PROGRAMMING

RESOURCES FOR GETTING BETTER AT PROGRAMMING

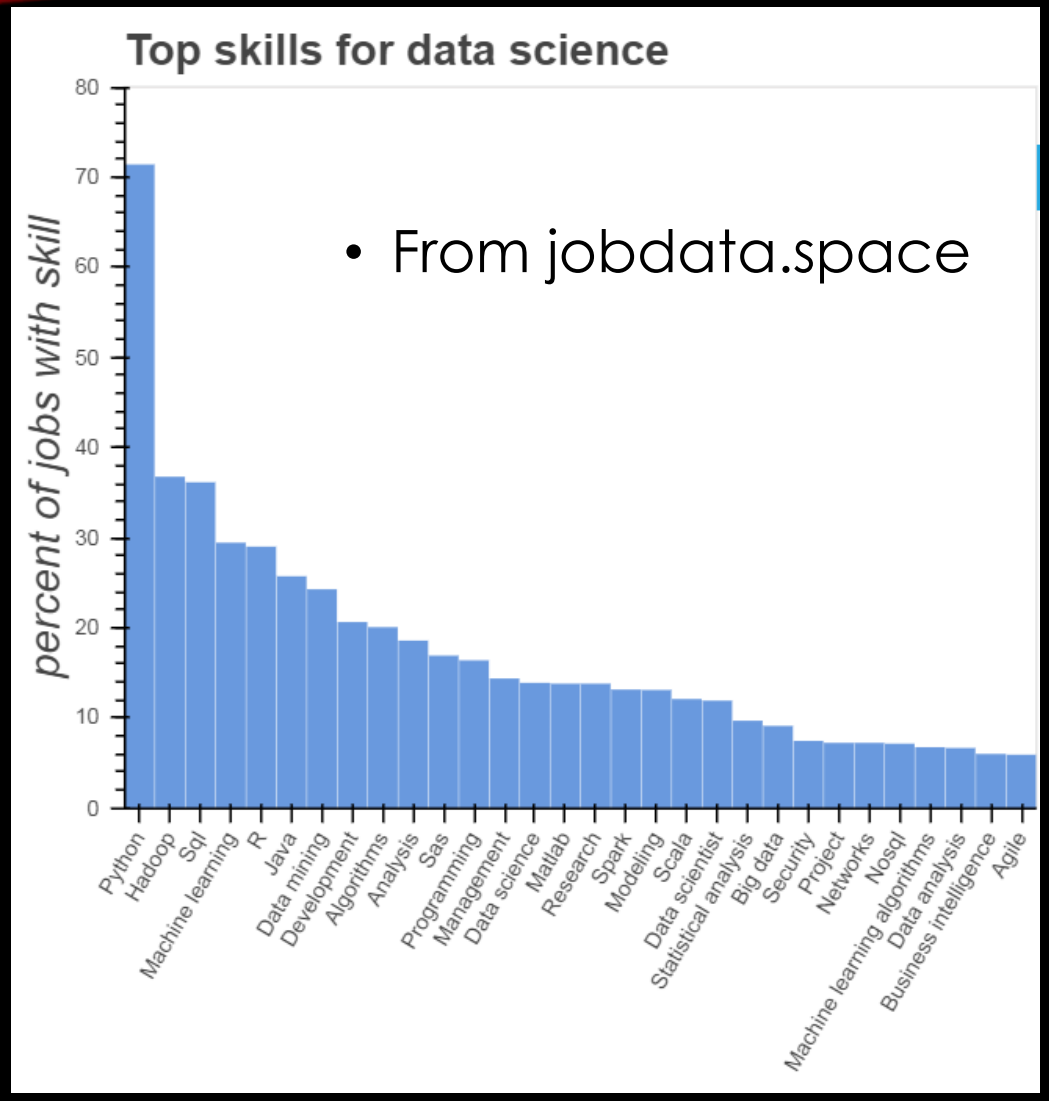
- Datacamp…
  - Good for beginners
- https://learnpythonthehardway.org/
  - Good for beginners
- https://www.codewars.com/
  - A ton of clever solutions to the problems, lots of new content constantly
- https://www.hackerrank.com/
  - Not as good as codewars in my opinion

RESOURCES TO LEARN MORE EDA TECHNIQUES

- www.kaggle.com kernels (competitions and datasets)
- www.drivendata.org (like kaggle, but geared toward social responsibility)
- Udacity
- Coursera
- DataCamp (has an EDA course)

