BC COMS 1016:
Intro to Comp Thinking & Data Science

—

Lecture 26 –
Final Project Discussion
Review

—

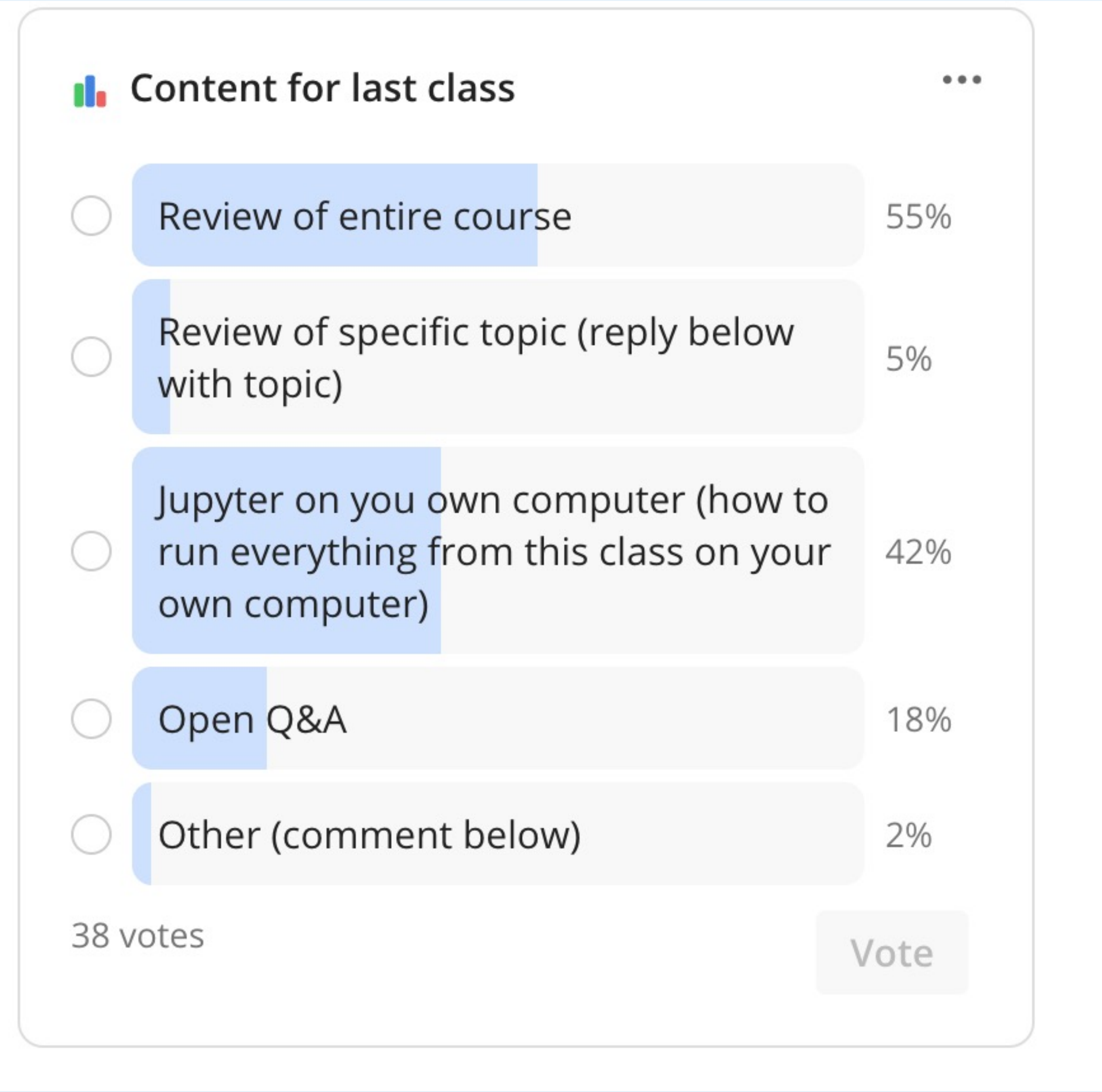
Jupyter on your Computer
Q&A

Announcements

- Homework 10 – Classification
  - Due Monday 05/04
- Course Evaluations:
  - Due XXX
- Project 3:
  - Due Monday 05/04

