Probability

Why do we need to look at probability? Because probability is fundamental to statistics. Okay, so that doesn't explain things, but it is true.

We use statistics to figure out how "probable" something is. Let's suppose we observe the following data:

	Placebo	Medicine
	156	147
	186	163
	145	162
	168	143
	149	147
$\bar{y}$	164.4	152.4

What is the probability that the observed change in blood pressure is due to chance? After all, we could have randomly picked 5 people who just happened to have a lower average blood pressure - the medicine could be irrelevant (not doing anything).

In other words, if we observe something (an event of some kind), we then figure out what the probability of that event. In this case, what's the probability that the medicine works (more accurately, we figure out the probability that the medicine does *not* work).

Another way of looking at this is to say, what is the probability of getting the above result completely due to chance? If this probability is really low, we say the medicine must be the cause. If the probability is high, then the medicine (probably) isn't working.

Obviously, we need to understand something about probability before we can answer this question.

Let's try another, non-biological, example:

We flip a coin 10 times. There are 11 possible results:

Number of heads	Number of tails
0	10
1	9
2	8
3	7
4	6
5	5
6	4
7	3
8	2
9	1
10	0

For which of these results would you think that the coin is unfair?

Most likely if you got one of the results near the top or bottom of these columns you'd think the coin is unfair.

Not convinced? Suppose I give you a dollar for every head, and you give me a dollar for every tail, and we flip the coin 10 times (my coin!) and we get 0 heads and 10 tails. Do you really think I'm being honest??

Although it's not a biological example, when we can figure out the probability of getting 8 tails and 2 heads, we will be done with basic probability.

So, to summarize, we need probability to answer questions about our data: what is the probability our results are due to chance?

Let's do some very basic probability. We won't learn a lot of probability here (probability is it's own mathematical discipline!), but we'll learn enough so we can understand what some of the statistical procedures we will learn are doing.

So let's define probability:

Probability is a number that describes the chance of an event happening. We note probability by:

> $Pr\{E\}$, or sometimes just $P\{E\}$.
>
> This means *the probability of the event E.*

So what is E? Here are some examples:

E_1: Heads, when tossing a coin.

E_2: the # 6, when rolling a dice.

E_3: the # 7, when rolling two dice.

E_4: the Orioles winning the World Series.

Let's look at the answers, and then we'll explain how we got them:

$Pr\{E_1\} = 1/2$

$Pr\{E_2\} = 1/6$

$Pr\{E_3\} = 1/6$

$Pr\{E_4\} = ??$

Before we go on, we need to make sure of a few things. Notice that any probability must always be between 0 and 1 (inclusive). More formally we say:

$$0 \leq Pr\{E\} \leq 1$$

If you ever calculate a probability outside this range, *you have made a mistake!!*

Let's finally define probability. We'll use a somewhat more old fashioned definition of probability, but it's still valid and is easier to understand:

$$Pr\{E\} = \frac{\textit{the number of ways the event } (E) \textit{ can occur}}{\textit{the number of possible outcomes}}$$

So let's apply this to the probabilities we calculated above:

$Pr\{E_1\} = 1/2$

Because there is only one way of getting a head, but two possible outcomes.

$Pr\{E_2\} = 1/6$

Again, because there is only one way of getting a six, but there are six possible outcomes (the numbers 1 through 6).

$Pr\{E_3\} = 1/6$

This one is trickier. First we need to figure out how many different ways you can get a 7:

dice # 1	dice # 2	sum
1	6	7
2	5	7
3	4	7
4	3	7
5	2	7
6	1	7

In other words, there are six different ways of getting a 7.

How many possible outcomes are there? Each dice has six possibilities, so we multiply (more below). We have $6 \times 6 = 36$ different possible outcomes.

So we have $6/36 = 1/6$.

$Pr\{E_4\} =??$

This one's actually quite difficult, and we won't illustrate this. But we can make a few comments. For many years the Orioles were one of the worst teams in baseball, in which case we could assume the numerator is essentially 0, and we don't care about the denominator. In those years in which they're playing well, the numerator isn't 0, and the calculation would be rather more difficult.

A totally naive way would be to assume that every team is doing equally well. Since there are 30 teams in baseball (in 2015) we could just do 1/30 (assuming each team is doing equally well, which is absurd). This is a rather *stupid* way do do things, and there's a reason that professional sports statistics are so complicated.

