What are statistics?

The term was popularised in United States by Mark Twain (among others), who attributed it to the British Prime Minister Benjamin Disraeli: "There are three kinds of **lies: lies, damned lies, and statistics.**"

History

I'm not going into the history, unusually, as there is too little time. However, if you are interested, I would recommend the following people to look into.

- Thomas Bayes: Wikipedia page:
- Francis Galton: Wikipedia page:
- Karl Pearson: Wikipedia page:

Two types of statistics

- **Descriptive** (dealing with populations)
- **Probabilistic** (dealing with samples of or individuals within populations)

Descriptive

Categorical

use graphs to show descriptive statistics

Numeric

graphs to show the difference of presenting numerical data how categorising numeric data is done and how it can be misleading

Relationships

Numeric

two variable scatter graphs. linear, non-linear and unrelated
importance of scales

Categorical

categorised bar charts

Probabilistic

Either numeric or categorical

Can present data as before, but not very clear. What confidence do we have that we are right

(use scatter graph)

Need to measure the data to determine patterns

Ask question of it, but answers to questions only can be probabilistically correct.

Descriptive

First, measurements of the "shape" of the data.

measures of centre

Mean = Average

Median = Middle value

Mode = Most frequent number

Range = Max - Min

EXAMPLE:

8, 9, 10, 10, 10, 11, 11, 11, 12, 13

Mean = 105 / 10 = 10.5

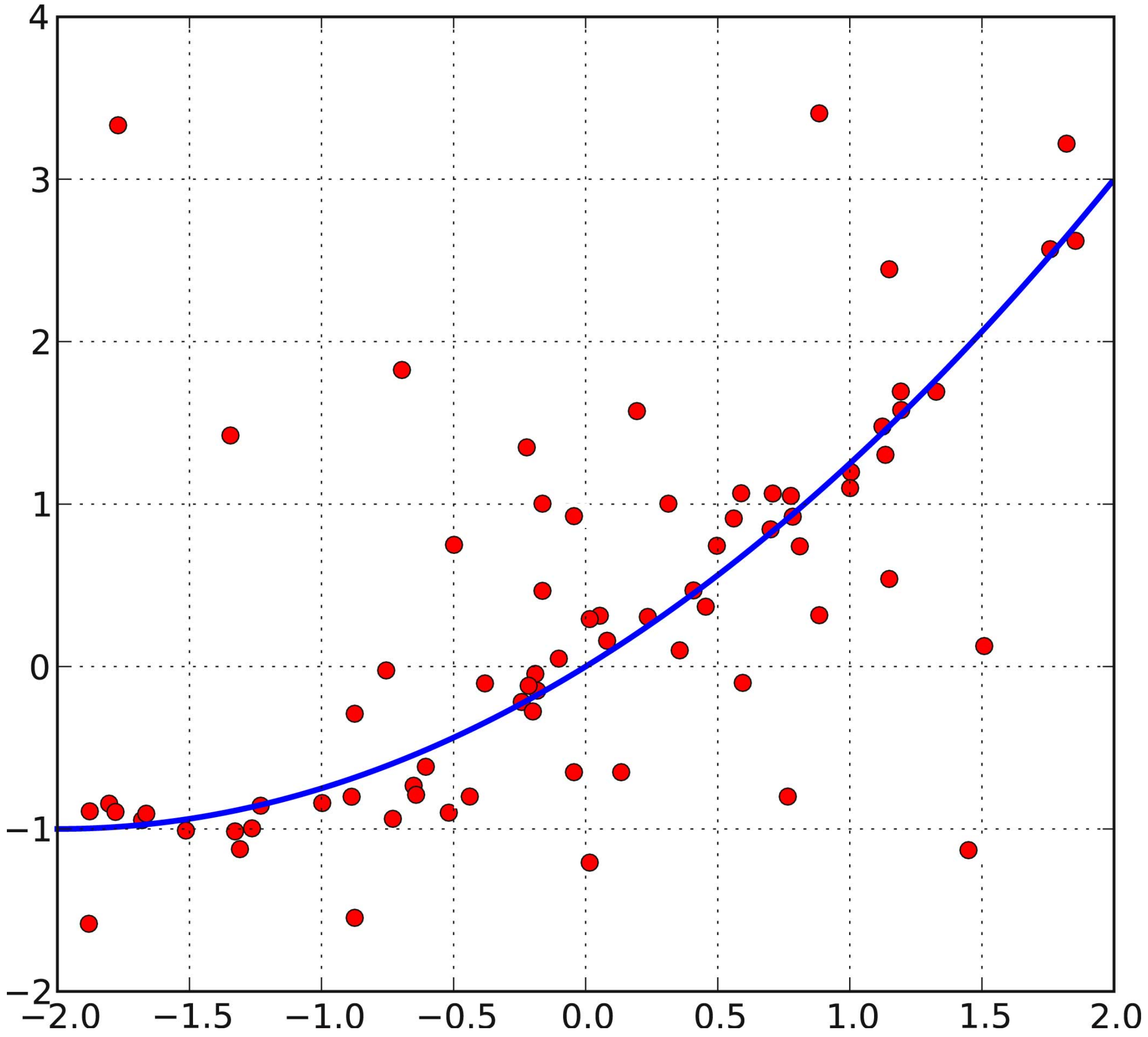
Median = 5.5th value: (10+11)/2 = 10.5

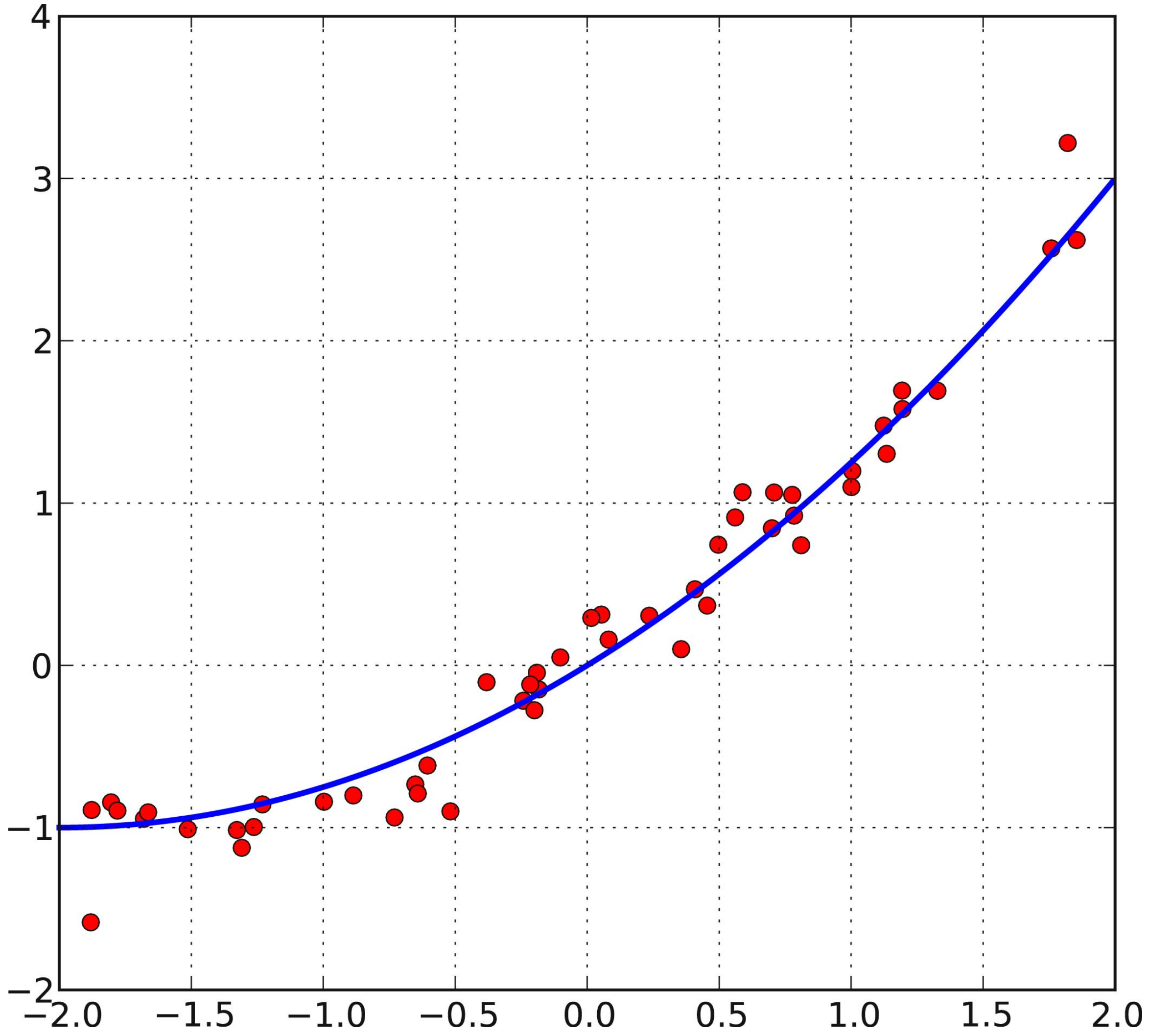
Mode = 10 and 11

Range = 13-8 = 5

Measures of centre give clues as to both the size of distribution and shape. Give clues to skewed data.

Measures of variation (how spread out the data is)





More spread out data has more "variation" than less spread out data.
More variation = less confidence in our conclusions of the meaning of the data.

Intro to Statistics on YouTube

Rudimentary measures of variation

Range (very crude)

Min and Max

Quartiles (3 quartiles)

2nd quartile = median, 1st quartile = median of lower half, 3rd quartile = median of upper half

Min, Max, and 3 quartiles = 5 number summary (box plot)

3rd Q – 1st Q = inter-quartile range

Common measures of variation

Variance Measures the average of the distance from each data point to the mean.

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

Why divide by n-1? Don't worry about it, it works better this way. Standard Deviation (s) = Square root of the Variance (s2)