Outline for Today's Class

- Course Project
- Review
- Jupyter on your computer
- Q&A



Final Project

Final Project – Real World Data Science

- Explore a real world dataset from multiple tables
  - Choose from 6 datasets
- Ask 2 questions that the dataset can help answer
  - Hypothesis Testing
  - Prediction
- Use methods covered in in the class to answer these questions

6 Datasets to choose from

- US-wealth
- airbnb
- contraceptive-data
- example
- fma-analysis
- hr-dataset
- police-scorecard

Final Project

We will provide:

1. An overview and description of the dataset
2. A preview section with code to read in all the datasets relevant to your specific project
3. A Research Report section which contains the outline for the content of your final project.

Final Project - Report Section

1. Introduction:

   250-300 word background

2. Hypothesis Testing and Prediction Questions

   State the questions and how you plan to answer them

3. Exploratory Data Analysis
   1. Visualize!
4. Hypothesis Testing
5. Prediction
6. Conclusion

Final Project - Proposal Deadline – Thursday 05/05

1. Introduction:

   250-300 word background

2. Hypothesis Testing and Prediction Questions

   State the questions and how you plan to answer them

3. Exploratory Data Analysis
   1. Visualize!
4. Hypothesis Testing
5. Prediction
6. Conclusion

The earlier you submit the proposal the better so we can give you more feedback

Final Project – Notes/Recommendations

Randomization

Question: **How do we make our random simulations reproducible?**

Answer: random.np.seed(), random.seed()

Classifiers Evaluation

Evaluating a classifier – Train/test splits

- Split data into training and test splits
- Question: Why?
- Answer: To see how well our classifier generalizes to unseed data

Evaluation metrics

- Scenario: Classifier to predict if a student will pass or fail, binary classification
- Question: What should our features be?
  - Answer: scores on homeworks,projects, labs
- Model performs 95% accuracy on test set
  - Is this a good performace?

Evaluation Metrics

- Model performs 95% accuracy on test set
- Is this good performance?

- Depends:
  - If 90% of students passed, 95% accuracy isn't that great, its not much better than just saying everyone passed (this would get us 90% accuracy)
  - If 50% of the students passed, 95% accuracy seems good

- Final Project: make sure to contextualize your model's performance with a majority baseline (always predicting one of the classes)

