STAT/SOC/CSSS 221
Statistical Concepts and Methods for the Social Sciences

Introduction to Probability

Christopher Adolph

Department of Political Science

and

Center for Statistics and the Social Sciences

University of Washington, Seattle

Aside on mathematical notation

$\Pr(A)$ — the probability that event A happens out of a discrete set of possible events

Aside on mathematical notation

$\Pr(A)$	the probability that event A happens out of a discrete set of possible events
$\mathrm{P}(A)$	the probability that event A happens out of a continuum of possible events

Aside on mathematical notation

$\Pr(A)$	the probability that event A happens out of a discrete set of possible events
$\mathrm{P}(A)$	the probability that event A happens out of a continuum of possible events
$\Pr(\text{not } A)$	the probability something other than A happens

Aside on mathematical notation

$\Pr(A)$	the probability that event A happens out of a discrete set of possible events
$\mathrm{P}(A)$	the probability that event A happens out of a continuum of possible events
$\Pr(\text{not } A)$	the probability something other than A happens
$\Pr(A \cup B)$	the probability either A or B (or both) happen ($\cup$ stands for union)

Aside on mathematical notation

$\Pr(A)$	the probability that event A happens out of a discrete set of possible events
$\mathrm{P}(A)$	the probability that event A happens out of a continuum of possible events
$\Pr(\text{not } A)$	the probability something other than A happens
$\Pr(A \cup B)$	the probability either A or B (or both) happen ($\cup$ stands for union)
$\Pr(A \cap B)$	the probability A and B happen *at the same time* ($\cap$ stands for intersection)

Aside on mathematical notation

$\Pr(A)$	the probability that event A happens out of a discrete set of possible events
$\mathrm{P}(A)$	the probability that event A happens out of a continuum of possible events
$\Pr(\text{not } A)$	the probability something other than A happens
$\Pr(A \cup B)$	the probability either A or B (or both) happen ($\cup$ stands for union)
$\Pr(A \cap B)$	the probability A and B happen *at the same time* ($\cap$ stands for intersection)
$\Pr(A\|B)$	the probability A happens given that B is certain to happen

Two examples

Time for a second opinion? During a routine checkup, your doctor tells you some bad news: you tested positive for a rare disease. The disease affects 1 in 10,000 people, and the test has a 99% effectiveness rate. What are the chances you have the disease?

Concepts applied: Sample spaces. Conditional probability.

Two examples

Time for a second opinion? During a routine checkup, your doctor tells you some bad news: you tested positive for a rare disease. The disease affects 1 in 10,000 people, and the test has a 99% effectiveness rate. What are the chances you have the disease?

Concepts applied: Sample spaces. Conditional probability.

A statistician plays the lottery What are your chances of winning the lottery?

Concepts applied: Complex events. Independence. Joint probability. Expected value.

Essential concepts

Event Any specific outcome that might occur. Mutually exclusive with other events.

Example: A World Cup Soccer game between the US and England could end in a victory for the US.

Essential concepts

Event Any specific outcome that might occur. Mutually exclusive with other events.

Example: A World Cup Soccer game between the US and England could end in a victory for the US.

Sample space A set of *all* the events that might occur.

Example: A World Cup Soccer game between the US and England could end in three ways: Win, Loss, Tie

Essential concepts

Event Any specific outcome that might occur. Mutually exclusive with other events.

Example: A World Cup Soccer game between the US and England could end in a victory for the US.

Sample space A set of *all* the events that might occur.

Example: A World Cup Soccer game between the US and England could end in three ways: Win, Loss, Tie

Probability A number, between 0 and 1, assigned to each of the possible outcomes in the sample space.

Example: $\mathrm{Pr}(\mathrm{Win}) = 0.3, \quad \mathrm{Pr}(\mathrm{Loss}) = 0.5, \quad \mathrm{Pr}(\mathrm{Tie}) = 0.2$

Axioms of Probability

An axiom is an assumption we cannot prove, and must make to get started in a field of mathematics.

Probability theory relies on just three axioms:

1. Each event has a probability of 0 or more. (Anything could happen.)

Axioms of Probability

An axiom is an assumption we cannot prove, and must make to get started in a field of mathematics.

Probability theory relies on just three axioms:

1. Each event has a probability of 0 or more. (Anything could happen.)
2. The total probability of all the events in the sample space is 1. (Something must happen.)

Axioms of Probability

An axiom is an assumption we cannot prove, and must make to get started in a field of mathematics.

Probability theory relies on just three axioms:

1. Each event has a probability of 0 or more. (Anything could happen.)
2. The total probability of all the events in the sample space is 1. (Something must happen.)
3. If two events A and B cannot simultaneously occur, the probability that either occurs is $\Pr(A) + \Pr(B)$

Approaches to Probability

Frequency Interpretation We observe how often an event occurs in a set of trials. The ratio of successes to trials show reflect the long-term probability of the event.

Approaches to Probability

Frequency Interpretation We observe how often an event occurs in a set of trials. The ratio of successes to trials show reflect the long-term probability of the event.

Theoretical interpretation The above suggests there is also some true, usually unknown, probability that an event occurs in nature.

Approaches to Probability

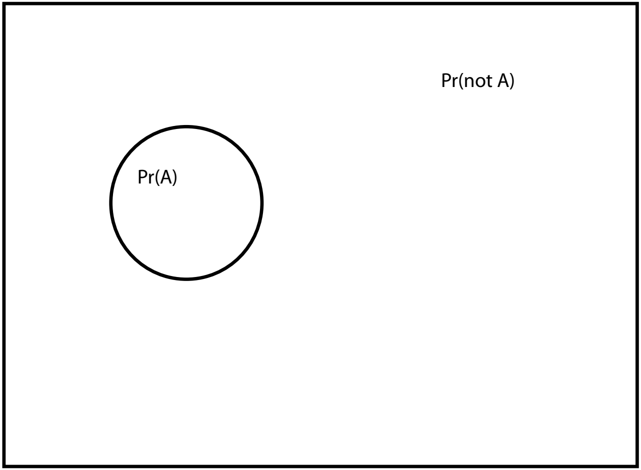
Frequency Interpretation We observe how often an event occurs in a set of trials. The ratio of successes to trials show reflect the long-term probability of the event.

Theoretical interpretation The above suggests there is also some true, usually unknown, probability that an event occurs in nature.

Subjective Interpretation Probabilities may also reflect personal beliefs about the likelihood of an event.

