YData: An Introduction to Data Science

Lecture 39: Review

Elena Khusainova & John Lafferty
Statistics & Data Science, Yale University
Spring 2021

Credit: data8.org

Announcements

- TAs will have review sessions next week
- TAs and instructors will have office hours next week
- pandas won't be on the final exam

Review

1. Distributions
2. Regression
3. Probability
4. Central Limit Theorem
5. Inference: estimating parameters
6. Hypotheses testing
7. Classification

Distributions

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 (see Chapter 13.1)
- Median is not affected by outliers
- Mean of 5, 7, 8, 8 = (5+7+8+8)/4 (see Chapter 14.1)
  = 5*0.25 + 7*0.25 + 8*0.5
- Mean depends on all the values; 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	mean	square of	deviations from	average
5	4	3	2	1

Measures roughly how far off the values are from average

(see Chapter 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 (see Chapter 14.2)

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 $\pm$ 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 $\pm$ 3 SDs"

(see Chapter 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%

(see Chapter 14.3)

Regression

Standard Units z

"average $\pm$ z SDs" (see Chapter 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

(see Chapter 15.1)

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
- (see Chapter 15.1)

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
- (see Chapter 15.2)

Regression Estimate, Method I

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

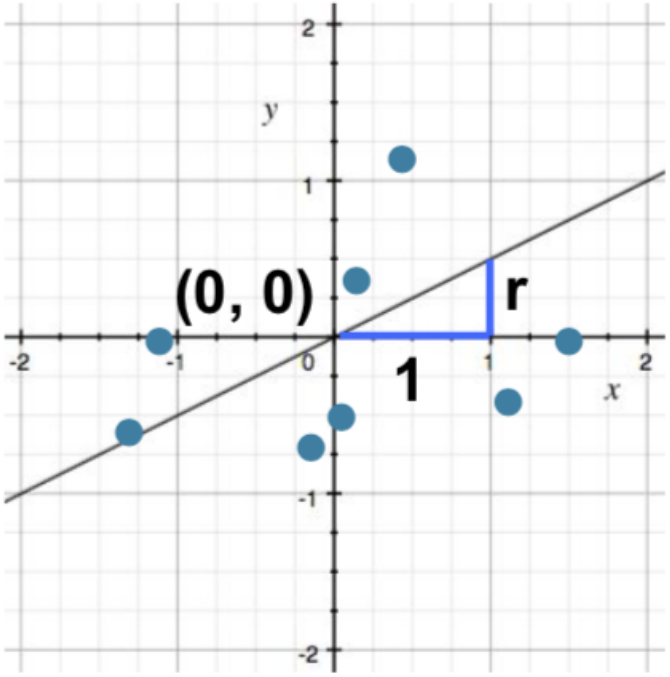
If the scatter of midterm & final scores for students looks like a typical oval with correlation 0.75, then...

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

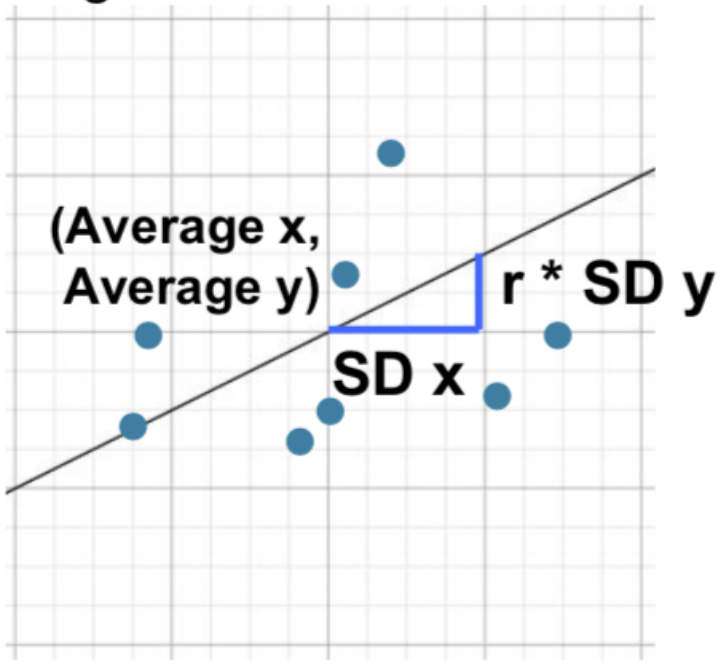
90 points corresponds to 2 standard units on midterm,
so estimate 0.75 * 2 = 1.5 standard units on final.
So estimated final score = 1.5 * 12 + 50 = 68 points

