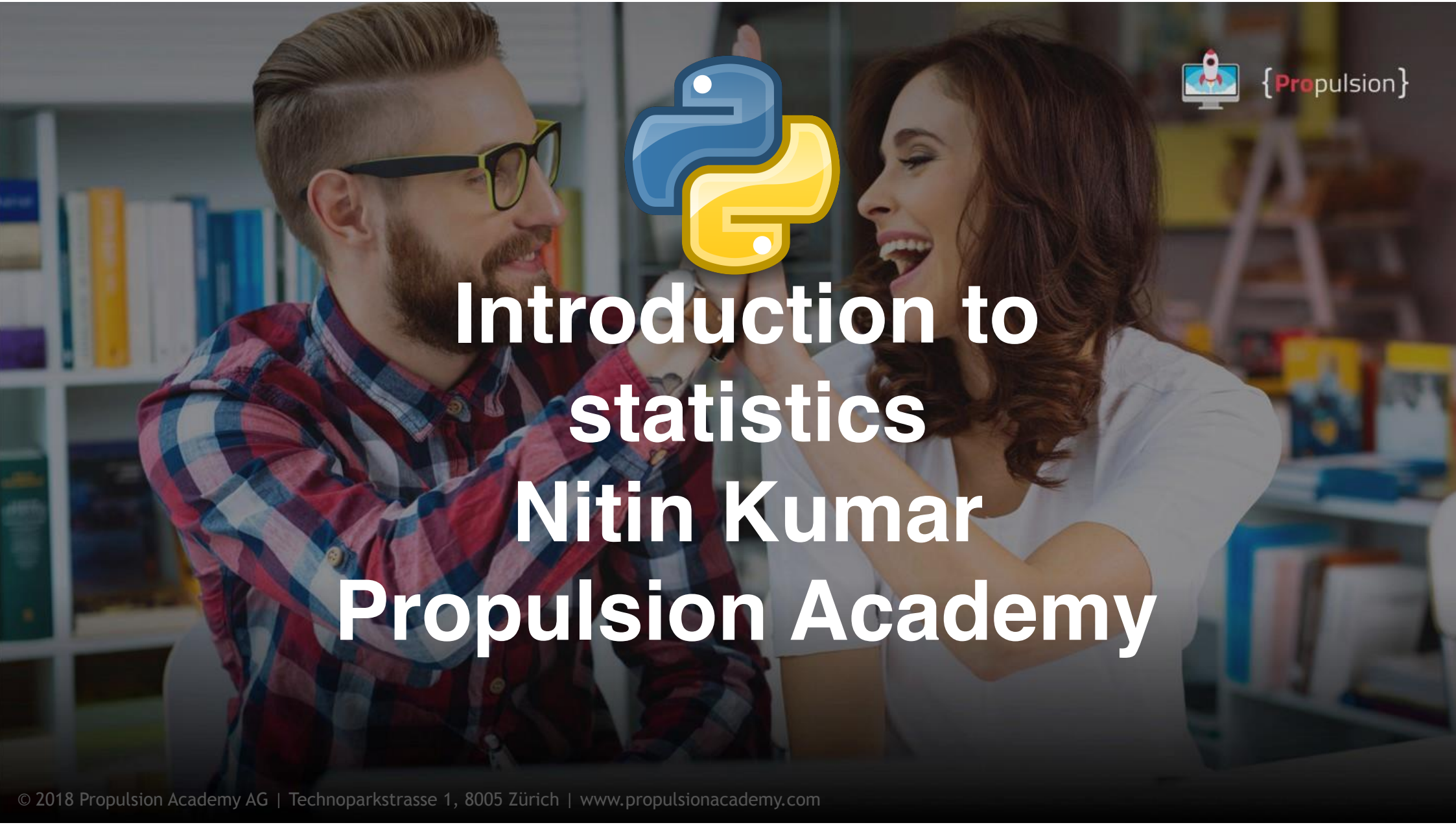


Today's plan

All about distributions:

- Central tendency - Exercise
- Distributions and parameters
- Probability Density Functions - Exercise
- Bootstrapping - Exercise
- Hypothesis testing
- t-test - Exercise
- Power analysis
- A/B testing

Central tendency

Central tendency: a single value that attempts to describe (one variable in) the data. (mean, median, mode).

Mean $\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$

$$\bar{x} = \frac{\sum x}{n}$$

Central tendency

Median
The “middle” score, such that there are an equal number of scores that are higher and scores that are lower.

14 45 45 45 55 **56** 56 65 87 89 92

If you have an even number of scores, take the mean of the middle two.

Central tendency

Mode
The most common score.

14 45 45 45 55 **56** 56 65 87 89 92

Calculating central tendency - exercise

https://vincentarelbundock.github.io/Rdatasets/datasets.html
“baseball” dataset -> download .csv and read html page for information on variables

