisson distributed with parameter $np$. 
To prove the proposition, first note that

$$\lim_{n \to \infty} P(X_n = 0) = \lim_{n \to \infty} (1 - p_n)^n = e^{-\lambda}$$

by the previous proposition.
For $k > 0$, we have

$$
\lim_{n \to \infty} P(X_n = k) = \lim_{n \to \infty} \binom{n}{k} \left( \frac{np_n}{n} \right)^k \left( 1 - \frac{np_n}{n} \right)^{n-k}
$$

$$
= \lim_{n \to \infty} \frac{n(n-1) \cdots (n-k+1)}{n^k} \frac{(np_n)^k}{k!} \left( 1 - \frac{np_n}{n} \right)^n \left( 1 - \frac{np_n}{n} \right)^{-k}
$$

$$
= 1 \cdot \frac{\lambda^k}{k!} e^{-\lambda} \cdot 1,
$$

which proves the proposition.
On the history of the Poisson distribution, as Wikipedia puts it:

“The distribution was first introduced by Siméon Denis Poisson (1781-1840) and published, together with his probability theory, in 1837 in his work *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*. The work theorized about the number of wrongful convictions in a given country by focusing on certain random variables $N$ that count, among other things, the number of discrete occurrences (sometimes called ‘events’ or ‘arrivals’) that take place during a time-interval of given length. The result had already been given in 1711 by Abraham de Moivre in *De Mensura Sortis seu; de Probabilitate Eventuum in Ludis a Casu Fortuito Pendentibus*. This makes it an example of Stigler’s law and it has prompted some authors to argue that the Poisson distribution should bear the name of de Moivre.
Again from Wikipedia: “A further practical application of this distribution was made by Ladislaus Bortkiewicz in 1898 when he was given the task of investigating the number of soldiers in the Prussian army killed accidentally by horse kicks; this experiment introduced the Poisson distribution to the field of reliability engineering.”
EXAMPLE

Up to 2000, the yearly average number of shark attacks off the coast of Florida was 2. In 2001, there were 6 shark attacks there. Is this unusual?
If we take the number $X$ of shark attacks per year to be Poisson distributed and use the average of 2 to approximate the mean, then $X \sim \text{Poisson}(2)$ approximately.

A professor from a university in Pennsylvania, argued that this number was not that unusual, but

$$P(X \geq 6) \approx 0.017.$$ 

Most people would probably regard that as unusual.
As an example, consider a 1939 paper by the psychologist B.F. Skinner, modeling alliteration in the Sonnets of Shakespeare.

He found the following numbers of s sounds per line, out of 1400 lines in the 100 sonnets that he analyzed.

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>702</td>
<td>501</td>
<td>161</td>
<td>29</td>
<td>7</td>
</tr>
<tr>
<td>Expected</td>
<td>716</td>
<td>480</td>
<td>161</td>
<td>36</td>
<td>6</td>
</tr>
</tbody>
</table>

“Expected” is what one would expect on average under a Poisson distribution with $\lambda$ equal to the sample mean of these counts.

The agreement is actually quite good! (One could test it with statistical techniques.)
The **gamma function** \( \Gamma : \mathbb{R} \rightarrow \mathbb{R} \) is defined by:

\[
\Gamma(x) = \int_{0}^{\infty} y^{x-1} e^{-y} \, dy.
\]

In the homework, you will explore some of the basic properties of the gamma function, such as:

\[
\Gamma(n) = (n - 1)!
\]

for all positive integers \( n \), and

\[
\Gamma(1/2) = \sqrt{\pi}.
\]
The **chi-square** distribution family $\chi^2(k)$ is a 1-parameter family of continuous distributions, where the **degrees of freedom** $k$ is a positive real number.

A continuous random variable $X$ is **chi-square distributed** with $k$ degrees of freedom (written $X \sim \chi^2(k)$) if its probability density function is

$$f(x) = \begin{cases} \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k-2)/2} e^{-x/2} & \text{for } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

An interactive version is at

http://www.distributome.org/V3/calc/ChiSquareCalculator.html
The \( t \) distribution family \( t(k) \) is a 1-parameter family of continuous distributions, where the **degrees of freedom** \( k \) is a positive real number.

A continuous random variable \( X \) is **\( t \) distributed** with \( k \) degrees of freedom (written \( X \sim t(k) \)) if its probability density function is

\[
f(x) = \frac{\Gamma((k + 1)/2)}{\Gamma(k/2)\sqrt{k\pi}} \left(1 + \frac{x^2}{k}\right)^{-(k+1)/2}.
\]

An interactive version is at:

http://www.distributome.org/V3/calc/StudentCalculator.html
The \( F \) distribution family \( F(k, \ell) \) is a 2-parameter family of continuous distributions, where the **numerator degrees of freedom** \( k \) and the **denominator degrees of freedom** \( \ell \) are both positive real numbers.

