Unit 8: Final Review

1. Final Exam Review

Sta 101 - Spring 2019

Duke University, Department of Statistical Science

Dr. Ellison

Slides posted at
https://www2.stat.duke.edu/courses/Spring19/sta101.001/

Outline

Coming up…

- Don't forget to **ask/answer 2 questions on Piazza** before the final exam… part of your participation grade!
- Monday and Wednesday Lectures: Final Exam Review
- TA office hours officially end after 4/24!
- Extra Review Session 4/27 Saturday 1-3pm Old Chem 203
- **Final Exam:**
  - Monday 4/29 2:00pm-5:00pm
  - French Science 2231

Final Exam

- **Material Covered:**
  - Units 1-7
  - Slightly higher emphasis on Units 6 and 7
- **What to bring:**
  - Cheat sheet
    - 1 page (8.5" by 11")
    - Front/back ok
    - CAN be typed
  - Calculator (no phones)
- **Provided:**
  - Z-tables, t-tables, Chi-Squared-tables
- **Exam**
  - 10 MC (worth 10 points)
  - 10 TF (worth 20 points)
  - 5 fill in the blank (10 points)
  - 5 matching (5 points)
  - 5 Short answers (55 points)

1

Final Review Suggestions

- **Short answer *review*:**
  - Make sure you understand how to do the application exercises.
  - Review Problem Sets (graded)
- **Short answer *practice*:**
  - Practice test
  - Suggested practice problems in the book
  - Lab Review tomorrow

1

Final Review Suggestions

- **Concept review:**
  - Video notes:
  - Lecture slides **(has material not in the videos/book)**:
  - Readiness Assessments+Performance Assessments:
    - Why are all the other options wrong?
  - What to think about (among other things):
    - Interpretations of analyses (WORDING IS IMPORTANT)
    - Conclusions we would make (WORDING IS IMPORTANT)
    - Relationships between different analyses
    - Know exact definitions
    - FOCUS ON THE WHY BEHIND ANALYSES
    - If there's an equation/analysis, make sure you know how to put that equation/analysis into words in the context of the problem.
    - Common misconceptions (lecture notes)
    - What test to use under certain conditions

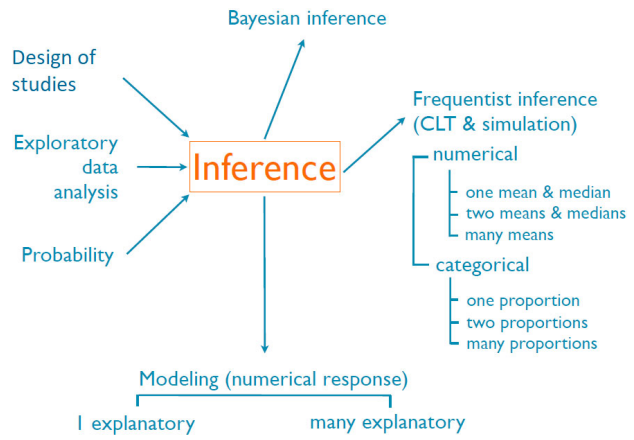
1

Outline

1. Housekeeping

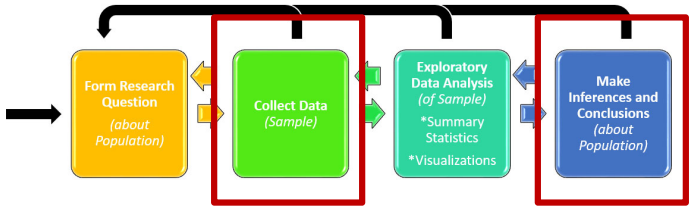
2. Review

Outline



Outline

Course Roadmap



How does our data collection process change the type of conclusions we can make?

What are good ways to collect data?

Case Study: How many issues are there with this analysis and conclusion below?

Students at a local magnate high school want to test if drinking caffeine before the SAT causes a student to score higher on the exam. 80 students that attend the magnate school took the SAT at the testing facility one morning. They decide to assign the first 40 students that arrived at the testing facility to the "coffee group" and offer them coffee before the exam. They assigned the last 40 students to arrive at the testing facility to the "no coffee group" and did not offer them any coffee.