Regression Line

Standard Units



Original Units



Slope and Intercept

$$\text{estimate of } y = \text{slope} \times x + \text{intercept}$$

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

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

(see Chapter 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 $= 0.64 \times$ 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
- (see Chapter 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:
  - Average of residuals = 0

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

(see Chapter 15.5, 15.6)

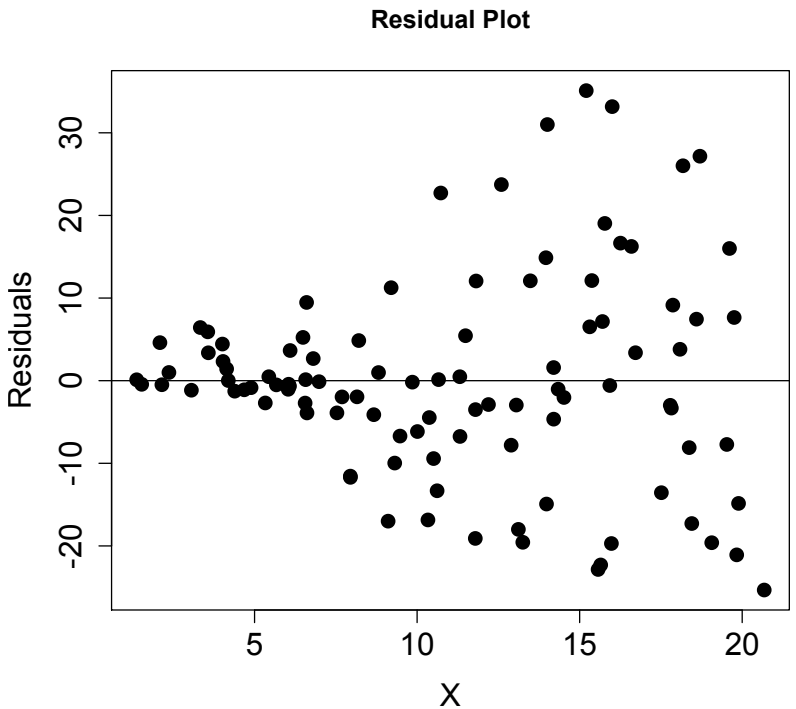
Discussion question

The least-squares regression line is

a. the line that makes the sum of the squares of the vertical distances of the data points to the line as small as possible
b. the line that best splits the data in half, with half of the points above the line and half below the line.
c. the line that makes the square of the correlation in the data as large as possible.
d. all of the above
e. a and b
f. a and c
g. b and c

Discussion question

Suppose a Least-squares linear model was fit on explanatory variable X and response variable Y, with the residuals plotted in the figure below against X. What linear model assumption appears to be violated given the residual plot below?



Probability

Equally Likely Outcomes

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

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

(see Chapter 9.5)

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) $\times$ 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)

(see Chapter 9.5)

Updating 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**: (see Chapter 18.1, 18.2)

Central Limit Theorem

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

(see Chapter 14.4)

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)/$\sqrt{\text{(sample size)}}$
- If the sample is large, it's roughly bell shaped by CLT

(see Chapter 14.5)

Accuracy of Random Sample Mean

- Greater if SD of sample mean is smaller
- Does not 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

(see Chapter 14.5)

Application to Proportions

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

  $= 4$ SDs of the sample proportion
  $= 4 \times$ (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