Evaluation Metrics

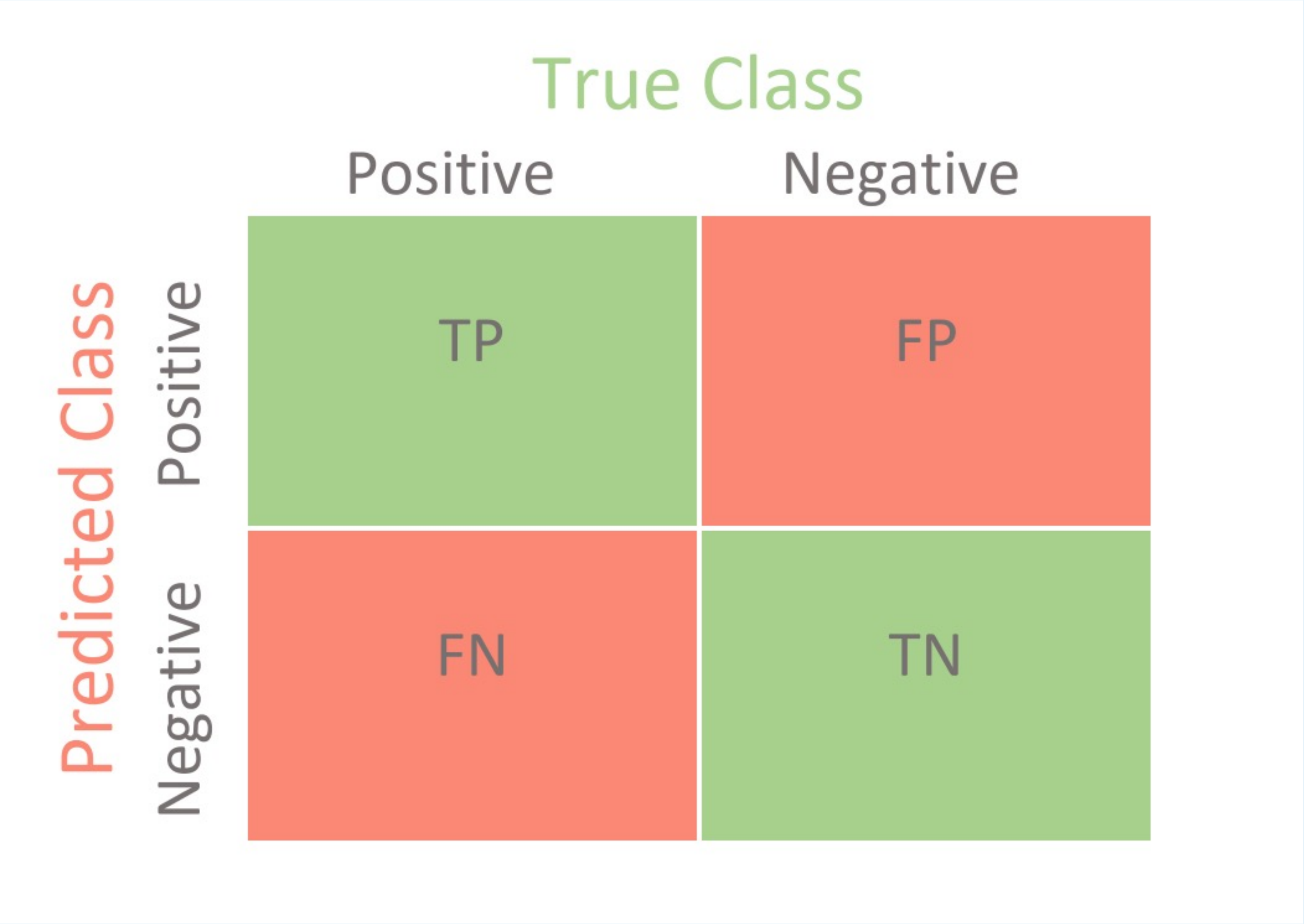


Image from towardsdatascience

Evaluation Metrics

- Accuracy: $\frac{TP}{TP+FP+FN+TN}$
- Precision: $\frac{TP}{TP+FP}$
- Recall/Sensitivity: $\frac{TP}{TP+FN}$

Course Review

Course Outline

- Computation: Python and Tables
- Exploration
  - Discover patterns in data
  - Articulate insights (visualizations)
- Inference
  - Make reliable conclusions about the world
  - Probability & Statistics
- Prediction
  - Informed guesses about unseen data
  - Machine Learning: Regression & Classification

Computation in Python

- Textbook sections
  - General features and Table methods: 3.1 - 9.3, 17.3
  - sample_proportions: 11.1
  - percentile: 13.1
  - np.average, np.mean, np.std: 14.1, 14.2
  - minimize: 15.4

Exploring Data

Describing Data

- Qualitative:
  - Visualizing Distributions: Chapter 7
- Quantitative
  - Center and spread: 14.1-14.3
  - Linear trend and non-linear patterns: 8.1, Chapter 15

Inference

General Concepts

- Study, experiment, treatment, control, confounding, randomization, causation, association: Chapter 2
- Distribution: 7.1, 7.2
- Sampling, probability sample: 10.0
- Probability distribution, empirical distribution, law of averages: Chapter 10
- Population, sample, parameter, statistic, estimate: 10.1, 10.3
- Model: every null and alternative hypothesis; 16.1