A continuous random variable \( X \) is \( F \) **distributed** with \( k \) and \( \ell \) degrees of freedom (written \( X \sim F(k, \ell) \)) if its probability density function is

\[
f(x) = \begin{cases} 
\frac{\Gamma((k+\ell)/2)}{\Gamma(k/2)\Gamma(\ell/2)} \left( \frac{k}{\ell} \right)^{k/2} \frac{x^{(k-2)/2}}{(1+(k/\ell)x)^{(k+\ell)/2}} & \text{if } x \geq 0 \\
0 & \text{otherwise.}
\end{cases}
\]

An interactive version is at:

http://www.distributome.org/V3/calc/FCalculator.html
PROPOSITION

Let $Z \sim N(0, 1)$. Then $Z^2 \sim \chi^2(1)$. 
Proof We know that $P(Z^2 < 0) = 0$, so let $z > 0$. Denoting the cumulative distribution function of $Z^2$ by $F$, we have

$$F(z) = P(Z^2 \leq z)$$
$$= P(-\sqrt{z} \leq Z \leq \sqrt{z})$$
$$= \int_{-\sqrt{z}}^{\sqrt{z}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \, dt$$
$$= 2 \int_{0}^{\sqrt{z}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \, dt.$$ 

Let $u = t^2$. The limits of integration become 0 and $z$, and we find that

$$F(z) = \int_{0}^{z} \frac{1}{\sqrt{2\pi u}} e^{-u/2} \, du.$$
Differentiating this with respect to $z$, we get that the probability density function of $Z^2$ equals

$$f(z) = \begin{cases} \frac{1}{\sqrt{2\pi}z}e^{-z/2} & \text{if } z \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

which is the probability density function of a $\chi^2(1)$ distribution.
**PROPOSITION**

Let \( U_1, \ldots, U_k \) be independent and have \( \chi^2 \) distributions with \( n_1, \ldots, n_k \) degrees of freedom. Then

\[
U = U_1 + \cdots + U_k \sim \chi^2(n_1 + \cdots + n_k).
\]

The usual proof of this proposition uses moment-generating functions.
**PROPOSITION**

Let $U \sim N(0,1)$, let $V \sim \chi^2(k)$, and suppose that $U$ and $V$ are independent. Define

$$T = \frac{U}{\sqrt{V/k}}.$$ 

Then $T \sim t(k)$.

The proof of this proposition uses multivariate techniques that are beyond the scope of this class.
**PROPOSITION**

Let \( U \sim \chi^2(k) \), let \( V \sim \chi^2(\ell) \), and suppose that \( U \) and \( V \) are independent. Define

\[
F = \frac{U}{k} / \frac{V}{\ell}.
\]

Then \( F \sim F(k, \ell) \).

The proof of this proposition uses multivariate techniques that are beyond the scope of this class.
Proving that the test statistics from the usual types of tests have their advertised distributions under the null hypothesis usually involves the previous propositions.

Where might these sorts of things go?

In mathematical statistics, we look at questions such as:

- Where do estimators come from?
- Where do test statistics come from?
- What makes a good estimator?
- What makes a good test statistic?

Considerations such as these allow one to have a better sense of what to do when “out of the box” statistics aren’t available, and also when assumptions go awry (as they usually do) in the usual statistical inference processes.
Went over example project.
Central Limit Theorems

by James Bernhard
A central limit theorem gives the distribution of sums or means of certain classes of random variables, in the limit as the number of terms in the sum or mean tends to infinity.

These theorems have a long and somewhat tangled history.

According to Wikipedia [W], as Pólya puts it, “the” theorem goes back to Laplace, the first rigorous proof is due to Chebyshev, and its “sharpest” formulation is in an article by Liapounov.
However, the history is more complicated than this:

- In the 1700’s, De Moivre proved a version for binomially distributed random variables.
- In the early 1800’s, Laplace used approximations resembling a central limit theorem.
- Others in the 1800’s contributed to the development, such as Poisson, Cauchy, and Dirichlet.
- “the” central limit theorem nowadays was first formulated and rigorously proved by Liapounov in 1900.
Theorem (Liapounov Central Limit Theorem)

Let $X_1, X_2, \ldots$, be independent and identically distributed random variables with finite expected value $\mu$ and finite, nonzero variance $\sigma^2$. Also, let

$$\bar{X}_n = \frac{X_1 + X_2 \cdots + X_n}{n},$$

and let $\Phi(z)$ denote the cumulative distribution function of a standard normal distribution:

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Then

$$\lim_{n \to \infty} P\left(a \leq \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \leq b\right) = \Phi(b) - \Phi(a)$$

for all $a, b \in \mathbb{R}$. 
We prove a slightly weaker version that assumes that the moment-generating functions of the IID random variables exist.