(see Chapter 14.6)

Inference

Inference

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

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** (see Chapter 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.
- (see end of Chapter 13.2)

Hypotheses testing
Regression Inference

Hypothesis testing

- **Null hypothesis**
  - A well defined chance model about how the data were generated
  - We can simulate data under the assumptions of this model – "under the null hypothesis"
- **Alternative hypothesis**
  - A different view about the origin of the data
- **Test Statistic**
  - A statistic that helps you decide between the two hypotheses, based on its empirical distribution under the null
- (see Chapter 11.3)

Prediction Under the Null Hypothesis

- Simulate the test statistic under the null hypothesis; draw the histogram of the simulated values
- This displays the **empirical distribution of the statistic under the null hypothesis**
- It is a prediction about the statistic, made by the null hypothesis
  - It shows all the likely values of the statistic
  - Also how likely they are (assuming the null hypothesis is true)

Resolve choice between null and alternative hypotheses

- Compare the **observed test statistic** and its empirical distribution under the null hypothesis
- If the observed value is not consistent with the distribution, then the test favors the alternative – "rejects the null hypothesis"

Discussion question

In a hypothesis test about an unknown parameter, the test statistic...

a. is the value of the unknown parameter under the null hypothesis.
b. measures the compatibility between the null and alternative hypotheses.
c. is the value of the unknown parameter under the alternative hypothesis.
d. measures the compatibility between the null hypothesis and data.
e. None of the above.

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.

- (see Chapter 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

- (see Chapter 11.3)

Conventions About Inconsistency

- **"Inconsistent"**: The test statistic is in the tail of the empirical distribution under the null hypothesis

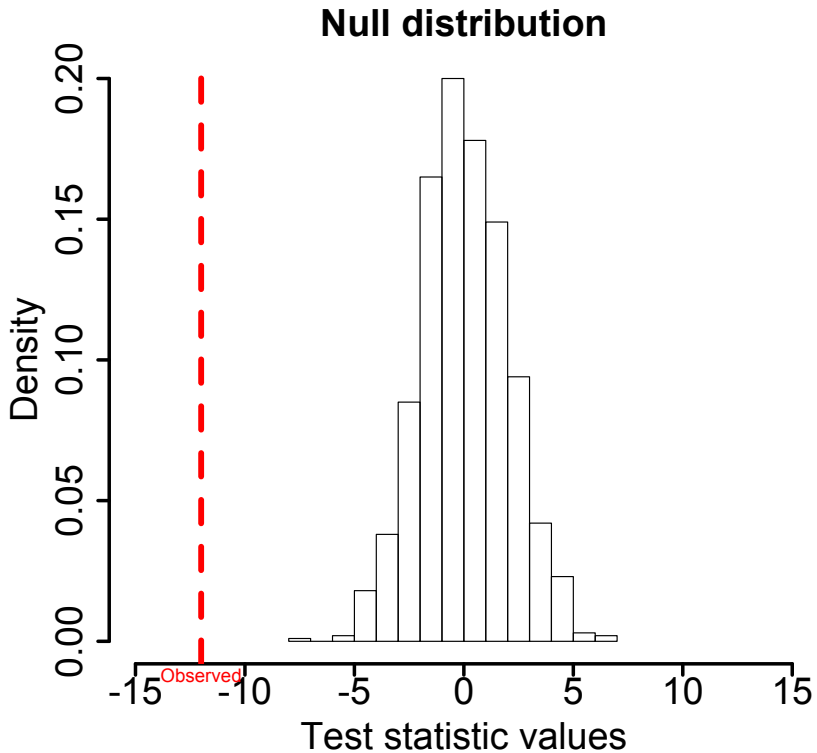
- **"In the tail," first convention**:
  - The area in the tail is less than 5%
  - The result is "statistically significant"

- **"In the tail," second convention**:
  - The area in the tail is less than 1%
  - The result is "highly statistically significant"