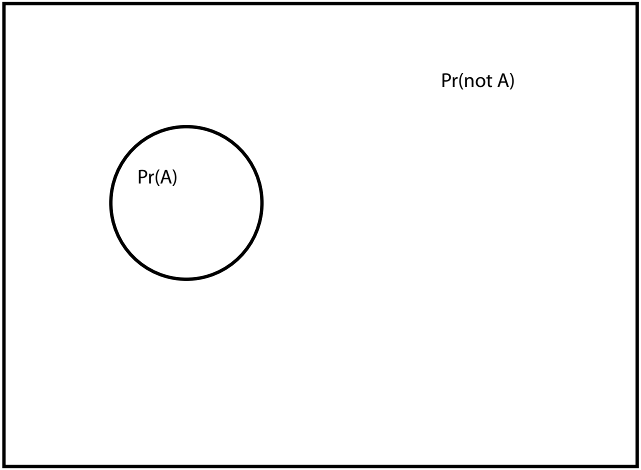
We'll mostly rely on the frequency interpretation, but will use all three at different points in the course

Visualizing sample spaces and events with Venn Diagrams



The entire rectangle is the *sample space*, and has area, or total probability, of 1

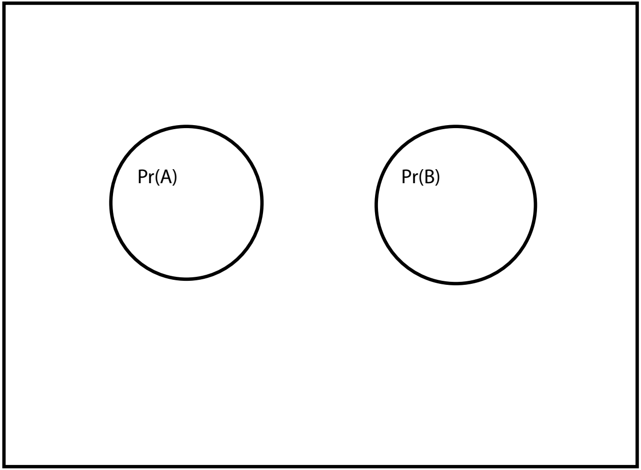
Visualizing sample spaces and events with Venn Diagrams



The entire rectangle is the *sample space*, and has area, or total probability, of 1

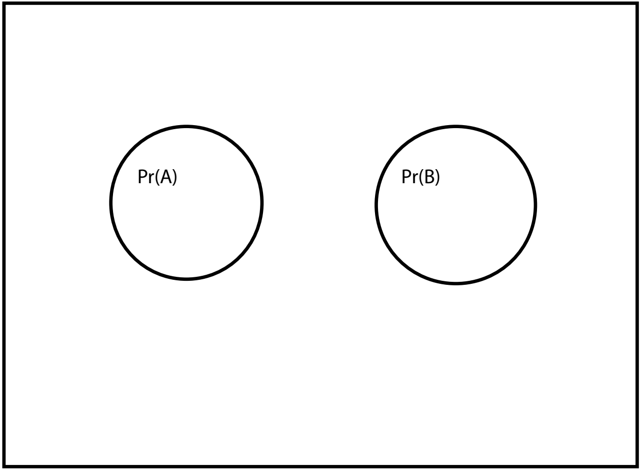
The area of the small circle represents the probability that event A happens

Visualizing sample spaces and events with Venn Diagrams



If the circles for two different events don't overlap, they are disjoint

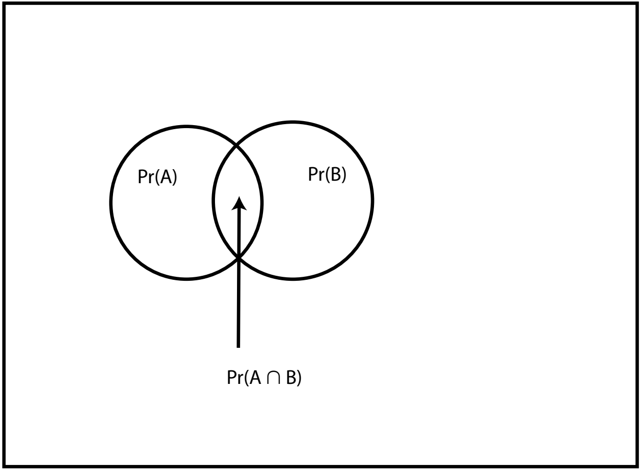
Visualizing sample spaces and events with Venn Diagrams



If the circles for two different events don't overlap, they are disjoint

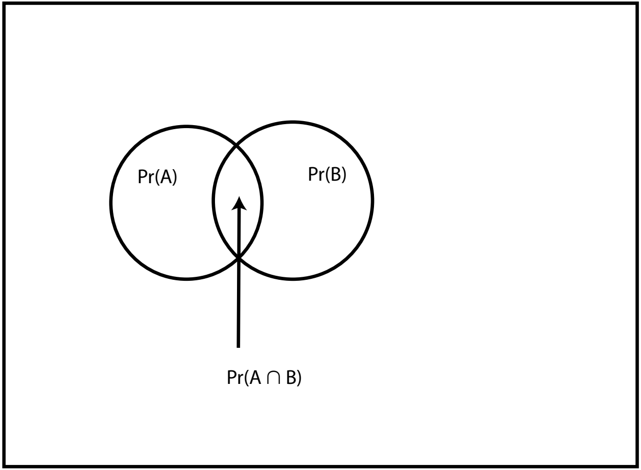
We say these events are mutually exclusive

Visualizing sample spaces and events with Venn Diagrams



If the circles overlap, both events could happen at once

Visualizing sample spaces and events with Venn Diagrams



If the circles overlap, both events could happen at once

The probability both happen is the *joint* probability of $A \cap B$

Tools to solve probability problems

Complement rule $\Pr(A) = 1 - \Pr(\text{not } A)$

Tools to solve probability problems

Complement rule $\Pr(A) = 1 - \Pr(\text{not } A)$

Addition rule (generally) $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$

Tools to solve probability problems

Complement rule $\Pr(A) = 1 - \Pr(\text{not } A)$

Addition rule (generally) $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$

Addition rule (mutually exclusive events) $\Pr(A \cup B) = \Pr(A) + \Pr(B)$ if $\Pr(A \cap B) = 0$

Tools to solve probability problems

Complement rule $\Pr(A) = 1 - \Pr(\text{not } A)$

Addition rule (generally) $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$

Addition rule (mutually exclusive events) $\Pr(A \cup B) = \Pr(A) + \Pr(B)$ if $\Pr(A \cap B) = 0$

Conditional probability $\Pr(A|B) = \Pr(A \cap B)/\Pr(B)$

Tools to solve probability problems

Complement rule $\Pr(A) = 1 - \Pr(\text{not } A)$

Addition rule (generally) $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$

Addition rule (mutually exclusive events) $\Pr(A \cup B) = \Pr(A) + \Pr(B)$ if $\Pr(A \cap B) = 0$

Conditional probability $\Pr(A|B) = \Pr(A \cap B)/\Pr(B)$

Independence of events Events are independent *only if* $\Pr(A|B) = \Pr(A)$

Tools to solve probability problems

Complement rule $\Pr(A) = 1 - \Pr(\text{not } A)$

Addition rule (generally) $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$

Addition rule (mutually exclusive events) $\Pr(A \cup B) = \Pr(A) + \Pr(B)$ if $\Pr(A \cap B) = 0$

Conditional probability $\Pr(A|B) = \Pr(A \cap B)/\Pr(B)$

Independence of events Events are independent *only if* $\Pr(A|B) = \Pr(A)$

Joint probability (independent events) $\Pr(A \cap B) = \Pr(A)\Pr(B)$

Tools to solve probability problems

Complement rule $\Pr(A) = 1 - \Pr(\text{not } A)$

Addition rule (generally) $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$

Addition rule (mutually exclusive events) $\Pr(A \cup B) = \Pr(A) + \Pr(B)$ if $\Pr(A \cap B) = 0$

Conditional probability $\Pr(A|B) = \Pr(A \cap B)/\Pr(B)$

Independence of events Events are independent *only if* $\Pr(A|B) = \Pr(A)$

Joint probability (independent events) $\Pr(A \cap B) = \Pr(A)\Pr(B)$

Joint probability (generally) $\Pr(A \cap B) = \Pr(A)\Pr(B|A)$

Time for a Second Opinion?