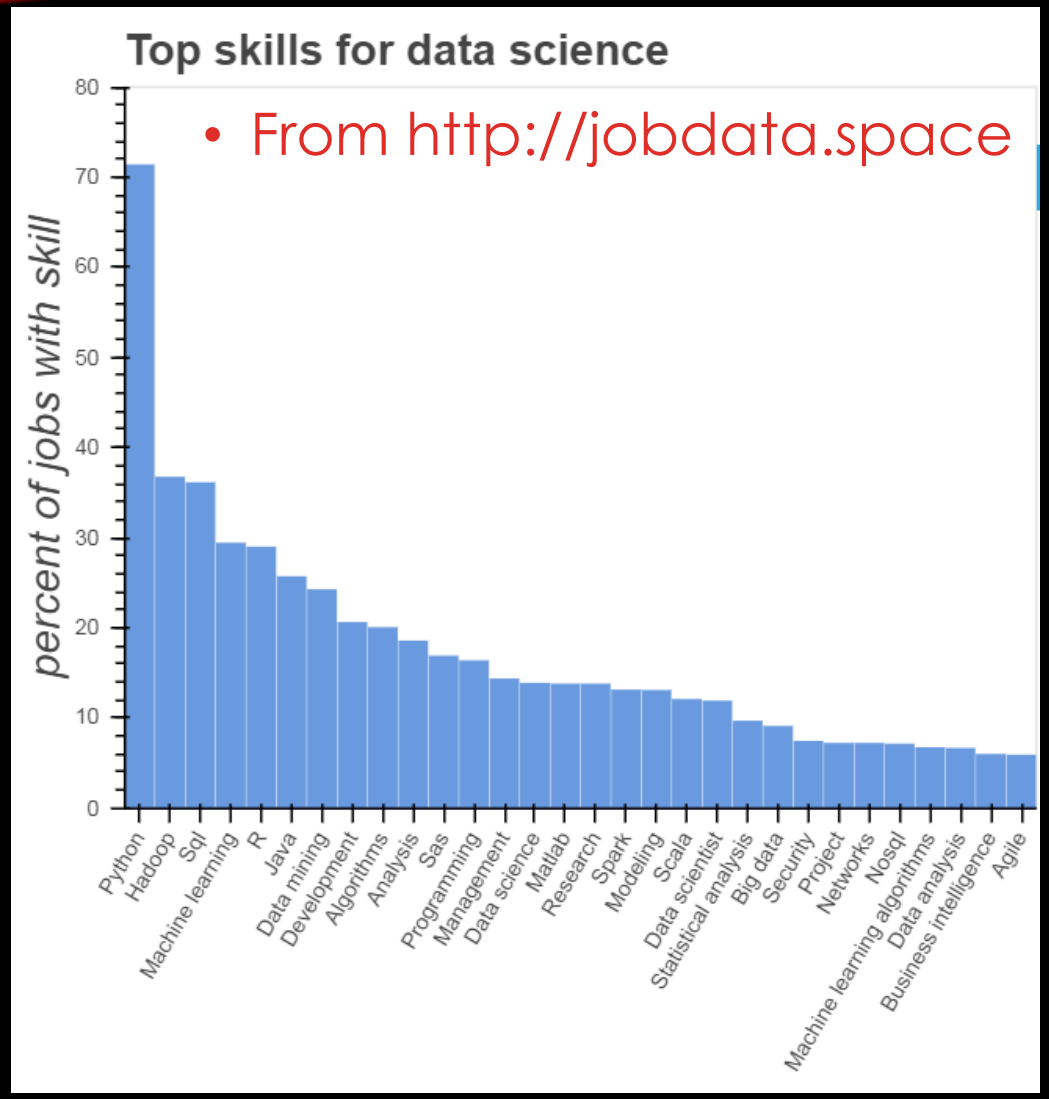
- Here's code for some work I did on a Kaggle competition: http://ngeorge.us/kaggle-bimbo/
- https://www.kaggle.com/wordsforthewise/grupo-bimbo-inventory-demand/bimbo-mexico-walmart

R OR PYTHON?

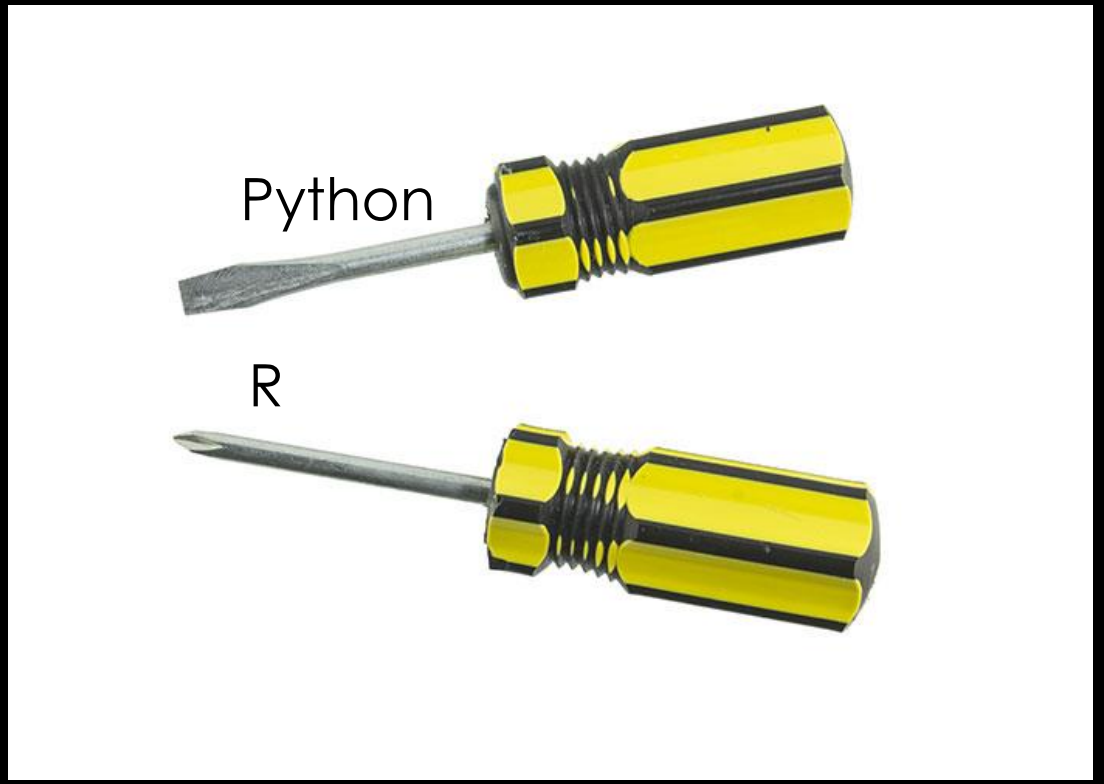


- From jobdata.space

R OR PYTHON?