Hypothesis testing: illustrations

- Null hypothesis: **Population average = 0**
  (Or some other assumption about the population)
  Recall Swain vs. Alabama, Mendel purple flowering plan (Lec 16), or Jury Selection in Alameda County (Lec 17)

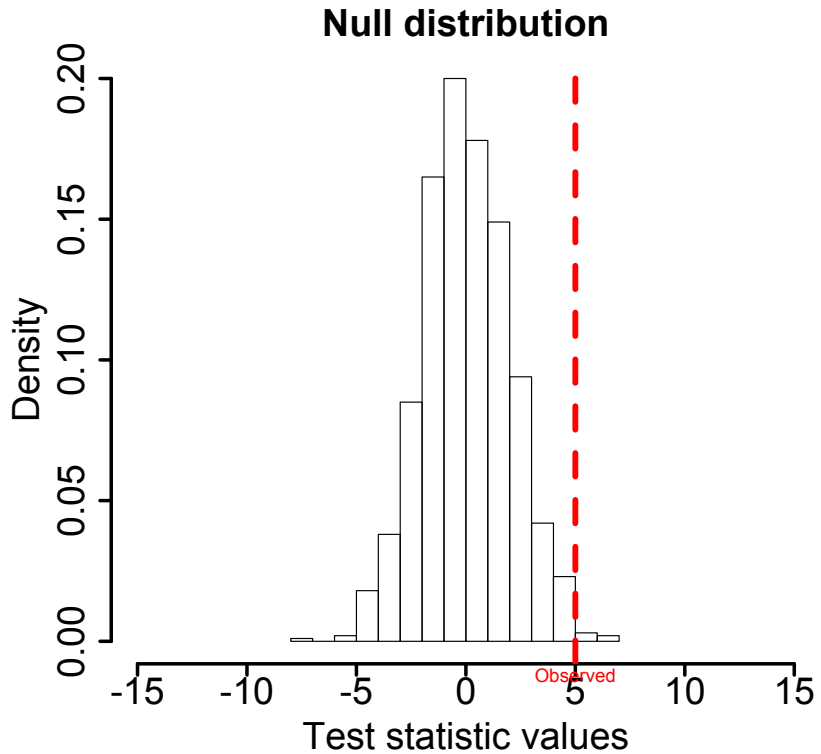
- Alternative hypothesis: **Population average $<$ 0**



Hypothesis testing: illustrations

- Null hypothesis: **Population average = 0**
  (Or some other assumption about the population)
  Recall Swain vs. Alabama, Mendel purple flowering plan (Lec 16), or Jury Selection in Alameda County (Lec 17)

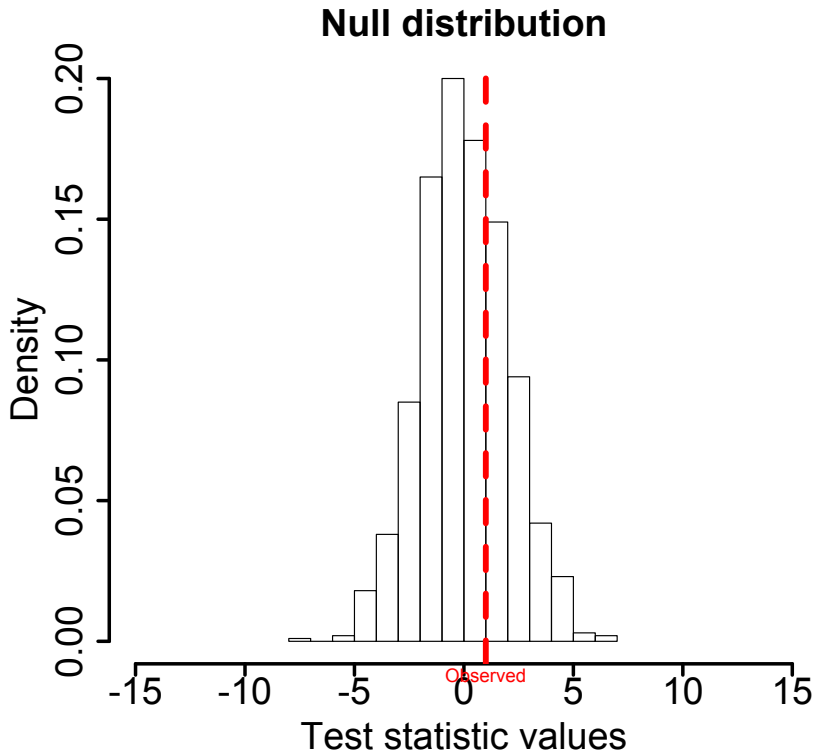
- Alternative hypothesis: **Population average > 0**