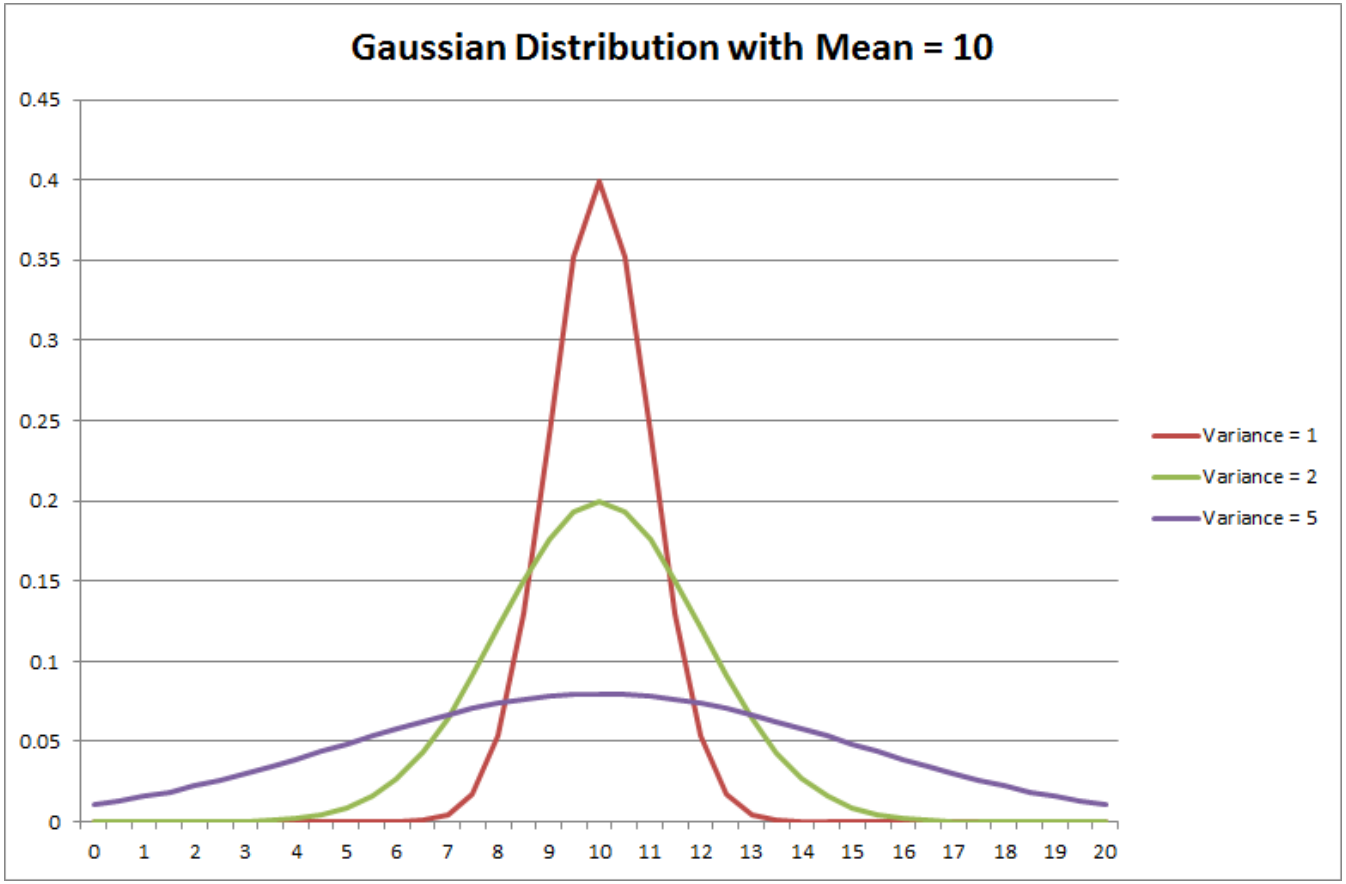
Choose 5 variables in the data and for each:

- Calculate the mean, median, and mode.
- Which do you think is the best choice to represent that variable? Why?
- Are any of the measures of central tendency misleading? Why?

Bonus: produce a histogram of the data, with vertical lines showing the mean, median, and mode.

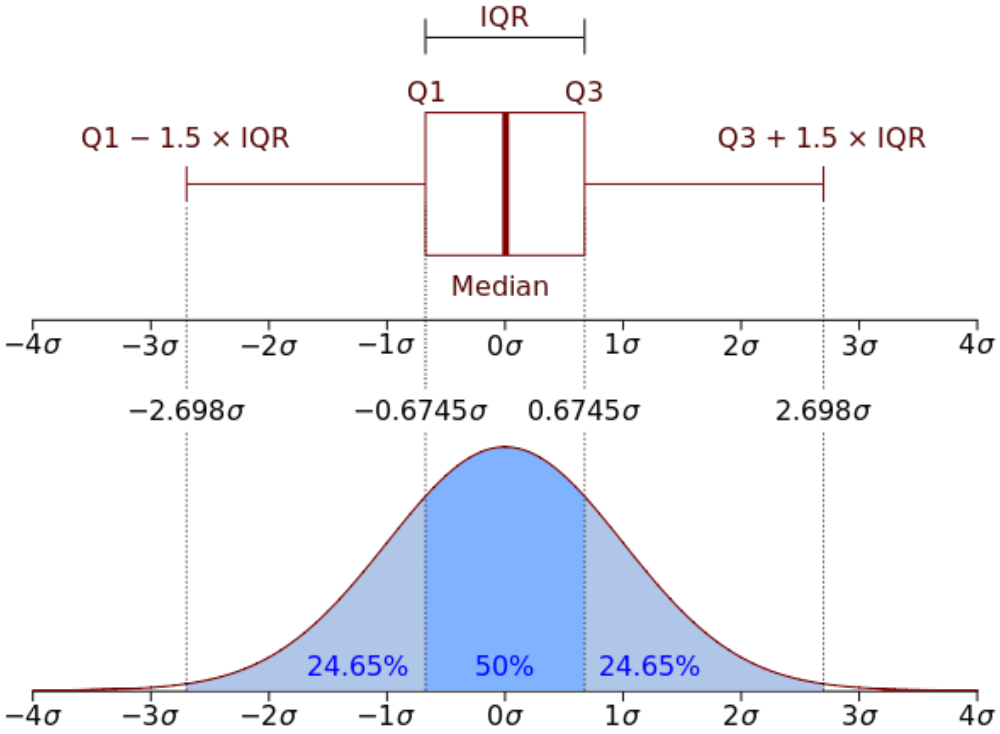
Variance

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$



Interquartile Range

- The middle 50% of the data
- The difference between the 75th percentile (Q3) and 25th percentile (Q1)
- Can define outliers as below Q1 – 1.5*IQR or above Q3 + 1.5*IQR.