Goal of Inference

- To make conclusions about unknown features of the population or model, based on assumptions of randomness

Inference: Estimation

Estimating a Numerical Parameter

- Question: What is the value of the parameter?
- Terms: **predict, estimate, construct a confidence interval, confidence level**
- Answer: Between x and y, with 95% confidence
- Method (13.2, 13.3):
  - **Bootstrap the sample**; compute estimate
  - Repeat; draw empirical histogram of estimates
  - Confidence interval is "middle 95%" of estimates
- Can replace 95% by other confidence level (not 100%)

Meaning of "95% Confidence"

- You'll never get to know whether or not your constructed interval contains the parameter.
- The confidence is in the process that generates the interval.
- The process generates a good interval (one that contains the parameter) about 95% of the time.
- End of 13.2

Reasons to use a confidence interval

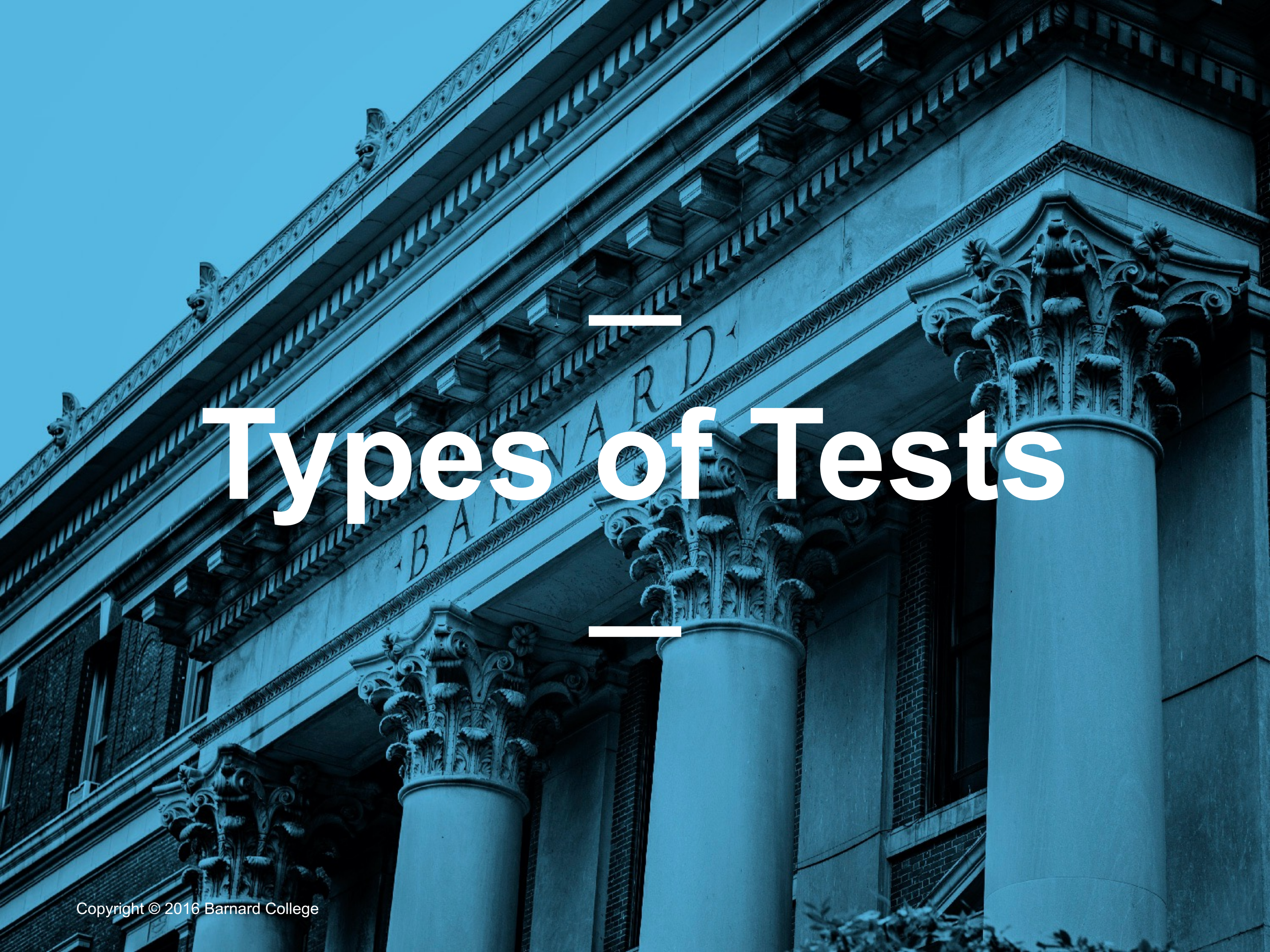
- To **estimate** a numerical parameter: 13.3
  - Regression **prediction**, if regression model holds: Predict *y* based on a new *x*: 16.3