$$s = \sqrt{s^2} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Measures of relationship

Relationships between 2 variables (scatterplot)

measure **strength** and **nature**.

Strength of a relation between two variables is measure by **correlation**:

$$r = \frac{\sum \frac{(x_i - \bar{x})}{s_x} \frac{(y_i - \bar{y})}{s_y}}{n - 1}$$

Don't worry about this math, you don't need to calculate it. Software does this for you.

You do, however, need to be able to interpret it.

Correlation is always between $-1 \le r \le 1$

positive ($r \le 1$) means a positive relationship

[show scatter plot] increasing linear relationship

negative ($r \le 0$) means a negative relationship

[show scatter plot] decreasing linear relationship

the higher or lower the value, the **stronger** the relationship is to the linear.

Correlation close to 0 shows a weak relationship.

Probability

- A notion of probability
- Formalising probability
- Calculating probability
- Conditional probability
- Bayes' Theorm

A notion of probability

With descriptives, we have been dealing with certainty, with describing a whole dataset. We have been measuring whole populations.

This is rarely the case. Mostly we are dealing with samples and we need to know, from the sample, what is the likelihood of a single event occurring.

Examples

To complicate matters there are two approaches to probability:

relative frequency and **a priori classical approach**

Relative Frequency:

1st, just because something is random does not mean we cannot study it and find patterns in it. Randomness doesn't mean there is no pattern.

E.G. dice. If you roll a die, what face that comes up is random, but if you roll a die thousands of times, there will be a pattern. For example, the 4 will occur, on average, once every 6 throws.

The probability of an event is the *proportion** of times it occurs.*

This is the relative frequency approach.

Limitations of the RF approach:

RF is empirical, but, in theory, we would like to observe the events an infinite number of times. Impractical, impossible!

RF doesn't tell us why these proportions are they way they are.

Example. Students from one course get a higher proportion of good grades than the students from a second course. They proportions say something is different, but not what the reasons are.

Formalising probability

Formalising our notions of outcomes and events. When we talk about probability, we are talking about the probability of an outcome or an event occurring.