During a routine checkup, your doctor tells you some bad news:
you tested positive for a rare disease.

The disease affects 1 in 10,000 people,
and the test has a 99% effectiveness rate.

What are the chances you have the disease?

Steps to solve a probability problem

1. Identify the possible events
2. Identify the quantity of interest in terms of probability of specific events
3. Collect all the probabilities you know
4. Use the rules of probability to calculate what you want to know from what you do know

Identify the possible events

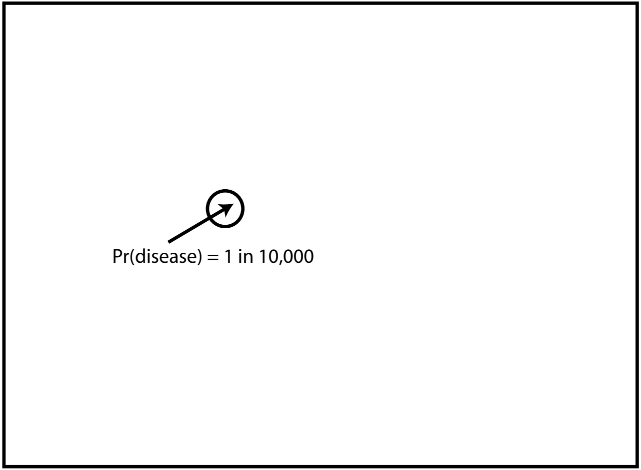
You can either have the disease or not have the disease

You can either test positive or negative

This leads to four possible combinations which comprise the whole sample space:

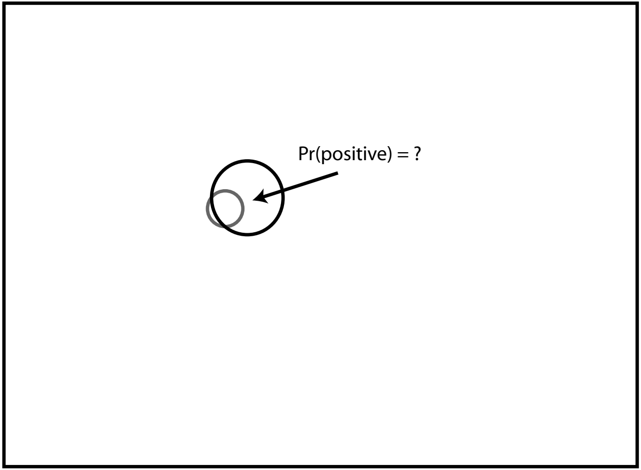
- have disease and positive
- have disease and negative
- no disease and positive
- no disease and negative

Venn Diagram of disease and test events



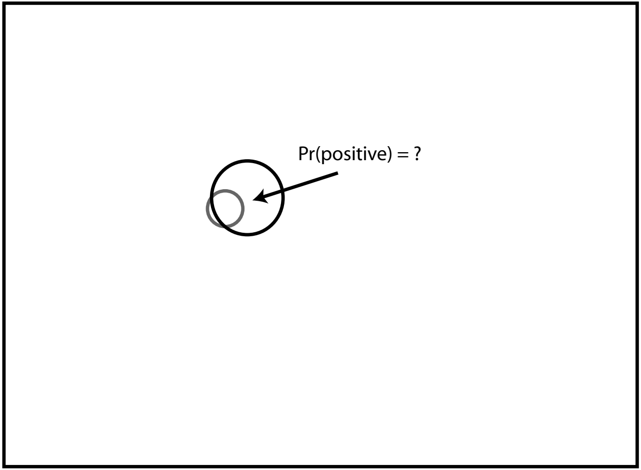
Let's start with the event of having the disease

Venn Diagram of disease and test events



Testing positive is a separate, possibly joint event with having the disease

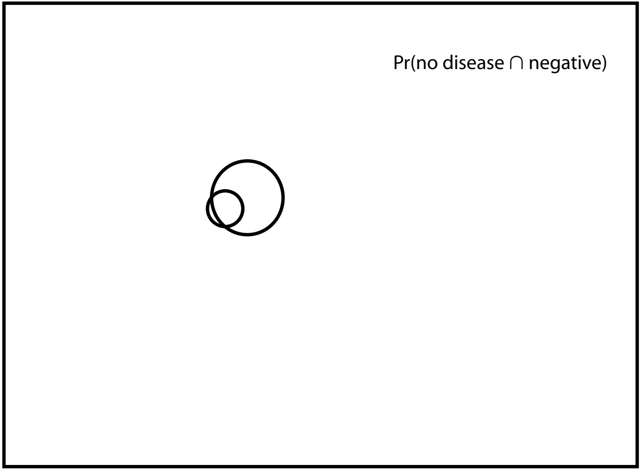
Venn Diagram of disease and test events



Testing positive is a separate, possibly joint event with having the disease

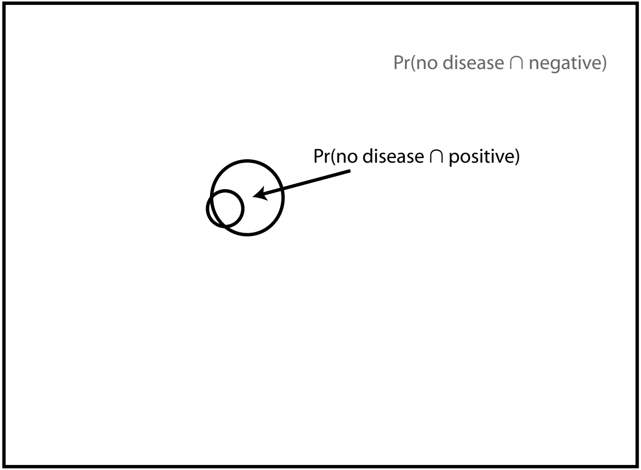
We don't yet know the marginal probability of testing positive

Venn Diagram of disease and test events



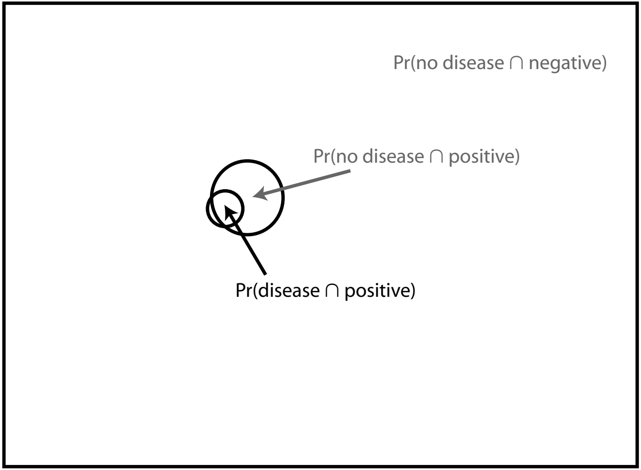
The area outside both circles is the probability of neither having the disease nor testing positive