- To **test** whether or not a numerical parameter is equal to a specified value: 13.4
  - In the regression model, used for testing whether the slope of the true line is 0: 16.2

Inference: Testing

Testing Hypotheses

- **Null:** A completely specified chance model, under which you can simulate date.
  - Need to say exactly what is due to chance, and what the hypothesis specifies.
- **Alternative:** The null isn’t true
  - something other than chance is going on; might have a direction
- **Test Statistic:** A statistic that helps decide between the two hypotheses, based on its empirical distribution under the null
- 11.3



Types of Tests

Hypothesis Testing Review

1 Sample: One Category *(e.g. percent of black male jurors)*
- Test Statistic: empirical_percent, abs(empirical_percent - null_percent)
- How to Simulate: sample_proportions(n, null_dist)

1 Sample: Multiple Categories *(e.g. ethnicity distribution of jury panel)*
- Test Statistic: tvd(empirical_dist, null_dist)
- How to Simulate: sample_proportions(n, null_dist)

1 Sample: Numerical Data *(e.g. scores in a lab section)*
- Test Statistic: empirical_mean, abs(empirical_mean - null_mean)
- How to Simulate: population_data.sample(n, with_replacement=False)

2 Samples: Numerical Data *(e.g. birth weights of smokers vs. non-smokers)*
- Test Statistic: group_a_mean - group_b_mean,
  - group_b_mean - group_a_mean, abs(group_a_mean - group_b_mean)
- How to Simulate: empirical_data.sample(with_replacement=False)

The P-value

- The chance, **under the null hypothesis**, that the test statistic comes out equal to the one in the sample or more in the direction of the alternative
- If this chance is small, then:
  - If the null is true, something very unlikely has happened.
  - Conclude that the data support the alternative hypothesis more than they support the null.
- 11.3

Error Probability

- Even if the null is true, your random sample might indicate the alternative, just by chance
- The **cutoff** for P is the chance that your test makes the wrong conclusion when the null hypothesis is true
- Using a small cutoff limits the probability of this kind of error
- 11.4

Prediction

Regression

- Regression model 16.1
- Bootstrap confidence interval for the true slope 16.2
  - Use of this interval to test if the true slope is 0
- Bootstrap prediction interval for *y* at a given value of *x* 16.3

Classification

- Binary classification based on attributes 17.1
  - *k*-nearest neighbor classifiers
- Training and test sets 17.2
  - Why these are needed
  - How to generate them
- Implementation: 17.4
  - Distance between two points
  - Class of the majority of the *k* nearest neighbors
- Accuracy: Proportion of test set correctly classified 17.5

Data Science

Jupyter on your own

Installing Anaconda

- Installing Anaconda - https://www.anaconda.com/products/distribution
- Open up the Anaconda Navigator
- Launch a new notebook
- Install “datascience” package:
  - pip install datascience