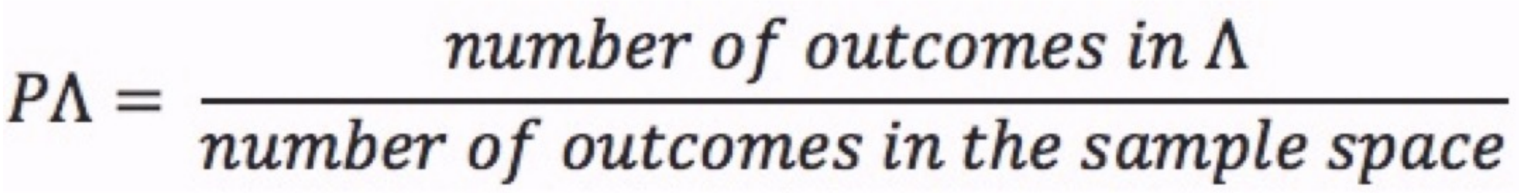
First, some definitions.

For all observable phenomena in the world, there are any number of possible recordable observations. All these recordable observations will have a number of possible **outcomes** that can be recorded. E.g. a coin toss has 2 possible outcomes, heads or tails. The throw of a die, 6.

When you have defined all of the possible outcomes for a variable, these possible outcomes is called the **sample space**.

An **Event** in that sample space is any combination of outcomes, e.g. the roll of an even number on a die (which would be a combination of three possible outcomes).

a priori classical probability:



$$P\Lambda = \frac{number\ of\ outcomes\ in\ \Lambda}{number\ of\ outcomes\ in\ the\ sample\ space}$$

PɅ= (number of outcomes in Ʌ)/(number of outcomes in the sample space)

The total number of outcomes in an event (Ʌ) divided by the total number of outcomes in the entire sample space.

Calculating probabilities

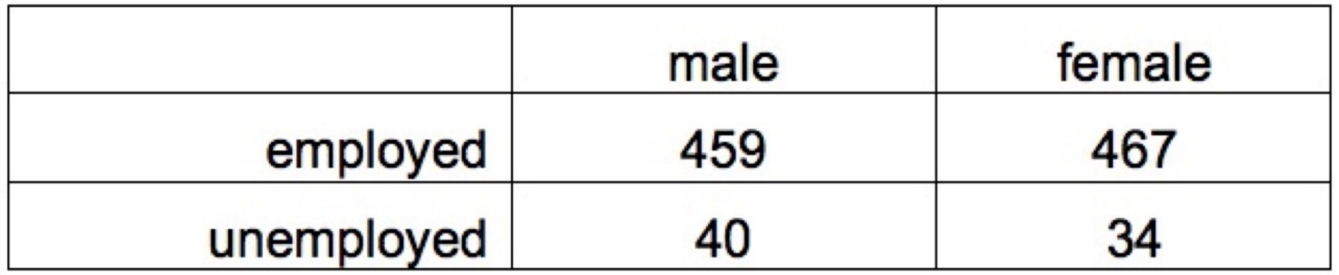
How do we calculate the probability of a given event?

One way is with what we call a **contingency table**.

To calculate the probability of something occurring, you need to know how many outcomes there are of that event. You also need to know what the size of the sample space is.

Example:

You survey 1,000 people recording if they are male or female, and asking them if they are employed. And you get these results:



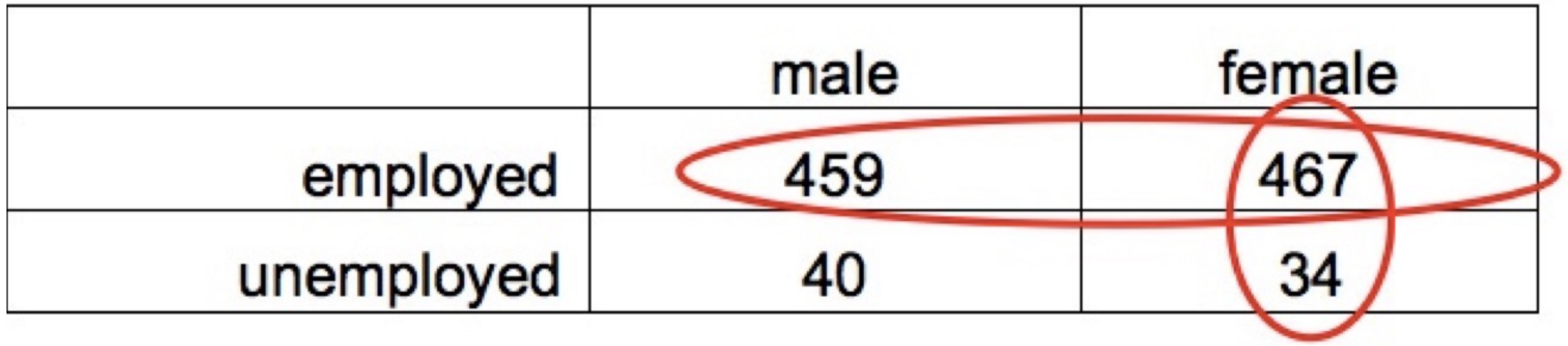
	male	female
employed	459	467
unemployed	40	34

So, if 459 people were male and employed, what is the probability of someone being male and employed?

$$Probability = \frac{459}{1000} = .459$$

But you can extend your understanding of the situation by adding up rows or columns. For instance, adding male and female employed people you get 926 employed people. So the probability of being employed is .926.

BUT, what if you want to know the probability of someone being female and employed? You can't just add up all the females and all the employed people, as you would be counting employed-females twice.



	male	female
employed	459	467
unemployed	40	34

You can't do that so you need to subtract this from the calculation. This is what is called the General Addition Rule.

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Conditional probability

Conditional probability is simply the probability of an event occurring given another event occurring. E.g. what is the probability of someone having children if they are also married.

$$P(A\,|B) = \frac{P(A \text{ and } B)}{P(B)}$$

The probability of A given B, or the probability of A ba B

$$P(A\,|B) = \frac{\textit{number of outcomes where } A \textit{ and } B \textit{ occur}}{\textit{number of outcomes where only } B \textit{ occurs}}$$

Bayes' Theorem

Many, many, many events in the Humanities are conditional. One form of conditional probability that is used a lot in Digital Humanities is Bayes' Theorem, or Bayesian probability.