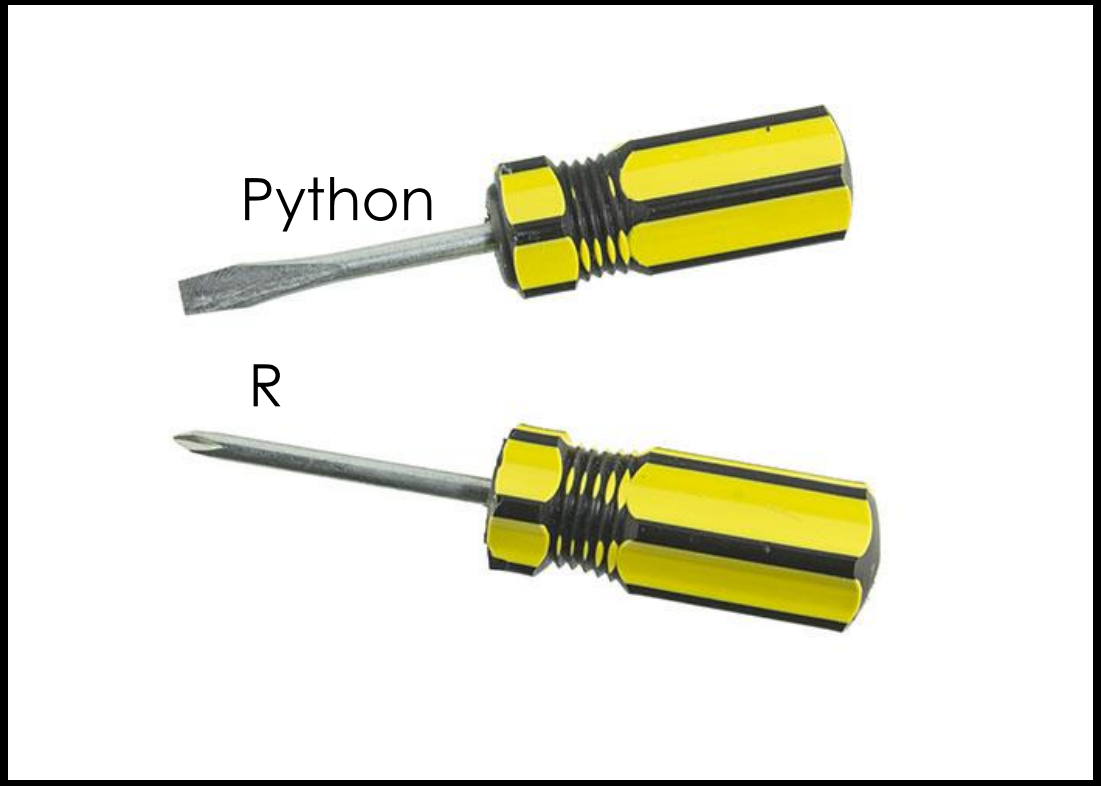


- From http://jobdata.space



R OR PYTHON?

- Python is faster
- R has more classic statistics built-in
- Both have a strong community



PANDAS INTRO

- See Jupyter notebook

WARMUP EXCERCISE

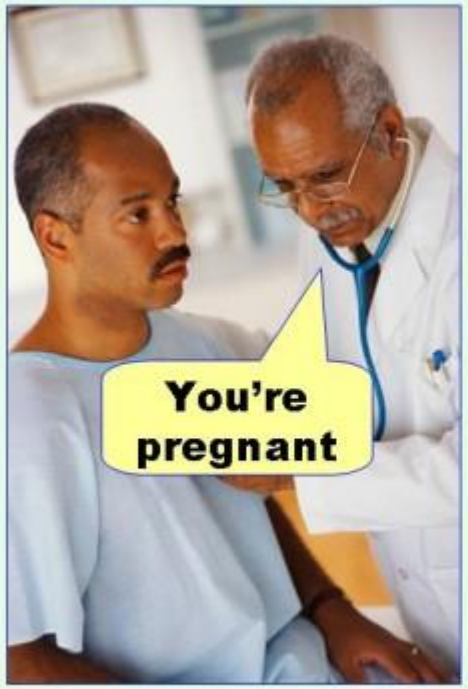
Weird Al's 'UHF'

MY TYPICAL DEVELOPMENT WORKFLOW

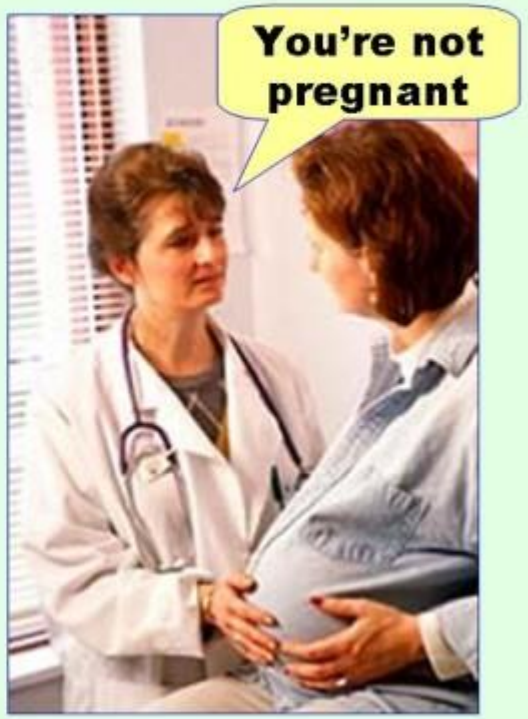
- Use atom ide for writing code
  - Use a linter to detect code problems (flake8 for Python)
    - First need to install flake8: conda install flake8 or pip install flake8
  - atom-pair for pair programming
- Run code in Ipython
  - `paste` magic command
  - `run filename.py`
- Demo

HYPOTHESIS TESTING

Type I error
(false positive)



Type II error
(false negative)



INTRO TO HYPOTHESIS TESTING

- What is a hypothesis?