{Propulsion}

Probabilities

“Probability is the measure of the likelihood that an event will occur”

Probability of seeing a girl on any random day in full stack class (P_{girl}) **= #girls / #students**

Warning :- This is the frequentist paradigm

https://en.wikipedia.org/wiki/Probability

{Propulsion}

Probabilities

“Probability is the measure of the likelihood that an event will occur”

Probability of seeing a boy on any random day in full stack class (P_{boy}) **=**

Warning :- This is the frequentist paradigm

https://en.wikipedia.org/wiki/Probability

{Propulsion}

Probabilities

“Probability is the measure of the likelihood that an event will occur”

Probability of seeing a boy on any random day in full stack class (P_{boy}) **= 1 - P_{girl}**

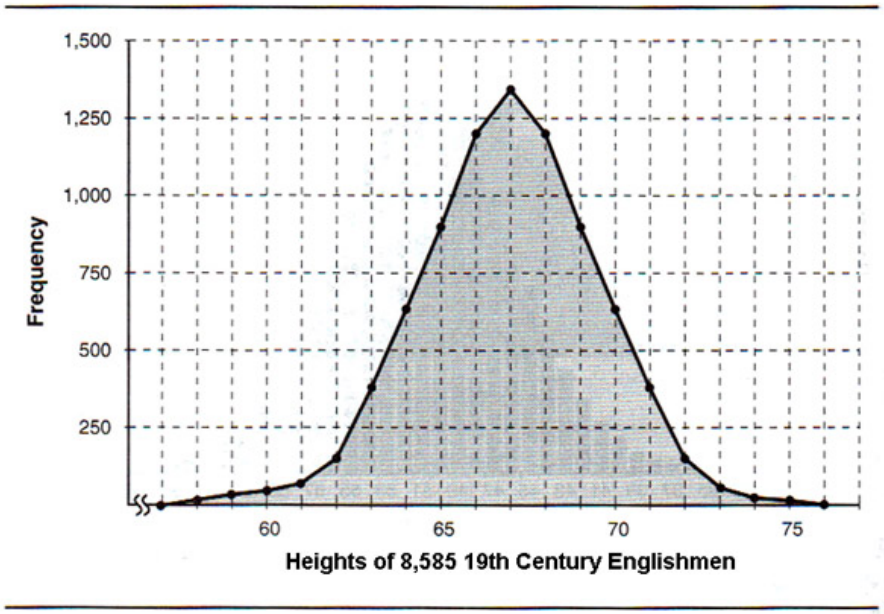
Warning :- This is the frequentist paradigm

https://en.wikipedia.org/wiki/Probability

Distributions

Data is drawn from some data generating process.
There are common **distributions** that can be described mathematically.
A distribution has **parameters** that change its shape depending on their values.