Bayes Theorem helps when you already know the probability of one event vs. a second event, but you really want to know the probability of the second event given the first.

You have P(B|A), but you want P(A|B)

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B|A) \times P(A) + \; P(B|A^c) \times P(A^c)}$$

Example:

In law firm, 64% of lawyers are male, 36% female. Of these, 44% of male lawyers are corporate lawyers, while 24% of female lawyers are corporate lawyers. If you choose a corporate lawyer at random from the law firm, what is the probability that they will be male?

A = event that lawyer is male

B = event that lawyer is corporate

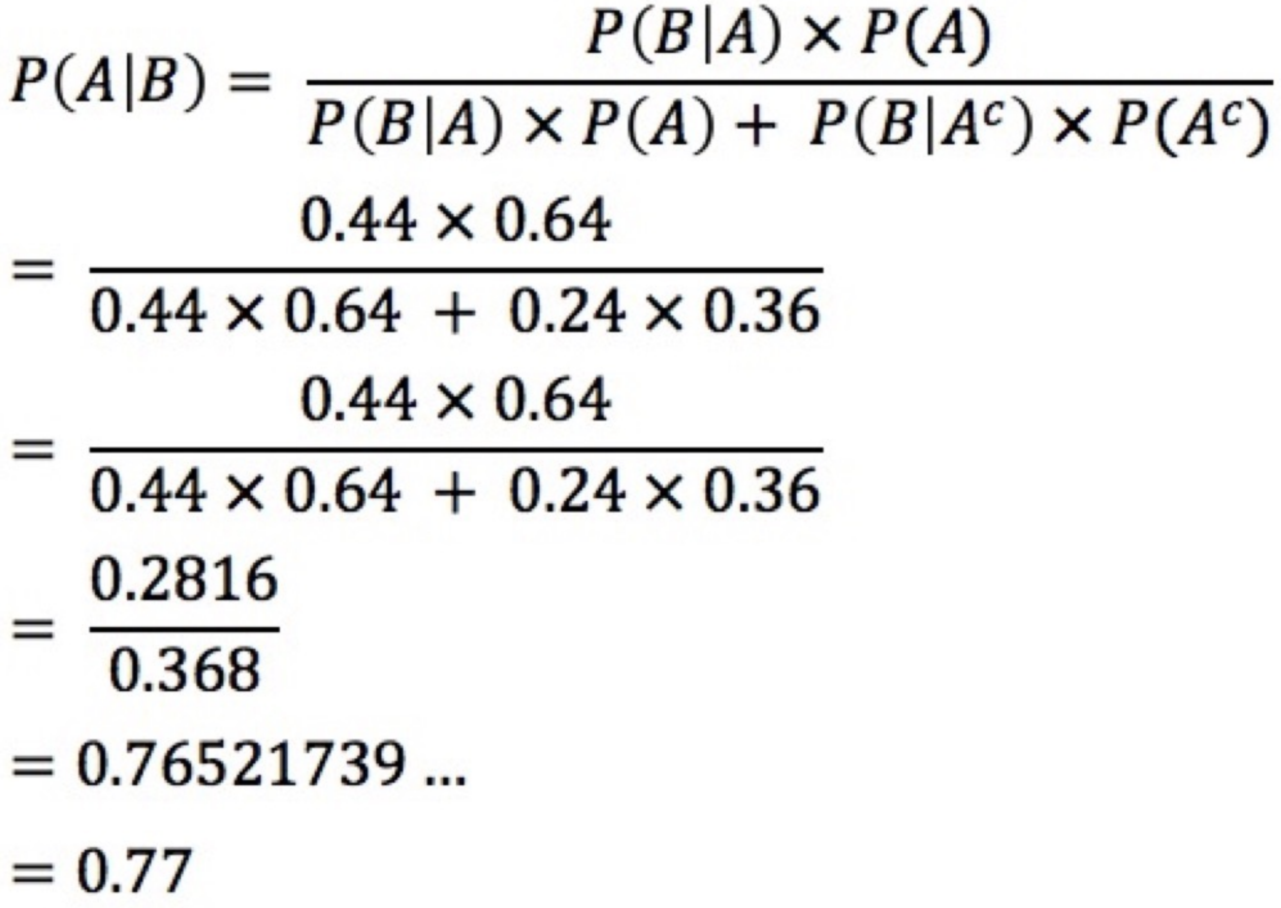
P(A) = probability that lawyer is male = 0.64

P(Ac) = probability that lawyer is female = 0.36

P(B|A) = probability that a lawyer is a male corporate lawyer = 0.44

P(B|Ac) = probability that a lawyer is a female corporate lawyer = 0.24

P(A|B) = probability that a lawyer is male given that they are a corporate lawyer



$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B|A) \times P(A) + P(B|A^c) \times P(A^c)}$$

$$= \frac{0.44 \times 0.64}{0.44 \times 0.64 + 0.24 \times 0.36}$$

$$= \frac{0.44 \times 0.64}{0.44 \times 0.64 + 0.24 \times 0.36}$$

$$= \frac{0.2816}{0.368}$$

$$= 0.76521739 \ldots$$

$$= 0.77$$

Probability Distributions:

- Discrete random variables
- The binomial distribution

Continuous random variables

The normal distribution

We have been describing data, data that had already been collected. We can describe or show these data in a variety of ways, but data always comes from a **variable**.

We finally arrive at **Statistical Inference**. Statistical Inference is where we go beyond describing the data, and its patterns, and use statistics say something about the world. To do this, we need to study the underlying variable.