Hypothesis testing: illustrations

- Null hypothesis: **Population average = 0**
  (Or some other assumption about the population)
  Recall Swain vs. Alabama, Mendel purple flowering plan (Lec 16), or Jury Selection in Alameda County (Lec 17)

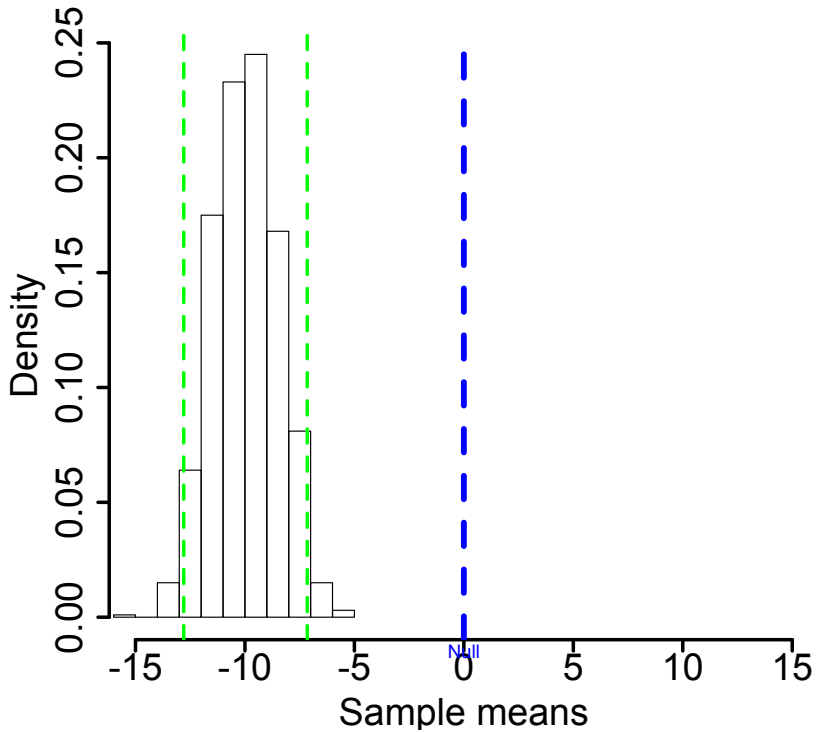
- Alternative hypothesis: **Population average $>$ 0**



Hypothesis testing with confidence intervals

- Null hypothesis: **Population average = 0**
- Alternative hypothesis: **Population average $\neq$ 0**