Normal
distribution

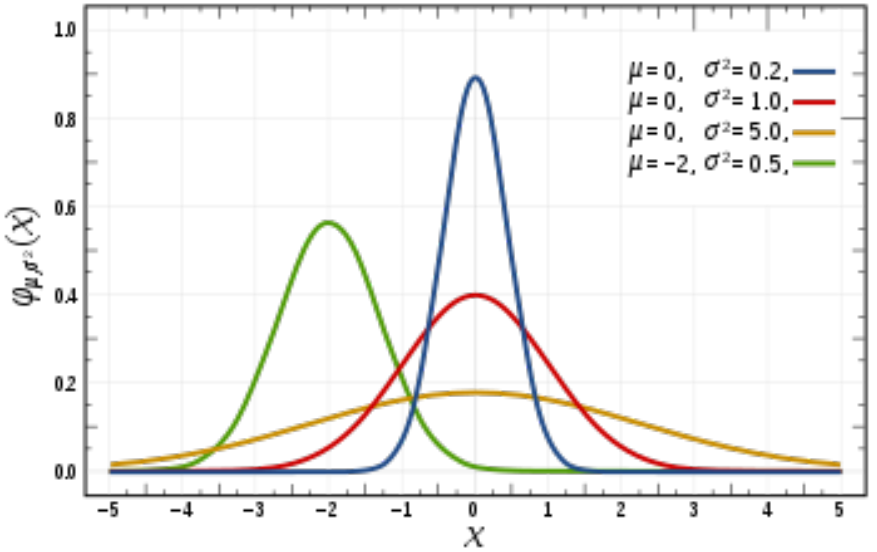


http://www.gvglass.info/EDUC5716/Lesson-3/

Distributions

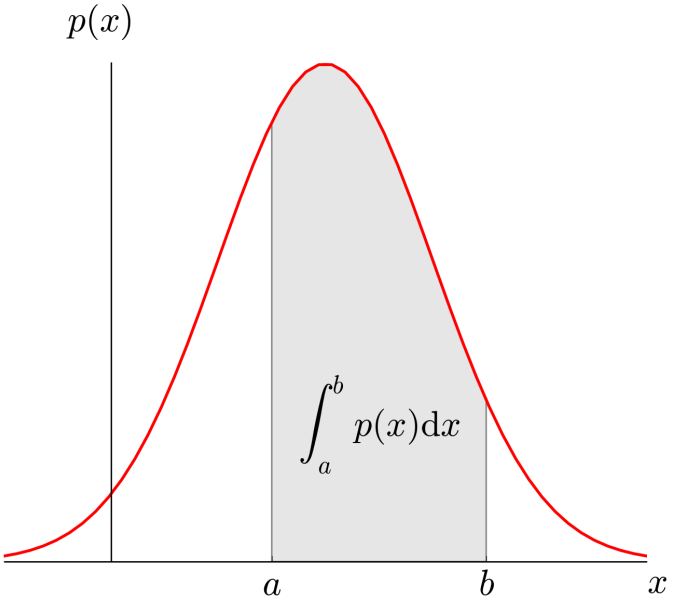
Data is drawn from some data generating process.
There are common **distributions** that can be described mathematically.
A distribution has **parameters** that change its shape depending on their values.

Normal
distribution



Probability Density Function (PDF)

The probability density function (PDF) of a distribution describes how likely it is that a draw from that distribution will have a particular value.