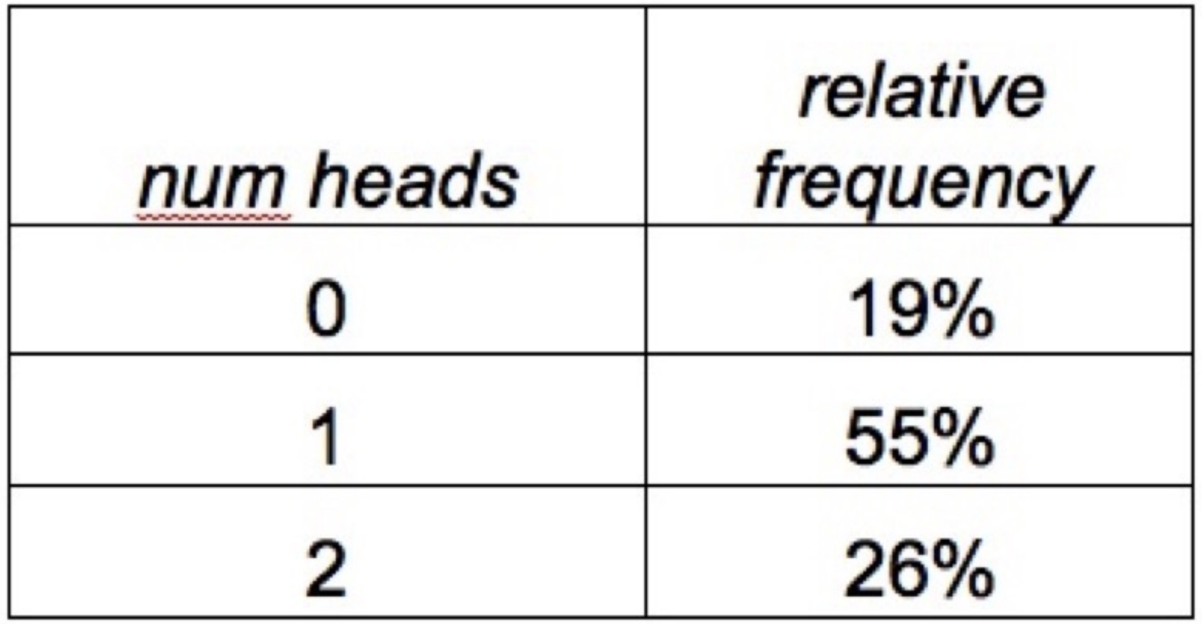
The principle of statistical inference is, rather than study data which shows us what a variable did do, we will study probability to try and infer what a variable can do.

Discrete random variables

Start by defining a random variable.
Why not a coin toss? Take a coin, toss it twice, and count the number of heads. You have four possibilities.

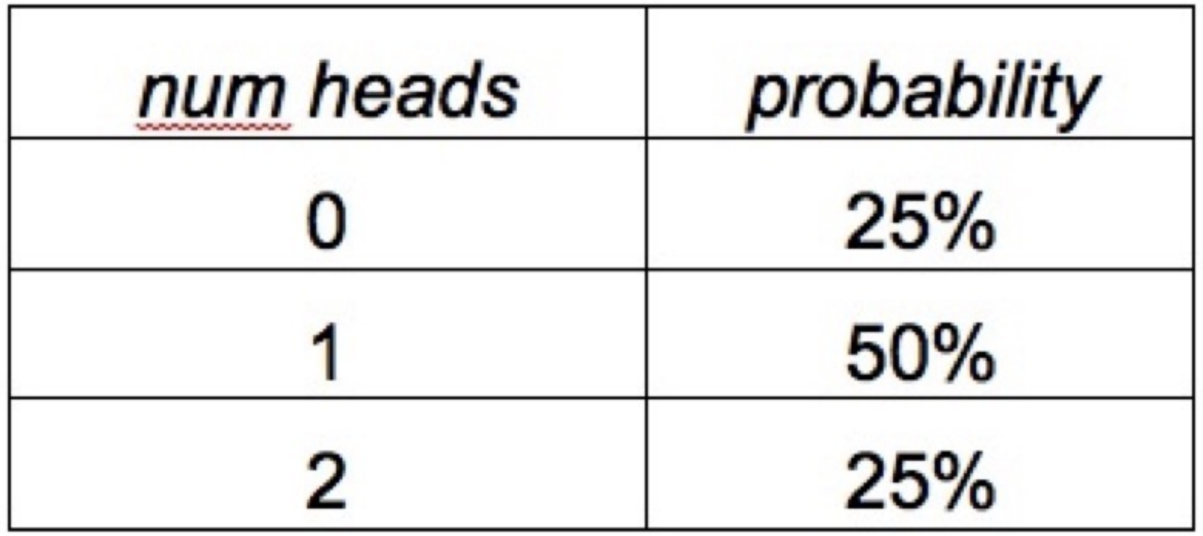
tail, tail = 0
tail, head = 1
head, tail = 1
head, head = 2

We could treat this as a data problem, do the double toss 100 times and describe the data and its expressed probabilities. Even do some graphs.



num heads	*relative frequency*
0	19%
1	55%
2	26%

But what if we don't collect any data, but just think about what variable we are looking at, and what values it could take?



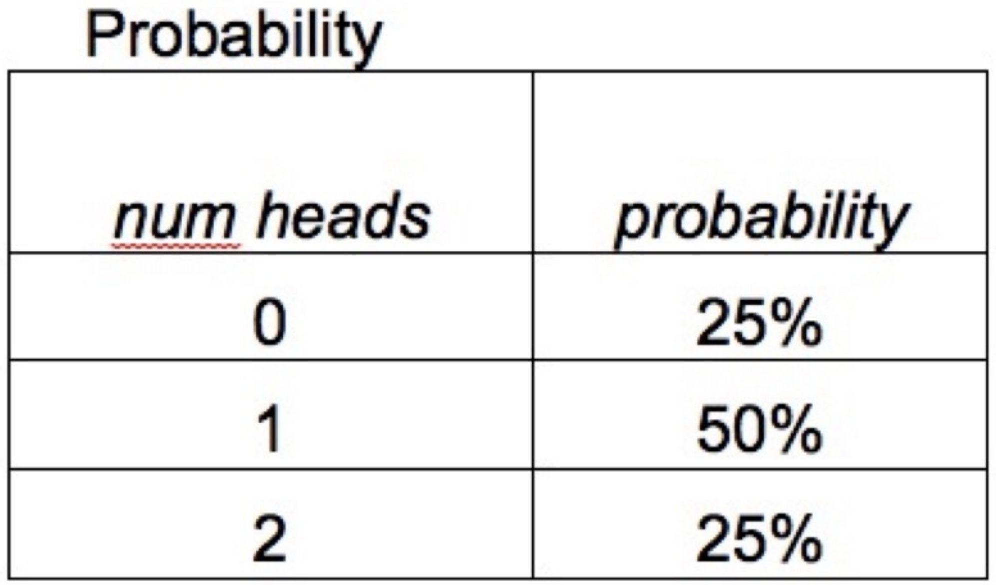
num heads	*probability*
0	25%
1	50%
2	25%

Somewhat strangely, when we talk about what a variable could do, rather than what it did do, we call that a **random variable**.