INTRO TO HYPOTHESIS TESTING

- What is a hypothesis?
  - A proposed explanation for a phenomenon.

INTRO TO HYPOTHESIS TESTING

- What is a typical example of the null hypothesis?

INTRO TO HYPOTHESIS TESTING

- What is a typical example of the null hypothesis?
  - Often we are measuring the means of something, and the null hypothesis is that two means are the same.
  - Abbreviated H_0

INTRO TO HYPOTHESIS TESTING

- What is a typical example of the alternative hypothesis?

INTRO TO HYPOTHESIS TESTING

- What is a typical example of the alternative hypothesis?
  - Again, when measuring the means of something (average height of American vs Mexican corn, for example), the alternative hypothesis is that the means are not the same.
  - Typically abbreviated H_a

INTRO TO HYPOTHESIS TESTING

- Null hypothesis – means are the same
  - Can use with regression too – null is coefficient is no different than from 0
- Alternative
  - Means are different, or coefficient is different from 0

F-, T-, AND Z-TESTS

- T-test
  - Used for: ____

F-, T-, AND Z-TESTS

- T-test
  - Used for comparing means (or another summary statistic) of 1 group to a number, or between two groups

F-, T-, AND Z-TESTS

- T-test
  - Used for comparing means (or another summary statistic) of 1 group to a number, or between two groups
- Z-test
  - Used for: _____

F-, T-, AND Z-TESTS

- T-test
  - Used for comparing means (or another summary statistic) of 1 group to a number, or between two groups
- Z-test
  - Used for same thing as a t-test, but typically for large samples and known population variance. T-test is for unknown population variance, so we use sample standard deviation (sqrt of variance).

F-, T-, AND Z-TESTS

- T-test
  - Used for comparing means (or another summary statistic) of 1 group to a number, or between two groups
- Z-test
  - Used for same thing as a t-test, but typically for large samples and known population variance
- F-test
  - Used for: _____

F-, T-, AND Z-TESTS

- T-test
  - Used for comparing means (or another summary statistic) of 1 group to a number, or between two groups
- Z-test
  - Used for same thing as a t-test, but typically for large samples and known population variance
- F-test
  - Used for testing if different levels of a factor have a significant effect on a summary statistic

F-, T-, AND Z-TESTS

- T-test
  - Example: test if average height of everyone in the class is different from 5'10"
  - If just testing if height is different, is it 1-sided or 2-sided?
  - If testing if average height is above 5'10", 1-sided or 2-sided?
- Z-test
  - Used for same thing as a t-test, but typically for large samples and known population variance
- F-test
  - Example: group people by how many gallons of milk we drank as kids, test if amount of milk we drank has a significant effect on average height

F-, T-, AND Z-TESTS

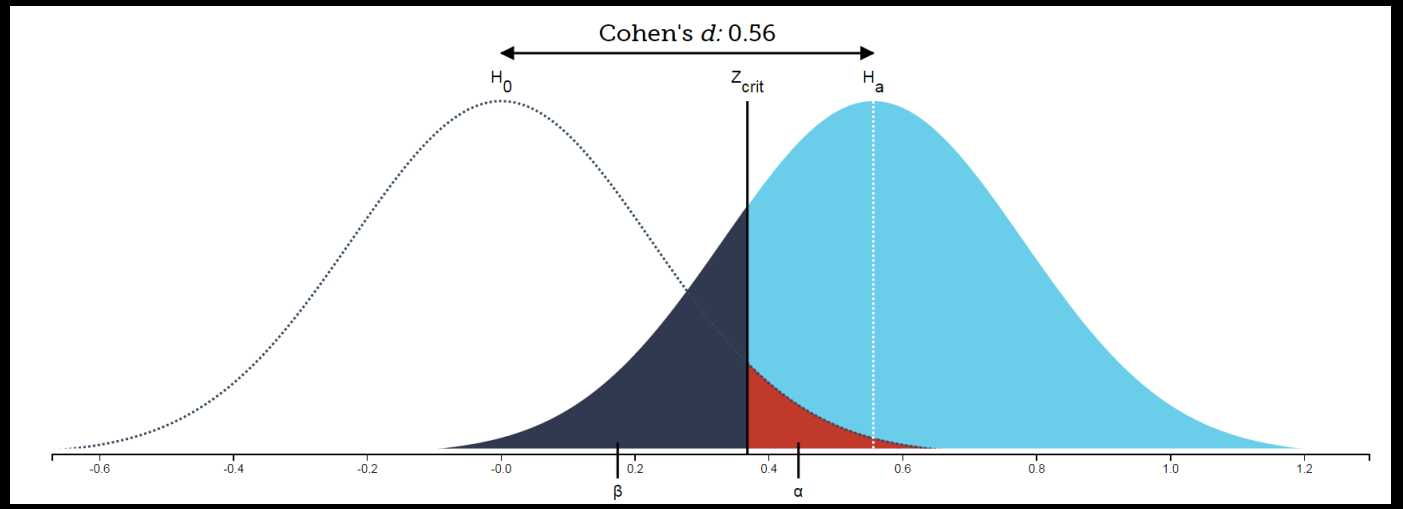
- Calculate some test statistic (F- or t- or Z-statistic), compare to significance level (almost always 0.05)
  - T-test can be used for comparing 1 sample to a value

    or
  - 2 samples

  - Note: if making multiple comparisons (comparing means between more than 2 groups) can use the Bonferroni correction, or something similar