More instructions on setting up Jupyter

- Resources from Brian Mailloux:
  - https://edblogs.columbia.edu/eescx3050-001-2015-3/category/classes/class-1-intro/
  - https://youtu.be/FOJG3PxqWV0

Why Data Science

- Unprecedented access to data means that we can make new discoveries and more informed decisions
- Computation is a powerful ally in data processing, visualization, prediction, and statistical inference
- People can agree on evidence and measurement
- Data and computation are everywhere: understanding and interpreting are more important than ever

Limitations of Data Science

- Evidence and measurements are critical ingredients for good decision-making
  - ...but they're not enough by themselves!

- Data science is a powerful complement to qualitative analysis
  - but it's not a replacement!

How to Analyze Data

- Begin with a question from some domain, make reasonable assumptions about the data and a choice of methods.

- Visualize, then quantify!

- *Perhaps the most important part*: Interpretation of the results in the language of the domain, without statistical jargon.

How *Not* to Analyze Data

- Begin with a question from some domain, make reasonable assumptions about the data and a choice of methods.
- Visualize, then quantify!
- *Perhaps the most important part*: Interpretation of the results in the language of the domain, without statistical jargon.

How to Analyze Data after BC COMS 1016

- Begin with a question from some domain, make reasonable assumptions about the data and a choice of methods.
- Visualize, then quantify! **Do both using computation**
- *Perhaps the most important part*: Interpretation of the results in the language of the domain, without statistical jargon.

- Table manipulation using Python
- Working with whole distributions, not just means
- Decisions based on sampling: assessing models
- Estimation based on resampling
- Understanding sampling variability
- Prediction


Thank you!

Course Review
(in depth)

Course Outline

- Computation: Python and Tables
- Exploration
  - Discover patterns in data
  - Articulate insights (visualizations)
- Inference
  - Make reliable conclusions about the world
  - Probability & Statistics
- Prediction
  - Informed guesses about unseen data
  - Machine Learning: Regression & Classification

Computation in Python

- Textbook sections
  - General features and Table methods: 3.1 - 9.3, 17.3
  - sample_proportions: 11.1
  - percentile: 13.1
  - np.average, np.mean, np.std: 14.1, 14.2
  - minimize: 15.4

Exploring Data

Describing Data

- Qualitative:
  - Visualizing Distributions: Chapter 7
- Quantitative
  - Center and spread: 14.1-14.3
  - Linear trend and non-linear patterns: 8.1, Chapter 15

Measures of Center

- Median
  - 50th percentile, where
  - *p*th percentile = smallest value on list that is at least as large as *p*% of the values 13.1
  - Median is not affected by outliers
- Mean/Average
  - Depends on all the values
  - smoothing operation
  - center of gravity of histogram
    - if histogram is skewed, mean is pulled away from median towards the tail

Measure of Spread

- Standard deviation (SD) measures roughly how far the data are from their average

- SD = root mean square of deviations from average

Steps: 5 4 3 2 1

Chebyshev's Bounds

Range	Proportion
average ± 2 SDs	at least 1 - 1/4 (75%)
average ± 3 SDs	at least 1 - 1/9 (88.888...%)
average ± 4 SDs	at least 1 - 1/16 (93.75%)
average ± 5 SDs	at least 1 - 1/25 (96%)

True no matter what the distribution looks like

Bounds and Normal Approximations

Percent in Range	All Distributions	Normal Distributions
Average +- 1 SD	At least 0%	About 68%
Average +- 2 SDs	At least 75%	About 95%
Average +- 3 SDs	At least 88.888…%	About 99.73%

Standard Units Z

"average ± SDs" 14.2

- *z* measures "how many SDs above average"
- Almost all standard units are in the range (-5, 5)
- To convert a value to standard units:

$$z = \frac{\text{value - average}}{\text{SD}}$$

The Correlation Coefficient *r*

- Measures ***linear*** association
- Based on standard units; pure number with no units
- *r* is not affected by changing units of measurement
- $-1 \leq r \leq 1$
- *r* = 0: No linear association; *uncorrelated*
- *r* is not affected by switching the horizontal and vertical axes
- Be careful before you use it
- 15.1