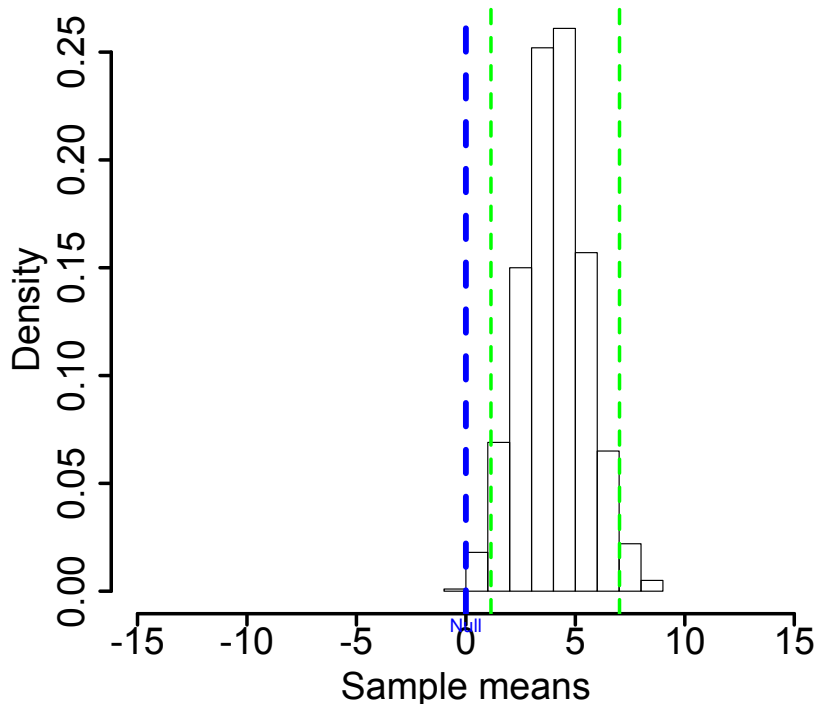
Vertical green dashed lines indicate approximate 95% confidence bounds using bootstrap samples.



Hypothesis testing with confidence intervals

- Null hypothesis: **Population average = 0**
- Alternative hypothesis: **Population average $\neq$ 0**

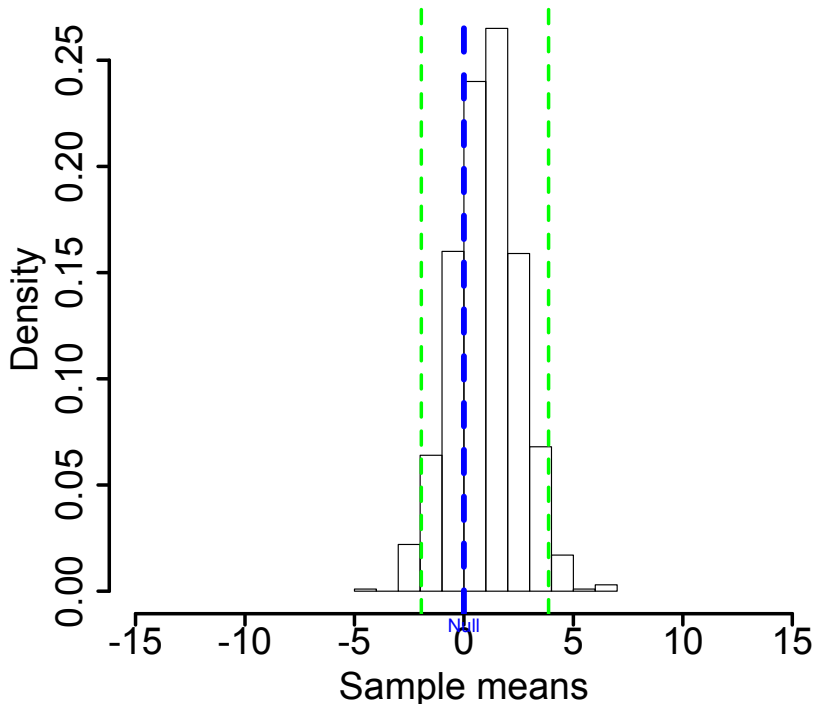
Vertical green dashed lines indicate approximate 95% confidence bounds using bootstrap samples.



Hypothesis testing with confidence intervals

- Null hypothesis: **Population average = 0**
- Alternative hypothesis: **Population average $\neq$ 0**

Vertical green dashed lines indicate approximate 95% confidence bounds using bootstrap samples.