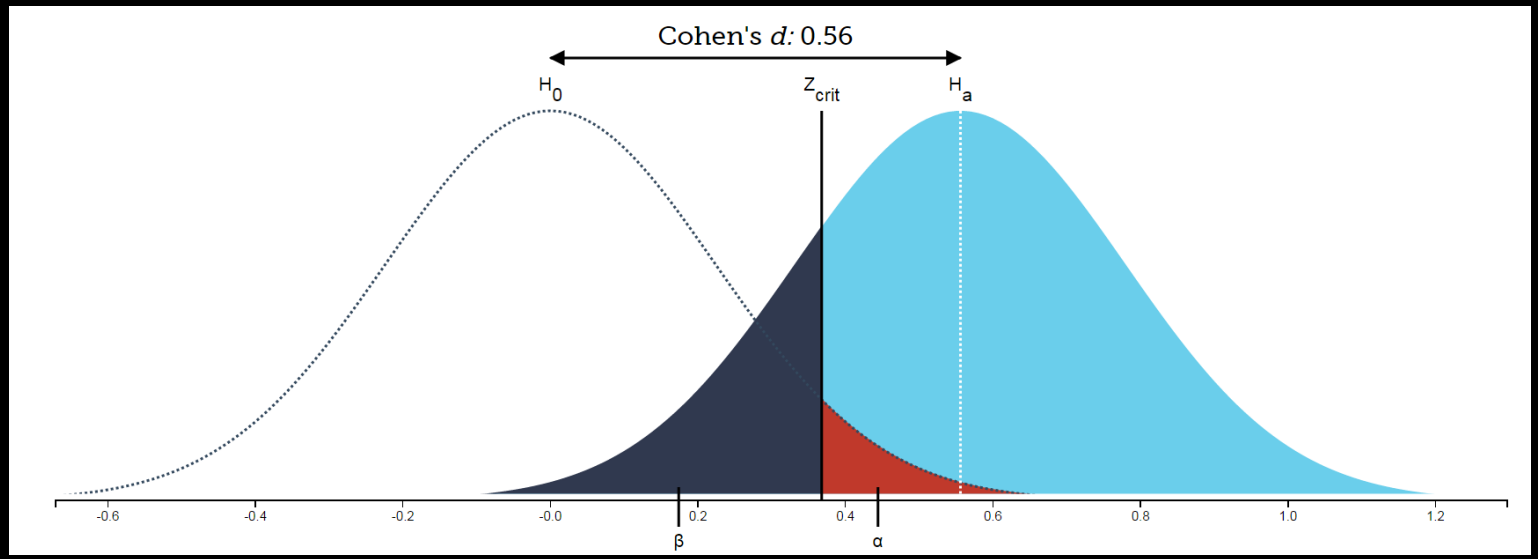
INTRO TO HYPOTHESIS TESTING

- Calculate some test statistic (F- or t- or Z-statistic), compare to significance level (almost always 0.05)
  - T-test can be used for comparing 1 sample to a value (one-sample t-test)
    or
  - 2 samples (paired t-test)
  - http://rpsychologist.com/d3/NHST/



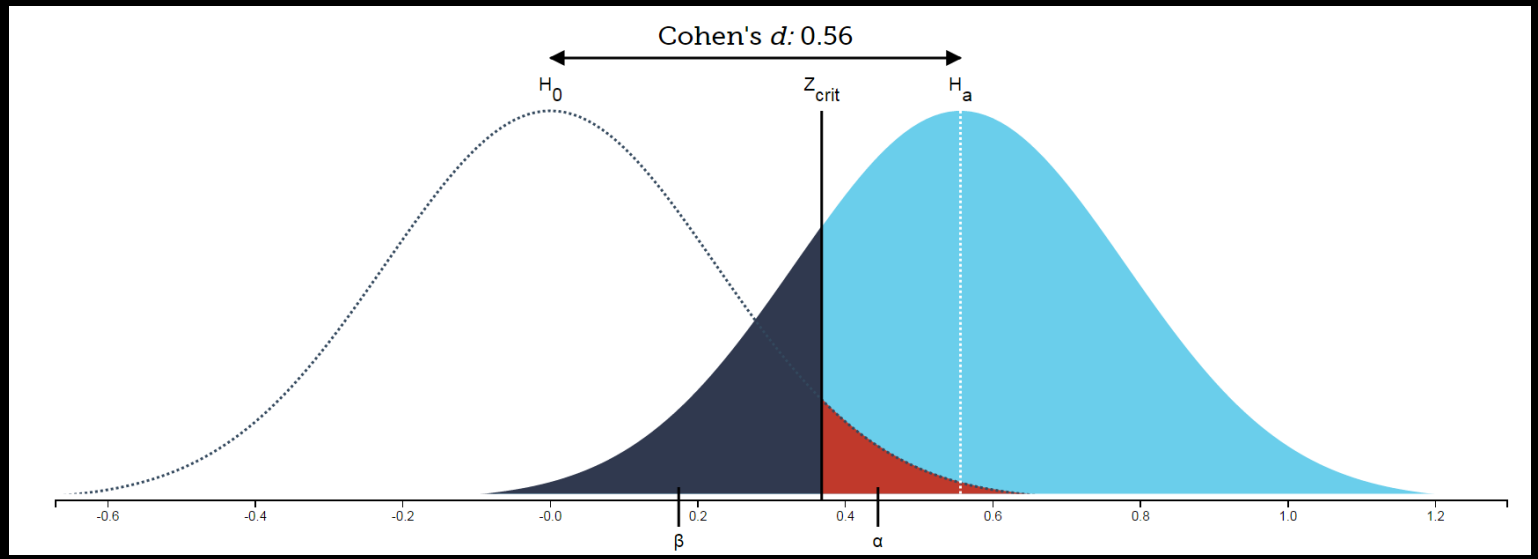
INTRO TO HYPOTHESIS TESTING

- Calculate some test statistic (F- or t-statistic), compare to significance level
  - Relies on the central limit theorem – says that sampling a population many times and taking a test stat will result in a normal distribution
  - http://rpsychologist.com/d3/NHST/