Venn Diagram of disease and test events



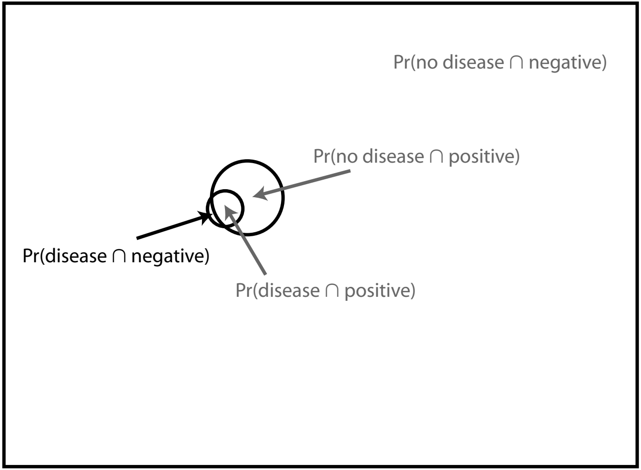
This large crescent shape is the probability of testing positive even though you don't have the disease

Venn Diagram of disease and test events



This intersection is the probability of testing positive and having the disease

Venn Diagram of disease and test events



And this crescent is the probability of having the disease but failing to detect it

Identify the quantity of interest

We want to know the possibility you have the disease *given* a positive test result

The simple probability of having a disease isn't enough
—you're worried because you have new information:
you've had a positive test

In formal terms, we want to find:

$$\Pr(\text{disease}|\text{positive})$$

This is not the same as either $\Pr(\text{disease})$ or $\Pr(\text{positive})$

How do we find it? Let's start with what we know

Collect all the probabilities you know

We know how likely a random person is to have the disease:

$$\begin{aligned}\Pr(\text{disease}) &= \text{ 1 in 10,000} \\ &= 0.01\% \\ &= 0.0001\end{aligned}$$

Collect all the probabilities you know

We know how likely a random person is to have the disease:

$$\begin{aligned} \Pr(\text{disease}) &= \text{1 in 10,000} \\ &= 0.01\% \\ &= 0.0001 \end{aligned}$$

From this, we can calculate the complement:

$$\begin{aligned} \Pr(\text{no disease}) &= 1 - \Pr(\text{disease}) \\ &= \text{9,999 in 10,000} \\ &= 99.99\% \\ &= 0.9999 \end{aligned}$$

Collect all the probabilities you know

We also know some conditional probabilities. The test is “99% effective”, meaning it has a 99% probability of detecting the correct disease status, so:

$$\begin{aligned}\Pr(\text{positive}|\text{disease}) &= 0.99\\ &= 99 \text{ in } 100\end{aligned}$$

Collect all the probabilities you know

We also know some conditional probabilities. The test is "99% effective", meaning it has a 99% probability of detecting the correct disease status, so:

$$\begin{aligned}
\Pr(\text{positive}|\text{disease}) &= 0.99 \\
&= 99 \text{ in } 100 \\
\Pr(\text{negative}|\text{disease}) &= 0.01 \\
&= 1 \text{ in } 100
\end{aligned}$$

Collect all the probabilities you know

We also know some conditional probabilities. The test is "99% effective", meaning it has a 99% probability of detecting the correct disease status, so:

$$\begin{aligned}
\Pr(\text{positive}|\text{disease}) &= 0.99 \\
&= 99 \text{ in } 100 \\
\Pr(\text{negative}|\text{disease}) &= 0.01 \\
&= 1 \text{ in } 100 \\
\Pr(\text{positive}|\text{no disease}) &= 0.01 \\
&= 1 \text{ in } 100
\end{aligned}$$

Collect all the probabilities you know

We also know some conditional probabilities. The test is "99% effective", meaning it has a 99% probability of detecting the correct disease status, so:

$$
\begin{aligned}
\Pr(\text{positive}|\text{disease}) &= 0.99 \\
&= 99 \text{ in } 100 \\
\Pr(\text{negative}|\text{disease}) &= 0.01 \\
&= 1 \text{ in } 100 \\
\Pr(\text{positive}|\text{no disease}) &= 0.01 \\
&= 1 \text{ in } 100 \\
\Pr(\text{negative}|\text{no disease}) &= 0.99 \\
&= 99 \text{ in } 100
\end{aligned}
$$

What we still don't know

$$\begin{aligned} \Pr(\text{disease}|\text{positive}) &= \textbf{?} \\ \Pr(\text{no disease}|\text{positive}) &= \textbf{?} \end{aligned}$$

How are we going to calculate these?

What we still don't know

$$\begin{aligned} \Pr(\text{disease}|\text{positive}) &= \textbf{?} \\ \Pr(\text{no disease}|\text{positive}) &= \textbf{?} \end{aligned}$$

How are we going to calculate these?

We have a formula for conditional probabilities:

$$\text{conditional probability} = \frac{\text{joint probability}}{\text{marginal probability}}$$

What we still don't know

$$\begin{aligned} \Pr(\text{disease}|\text{positive}) &= \textbf{?} \\ \Pr(\text{no disease}|\text{positive}) &= \textbf{?} \end{aligned}$$

How are we going to calculate these?

We have a formula for conditional probabilities:

$$\begin{aligned} \text{conditional probability} &= \frac{\text{joint probability}}{\text{marginal probability}} \\ \Pr(A|B) &= \frac{\Pr(A \cap B)}{\Pr(B)} \end{aligned}$$

So we need to calculate some joint probabilities
(the areas in the Venn Diagrams)

Use the rules of probability to calculate the missing probabilities

2×2 table for disease test example

	disease	no disease	
positive			
negative			

One way to calculate missing probabilities is to make a contingency table, and see if we can fill in the blanks

Use the rules of probability to calculate the missing probabilities

2 $\times$ 2 table for disease test example

	disease	no disease	
positive			
negative			
			1,000,000

Let's choose a large "sample" size, and then compute appropriate frequencies

That is, we'll work with the frequency interpretation of probability

Use the rules of probability to calculate the missing probabilities

2 $\times$ 2 table for disease test example

	disease	no disease	
positive			
negative			
	100	999,900	1,000,000

If the $\Pr(\text{disease})$ is 1 in 10,000, then out of a sample of 1 million, about 100 should have the disease.