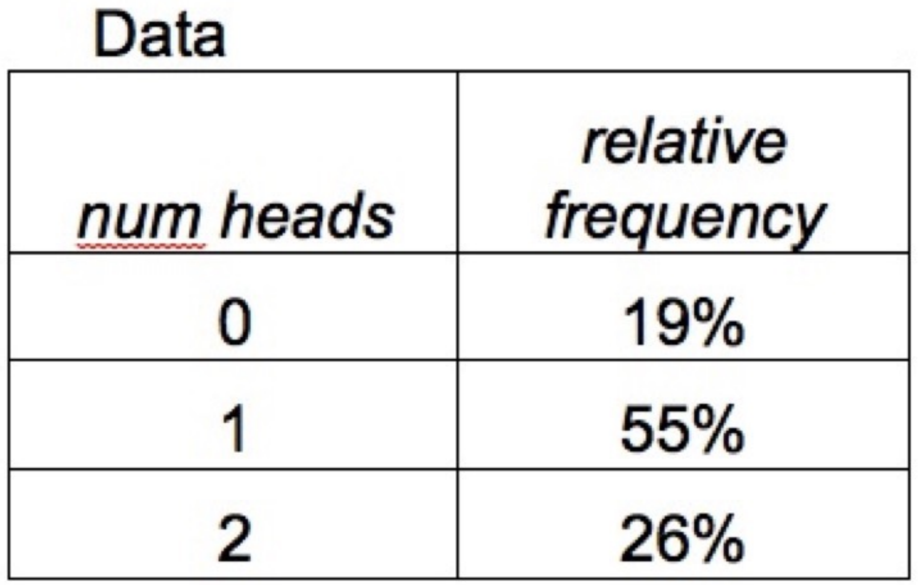
- A **random variable** X can assume a value for every outcome in a sample space.
- X is **discrete** if all outcomes can be put in a list of separate items.

What we just did with the coin toss variable was to assign a **probability distribution**. A Probability Distribution is a function p(x) that assigns a probability to each outcome.

Our probability distribution looks very like our data, but our data will vary from sample to sample. Our PD is an ideal.

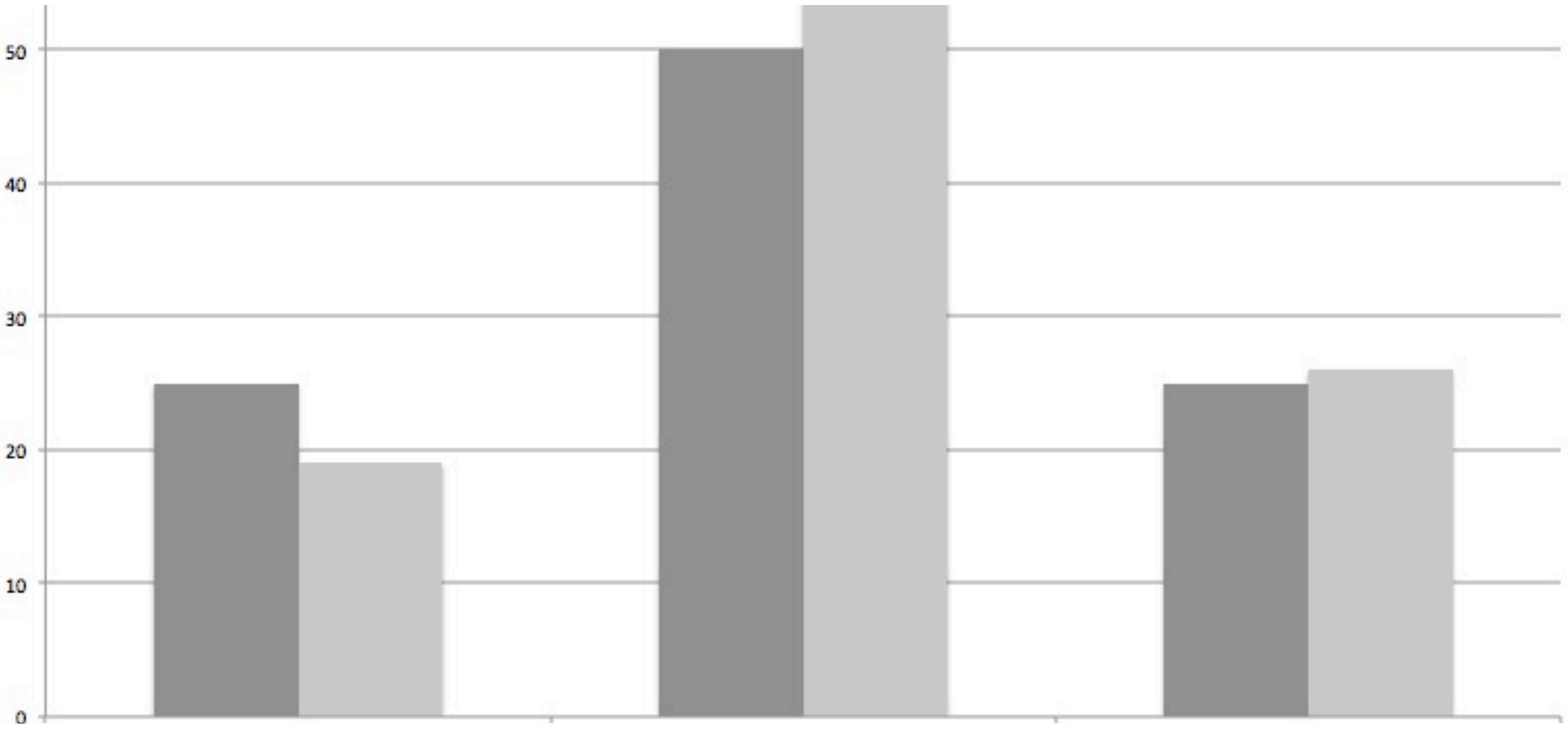

Probability

num heads	*probability*
0	25%
1	50%
2	25%


Data

num heads	*relative frequency*
0	19%
1	55%
2	26%

We can also draw a histogram which will show not what the variable did, but what, ideally, the variable can do.