After scores were released, they were able to collect 36 SAT scores from the "coffee group" and 31 SAT scores from the "no coffee group." They conducted an independent means test between the two groups using this data and found a p-value=0.03. **They used this to conclude that drinking coffee before the SAT causes high school students to score higher on the SAT.**

a) This analysis and conclusion is fine.
b) 1 issue
c) ≤ 3 issues
d) ≤ 5 issues
e) >5 issues

2

Case Study: How many issues are there with this analysis and conclusion below?

Students at a local magnate high school want to test if drinking caffeine before the SAT causes a student to score higher on the exam. 80 students that attend the magnate school took the SAT at the testing facility one morning. They decide to assign the first 40 students that arrived at the testing facility to the "coffee group" and offer them coffee before the exam. They assigned the last 40 students to arrive at the testing facility to the "no coffee group" and did not offer them any coffee.

After scores were released, they were able to collect 36 SAT scores from the "coffee group" and 31 SAT scores from the "no coffee group." They conducted an independent means test between the two groups using this data and found a p-value=0.03. **They used this to conclude that drinking coffee before the SAT causes high school students to score higher on the SAT.**

A lot! Let's unpack all the issues with this inference they are making!

2

Is this an observational study or a random experiment?

Students at a local magnate high school want to test if drinking caffeine before the SAT causes a student to score higher on the exam. 80 students that attend the magnate school took the SAT at the testing facility one morning. They decide to assign the first 40 students that arrived at the testing facility to the "coffee group" and offer them coffee before the exam. They assigned the last 40 students to arrive at the testing facility to the "no coffee group" and did not offer them any coffee.

After scores were released, they were able to collect 36 SAT scores from the "coffee group" and 31 SAT scores from the "no coffee group." They conducted an independent means test between the two groups using this data and found a p-value=0.03. They used this to conclude that drinking coffee before the SAT causes high school students to score higher on the SAT.

a.) observational study
b.) random experiment

2

Is this an observational study or a random experiment?

Students at a local magnate high school want to test if drinking caffeine before the SAT causes a student to score higher on the exam. 80 students that attend the magnate school took the SAT at the testing facility one morning. They decide to assign the first 40 students that arrived at the testing facility to the "coffee group" and offer them coffee before the exam. They assigned the last 40 students to arrive at the testing facility to the "no coffee group" and did not offer them any coffee.

After scores were released, they were able to collect 36 SAT scores from the "coffee group" and 31 SAT scores from the "no coffee group." They conducted an independent means test between the two groups using this data and found a p-value=0.03. They used this to conclude that drinking coffee before the SAT causes high school students to score higher on the SAT.

a.) observational study
b.) random experiment

Is this an observational study or a random experiment?

Students at a local magnate high school want to test if drinking caffeine before the SAT causes a student to score higher on the exam. 80 students that attend the magnate school took the SAT at the testing facility one morning. They decide to **assign** the first 40 students that arrived at the testing facility to the "coffee group" and offer them coffee before the exam. They assigned the last 40 students to arrive at the testing facility to the "no coffee group" and did not offer them any coffee.

After scores were released, they were able to collect 36 SAT scores from the "coffee group" and 31 SAT scores from the "no coffee group." They conducted an independent means test between the two groups using this data and found a p-value=0.03. They used this to conclude that drinking coffee before the SAT causes high school students to score higher on the SAT.

a.) observational study
b.) random experiment

Why?
- They use **assignment** to groups, but the assignment is *not RANDOM*!
- A random experiment, must have **random assignment** of subjects to groups.

Thus, can we make the conclusion below, then?

Students at a local magnate high school want to test if drinking caffeine before the SAT causes a student to score higher on the exam. 80 students that attend the magnate school took the SAT at the testing facility one morning. They decide to **assign** the first 40 students that arrived at the testing facility to the "coffee group" and offer them coffee before the exam. They assigned the last 40 students to arrive at the testing facility to the "no coffee group" and did not offer them any coffee.

After scores were released, they were able to collect 36 SAT scores from the "coffee group" and 31 SAT scores from the "no coffee group." They conducted an independent means test between the two groups using this data and found a p-value=0.03. They used this to conclude that drinking coffee before the SAT causes high school students to score higher on the SAT.

a.) yes
b.) no

Thus, can we make the conclusion below, then?