These simple probabilities fill in the *margins* of the table, and so simple probabilities are often called *marginal probabilities*

Use the rules of probability to calculate the missing probabilities

2 × 2 table for disease test example

	disease	no disease	
positive	99		
negative	1		
	100	999,900	1,000,000

We know from the conditional probability of test results given disease that 99 of the 100 disease cases should be detected

This helps us fill in the joint probabilities of disease and test results

Use the rules of probability to calculate the missing probabilities

2×2 table for disease test example

	disease	no disease	
positive	99	9,999	
negative	1	989,901	
	100	999,900	1,000,000

Likewise, we can fill in the fraction of non-disease cases the test should correctly identify

Use the rules of probability to calculate the missing probabilities

2 × 2 table for disease test example

	disease	no disease	
positive	99	9,999	10,098
negative	1	989,901	989,902
	100	999,900	1,000,000

That just leaves the marginal totals of positive and negative results, which we find by adding up the rows

Use the rules of probability to calculate the missing probabilities

2 $\times$ 2 table for disease test example

	disease	no disease	
positive	0.000099	0.009999	0.010098
negative	0.000001	0.989901	0.989902
	0.0001	0.99999	1.0

Dividing through by the grand sum of the table converts all the entries to probabilities

Note that unlikely last week, we want to divide by the overall sum, not the column sums. Our goal is to find the probabilities of each combination of events

Use the rules of probability to calculate the missing probabilities

2 × 2 table for disease test example

	disease	no disease	
positive			Pr(positive)
negative			Pr(negative)
	Pr(disease)	Pr(no disease)	Pr(any event)

Let's think about what we've done in terms of probabilities

The margins of the table have the simple probabilities of each event

Use the rules of probability to calculate the missing probabilities

2 × 2 table for disease test example

	disease	no disease	
positive	$\Pr(\text{disease} \cap \text{positive})$	$\Pr(\text{no disease} \cap \text{positive})$	$\Pr(\text{positive})$
negative	$\Pr(\text{disease} \cap \text{negative})$	$\Pr(\text{no disease} \cap \text{negative})$	$\Pr(\text{negative})$
	$\Pr(\text{disease})$	$\Pr(\text{no disease})$	$\Pr(\text{any event})$

And the cells of the table show the joint probability of each combination of events

Use the rules of probability to calculate the missing probabilities

2 $\times$ 2 table for disease test example

	disease	no disease	
positive	$\Pr(\text{disease} \cap \text{positive})$	$\Pr(\text{no disease} \cap \text{positive})$	$\Pr(\text{positive})$
negative	$\Pr(\text{disease} \cap \text{negative})$	$\Pr(\text{no disease} \cap \text{negative})$	$\Pr(\text{negative})$
	$\Pr(\text{disease})$	$\Pr(\text{no disease})$	$\Pr(\text{any event})$

We can use these cells to compute any conditional probability we want:

$$\text{conditional probability} \quad = \quad \frac{\text{joint probability}}{\text{marginal probability}}$$

Use the rules of probability to calculate the missing probabilities

2 × 2 table for disease test example

	disease	no disease	
positive	$\Pr(\text{disease} \cap \text{positive})$	$\Pr(\text{no disease} \cap \text{positive})$	$\Pr(\text{positive})$
negative	$\Pr(\text{disease} \cap \text{negative})$	$\Pr(\text{no disease} \cap \text{negative})$	$\Pr(\text{negative})$
	$\Pr(\text{disease})$	$\Pr(\text{no disease})$	$\Pr(\text{any event})$

We can use these cells to compute any conditional probability we want:

$$\text{conditional probability} = \frac{\text{joint probability}}{\text{marginal probability}}$$

$$\Pr(\text{disease}|\text{positive}) = \frac{\Pr(\text{disease} \cap \text{positive})}{\Pr(\text{positive})}$$

Use the rules of probability to calculate the missing probabilities

2 × 2 table for disease test example

	disease	no disease	
positive	0.000099	0.009999	0.010098
negative	0.000001	0.989901	0.989902
	0.0001	0.99999	1.0

$$\Pr(\text{disease}|\text{positive}) = \frac{\Pr(\text{disease} \cap \text{positive})}{\Pr(\text{positive})}$$

Use the rules of probability to calculate the missing probabilities

2 × 2 table for disease test example

	disease	no disease	
positive	0.000099	0.009999	0.010098
negative	0.000001	0.989901	0.989902
	0.0001	0.99999	1.0

$$\begin{aligned}\Pr(\text{disease}|\text{positive}) &= \frac{\Pr(\text{disease} \cap \text{positive})}{\Pr(\text{positive})} \\ &= \frac{0.000099}{0.010098} = 0.0098\end{aligned}$$

Use the rules of probability to calculate the missing probabilities

2 × 2 table for disease test example

	disease	no disease	
positive	0.000099	0.009999	0.010098
negative	0.000001	0.989901	0.989902
	0.0001	0.99999	1.0

If you randomly test people for the disease, and get a positive result, there are 99 chances in 100 that the person doesn't have the disease

And only a 1 in 100 chance of the person having the disease, despite the positive test!

Alternative Presentation: Probability Trees

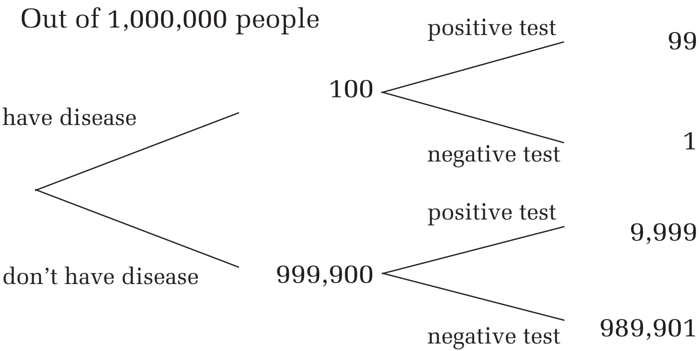
Many students find conditional probabilities easier to understanding using tree diagrams

We will look at this problem a second way using a tree

Each node in the tree represents a random variable

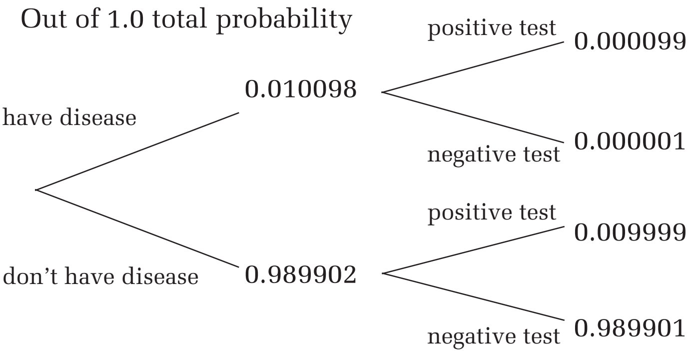
Each branch represents a possible value of that variable

Tracing out each branch to the tip shows the joint probability that a set of variables come out a certain way



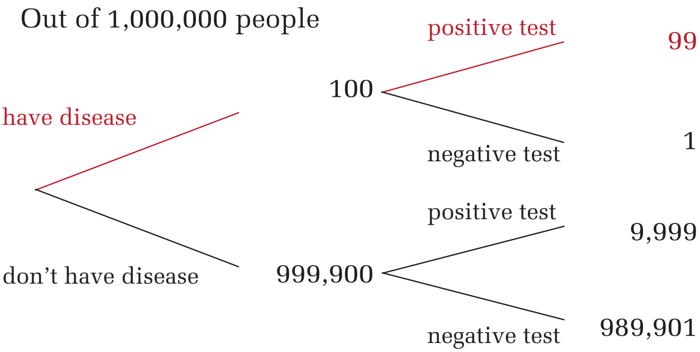
Again suppose we took 1,000,000 people, and their outcomes followed exactly the joint and marginal probabilities we determined for each event.