INTRO TO HYPOTHESIS TESTING

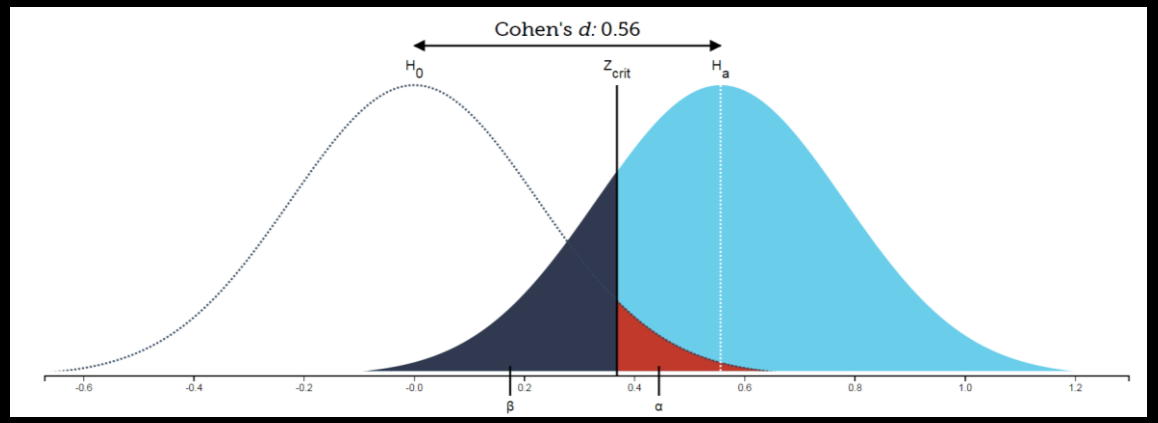
- http://rpsychologist.com/d3/NHST/
- This is a one-sample, one-sided Z-test
- H_0 could be that the class is 5'10" on average, H_a that the class is above 5'10"



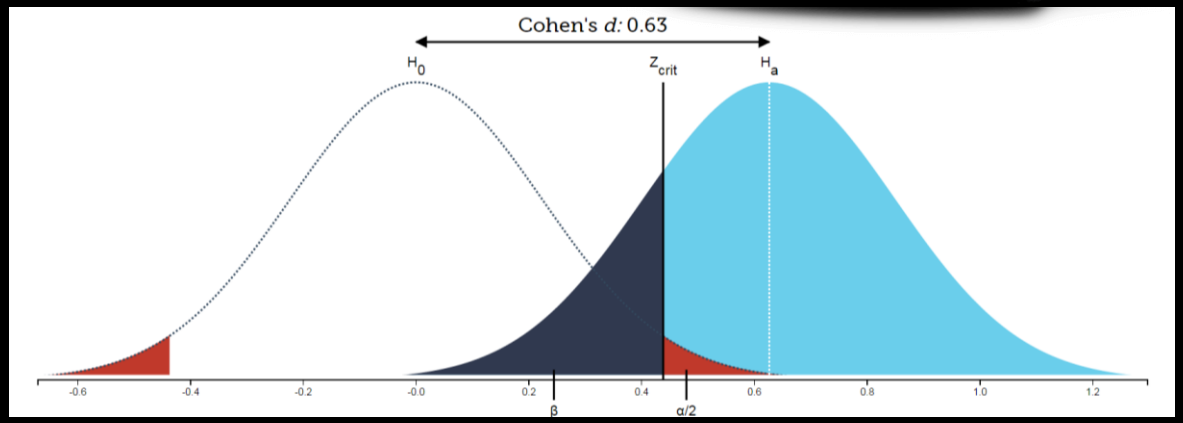
INTRO TO HYPOTHESIS TESTING

- http://rpsychologist.com/d3/NHST/
- What if we want to know if the top speed of a Tesla Model S is different from a Tesla Model X? Would it be a 1- or 2-sided t- (or Z-) test?

1-sided



2-sided



ONE-SAMPLE T-TEST

- Test if mean is different from a specified value
  - E.g., if pollution in the air is higher than a critical value
  - Z-test is the same, but s is population standard deviation rather than the sample standard deviation

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

ONE-SAMPLE T-TEST

$$t = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{N}}}$$

Z-test

$$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}.$$

t-test

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}.$$

ONE-SAMPLE T-TEST

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

Z-test (normal distribution, or bell curve)

Normal Probability Density Function

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

t-test (uses the gamma distribution, and number of samples (v – degrees of freedom))

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

T-TESTS

- Paired t-test
- E.g. if air pollution in NYC is worse (higher concentrations of pollutants) than Denver.
- Test if significant difference between two sample means

$$t' = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

F-TESTS

- ANOVA provides significance test
- Test stat is F-ratio:

  Variance between groups / variance within groups

  Or

  Systematic variance / random variance

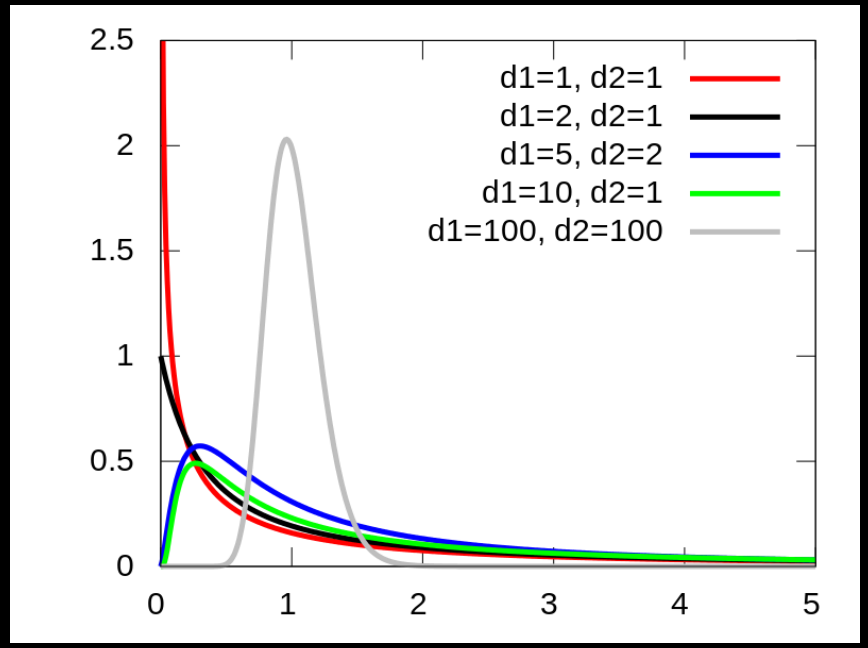
INTRO TO HYPOTHESIS TESTING

- How do we check for significance?
  - Compare F-statistic to an F-distribution: F(df1, df2)
  - The F-distribution depends on number of samples and groups
    - Which df is for number of groups and which the number of total samples?

F-distribution(uses the beta distribution, and number of groups and samples (d1, d2 – degrees of freedom))

$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1\,x)^{d_1}\;d_2^{d_2}}{(d_1\,x+d_2)^{d_1+d_2}}}}{x\,\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

$$= \frac{1}{\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1+\frac{d_1}{d_2}\,x\right)^{-\frac{d_1+d_2}{2}}$$



F-DISTRIBUTIONS IN R AND PYTHON

- In R, many statistical distributions are built-in. The f distribution is df().
- In Python, most classical stats functions are in the statsmodels or scipy.stats packages. These are in general less wieldy than the functions in R.

- Key difference between Python and R here:
  - R is more suited for classical statistics, like f- and t-tests, simpler regressions, and other distributions are built-in and fairly easy to use
  - Python has these tools in packages and they are less easy to use than in R
    - Note: we can call R functions from Python, and vice-versa.

T-TESTS

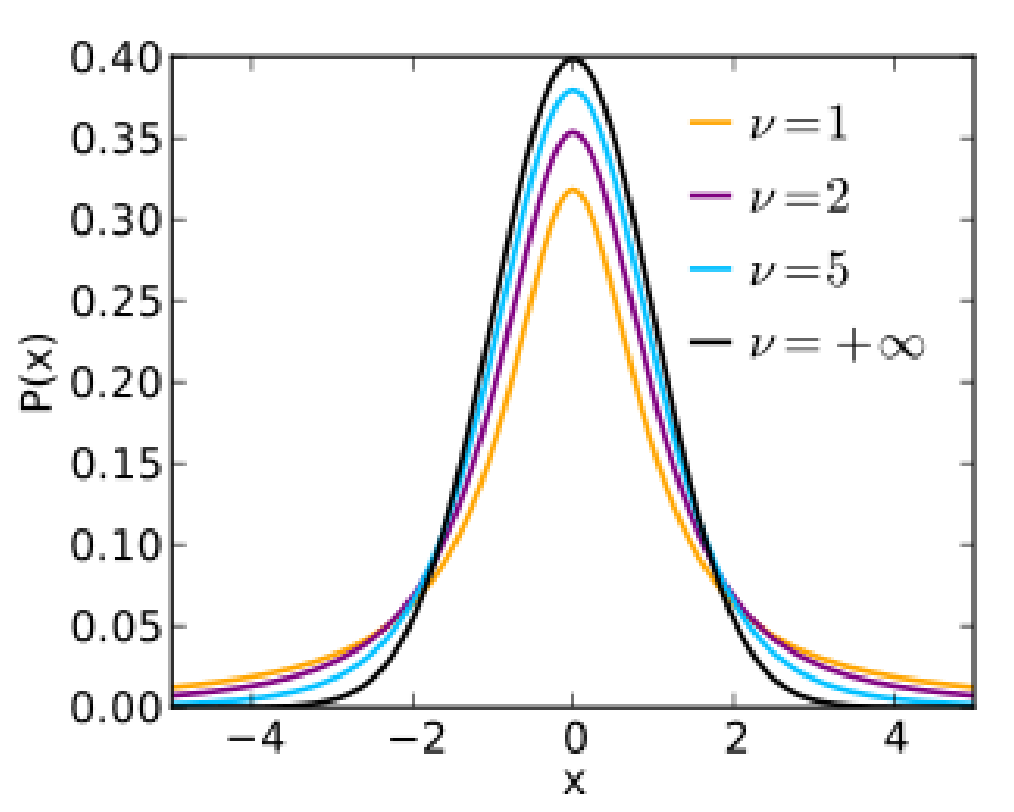
- What's the difference between a t-/Z-test and an F-test?