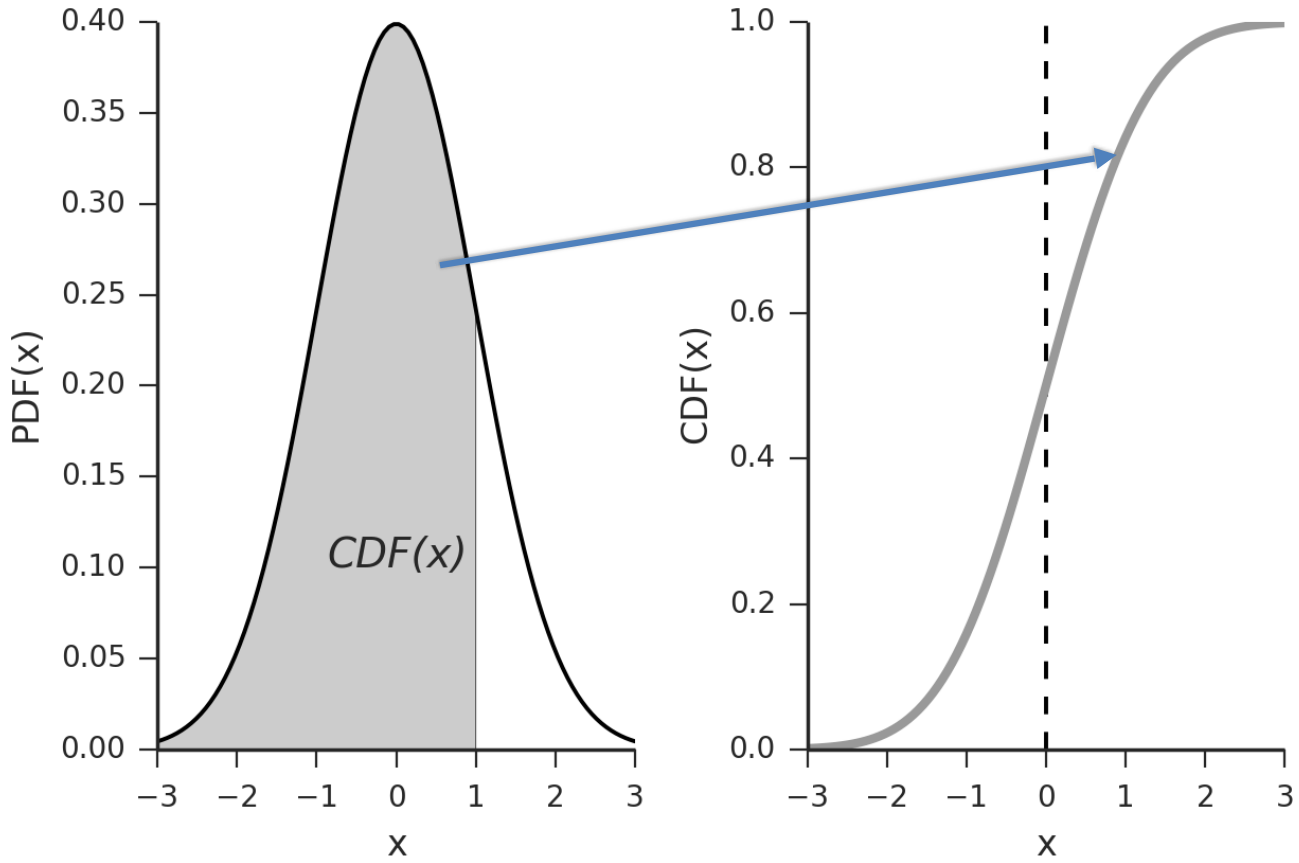


http://work.thaslwanter.at/Stats/html/statsDistributions.html

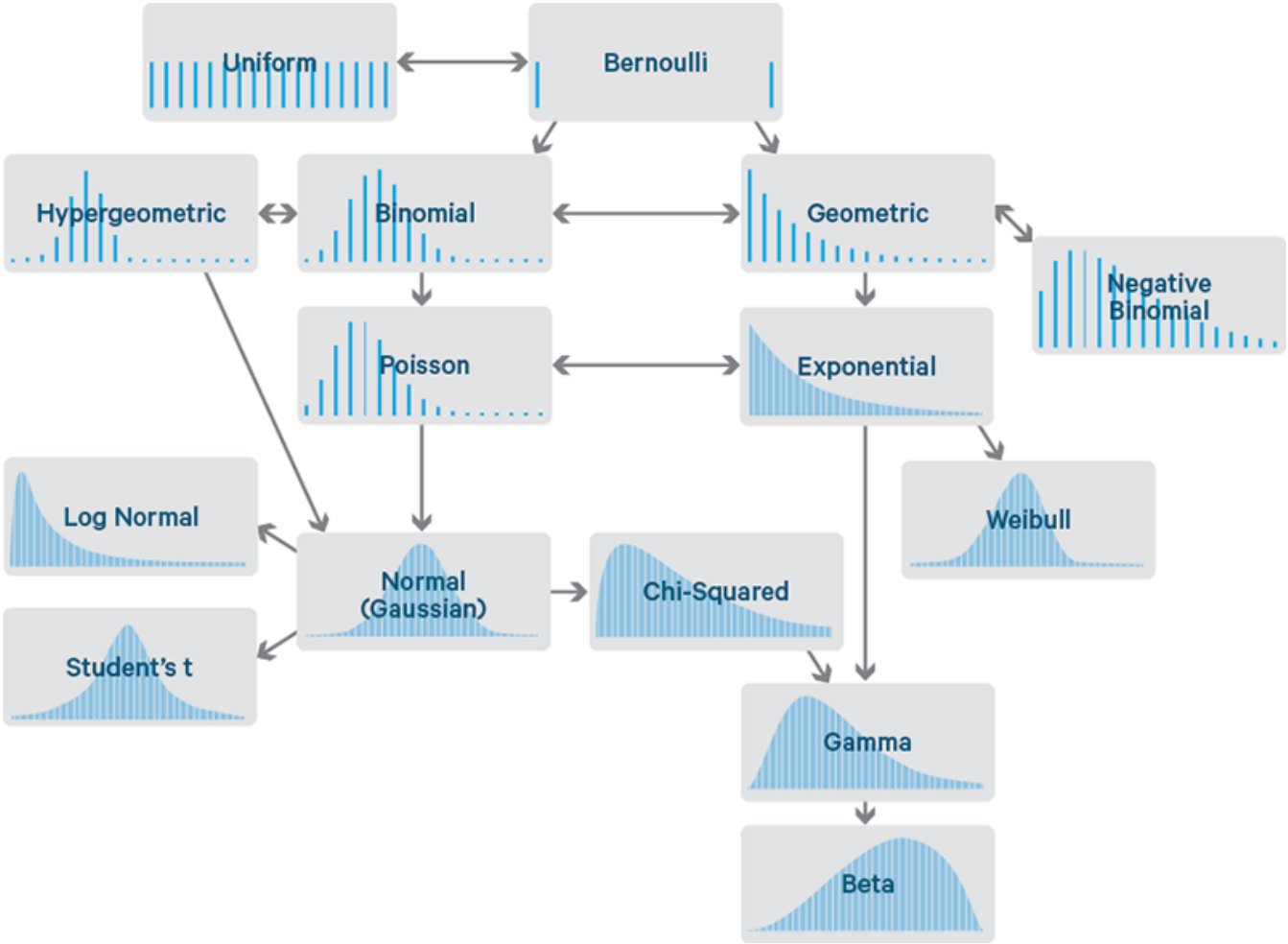
- $PDF(x) \geq 0 \, \forall \, x \in \mathbb{R}$ Never negative
- $\int_{-\infty}^{\infty} PDF(x)dx = 1$ Sums to 1 over all values

Integral of a particular range describes the probability of a drawn value falling in that range

Cumulative Distribution Function (CDF)



Distributions



https://blog.cloudera.com/blog/2015/12/common-probability-distributions-the-data-scientists-crib-sheet/

Distributions - exercise

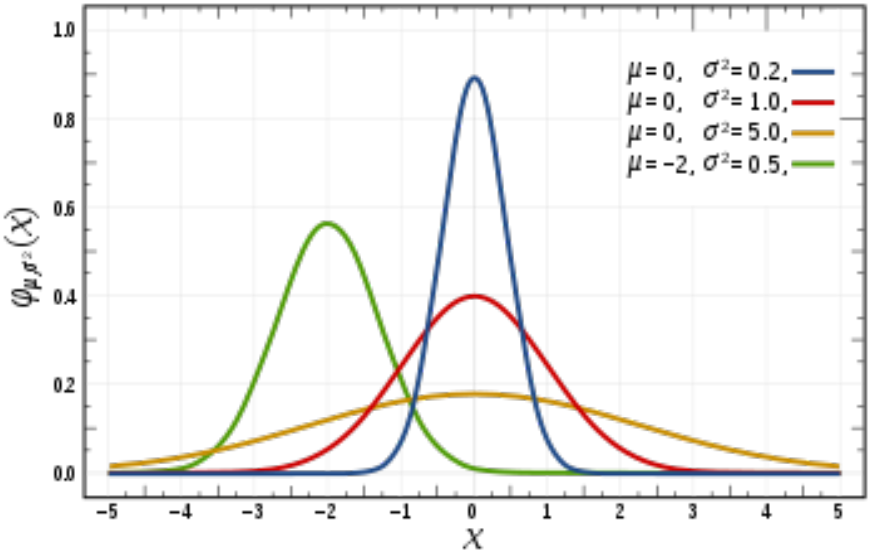
Let's sample from some normal distributions

- Generate this plot of normal distributions with different parameters

Get started with:

- Python `numpy.random.normal()`

First write out what you're going to do!