We would have the above tree, where numbers at the right show the total people with all the conditions on the corresponding branch



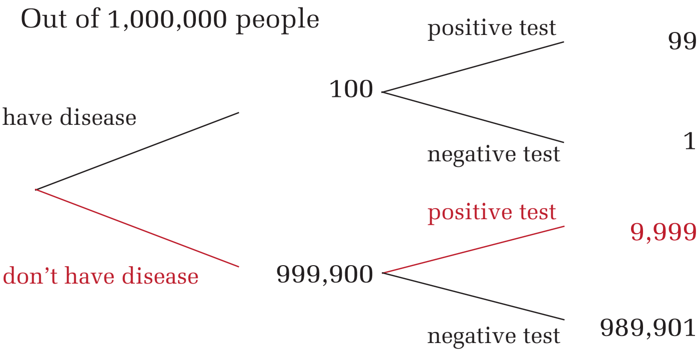
If we divide through by 1,000,000, we get the underlying probabilities.

Let's leave them aside for the moment, and find the probability of disease given a positive diagnosis, $\Pr(\text{disease}|\text{positive})$ using frequencies out of one million



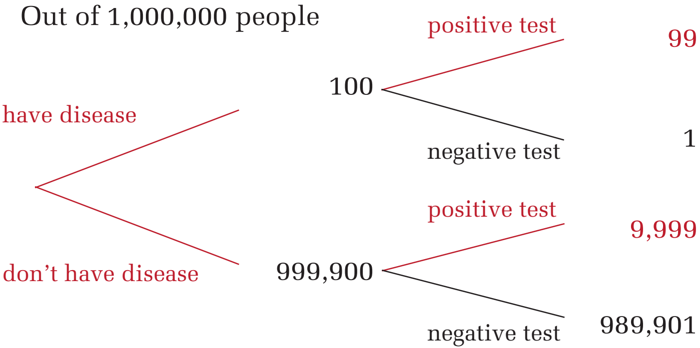
How many people out of 1,000,000 received positive diagnoses?

The red path above highlights one way: 99 people in 1,000,000 will get a positive diagnosis and have the disease



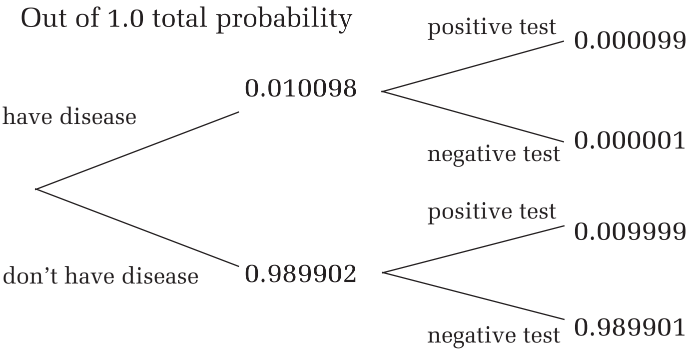
How many people out of 1,000,000 received positive diagnoses?

But there is another way: 9,999 people out of 1,000,000 will test positive even without the disease



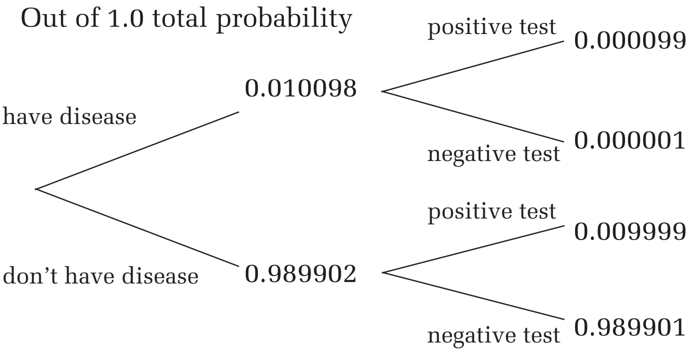
A total of $99 + 9,999$ receive positive diagnoses, but only 99 of these people have the disease.

$$\Pr(\text{disease}|\text{positive}) = \frac{99}{99 + 9,999} = 0.0098 \approx \text{ 1 percent probability}$$



We could also solve this tree using the probabilities themselves, and the formula for conditional probability

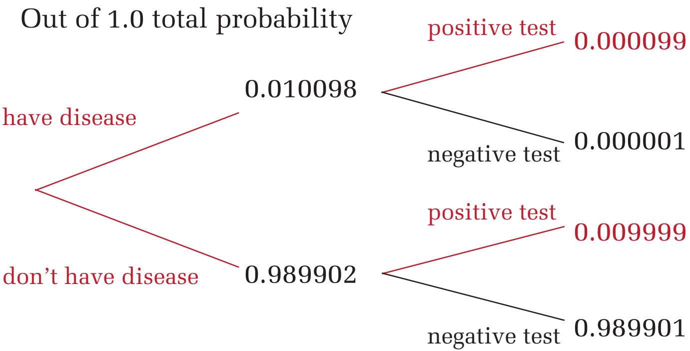
$$\text{conditional probability} = \frac{\text{joint probability}}{\text{marginal probability}}$$



We could also solve this tree using the probabilities themselves, and the formula for conditional probability

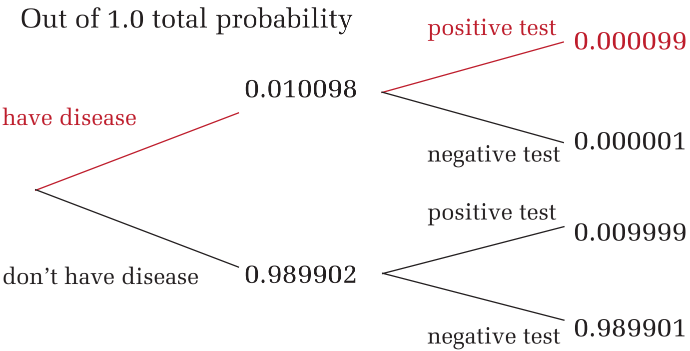
$$\text{conditional probability} = \frac{\text{joint probability}}{\text{marginal probability}}$$

$$\Pr(\text{disease}|\text{positive}) = \frac{\Pr(\text{disease} \cap \text{positive})}{\Pr(\text{positive})}$$



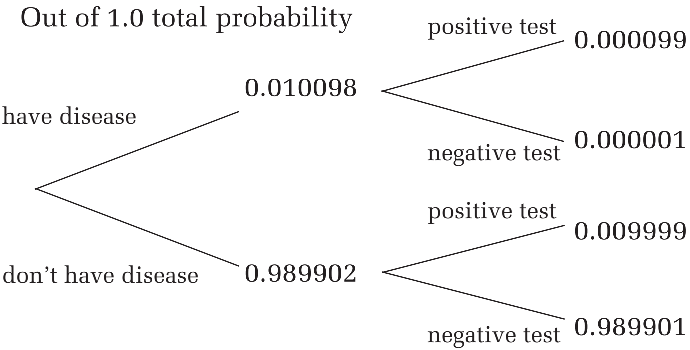
What is the marginal probability of a positive diagnosis, $\Pr(\text{positive})$?

