DATA8

Berkeley
UNIVERSITY OF CALIFORNIA

Summer 2021

Course Overview

Conclusion

Big Picture of Course Contents

1. Computation: Python and tables
2. Describing data
   a. By visualizing
   b. By quantifying
3. Probability
4. Inference
5. Prediction

Review In Detail

2. Computation

1. Computation

- 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

1. Computation: Big Ideas

- How to organize data: tables
- How to work with data: table operations -- the next step beyond Excel
- How to compute: solve complex tasks by combining small building blocks

- Data frames (R, pandas) = tables
- Databases = tables + table operations + concurrency
- Spark = tables + table operations, for big data

2. Describing Data

2. Describing Data

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

2. Describing Data: Big Ideas

- Visualization for exploring data and forming hypotheses: can we spot any patterns or trends? what questions does it raise?
- Statistics (mean, median, SD, ...) summarize data
- Also, helps us communicate to others

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 of 5, 7, 8, 8 = (5+7+8+8)/4 14.1

  = 5*0.25 + 7*0.25 + 8*0.5
- Mean 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)

=

root 5	mean 4	square of 3	deviations from 2	average 1

Measures roughly how far off the values are from average

- 14.2

Chebychev’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 ± z SDs	at least 1 - 1/z^2

no matter what the distribution looks like

How Big are Most of the Values?

No matter what the shape of the distribution,
the bulk of the data are in the range “average ± a few SDs”

If a histogram is bell-shaped, then

- the SD is the distance between the average and the points of inflection on either side
- Almost all of the data are in the range “average ± 3 SDs”

14.2, 14.3

Bounds and normal approximations

Percent in Range	All Distributions	Normal Distribution
average $\pm$ 1 SD	at least 0%	about 68%
average $\pm$ 2 SDs	at least 75%	about 95%
average $\pm$ 3 SDs	at least 88.888...%	about 99.73%

Standard Units z

"average ± *z* 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:

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

Definition of *r*

Correlation Coefficient (*r*) =

average of	product of	*x* in standard units	and	*y* in standard units

Measures how clustered the scatter is around a straight line

The Correlation Coefficient *r*

- Measures ***linear*** association
- $-1 \leq r \leq 1$
- $r = 0$: No linear association; *uncorrelated*
- Be careful before you use it
- 15.1

3. Probability

3. Probability

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

Equally Likely Outcomes

- **If all outcomes are assumed equally likely**, then probabilities are proportions of outcomes:

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

$$= \text{proportion of outcomes that make A happen}$$

- 9.5

Probability: Exact Calculations

- Probabilities are between 0 (impossible) and 1 (certain)
- P(event happens) = 1 - P(the event doesn't happen)
- Chance that two events *A* and *B* both happen

= P(*A* happens) x P(*B* happens given that *A* has happened)

- If event *A* can happen in *exactly one* of two ways, then

  P(*A*) = P(first way) + P(second way)
- 9.5

Conditional Probabilities

- Start with **prior probabilities** of two classes; priors can be **subjective**
- Known: **likelihood** of data, given each of the classes
- Acquire data according to these likelihoods
- Update the prior probabilities by finding **posterior probabilities** of the two classes, **given the data**
- Tree diagrams and **Bayes' Rule:** 18.1, 18.2

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

Random Sample Mean

- Fix a sample size
- Draw ***all possible random samples*** of that size
- Compute the mean of each sample
- You'll end up with a lot of means
- The distribution of those is the *probability distribution of the sample mean*
- It's centered at the population mean
- SD = (population SD)/√(sample size) 14.5
- If the sample is large, it's roughly bell shaped by CLT

Accuracy of Random Sample Mean

- Greater if SD of sample mean is smaller
- Doesn’t depend on population size
- Increases as sample size increases, because SD of sample mean decreases
- For 3 times the accuracy, you have to multiply the sample size by a factor of $3^2 = 9$
- **Square Root Law:** If you multiply sample size by a factor, accuracy goes up by the square root of the factor
- 14.5

Application to Proportions

- Fact: **SD of 0-1 population ≤ 0.5** 14.6
- Total width of 95% CI for population proportion:

  = 4 SDs of the sample proportion

  $= 4 \times (\text{SD of 0-1 population})/\sqrt{\text{(sample size)}}$

  $\leq 4 \times 0.5/\sqrt{\text{(sample size)}}$

  $= 2 / \sqrt{\text{(sample size)}}$

- So if you know the desired width of the interval, you can solve for (an overestimate of) the sample size

4. Inference

4. 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

4. Inference: Big Ideas

- Suppose you found what looks like a pattern in the data. Does it reflect something real about the world? Or could it be due to just chance?
- Random sampling is our friend: it ensures we can draw inferences about population from the sample
  - Law of large numbers: the distribution (shape of histogram) of a sample will look similar to the distribution (shape of histogram) of the population

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; it rarely goes awry.
- The process generates a good interval (one that contains the parameter) about 95% of the time.
- End of 13.2

Main Uses of Confidence Intervals

- 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

Tests of 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 you 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

An 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

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)

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.