Let's do another example and introduce some new/different notation:

For a single dice, what is the probability of getting a 2, 3, or 4? Let Y = our result and remember that Y is our random variable. Now we can write our event as follows:

$$E = 2 \leq Y \leq 4$$

(Note the $\leq$ symbol, which is not the same as $<$)

There are three possible ways to get what we want (2, 3, or 4). We have six possible outcomes (the numbers 1-6), so we can write our probability as follows::

$$Pr\{E\} = Pr\{2 \leq Y \leq 4\} = 3/6 = 1/2$$

We'll use this kind of notation involving Y a lot.

What if we want to combine the probabilities of different events? Again, we're just introducing the basics, so let's do an example.

Suppose you wanted to figure out the probability of getting a 6 with a dice and heads with a coin.

You can count up all the possible outcomes, and do it that way, or you can *multiply* some probabilities (but see note below about independence).

For instance:

Let Y_1 = roll with dice.

Let Y_2 = outcome with coin.

Then we have:

$Pr\{Y_1 = 6\} = 1/6$

$Pr\{Y_2 = \text{heads}\} = 1/2$

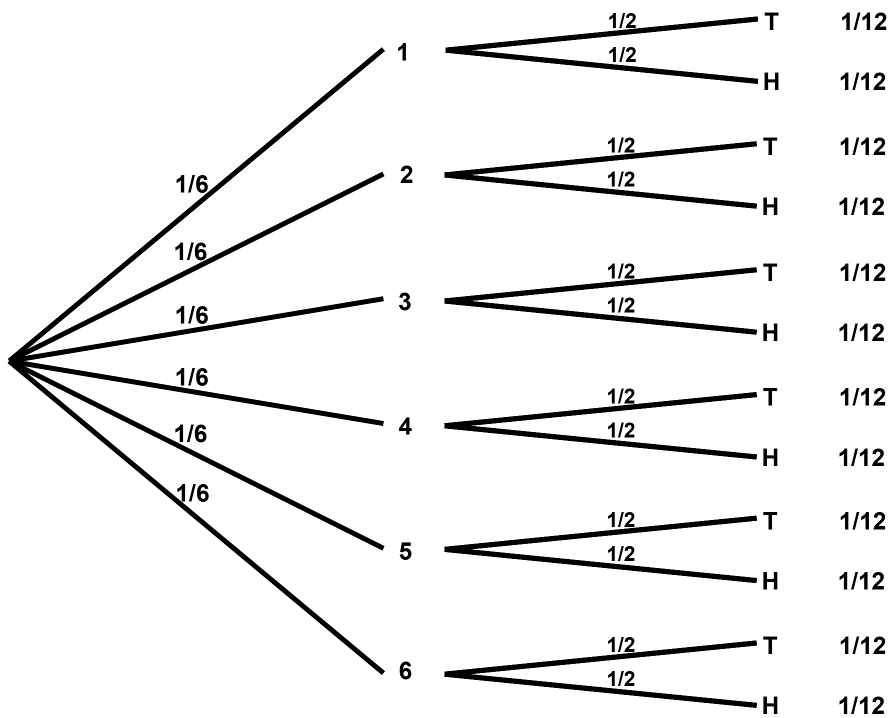
So, assuming Y_1 and Y_2 are *independent* we have:

$$Pr\{Y_1 = 6 \text{ and } Y_2 = \text{heads}\} = 1/6 \times 1/2 = 1/12$$

Let's discuss the idea of *independence* for a moment. We assumed our events (dice and coin) do not influence each other. What I roll on my dice has no effect on what I get on my coin (and vice-versa). This is important - if the events are not independent, then we can't multiply to get our probabilities. There are ways to figure this out, but we won't learn them here.

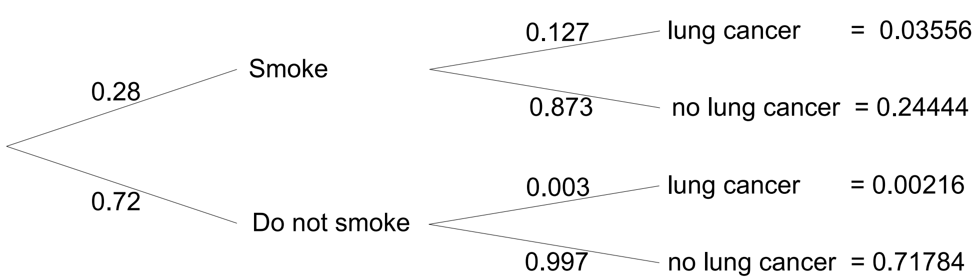
To be sure we understand independence, let's assume for a minute that we can connect the dice to the coin in such a way so that when you roll a 6, you *always* get heads. Obviously the multiplication rule is no longer valid (you should make sure you understand the probability is now 1/6, and not 1/12). The multiplication rule only works if the events are independent.