The sum of the probabilities in red: $\Pr(\text{positive}) = 0.010098$

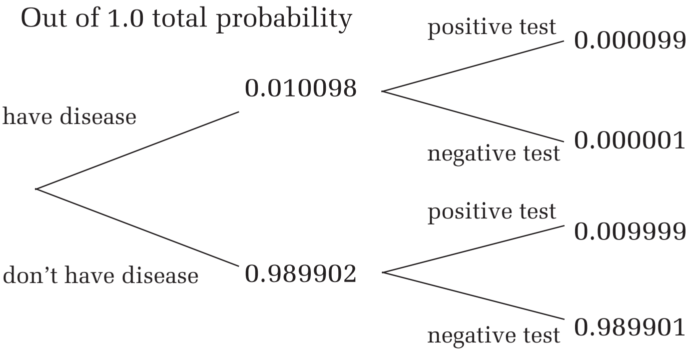


What is the joint probability of a positive diagnosis and a positive test?

The probability in red: $\Pr(\text{disease} \cap \text{positive}) = 0.000099$



$$\Pr(\text{disease}|\text{positive}) = \frac{\Pr(\text{disease} \cap \text{positive})}{\Pr(\text{positive})}$$



$$
\begin{aligned}
\Pr(\text{disease}|\text{positive}) &= \frac{\Pr(\text{disease} \cap \text{positive})}{\Pr(\text{positive})} \\
&= \frac{0.000099}{0.010098} = 0.0098 \\
&\approx \text{1 percent probability}
\end{aligned}
$$

Things to ponder

1. It looks like a test for a rare disease would need to be staggeringly accurate before we trust it

   But remember we assumed the test was administered at random

   What if your doctor already suspected you had the disease. Would you still discount a positive result?

Things to ponder

1. It looks like a test for a rare disease would need to be staggeringly accurate before we trust it

   But remember we assumed the test was administered at random

   What if your doctor already suspected you had the disease. Would you still discount a positive result?

2. We can present probabilities as proportions (0.001) or as ratios (1 in 1000). Which do you find easier to understand?

Things to ponder

1. It looks like a test for a rare disease would need to be staggeringly accurate before we trust it

   But remember we assumed the test was administered at random

   What if your doctor already suspected you had the disease. Would you still discount a positive result?

2. We can present probabilities as proportions (0.001) or as ratios (1 in 1000). Which do you find easier to understand?

   This example was administered as a test question to a panel of doctors

   Using proportions, *most* doctors got the answer badly wrong (Kahneman & Tversky)

   Using ratios to understand the problem, most got it right! (Gigerenzer)

A statistician plays the lottery

Many US states run lotteries involving a daily drawing of 6 numbered balls from an urn, without replacement

In WA state, lottery players pick six unique integers from 1 to 49

A player who picks all 6 numbers correctly wins the jackpot

If no one picks all 6 numbers, the jackpot rolls over to the next drawing

In WA and other states, partial matches win smaller prizes;
we will neglect these to keep our example simple

A statistician plays the lottery

Questions a statistician asks about a lottery:

1. What is the probability of winning the jackpot from buying a single ticket?
2. What is the expected return on a single ticket?
3. Can I increase my expected return using a strategy?
4. Based on the above, should I play the lottery?

Probability of winning the jackpot

What is the sample space? All possible combinations of 6 different numbers chosen at random between 1 and 49.

What is the probability of matching such a number?

Start by identifying the event of interest:

Probability of winning the jackpot

What is the sample space? All possible combinations of 6 different numbers chosen at random between 1 and 49.

What is the probability of matching such a number?

Start by identifying the event of interest: Matching all 6 selected numbers

A complex event, consisting of six sequentially sub-events

Let's denote a successful match of the nth number as m_n

Probability of winning the jackpot

To have a “success”, we must see the following six events in order:

- Match the first draw (a random selection out of 49 numbers) to any of our 6 picks.

  We have six chances in 49 to get this right, so our probability is:

$$\Pr(m_1) = 6/49 = 0.1224$$

Probability of winning the jackpot

To have a "success", we must see the following six events in order:

- Match the first draw (a random selection out of 49 numbers) to any of our 6 picks.

  We have six chances in 49 to get this right, so our probability is:

  $$\Pr(m_1) = 6/49 = 0.1224$$

- Match the second draw (randomly chosen from 48 remaining numbers) to any of our remaining five picks.

  Assuming we got the first number, we have five chances in 48 to get this right:

  $$\Pr(m_2|m_1) = 5/48 = 0.1042$$

Probability of winning the jackpot

- Match the third draw (randomly chosen from 47 remaining numbers) to any of our remaining four picks.

  Assuming we got the first and second numbers, we have four chances in 47 to get this right:

$$\Pr(m_3|m_1, m_2) = 4/47 = 0.0851$$

Probability of winning the jackpot

- Match the third draw (randomly chosen from 47 remaining numbers) to any of our remaining four picks.

  Assuming we got the first and second numbers, we have four chances in 47 to get this right:

$$\Pr(m_3|m_1, m_2) = 4/47 = 0.0851$$

- Match the fourth draw (randomly chosen from 46 remaining numbers) to any of our remaining three picks.

  Assuming we got the first, second, and third numbers, we have three chances in 46 to get this right:

$$\Pr(m_4|m_1, m_2, m_3) = 3/46 = 0.0652$$

Probability of winning the jackpot

- Match the fifth draw (randomly chosen from 45 remaining numbers) to any of our remaining two picks.

  Assuming we got the first, second, third, and fourth numbers, we have two chances in 45 to get this right:

$$\Pr(m_5|m_1, m_2, m_3, m_4) = 2/45 = 0.0444$$

Probability of winning the jackpot

- Match the fifth draw (randomly chosen from 45 remaining numbers) to any of our remaining two picks.

  Assuming we got the first, second, third, and fourth numbers, we have two chances in 45 to get this right:

  $$\Pr(m_5|m_1, m_2, m_3, m_4) = 2/45 = 0.0444$$

- Match the sixth draw (randomly chosen from 44 remaining numbers) to our only remaining pick.

  Assuming we got the first, second, third, fourth, and fifth numbers, we have one chance in 44 to get this right:

  $$\Pr(m_6|m_1, m_2, m_3, m_4, m_5) = 1/44 = 0.0227$$

Probability of winning the jackpot

Note something interesting: matching numbers gets harder as we go, because our set of possible matches is shrinking.

To win the jackpot, we must match all six numbers, so we need the joint probability of the above events.

We use the general rule for calculating the joint probability of multiple events:

$$\Pr(\text{Jackpot}) = \Pr(m_1 \cap m_2 \cap m_3 \cap m_4 \cap m_5 \cap m_6)$$

Probability of winning the jackpot

Note something interesting: matching numbers gets harder as we go, because our set of possible matches is shrinking.