Using a CI for Testing

- Null hypothesis: **Population average = x**
- Alternative hypothesis: **Population average $\neq$ x**
- Cutoff for P-value: $p\%$
- Method:
  - Construct a (100-p)% confidence interval for the population average
  - If x is not in the interval, reject the null
  - If x is in the interval, can't reject the null

Discussion question

If we only have a 90% confidence interval for the population mean (μ), which is $(-0.2, 0.8)$. Based on this interval, we wish to test the hypotheses $H_0 : \mu = 0$ *vs.* $H_a : \mu \neq 0$ at a p-value cutoff of $\alpha = 0.05$. Determine which of the following statement is true.

a. We cannot make any decision since the confidence level we used to calculate the confidence interval is 90%, and we would need a 95% confidence interval.

b. We do not reject H_0, because the value 0 falls in the 90% confidence interval.

c. We reject H_0, because the value 0 falls in the 90% confidence interval.

d. We cannot make a decision since the confidence interval is so wide.

e. None of the above

Discussion question

A physics instructor is convinced that every test he writes has a population mean score of 78 ($\mu = 78$). Students who have enrolled in the course do not believe him, but are not sure if the population mean score is above or below 78. Suppose a random sample of students was taken from his large lecture course, and a 95% confidence interval was found to be [70.864, 77.136].

(a) State a null and alternative hypothesis test.

(b) Given the confidence interval provided, what would be your conclusion to the hypothesis test specified at the $\alpha = 0.05$ level of significance?

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.

- (see Chapter 11.1)

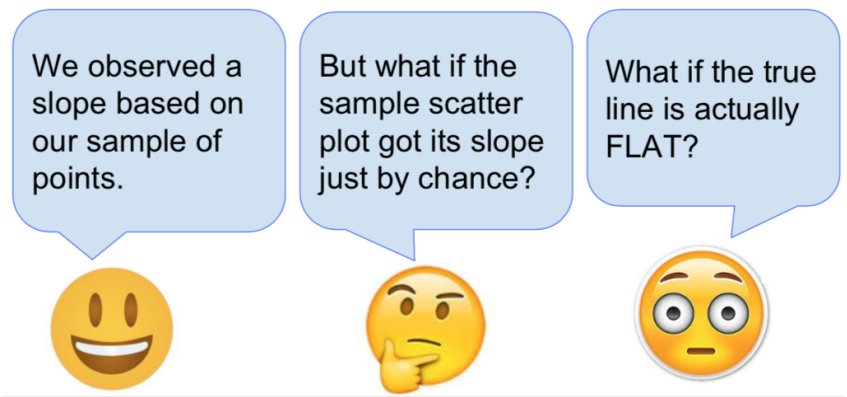
Data in Multiple Categories

- **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.

- (see Chapter 11.2)

Test Whether There Really is a Slope

- **Null hypothesis**: The slope of the true line is 0.
- **Alternative hypothesis**: No, it's not.
- Method:
  - Construct a bootstrap confidence interval for the true slope.
  - If the interval doesn't contain 0, reject the null hypothesis.
  - If the interval does contain 0, there isn't enough evidence to reject the null hypothesis.



Confidence Interval for True Slope

- **Bootstrap the scatter plot**
- **Find the slope of the regression line through the bootstrapped plot.**
- Repeat the two steps above many times
- Draw the empirical histogram of all the predictions.
- Get the "middle 95%" interval.
- That's an approximate 95% confidence interval for the slope of the true line.

A/B testing: Comparing Two Samples

- Previously, we only considered data from a single group

- Compare values of sampled individuals in Group A with values of sampled individuals in Group B.
  $\rightarrow$ Question: Do the two sets of values come from the same underlying distribution?
  $\rightarrow$ Answering this question by performing a statistical test is called A/B testing.

  Examples:
  (A) Birth weights of babies of mothers who smoked during pregnancy
  (B) Birth weights of babies of mothers who didn't

  (A) Control group
  (B) Treatment group

  Deflategate

A/B testing: Simulating Under the Null

- If the null is true, all rearrangements of the birth weights among the two groups are equally likely

- Plan:
  - Shuffle all the birth weights
  - Assign some to "Group A" and the rest to "Group B", maintaining the two sample sizes
  - Find the difference between the averages of the two shuffled groups
  - Repeat

Classification

Classification

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