Students at a local magnate high school want to test if drinking caffeine before the SAT causes a student to score higher on the exam. 80 students that attend the magnate school took the SAT at the testing facility one morning. They decide to **assign** the first 40 students that arrived at the testing facility to the "coffee group" and offer them coffee before the exam. They assigned the last 40 students to arrive at the testing facility to the "no coffee group" and did not offer them any coffee.

After scores were released, they were able to collect 36 SAT scores from the "coffee group" and 31 SAT scores from the "no coffee group." They conducted an independent means test between the two groups using this data and found a p-value=0.03. They used this to conclude that drinking coffee before the SAT **causes** high school students to score higher on the SAT.

a.) yes
b.) no

One reason why:
- Can't use causal language in your conclusion, because this is **not a random experiment!**

Issue # 1

Thus, can we make the conclusion below, then?

Students at a local magnate high school want to test if drinking caffeine before the SAT causes a student to score higher on the exam. 80 students that attend the magnate school took the SAT at the testing facility one morning. They decide to **assign** the first 40 students that arrived at the testing facility to the "coffee group" and offer them coffee before the exam. They assigned the last 40 students to arrive at the testing facility to the "no coffee group" and did not offer them any coffee.

After scores were released, they were able to collect 36 SAT scores from the "coffee group" and 31 SAT scores from the "no coffee group." They conducted an independent means test between the two groups using this data and found a p-value=0.03. They used this to conclude that drinking coffee before the SAT **causes** high school students to score higher on the SAT.

a.) yes
b.) no

One reason why:

- Can't use causal language in your conclusion, because this is **not a random experiment!**

Intuition:

- Students that arrive early to the exam might be more likely score higher (personality type). Without random assignment, we can't isolate the single effect of coffee over other potential **confounding factors**.

Issue # 1

2

Let's change the study slightly to be a random experiment. Can we make the conclusion below now?

Students at a local magnate high school want to test if drinking caffeine before the SAT causes a student to score higher on the exam. 80 students that attend the magnate school took the SAT at the testing facility one morning. They decide to **randomly assign** the 40 students that arrived "coffee group" and offer them coffee before the exam. They **randomly assigned** the other 40 students to the "no coffee group" and did not offer them any coffee.

After scores were released, they were able to collect 36 SAT scores from the "coffee group" and 31 SAT scores from the "no coffee group." They conducted an independent means test between the two groups using this data and found a p-value=0.03. They used this to conclude that drinking coffee before the SAT causes high school students to score higher on the SAT.

a.) yes
b.) no

2

Let's change the study slightly to be a random experiment. Can we make the conclusion below now?

Students at a local magnate high school want to test if drinking caffeine before the SAT causes a student to score higher on the exam. 80 students that attend the magnate school took the SAT at the testing facility one morning. They decide to **randomly assign** the 40 students that arrived "coffee group" and offer them coffee before the exam. They **randomly assigned** the other 40 students to the "no coffee group" and did not offer them any coffee.

After scores were released, they were able to collect 36 SAT scores from the "coffee group" and 31 SAT scores from the "no coffee group." They conducted an independent means test between the two groups using this data and found a p-value=0.03. They used this to conclude that drinking coffee before the SAT causes high school students to score higher on the SAT.

a.) yes
b.) no

2

Let's change the study slightly to be a random experiment. Can we make the conclusion below now?

Students at a local magnate high school want to test if drinking caffeine before the SAT causes a student to score higher on the exam. 80 students that attend the magnate school took the SAT at the testing facility one morning. They decide to **randomly assign** the 40 students that arrived "coffee group" and offer them coffee before the exam. They **randomly assigned** the other 40 students to the "no coffee group" and did not offer them any coffee.

After scores were released, they were able to collect 36 SAT scores from the "coffee group" and 31 SAT scores from the "no coffee group." They conducted an independent means test between the two groups using this data and found a p-value=0.03. They used this to conclude that drinking coffee before the SAT causes high school students to score higher on the SAT.

a.) yes
b.) no

Some reasons why:

- Population: ALL high school students

2

Let's change the study slightly to be a random experiment. Can we make the conclusion below now?