Definition of *r*

Correlation Coefficient (r) =

average of product of standard(x) and standard(y)

Steps: 4 3 2 1

Measures how clustered the scattered data are around a straight line

estimate of $y = r \cdot x$, when both variables are measured in standard units

Slope and Intercept

$$estimate\ of\ y = slope\ * x + intercept$$

$$\boldsymbol{slope\ of\ the\ regression\ line}$$

$$r\ * \frac{SD\ of\ y}{SD\ of\ x}$$

$$\boldsymbol{intercept\ of\ the\ regression\ line}$$

$$mean(y) - slope \times mean(x)$$

Regression Line

- Regression line is the "least squares" line
- Minimizes the root mean squared error of prediction, among all possible lines
- No matter what the shape of the scatter plot, there is one best straight line
  - but you shouldn't use it if the scatter isn't linear
- 15.3, 15.4

Residuals

- Error in regression estimate
- One residual corresponding to each point (*x*, *y*)
- **residual**

  = observed *y* - regression estimate of *y*

  = vertical difference between point and line
- No matter what the shape of the scatter plot:
  - Residual plot does not show a trend
  - Average of residuals = 0

Inference

General Concepts

- Study, experiment, treatment, control, confounding, randomization, causation, association: Chapter 2
- Distribution: 7.1, 7.2
- Sampling, probability sample: 10.0
- Probability distribution, empirical distribution, law of averages: Chapter 10
- Population, sample, parameter, statistic, estimate: 10.1, 10.3
- Model: every null and alternative hypothesis; 16.1

Goal of Inference

- To make conclusions about unknown features of the population or model, based on assumptions of randomness

Probability

- Probability theory:
  - Exact calculations
  - Normal approximation for mean of large random sample
  - Accuracy and sample size

Equally Likely Outcomes

Assuming all outcomes are equally likely, the chance of an event A is:

$$P(A) = \frac{\textit{number of outcomes that make A happen}}{\textit{total number of outcomes}}$$

Large Sample Approximation: CLT

Central Limit Theorem

If the sample is
- large, and
- drawn at random with replacement,

Then, *regardless of the distribution of the population,*

the probability distribution of the sample sum (or of the sample mean) is *roughly* bell-shaped

Inference: Estimation

Estimating a Numerical Parameter

- Question: What is the value of the parameter?
- Terms: **predict, estimate, construct a confidence interval, confidence level**
- Answer: Between x and y, with 95% confidence
- Method (13.2, 13.3):
  - **Bootstrap the sample**; compute estimate
  - Repeat; draw empirical histogram of estimates
  - Confidence interval is "middle 95%" of estimates
- Can replace 95% by other confidence level (not 100%)

Meaning of “95% Confidence”

- You’ll never get to know whether or not your constructed interval contains the parameter.
- The confidence is in the process that generates the interval.
- The process generates a good interval (one that contains the parameter) about 95% of the time.
- End of 13.2

Reasons to use a confidence interval

- To **estimate** a numerical parameter: 13.3
  - Regression **prediction**, if regression model holds: Predict *y* based on a new *x*: 16.3
- To **test** whether or not a numerical parameter is equal to a specified value: 13.4
  - In the regression model, used for testing whether the slope of the true line is 0: 16.2

Inference: Testing

Testing Hypotheses

- **Null:** A completely specified chance model, under which you can simulate date.
  - Need to say exactly what is due to chance, and what the hypothesis specifies.
- **Alternative:** The null isn’t true
  - something other than chance is going on; might have a direction
- **Test Statistic:** A statistic that helps decide between the two hypotheses, based on its empirical distribution under the null
- 11.3