Comparison of Probability (Theory) with
Relative Frequence (what actually happens)

We can also measure the distribution of probability distributions, just like we can relative frequencies. Here are some formulas.

If we have a discrete random variable X, with some outcomes {x1, x2, ... xn} and a probability distribution {p(x1), p(x2), ... p(xn)}, so the expected value of X is given by this formula:

$$E(X) = \sum_{i=1}^{n} x_i p(x_i)$$

Expected value is like the mean, actually we occasionally call it the mean of X.
The Variance of X is given by this formula:

$$VAR(X) = \sum_{i=1}^{n} (x_i - E(X))^2 p(x_i)$$

And the Standard Deviation of X by this formula:

$$SD(X) = \sqrt{VAR(X)}$$

The binomial distribution

This is the most common probability distribution.

Let's use and example:
Suppose you are sorting old beers for a party, and you are looking for beers past their sell-by date. You will discard these. You know, from past experience, that in your house, roughly 3% of beers are always past their sell-by date. So what is the probability that of the 20 beers you have to go through, two of them are too old?
X is the number of beers past their sell-by date.
Sample space (outcomes) can be anything from 0 to 20.

Binomial distribution of X is given by this formula:

$$P(X) = {}^{20}C_x(0.03)^x(0.97)^{20-x}$$

So let's say that we expect 2 defectives (x=2)

$$P(X) = {}^{20}C_2(0.03)^2(0.97)^{18}$$

$$= 0.0988 = 9.88\%$$

There are lots of binomial distributions as they all change with the values of x and p.

$$P(X) = {}^{n}C_{x}\, p^{x}(1-p)^{n-x}$$

The general binomial distribution. Only for discrete variables.

Continuous random variables

A random variable is continuous if its values are on a continuous spectrum. E.g. the time it takes to do some task, how far something goes when thrown, the temperature over time, etc.

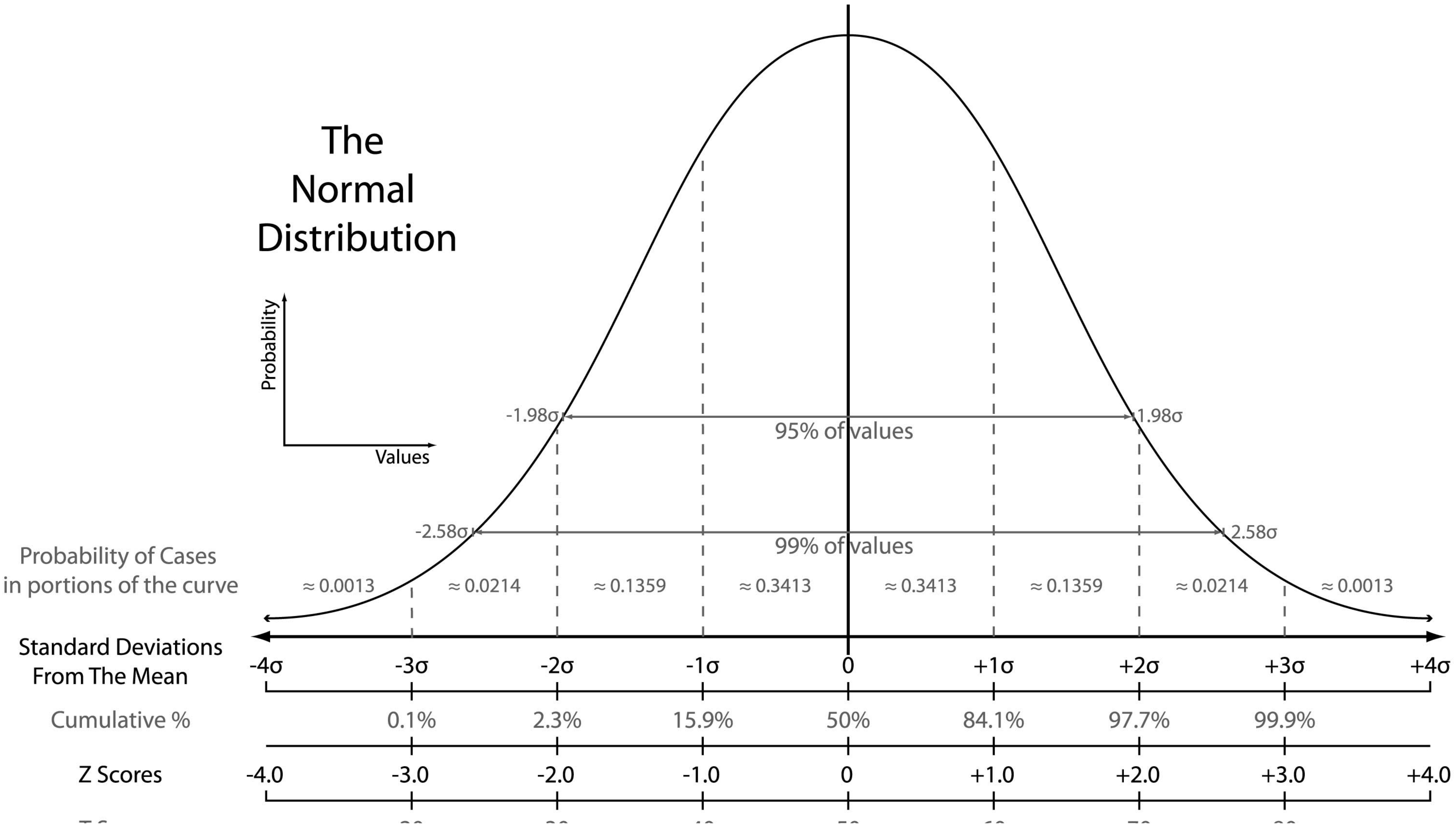
We deal with continuous random variables differently than discrete random variables, but we need not worry too much about that here.