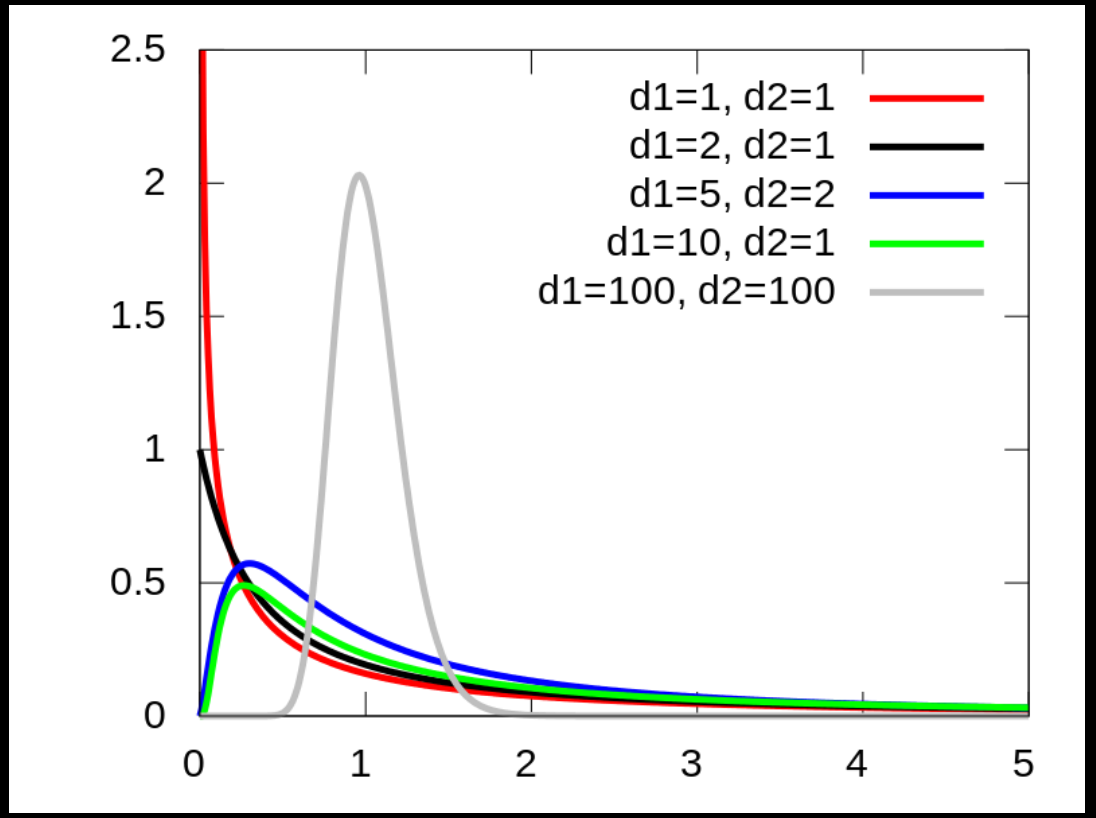
T-TESTS

- What's the difference between a t-/Z-test and an F-test?
  - T-tests are for comparing a property of 2-groups (e.g. avg height of men and women), f-tests are for comparing a property of multiple groups (e.g. avg height of 4 species of corn)

WHAT HAPPENS WITH MORE SAMPLES?

- Which lines have the most samples?





DESIGN OF EXPERIMENTS

RANDOMIZED DESIGN EXAMPLE

Group of granadilla seeds (randomly selected, mix from Nicaragua and Costa Rica)	Fertilizer level
1	100 ppm
2	200 ppm

- Only requires 2 groups

RANDOMIZED BLOCK

Origin of granadilla seeds	Group 1 fertilizer level (assigned randomly)	Group 2 fertilizer level (assigned randomly)
Nicaragua	200 ppm	100 ppm
Costa Rica	200 ppm	100 ppm

- Requires 4 groups

FACTORIAL DESIGN

Nicaragua/100ppm	Nicaragua/200ppm
Costa Rica/100ppm	Costa Rica/200ppm

- If we start adding more factors with 2 possibilities, what happens?

FACTORIAL DESIGN

Texas/100ppm	Texas/200ppm
Mexico/100ppm	Mexico/200ppm

- If we start adding more factors with 2 possibilities, what happens?
  - Size of the experiment grows exponentially
  - 2 factors -> square (4 groups, 2^2)
  - 3 factors -> cube (8 groups, 2^3)
  - n factors -> n-dimensional hypercube (2^n groups)
- Related to the ‘curse of dimensionality’ in machine learning

W2 LEARNER OUTCOMES

- What statistical tests can we use on these types of experiments?
- What is research/experimenter bias, and how does ANOVA help overcome this?
- What are the 3 DOE approaches?
  - Strengths/weaknesses for each?

W2 LEARNER OUTCOMES

- What statistical tests can we use on these types of experiments?
  - ANOVA, Tukey HSD, t- or Z-tests with Bonferroni or other corrections
- What is research/experimenter bias, and how does ANOVA help overcome this?
  - Tendency for experimenters to be biased by their expectations. People find what they want to see. Jan Hendrik Schon is a great example.
- What are the 3 DOE approaches?
  - Strengths/weaknesses for each?

Type	Method	Strengths	Weaknesses
Completely randomized	One primary factor randomly assigned to experimental units	• Easier for large number of treatments • Simple method	• Sometimes low accuracy due to natural variations, if experimental units are in groups (plots of land) • Can only do one type of treatment
Randomized block	Experimental units grouped into blocks according to known or suspected variation	• More accurate than completely randomized	• Possibility of interactions between blocks and treatments • Requires large blocks for large number of treatments
Factorial	Each combination of factors is assigned to an experimental group	• Less experimental error and confounding variables • Can study interactions between factors	• Requires many experimental units if there are many factors

OTHER RESOURCES/READINGS

- Course on experimental design: http://www2.hawaii.edu/~halina/603/
- AirBNB A/B testing: http://nerds.airbnb.com/experiments-at-airbnb/
- T-tests in Python: http://iaingallagher.tumblr.com/post/50980987285/t-tests-in-python

PAIR PROGRAMMING

- Why?
  - Decrease errors
  - Increase communication
  - Bring junior developers up to speed of seniors in the organization
  - Some companies exclusively use pair programming! (here's another link)
- You'll learn how to communicate your ideas
- Can help uncover some problems you may have

- Can work on assignments (W1-W3)