To win the jackpot, we must match all six numbers,
so we need the joint probability of the above events.

We use the general rule for calculating the joint probability of multiple events:

$$\begin{aligned}
\Pr(\text{Jackpot}) &= \Pr(m_1 \cap m_2 \cap m_3 \cap m_4 \cap m_5 \cap m_6) \\
&= \Pr(m_1) \times \Pr(m_2|m_1) \times \Pr(m_3|m_1, m_2) \times \Pr(m_4|m_1, m_2, m_3) \\
&\qquad \times \Pr(m_5|m_1, m_2, m_3, m_4) \times \Pr(m_6|m_1, m_2, m_3, m_4, m_5)
\end{aligned}$$

Probability of winning the jackpot

Note something interesting: matching numbers gets harder as we go, because our set of possible matches is shrinking.

To win the jackpot, we must match all six numbers,
so we need the joint probability of the above events.

We use the general rule for calculating the joint probability of multiple events:

$$\begin{aligned}
\Pr(\text{Jackpot}) &= \Pr(m_1 \cap m_2 \cap m_3 \cap m_4 \cap m_5 \cap m_6) \\
&= \Pr(m_1) \times \Pr(m_2|m_1) \times \Pr(m_3|m_1, m_2) \times \Pr(m_4|m_1, m_2, m_3) \\
&\qquad \times \Pr(m_5|m_1, m_2, m_3, m_4) \times \Pr(m_6|m_1, m_2, m_3, m_4, m_5) \\
&= \frac{6}{49} \times \frac{5}{48} \times \frac{4}{47} \times \frac{3}{46} \times \frac{2}{45} \times \frac{1}{44}
\end{aligned}$$

Probability of winning the jackpot

Note something interesting: matching numbers gets harder as we go, because our set of possible matches is shrinking.

To win the jackpot, we must match all six numbers,
so we need the joint probability of the above events.

We use the general rule for calculating the joint probability of multiple events:

$$\begin{aligned}
\Pr(\text{Jackpot}) &= \Pr(m_1 \cap m_2 \cap m_3 \cap m_4 \cap m_5 \cap m_6) \\
&= \Pr(m_1) \times \Pr(m_2|m_1) \times \Pr(m_3|m_1, m_2) \times \Pr(m_4|m_1, m_2, m_3) \\
&\qquad \times \Pr(m_5|m_1, m_2, m_3, m_4) \times \Pr(m_6|m_1, m_2, m_3, m_4, m_5) \\
&= \frac{6}{49} \times \frac{5}{48} \times \frac{4}{47} \times \frac{3}{46} \times \frac{2}{45} \times \frac{1}{44} \\
&= \frac{1}{13,983,816} = 0.0000000715
\end{aligned}$$

What is the expected return from a single ticket?

Suppose a lottery ticket costs one dollar

How much, on average, you expect to get back from the dollar you spend on a lottery ticket is the *expected value* of the ticket.

Let's denote the specific set of six number we chose as ticket_i, and the amount of money in the jackpot as J.

$$\mathrm{E}(\mathrm{ticket}_i) = J \times \Pr(\mathrm{ticket}_i \text{ wins})$$

Is the lottery a good investment?

Suppose the jackpot is 1 million dollars. Then

$$\mathrm{E}(\mathrm{ticket}_i) = 1,000,000 \times 0.0000000715 = 0.07$$

If you played the lottery many millions of times at a $1 million jackpot, you would expect to get back an average of 7 cents for every dollar in tickets purchased

To break even, the jackpot would have to be about $14 million each time

But even then, over any short run of lotteries, you would expect nothing.

The lottery is high risk, with most of the expected return coming from a low probability event

What is the expected return from a single ticket? (redux)

But wait! There are other players, and nothing prevents them from picking the same six numbers!

If we pick the winning numbers, but so do q other people, we will have to split the jackpot $q+1$ ways!

And the bigger the jackpot, the more people play...

So really,

$$\begin{aligned}\mathrm{E}(\text{ticket}_i) \quad = \quad J \times & \frac{\Pr(\text{ticket}_i \text{ wins})}{[1 \times \Pr(\text{no one else chooses ticket}_i)} \\ & +2 \times \Pr(\text{one other player chooses ticket}_i) \\ & \quad +3 \times \Pr(\text{two other players choose ticket}_i) \ldots]\end{aligned}$$

Split jackpots shrink our winnings a lot!

A strategy for the lottery

All lottery numbers are equally likely to appear.

No six numbers are more likely to win than any other.

So is there any strategy we can use to maximize our winnings?

A strategy for the lottery

All lottery numbers are equally likely to appear.

No six numbers are more likely to win than any other.

So is there any strategy we can use to maximize our winnings?

All numbers are equally likely to win,
but not all numbers are equally likely to split the jackpot

If we pick numbers no one else plays,
we win just as often as ever, but never have to split our winnings

A strategy for the lottery

Most people misunderstand the concept of a random number sequence.

The sequence

$$9, 15, 17, 20, 35, 37$$

is just as likely to appear as

$$1, 2, 3, 4, 5, 6$$

or

$$44, 45, 46, 47, 48, 49$$

But most lottery players would see these sequences as “unlikely”

(How would you *minimize* returns from the lottery?

A strategy for the lottery

Most people misunderstand the concept of a random number sequence.

The sequence

$$9, 15, 17, 20, 35, 37$$

is just as likely to appear as

$$1, 2, 3, 4, 5, 6$$

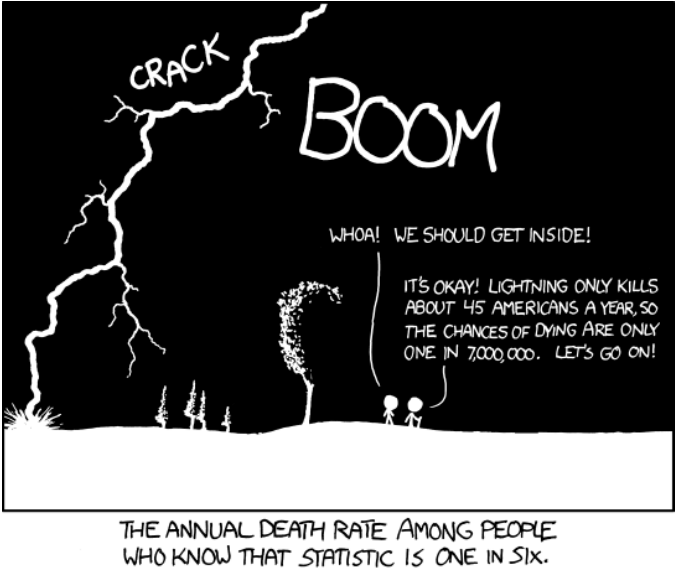
or

$$44, 45, 46, 47, 48, 49$$

But most lottery players would see these sequences as “unlikely”

(How would you *minimize* returns from the lottery?
Play numbers appearing in recent dates!)

Is the lottery worth playing?



http://xkcd.com/795/