Students at a local magnate high school want to test if drinking caffeine before the SAT causes a student to score higher on the exam. 80 students that attend the magnate school took the SAT at the testing facility one morning. They decide to **randomly assign** the 40 students that arrived "coffee group" and offer them coffee before the exam. They **randomly assigned** the other 40 students to the "no coffee group" and did not offer them any coffee.

After scores were released, they were able to collect 36 SAT scores from the "coffee group" and 31 SAT scores from the "no coffee group." They conducted an independent means test between the two groups using this data and found a p-value=0.03. They used this to conclude that drinking coffee before the SAT causes high school students to score higher on the SAT.

a.) yes
b.) no

Some reasons why:
- Population: ALL high school students
- Sample:
  - Students at a magnate school.

Issue # 2

2

Let's change the study slightly to be a random experiment. Can we make the conclusion below now?

Students at a local magnate high school want to test if drinking caffeine before the SAT causes a student to score higher on the exam. 80 students that attend the magnate school took the SAT at the testing facility one morning. They decide to **randomly assign** the 40 students that arrived "coffee group" and offer them coffee before the exam. They **randomly assigned** the other 40 students to the "no coffee group" and did not offer them any coffee.

After scores were released, they were able to collect 36 SAT scores from the "coffee group" and 31 SAT scores from the "no coffee group." They conducted an independent means test between the two groups using this data and found a p-value=0.03. They used this to conclude that drinking coffee before the SAT causes high school students to score higher on the SAT.

a.) yes
b.) no

Some reasons why:
- Population: ALL high school students
- Sample:
  - Students at a magnate school.
  - Sampling was not random.

Issue # 3

2

NEW

- *Non-response:* If only a non-random fraction of the randomly sampled people choose to respond to a survey such that the sample may be no longer be representative of the population
- *Voluntary response:* Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue since such a sample will also not be representative of the population
- *Convenience sample:* Individuals who are easily accessible are more likely to be included in the sample

5

Let's change the study slightly to be a random experiment. Can we make the conclusion below now?

Students at a local magnate high school want to test if drinking caffeine before the SAT causes a student to score higher on the exam. 80 students that attend the magnate school took the SAT at the testing facility one morning. They decide to **randomly assign** the 40 students that arrived "coffee group" and offer them coffee before the exam. They **randomly assigned** the other 40 students to the "no coffee group" and did not offer them any coffee.

After scores were released, they were able to collect 36 SAT scores from the "coffee group" and 31 SAT scores from the "no coffee group." They conducted an independent means test between the two groups using this data and found a p-value=0.03. They used this to conclude that drinking coffee before the SAT causes high school students to score higher on the SAT.

a.) yes
b.) no

Some reasons why:
- Population: ALL high school students
- Sample:
  - Students at a magnate school.
  - Sampling was not random.
  - Sample data had bias:
    - Convenience sampling.

Issue # 4

2

Let's change the study slightly to be a random experiment. Can we make the conclusion below now?

Students at a local magnate high school want to test if drinking caffeine before the SAT causes a student to score higher on the exam. 80 students that attend the magnate school took the SAT at the testing facility one morning. They decide to **randomly assign** the 40 students that arrived "coffee group" and offer them coffee before the exam. They **randomly assigned** the other 40 students to the "no coffee group" and did not offer them any coffee.

After scores were released, they were able to collect 36 SAT scores from the "coffee group" and 31 SAT scores from the "no coffee group." They conducted an independent means test between the two groups using this data and found a p-value=0.03. They used this to conclude that drinking coffee before the SAT causes high school students to score higher on the SAT.

a.) yes
b.) no

Some reasons why:

- Population: ALL high school students
- Sample:
  - Students at a magnate school.
  - Sampling was not random.
  - Sample data had bias:
    - Convenience sampling.
    - Non-response bias.

Issue # 5

2

6. Random sampling helps generalizability, random assignment helps causality (two or more variables)

NEW

If two or more variables in research question

	Random assignment	No random assignment	
ideal experiment			*most observational studies*
Random sampling	Causal conclusion, generalized to the whole population.	No causal conclusion, correlation statement generalized to the whole population.	Generalizability
No random sampling	Causal conclusion, only for the sample.	No causal conclusion, correlation statement only for the sample.	No generalizability
most experiments	Causation	Correlation	*bad observational studies*

10

Let's change the study slightly to be a random experiment. Can we make the conclusion below now?