PDF and CDF - exercise

Assume you have a population of people whose heights are described by

Height ~ N(170, 7)

- What percent of the population is between 170 and 175cm tall?
- What height is taller than 70% of the population?
- If someone is 168cm tall, what percent of the population is shorter than they are?
- Find an interval that includes 50% of the population
- Find an interval that includes 20% of the population
- Bonus - Draw a sample of 5 data points at once from this data and compute their mean. Are all these means = 170?

PDF and CDF - exercise (bonus)

- Bonus - Draw a sample of 5 data points at once from this data and compute their mean. Are all these means = 170?

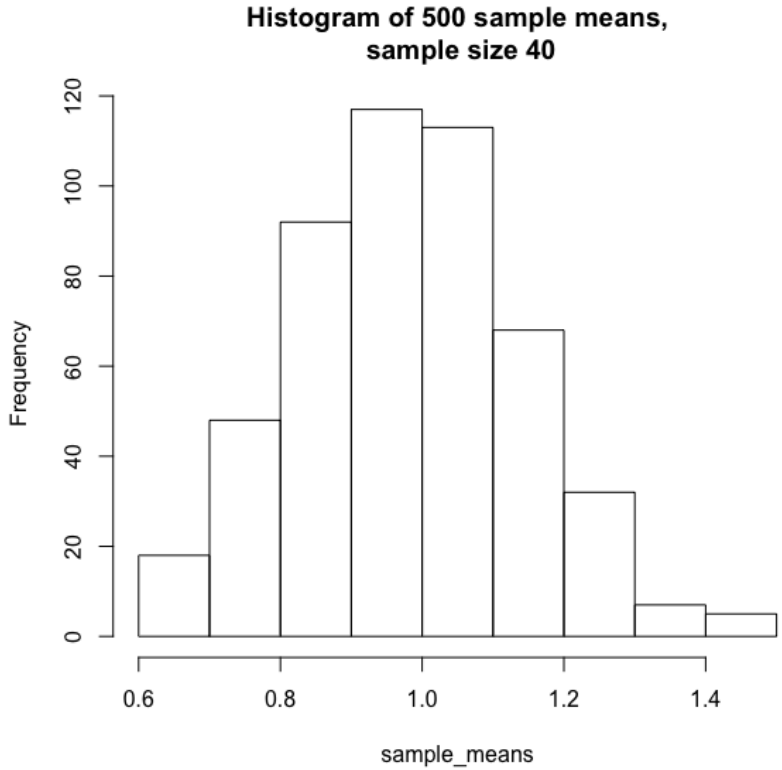
- Can we somehow find out the true mean of this data (should be 170)?

Bootstrapping

When we drew many samples, we had a new mean each time.
Where do these means fall 95% of the time?

(This should sound familiar…
~ in what interval do 95% of
draws from this distribution fall?)

-> Confidence interval for the mean



Bootstrapping

We can also do bootstrapping on empirical data for which we don't know the distribution.
Just like the sample mean of larger and larger samples from a known distribution looked more and more like the population mean, we will approximate the population statistic by taking many samples from our sample.

How? Sampling with replacement, from our sample.

Bootstrapping - Exercise

1. Take the sample `8, 12, 58, 94, 103, 115`, drawn from a population with unknown distribution
2. Resample with replacement 1000 times
3. Get the mean for each resample
4. Keep a vector of all of these resample means
5. Plot the resample means
6. Calculate a 95% confidence interval for the mean of this population

The SE of any sample statistic is the standard deviation (SD) of the sampling distribution for that statistic. And the 95% confidence limits of a sample statistic are well approximated by the 2.5th and 97.5th percentiles of the sampling distribution of that statistic.

More reading: https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading24.pdf

What if we have two distributions?

Hypothesis testing

We can use the t-test (and many other tests) for hypothesis testing: deciding whether a hypothesis about the data is true.

For a t-test of differences in group means:

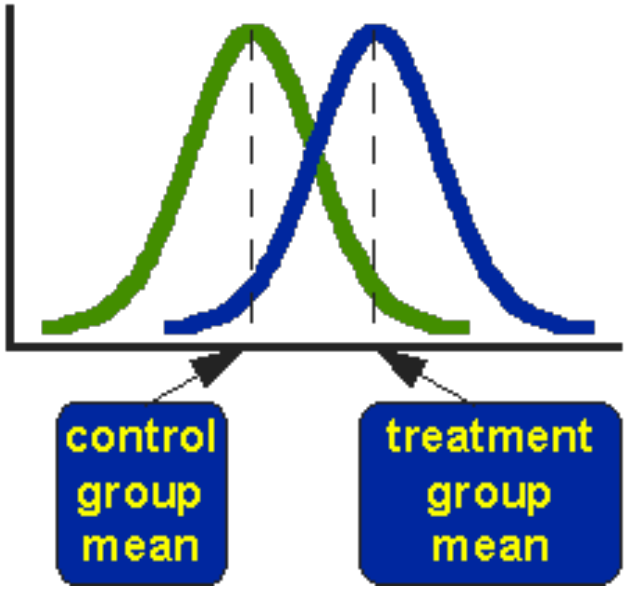
Null hypothesis (H_0): there is no difference between the groups (samples are drawn from the same distribution)

Alternative hypothesis (H_A): there is a difference between the groups (samples are drawn from different distributions)