Many textbooks use "probability trees" to combine probabilities, which are very convenient when you don't have a lot of outcomes, but can get out of hand very fast otherwise (as in our example):



But sometimes probability trees can be really useful for understanding what's happening. Let's use them to figure out the overall probability of lung cancer. This is obviously different in smokers versus nonsmokers.

In Europe, (the U.S. figures are a pain to find), about 28% of people smoke. Smokers have a 12.7% chance of getting lung cancer, while non-smokers have a 0.3% chance of getting lung cancer. So what's the overall percentage of people who get lung cancer (smokers and non smokers)? We can use a probability tree to find out:



Let's talk about conditional probability for a bit. In conditional probability what we're interested in is "what is the probability that Event A happens GIVEN that Event B has happened".

For example:

$$Pr\{rolling\ an\ 8\ with\ 2\ dice\ GIVEN\ that\ the\ first\ dice\ shows\ a\ 3\}$$

We usually write this as:

$$Pr\{rolling\ an\ 8\ with\ 2\ dice \mid that\ the\ first\ dice\ shows\ a\ 3\}$$

where the vertical bar (|) means *given*.

So lets solve the above problem:

> One die shows a 3
>
> This implies the other die MUST come up 5.
>
> So what is $Pr\{\text{rolling a 5}\}$? Easy, just 1/6.

But notice the following (one way of looking at independence):

> $Pr\{\text{rolling an 8 with 2 dice }\} = 5/36$
>
> (You can do the math yourself, just add up all the possible ways of getting an 8 with 2 dice and divide by 36).

This is *not* the same answer we got above, which incidentally tells us that the events "rolling an 8 with 2 dice" and "the first dice shows a 3" are not independent.

A good way of thinking about conditional probability is to *redefine* your universe:

> For example, in the above dice rolling experiment, we are no longer interested in all possible outcomes with two dice, *just those* in which the first dice you've rolled shows a 3.
>
> So you only look at the results where the first dice showed a 3, and ignore all others example, you're not interested in *anything* if the first die is a 2.

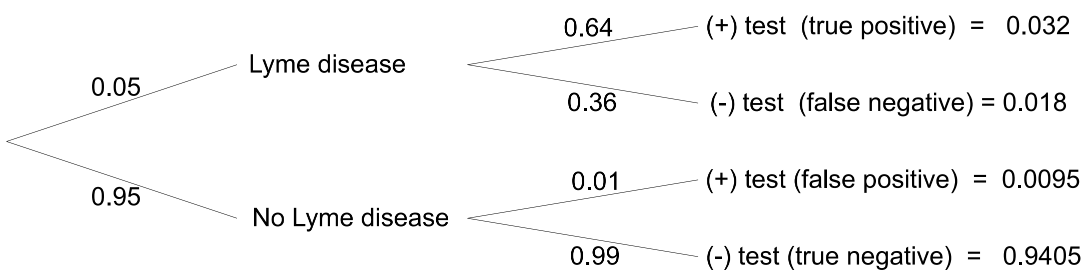
We should mention that we're being very informal here. There is a much more formal (and rigorous) approach to conditional probability. If you're really interested, a good place to look would be Wikipedia.

Let's do an example involving medical testing and Lyme disease. The screening test for Lyme disease is pretty atrocious and only catches 64% of patients who actually have Lyme disease. On the other hand, if someone doesn't have Lyme disease, then the test is accurate 99% of the time.

Before we can analyze this, we need to know how many people with symptoms similar to Lyme disease actually have Lyme disease. It turns out those figures are hard to find on a simple web search. So let's *unrealistically(?)* assume about 5% of the people who have similar symptoms actually have Lyme disease. Then we can then answer the following question:

> Suppose you go to the doctor and the screening test is positive. What is the probability of actually having Lyme disease?

Let's see if we can figure it out. First let's put our probability tree together:



Now what we want is the following:

$$Pr\{disease|(+)test\}$$

To do this, we need to look at all positive outcomes (that's our new "universe"). All positive outcomes are given on the probability tree above:

$$Pr\{(+)test\} = 0.032 + .0095 = .0415$$

And out of this, we're interested in true positives:

$$Pr\{disease|(+)test\} = \frac{0.032}{.0415} = 0.77$$

Which is pretty awful, when you think about it.

On the other hand, if the test comes back negative, what's the probability you don't have the disease? In other words, this time we want:

$$Pr\{no\ disease|(-)test\}$$

Now we need only at negative outcomes:

$$Pr\{(-)test\} = 0.018 + .9405 = .9585$$

And this time we're interested in true negatives:

$$Pr\{no\ disease|(-)test\} = \frac{0.9405}{.9585} = 0.98$$

That's much better. So if the test comes back negative, you have a 98% chance of not having Lyme disease.