So a few words about moment-generating functions are in order...
Things we need to know about moment-generating functions:

- Their definition: \( M_X(t) = E(e^{tX}) \).
- If \( X \sim N(\mu, \sigma^2) \) then \( M_X(t) = e^{\frac{t^2}{2}} \).
- The basic property that \( M_{aX+b}(t) = e^{bt}M_X(at) \).
- Another basic property that for IID \( X_1, \ldots, X_n \):
  \[
  M_{X_1+X_2+\ldots+X_n}(t) = M_X(t)^n.
  \]
**Proposition**

Let $\mathcal{F}$ denote the space of probability density functions for which all moments exist.

1. There is a one-to-one correspondence between cumulative distribution functions in $\mathcal{F}$ and their moment generating functions.

2. A sequence of moment generating functions converges to a limiting moment generating function if and only if its corresponding cumulative distribution functions in $\mathcal{F}$ converge to the limiting moment generating function’s corresponding cumulative distribution function in $\mathcal{F}$. 
Theorem (Central Limit Theorem)

Let $X_1, X_2, \ldots$ be independent and identically distributed random variables whose moment generating functions exist in a neighborhood of 0. Let $\mu$ denote the common mean of these random variables, and let $\sigma$ denote their common standard deviation. Let

$$X_n = \frac{X_1 + X_2 \cdots + X_n}{n},$$

and let $\Phi(z)$ denote the cumulative distribution function of a standard normal distribution:

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Then

$$\lim_{n \to \infty} P(a \leq \frac{X_n - \mu}{\sigma / \sqrt{n}} \leq b) = \Phi(b) - \Phi(a)$$
Idea of the proof (following [MM]): let

\[ Z_n = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}. \]

Then

1. Use the basic properties of moment-generating functions to compute the logarithm of the mgf of \( Z_n \).
2. Expand the mgf in terms of its moments.
3. Expand further using the Taylor series for \( \log(1 + x) \).
4. Show that the limit as \( n \to \infty \) of the logarithm of these moment-generating functions is \( t^2 / 2 \).
This shows the moment-generating functions converge to the moment-generating function of a standard normal distribution.

By the correspondence between moment-generating functions and cumulative distribution functions, this proves the result.

If we didn’t assume that all the moments exist, this correspondence wouldn’t hold, so this proof wouldn’t work.
In practice, people use central limit theorems to approximate the distribution of a sample mean with a normal distribution.

For example, if we’re rolling a ball down a ramp and would like to know the mean length that it rolls, we might roll it \( n \) times and compute the sample mean.

Suppose we were able to bound the standard deviation (and hence variance) empirically by \( \sigma < 0.25 \text{ cm} \).
Then if we want to estimate the mean to within 0.1 cm, the probability of this would be approximately:

- 0.7941 if \( n = 10 \),
- 0.9545 if \( n = 25 \),
- 0.9999 if \( n = 100 \).

In other words, as long as the central limit theorem is even a moderately reasonable approximation, then we are nearly certain to estimate the mean within 0.1 cm if we roll the ball 100 times.
Summary

*Central limit theorems* give the distribution of sums or means of certain classes of random variables, in the limit as the number of terms in the sum or mean tends to infinity.

The earliest results of this form go back to the 1700’s.

The traditional form, assuming only IID random variables and the existence of their mean and variance, goes back to Liapounov in 1900.

Central limit theorems form the foundation of laboratory science, as they explain why it is useful to average multiple measurements.
Bibliography


BASIC DEFINITIONS

DISCRETE RANDOM VARIABLES

COMBINATORIAL TECHNIQUES

MULTIPLE RANDOM VARIABLES

MARKOV CHAINS

CONTINUOUS RANDOM VARIABLES

SUMS OF RANDOM VARIABLES

STATISTICAL APPLICATIONS

FURTHER TOPICS

STUDENT PRESENTATIONS
BASIC DEFINITIONS

DISCRETE RANDOM VARIABLES

COMBINATORIAL TECHNIQUES

MULTIPLE RANDOM VARIABLES

MARKOV CHAINS

CONTINUOUS RANDOM VARIABLES

SUMS OF RANDOM VARIABLES

STATISTICAL APPLICATIONS

FURTHER TOPICS

STUDENT PRESENTATIONS

FRIDAY, DECEMBER 6
BASIC DEFINITIONS

DISCRETE RANDOM VARIABLES

COMBINATORIAL TECHNIQUES

MULTIPLE RANDOM VARIABLES

MARKOV CHAINS

CONTINUOUS RANDOM VARIABLES

SUMS OF RANDOM VARIABLES

STATISTICAL APPLICATIONS

FURTHER TOPICS

STUDENT PRESENTATIONS