- 11.2 (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

- A hypothesis test can help determine whether a difference or association is due to chance
- But it won't say ***why*** there is a difference …
- Unless the data are from an RCT 12.3
    - In that case we can infer that the treatment causes the difference

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

Regression to the Mean

- **estimate of $y = r \cdot x$**, when both variables are measured in standard units
- If $r = 0.6$, and the given x is 2 standard units, then:
  - The given x is 2 SDs above average
  - The prediction for y is 1.2 SDs above average
- On average (though not for each individual), regression predicts y to be closer to the mean than x is
- 15.2

Regression Estimate, Method I

A course has a midterm (average 70; standard deviation 10) and a really hard final (average 50; standard deviation 12)

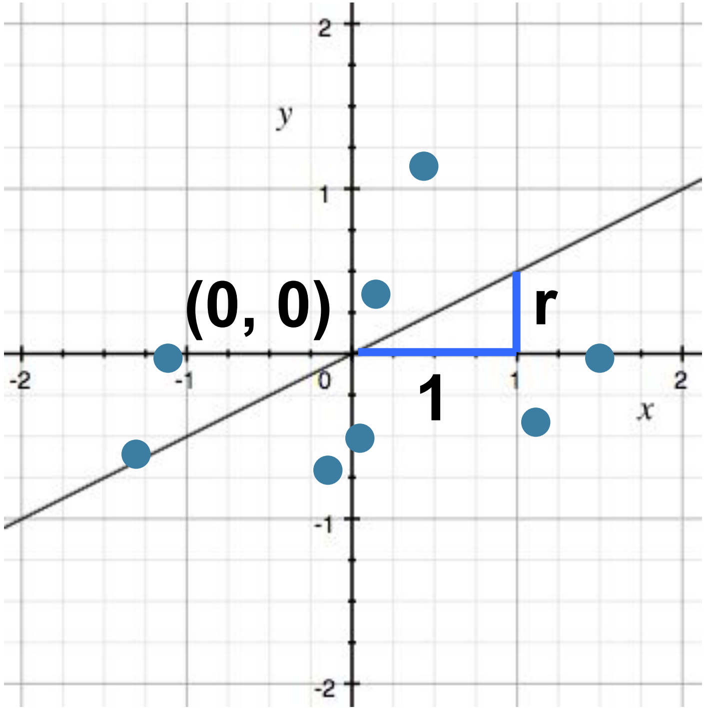
If the scatter of midterm & final scores for students looks like a typical oval with correlation 0.75, then...

What do you expect the average final score would be for a student who scored 90 on the midterm?

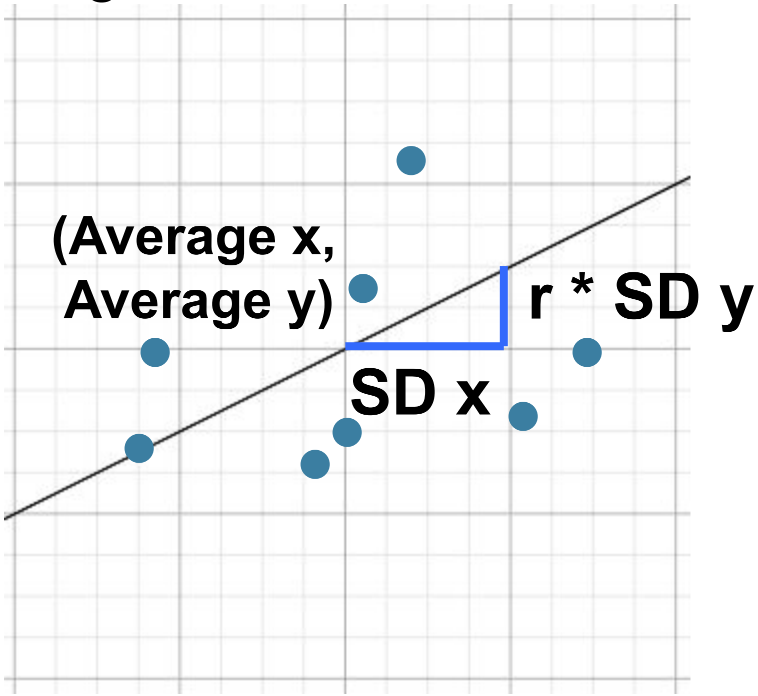
2 standard units on midterm,
so estimate 0.75 * 2 = 1.5 standard units on final.
So estimated final score = 1.5 * 12 + 50 = 68 points

Regression Line

Standard Units



Original Units



Slope and Intercept

estimate of y = slope * x + intercept

$$\textbf{slope of the regression line} = r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

$$\textbf{intercept of the regression line} = \text{average of } y - \text{slope} \cdot \text{average of } x$$

- 15.2

Regression Estimate, Method II

The equation of a regression line for estimating child's height based on midparent height is

estimated child's height = $0.64 \cdot$ midparent height + 22.64

Estimate the height of someone whose midparent height is 69 inches.

0.64*69 + 22.64 = 66.8 inches

Least Squares

- 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

$$\text{SD of residuals} = \sqrt{1 - r^2} \times \text{SD of } y$$

15.5, 15.6

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