So let's finally get back to coins and figure out the probability of getting 8 tails in 10 tosses. Let's try a totally naive approach and use the definition of probability.

Let's figure out the number of possible outcomes:

All possible ways of getting 0 tails:

HHHHHHHHHH

All possible ways of getting 1 tail:

THHHHHHHHH

HTHHHHHHHH

HHTHHHHHHH

etc.

All possilbe ways of getting 2 tails:

TTHHHHHHHH

THTHHHHHHH

THHTHHHHHH

etc.

Then you need to list all possible ways of getting 3 tails, then 4 tails, and so forth. You could do it this way, but you'd be at it for a *very* long time. Obviously we need to do something else.

Let's introduce the *binomial coefficient*. In a situation like this where we have a bunch of trials (tosses), we can figure out how many different possible ways we can get what we want by using something called the binomial coefficient:

$$\binom{n}{y} = \ {}_nC_y = \frac{n!}{y!(n-y)!}$$

Some texts (and most calculators) will use ${}_nC_j$, but just about everyone uses the large () symbols to indicate the binomial coefficient. The "!"symbol means *factorial*. This is defined as follows (for any positive integer, x):

$$x! = x(x-1)(x-2)...(2)(1)$$

For example, $3! = 3 \times 2 \times 1 = 6$ and $5! = 5 \times 4 \times 3 \times 2 \times 1$. This gets large *fast*; try 10! or 20!.

Although it may seem strange, it is also true that $0! = 1$. It doesn't make much sense but it does work. You'll just have to take that as a given for now.

So how does that help us? Well if we want to know how many different ways we can get 8 tails in 10 tosses, we can now plug this information into the binomial coefficient:

$n =$ the number of trials.

$y =$ the number of *successes* we're interested in.

If we have 10 tosses, that means $n = 10$, and if we define tails as a success, we have $y = 8$. Let's plug this into our formula:

$$\binom{10}{8} = \frac{10!}{8!(10-8)!} = 45$$

Don't just plug this in and use brute force to get an answer - often you can cancel a lot of terms and even do this by hand (without a calculator) after you've simplified it. In any case, the formula tells us that there are 45 different ways of getting 8 tails if we toss a coin 10 times.

We could now go through and calculate the number of different ways of getting 0 tails, 1 tail, 2 tails, etc., add these up and put this in our denominator. We would then have 45/(our calculated denominator).

This would work, and it's already much easier, but it is still a bit tedious, so let's multiply some probabilities instead:

The probability of getting a tail = 0.5, so if we want 8 tails, we can multiply (each coin toss is independent):

$$0.5 \times 0.5 \times 0.5 \times 0.5 \times 0.5 \times 0.5 \times 0.5 \times 0.5 = 0.5^8 = 0.00390625$$

But if we have 10 trials, that means we have 2 heads, so we multiply that in as well:

$$0.5 \times 0.5 = 0.5^2 = 0.25$$

Now if we multiply everything together, we get:

$$0.00390625 \times 0.25 = 0.0009765625$$

What we have now is the probability of *one way* of getting 8 heads and 2 tails, *but* there are 45 ways of getting 8 heads and 2 tails, so next we have to multiply this by 45:

$$45 \times 0.0009765625 = 0.04395$$

And finally we know that the probability of getting 8 heads in 10 tosses is 0.0439.

You might wonder why we couldn't just do 0.5^{10} above. That's because the probabilities might not always be 0.5 (see below for an example). This is actually an application of the *binomial distribution* formula, which is given as follows:

$$\binom{n}{y} p^y (1-p)^{n-y}$$

Where:

p = probability of success (in one trial).

n = number of trials.

y = number of successes.

In the coin example we're working on we have:

p = probability of success = 0.5

n = number of trials = 10

y = number of successes = 8

So we have, (much easier now):

$$Pr\{8 \textit{ tails in } 10 \textit{ tosses}\} = \binom{10}{8} 0.5^8 (1-0.5)^{10-8} = 0.04395$$

Let's finish up our survey of probability by using this formula to calculate the the probability of getting 3 people with red hair in a sample of 9. We'll use the U.K. for our example; 10% of people in the U.K. have red hair. So let's first figure out all our variables:

p = probability of success = probability of red hair = 0.1

n = number of trials = sample size = 9

y = number of successes = number of people with red hair = 3

Let's plug all this into our formula:

$$Pr\{3 \; people \; with \; red \; hair \; in \; a \; sample \; of \; 9\} = \binom{9}{3} 0.1^3 (1 - 0.1)^{9-3} = 0.0446$$

We'll see many more applications of the binomial distribution formula before we're done.