The P-value

- The chance, **under the null hypothesis**, that the test statistic comes out equal to the one in the sample or more in the direction of the alternative
- If this chance is small, then:
  - If the null is true, something very unlikely has happened.
  - Conclude that the data support the alternative hypothesis more than they support the null.
- 11.3

Error Probability

- Even if the null is true, your random sample might indicate the alternative, just by chance
- The **cutoff** for P is the chance that your test makes the wrong conclusion when the null hypothesis is true
- Using a small cutoff limits the probability of this kind of error
- 11.4

Testing Data in Two Categories

- **Null:** The sample was drawn at random from a specified distribution.
- Test statistic: Either count/proportion in one category, or distance between count/proportion and what you'd expect under the null; depends on alternative
- Method:
  - Simulation: Generate samples from the distribution specified in the null.
- 11.1 (Swain v. Alabama, Mendel)

Testing Data in Multiple Categories

- **Null:** The sample was drawn at random from a specified distribution.
- Test statistic: TVD between distribution in sample and distribution specified in the null.
- Method:
  - Simulation: Generate samples from the distribution specified in the null.
- 1.2 (Alameda county juries)

Comparing Two Numerical Samples

- **Null:** The two samples come from the same underlying distribution in the population.
- Test statistic: difference between sample means (take absolute value depending on alternative)
- **Method for A/B Testing:**
  - Permutation under the null: 12.2 (Deflategate), 12.1 (birth weight etc for smokers/nonsmokers), 12.3 (BTA randomized controlled trial)

One Numerical Parameter

- **Null:** parameter = a specified value.
- **Alternative:** parameter ≠ value
- Test Statistic: Statistic that estimates the parameter
- Method:
  - Bootstrap: Construct a confidence interval and see if the specified value is in the interval.
- 13.4, 16.2 (slope of true line)

Causality

- Tests of hypotheses can help decide that a difference is not due to chance
- But they don't say ***why*** there is a difference …
- Unless the data are from an RCT 12.3
  - In that case a difference that's not due to chance can be ascribed to the treatment

Prediction

Regression

- Regression model 16.1
- Bootstrap confidence interval for the true slope 16.2
  - Use of this interval to test if the true slope is 0
- Bootstrap prediction interval for *y* at a given value of *x* 16.3

Classification

- Binary classification based on attributes 17.1
  - *k*-nearest neighbor classifiers
- Training and test sets 17.2
  - Why these are needed
  - How to generate them
- Implementation: 17.4
  - Distance between two points
  - Class of the majority of the *k* nearest neighbors
- Accuracy: Proportion of test set correctly classified 17.5

Data Science

Why Data Science

- Unprecedented access to data means that we can make new discoveries and more informed decisions
- Computation is a powerful ally in data processing, visualization, prediction, and statistical inference
- People can agree on evidence and measurement
- Data and computation are everywhere: understanding and interpreting are more important than ever

Limitations of Data Science

- Evidence and measurements are critical ingredients for good decision-making
  - ...but they're not enough by themselves!
- Data science is a powerful complement to qualitative analysis
  - but it's not a replacement!

How to Analyze Data

- Begin with a question from some domain, make reasonable assumptions about the data and a choice of methods.
- Visualize, then quantify!
- *Perhaps the most important part*: Interpretation of the results in the language of the domain, without statistical jargon.

How *Not* to Analyze Data

- Begin with a question from some domain, make reasonable assumptions about the data and a choice of methods.
- Visualize, then quantify!
- *Perhaps the most important part*: Interpretation of the results in the language of the domain, without statistical jargon.

How to Analyze Data after BC COMS 1016

- Begin with a question from some domain, make reasonable assumptions about the data and a choice of methods.
- Visualize, then quantify! **Do both using computation**
- *Perhaps the most important part*: Interpretation of the results in the language of the domain, without statistical jargon.

- Table manipulation using Python
- Working with whole distributions, not just means
- Decisions based on sampling: assessing models
- Estimation based on resampling
- Understanding sampling variability
- Prediction


Thank you!