The normal distribution

What we do need to understand is that the most common form of a CRV is what we call the **normal distribution**.

Like the binomial distribution, there are infinitely many normal distributions.

However, each **normal distribution** is defined by its mean and its standard deviation. Why? Explain with a graph.



Normal distribution X with mean μ and standard deviation σ.

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

Standard Normal Distribution Z, mean of 0 and standard deviation of 1.

We use the SND(Z) to calculate probabilities of a value. Do this by transforming, using another formula, any normal distribution into the standard normal distribution, and then looking the probability up in a table (which you can easily find online).

The SND(Z) table gives you the probability that any value (x) will be less than it in the table. You can then use just arithmetic to calculate the probability of a value greater than, or between values. Very useful.

Null hypothesis, p-value and statistical significance

There are times when we are not interested in just one variable, but two, or the relationships between two groups. This is where comparative statistics come in.

In statistics, with all questions about the relationship between two groups, we begin with what we call the **Null Hypothesis (H0)**. H0 states that there is no difference between the two groups.

E.g. Let's say we are studying social media use. We would like to know if one group on Twitter interact more than another group – two continuous variables, the counts of interactions.

IN STATISTICS, BY DEFAULT YOU ALWAYS ASSUME THE NULL HYPOTHESIS (H0) IS TRUE, THAT THERE IS NO RELATIONSHIP BETWEEN THESE TWO VARIABLES, UNTIL YOU HAVE ENOUGH EVIDENCE TO REJECT THIS HYPOTHESIS (REJECT H0).

This is extremely important. Statistics do not prove a hypothesis, they give you enough confidence that you can reject the null hypothesis (H0).

Correspondingly, there is the **Alternative Hypothesis (denoted Ha or H1)**. I prefer H1. H1 states that there is a relationship between the two variables.

Obviously, as a researcher, you want H1 to be accepted. This is why you are doing the research. To find a relationship.

Notice that I say "accepted", not "proved". **Statistics don't prove anything**. You can only fail to reject H0 or have sufficient confidence in H1. The goal of all comparative statistics is to reject H0 because that is as close as you can possibly get to proving that there is a relationship. You can **never** be completely sure.

You do not prove a relationship, but your show that, given your data, H0 is highly unlikely.

However, there is always a chance, no matter how small, that you are wrong. If you incorrectly reject H0, no matter how good your statistics were, then we call this a **Type 1 Error**, also called a **False Positive**. The Type 1 Error is when the statistics give confidence that we can reject H0, but there is no relationship.

- α is the probability of making a Type 1 Error.
- The other error you can make is to fail to reject H0 when you should have rejected it. This is called, unsurprisingly, a **Type 2 Error**. It is what we also call a **False Negative**.
- β is the probability of making a Type 2 Error.

We also speak of **Power**, 1 – β, which is the probability of finding a relationship, or also the probability of not making a Type 2 Error. **Power** is good because it is a measure of how likely it is that you will succeed in finding the relationships you are looking for. Often researchers will do some Power calculations on trial data to make sure they are going in the right direction. However, the Power of the study will increase:

if the sample size increases,
if the difference between groups is greater,
if the effect between groups is greater,
if the data is more precise
if your standard deviation is smaller

How do we know if we can reject H0 or not? We use probability. In this case we use the **p-value**. P-value is the probability of obtaining a result at least as extreme as the current one, assuming H0 is correct.

P-value is a measure of how much the data disagrees with H0.
- low p = reject H0
- high p = fail to reject H0

What determines if p is low or high? This is where the **level of significance** (our earlier α) comes in. Recall α is the probability of making a Type 1 Error, or falsely rejecting H0.

If $p < \alpha$ we conclude there is statistically significant difference between groups. There is a relationship.

However, it is important to know that α is an arbitrary cut-off point – usually 5%. This is simply an agreement among researchers. Different research may require more or less confidence, hence a lower or higher α.

So if we accept a p of 5%, then these are the following statistical inferences:

If $p < 0.05$, there is a statistically significant difference between groups.

```
Reject H0.
```

If $p > 0.05$, there is no statistically significant difference between groups.

```
Fail to reject H0.
```

Some further caveats:

Statistical significance DOES NOT equal interpretive significance.

You may have a statistically significant difference that is interpretively trivial. In our example of Twitter groups above, we may have a statistically significant difference in the number of interactions between the two groups. But if the higher numbers in one group is due to them always Tweeting "OK" at the end of every interaction, it is not interpretively significant.

A $p < 0.05$ does not mean that the relationship is random or due to sampling error.

p-value is ONLY a measure of the degree of agreement between our data and H0. It says nothing about larger interpretations of the data.

Most importantly, the statistical results of a study can only give confidence for the acceptance or rejection of H0 IF THE ASSUMPTIONS OF THE STUDY ARE VALID.

Remember, p simply is a measure of the probability of H0 being true. It can say nothing about the validity of H1. That is another matter.

We do have measures of for testing H0. You don't need to know much about them, but you do need to know which one to use when.

- t-test when comparing 2 variables
- Analysis of Variance test (ANOVA) when comparing > 3 variables
- Chi-squared when comparing categorical data

How to lie with Statistics

A vast literature:

Best remains Darrel Huff's 1954 book, How to Lie with Statistics. (pdf)

Also a good slide show from UCSD (pdf)

Resources:

Wikipedia – Statistics

MIT Open Courseware: Introduction to Probability and Statistics

MIT Open Courseware: Statistics and Visualization for Data Analysis and Inference

Udacity introductory course on statistics

Statistical methods for studying literature using R

Good DH resources on Alan Liu's UCSD site:

Udemy has a huge number of good courses online for humanities. However, you have to pay a little for them.

Statistics for the Humanities downloadable book