If the p-value is less than .05, we **reject the null hypothesis**

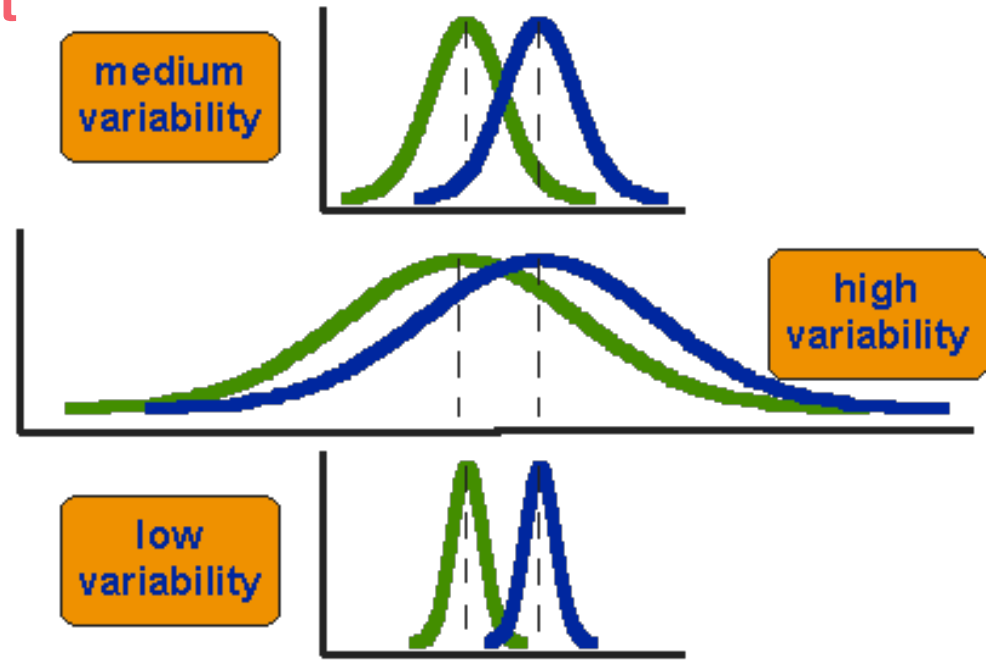
T-test

If we assume the underlying distribution is normal, we can do a **parametric** test: the t-test



http://socialresearchmethods.net/kb/stat_t.php

T-test



$$\text{t-value} = \frac{\text{difference between group means}}{\text{variability of groups}}$$

http://socialresearchmethods.net/kb/stat_t.php

T-test

$$\text{t-value} = \frac{\text{difference between group means}}{\text{variability of groups}}$$

if variances are equal

$$= t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}$$

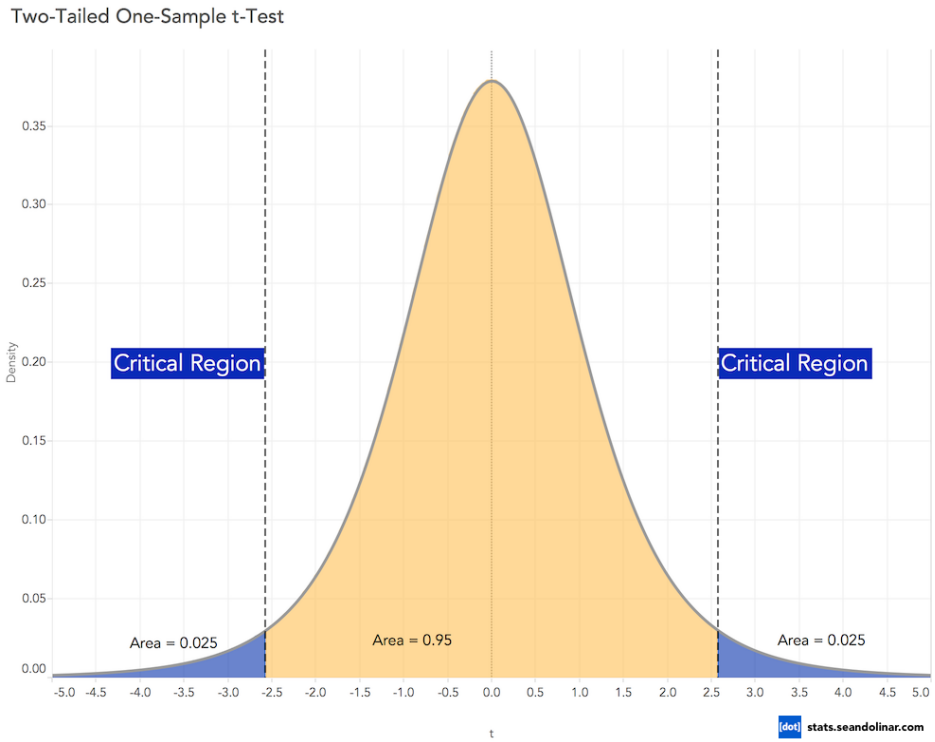
degrees of freedom = $n_1 + n_2 - 2$

if variances are unequal

$$= \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

T-test

- Look at the t-distribution with the degrees of freedom from your test
- Compare the t-value from your test to the distribution
- How likely is it that your t-value was drawn by chance? -> integrate PDF
- Critical value is a t-value with cumulative probability of $(1-\alpha)$
- Alpha is the chance that the difference in sample means would be observed by chance. Typically set at $\alpha = .05$. (In medical research or more critical applications, sometimes set much lower, $\alpha = .01$ or $.001$) - This is **Type I error**



T-test

T-test assumptions:

- Bivariate independent variable (data is in 2 groups)
- Continuous dependent variable
- Each observation of the dependent variable is *independent* of the other observations of the dependent variable (its probability distribution isn't affected by their values).
- Dependent variable has a normal distribution, with the same variance, σ^2, in each group

http://www.csic.cornell.edu/Elrod/t-test/t-test-assumptions.html

Hypothesis testing - exercise

What is the null hypothesis and alternative hypothesis for each of these research questions? Would you run a one-tailed or two-tailed test?

- Are there more sunny days per year in San Francisco than in Boston?
- Does a sandwich cost more in Zürich or in London?
- Do people who exercise every day eat more than people who don’t exercise?
- Do people in red cars drive at a different speed than people in blue cars?

Hypothesis testing - exercise

Say we run the appropriate one- or two-tailed t-test and get the following p-values. Interpret the result.

- Are there more sunny days per year in San Francisco than in Boston? ***p* = .30**
- Does a sandwich cost more in Zürich or in London? ***p* = .02**
- Do people who exercise every day eat more than people who don't exercise? ***p* = .43**
- Do people in red cars drive at a different speed than people in blue cars? ***p* = .04**

Hypothesis testing

When is hypothesis testing wrong?

		reality	
		H_0 = true	H_0 = false
conclusion	H_0 = true	OK	type II error
	H_0 = false	type I error	OK

false negative

false positive

http://www.vias.org/tmdatanaleng/cc_error_types.html

T-test exercise

Draw many (**1000**) sets of samples with **30** observations each, where either:

- both samples are from the same normal distribution **X ~ N(50,5)**

or

- each sample is from a different normal distribution,
  one from **X ~ N(50, 5)** and one from **X ~ N(45, 5)**

Run a t-test for each set of samples

Using this simulation, how often is there a type I or type II error?

Bonus: What happens if you change the sample sizes or the difference in means?
We'll talk about this "power" soon.

Statistics for product developers, what all do we need to know?

Power analysis

A **power analysis** is what we use to determine **sample sizes**, **effect sizes**, **significance level**, and **power**.

Power analysis – sample sizes

A **power analysis** is what we use to determine **sample sizes**, **effect sizes**, **significance level**, and **power**.

Sample size is just the number of observations in your data set. The larger your sample, the more it approaches the population distribution.

Power analysis – effect sizes

A **power analysis** is what we use to determine **sample sizes**, **effect sizes**, **significance level**, and **power**.

For comparisons of group means (e.g. t-test), **effect size** is the magnitude of the difference scaled by the standard deviation to standardize it.

For comparison of two group means,

$$\text{Cohen's } d = \frac{\mu_1 - \mu_2}{\sqrt{[(s_1^2 + s_2^2) / 2]}} = \frac{\text{difference in means}}{\text{pooled sd}}$$

While *p* was a measure of statistical significance, effect size can be more a measure of practical significance – is this difference large enough for us to care about?

Power analysis – significance level

A **power analysis** is what we use to determine **sample sizes**, **effect sizes**, **significance level**, and **power**.

Significance level (α) is the acceptable level of type I errors. It's the probability of saying there's a difference between groups, when in fact there is no true difference.

We choose the significance level by setting α.

		Reality	
		H0 = true (no diff)	**H0 = false (diff exists)**
Our conclusion	**H0 = true (no diff)**	**correct** **P | no diff = (1-α)**	**type II error** **(false negative)** **P | true diff = β**
	H0 = false (diff exists)	**type I error** **(false positive)** **P | no diff = α**	**correct** **P | true diff = (1-β)**

Power analysis – statistical power

A **power analysis** is what we use to determine **sample sizes**, **effect sizes**, **significance level**, and **power**.

Statistical power (1 - ß) is our ability to correctly determine that an effect exists (correctly reject the null hypothesis).

Basically, we calculated this in our simulation of hypothesis testing!

A common value to choose is .80 (so that ß = .20)

P(type I error) = α
P(type II error) = ß

		Reality	
		H0 = true (no diff)	**H0 = false (diff exists)**
Our conclusion	**H0 = true (no diff)**	**correct** **P | no diff = (1-α)**	**type II error** **(false negative)** **P | true diff = β**
	H0 = false (diff exists)	**type I error** **(false positive)** **P | no diff = α**	**correct** **P | true diff = (1-β)**

Power analysis

For a **power analysis,** from the variables:

- **sample size**
- **effect size**
- **significance level**
- **statistical power**

you need 3, and you calculate the 4th

When to do a power analysis:

Before running a study, you could do an analysis of the effect size, significance level, and statistical power you want in order to determine what sample size you should collect.

After running a study, you could use significance level, sample size, and effect size to determine the statistical power you have.

A priori power analysis

Before running a study, you could do an analysis of the effect size, significance level, and statistical power you want in order to determine what sample size you should collect. (Or to determine whether the study is worthwhile – it may not be, if the necessary sample size is impractically large.)

But before running a study, we don't know what the effect size will be!

- Estimate from previous literature
- Estimate using rules of thumb:
  e.g. for comparing two means, Cohen suggests the effect sizes:
  d = 0.2 small
  d = 0.5 medium
  d = 0.8 large

Effect size is measured on different scales for different types of tests: http://statmethods.net/stats/power.html

How do product developers use statistics?

A/B testing

A/B testing (split testing) is comparison of two versions of web pages to find out which version performs better

https://vwo.com/ab-testing/

Why do we do A/B testing

- Visitors to buy products

- Free visitors converted to paid ones

- Click on adds

https://vwo.com/ab-testing/

What can you test?

1. Headlines
2. Sub headlines
3. Paragraph Text
4. Testimonials
5. Call to Action text
6. Call to Action Button
7. Links
8. Images
9. Content near the fold
10. Social proof
11. Media mentions
12. Awards and badges

Advanced tests can include pricing structures, sales promotions, free trial lengths, navigation and UX experiences, free or paid delivery, and more.

https://vwo.com/ab-testing/