Students at a local magnate high school want to test if drinking caffeine before the SAT causes a student to score higher on the exam. 80 students that attend the magnate school took the SAT at the testing facility one morning. They decide to **randomly assign** the 40 students that arrived "coffee group" and offer them coffee before the exam. They **randomly assigned** the other 40 students to the "no coffee group" and did not offer them any coffee.

After scores were released, they were able to collect 36 SAT scores from the "coffee group" and 31 SAT scores from the "no coffee group." They conducted an independent means test between the two groups using this data and found a p-value=0.03. They used this to conclude that drinking coffee before the SAT causes high school students to score higher on the SAT.

a.) yes
b.) no

Some reasons why:

- Population: ALL high school students
- Sample:
  - Students at a magnate school.
  - Sampling was not random.
  - Sample data had bias:
    - Convenience sampling.
    - Non-response bias.

Issues # 2-5

The sample is not random and not representative of the population!

2

Let's change the random experiment slightly, what sampling technique was used to get the random sample?

Students at a local magnate high school want to test if drinking caffeine before the SAT causes a student to score higher on the exam. They randomly select 50 high schools (each with a diverse student body) through out the country. They decide to randomly assign half the SAT taking students from each school to the "coffee group" and offered them coffee. They randomly assigned the other half from each school to the "no coffee group" and did not offer them any coffee the morning of the exam.

They obtained SAT scores from all the students in their experiment. They conducted an independent means test between the two groups using this data and found a p-value=0.03.

1) Simple random sampling
2) Stratified sampling
3) Cluster sampling
4) Multistage sampling

2

Let's change the random experiment slightly, what sampling technique was used to get the random sample?

Students at a local magnate high school want to test if drinking caffeine before the SAT causes a student to score higher on the exam. They randomly select 50 high schools (each with a diverse student body) through out the country. They decide to randomly assign half the SAT taking students from each school to the "coffee group" and offered them coffee. They randomly assigned the other half from each school to the "no coffee group" and did not offer them any coffee the morning of the exam.

They obtained SAT scores from all the students in their experiment. They conducted an independent means test between the two groups using this data and found a p-value=0.03.

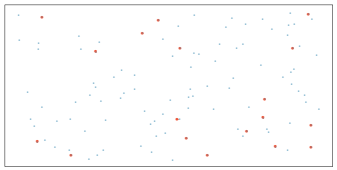
1) Simple random sampling
2) Stratified sampling
3) ***Cluster sampling***
4) Multistage sampling

2

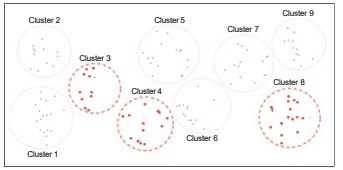
NEW

2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier

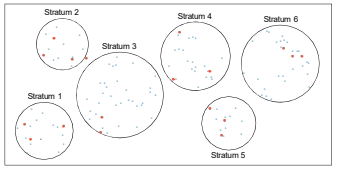
Simple random:
Drawing names from a hat



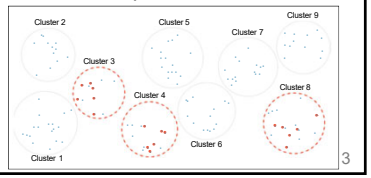
Cluster: heterogenous clusters
Sample all chosen clusters



Stratified: homogenous strata
Stratify to control for age group



Multistage: heterogeneous clusters
Random sample in chosen clusters



3

NEW

and cluster to make sampling easier

2. Ideally use a simple random sample, stratify to control for a variable,

	Sampling Method Name		
	Stratified Sampling	**Cluster Sampling**	**Multistage Sampling**
How to sample this way.	Step 1: Assign each observation in a population into groups. Each group is called a…		
	Stratum (Strata plural),	Cluster,	Cluster,
	where the objects in a given group are…		
	homogeneous.	heterogeneous.	heterogeneous.
	Step 2:Then, select…		
	ALL the groups.	a random selection of groups	a random selection of groups
	Step 3: Then from each of these selected groups, select…		
	a random sample of observations	ALL the observations	a random sample of observations

Let's make the experiment better…

The students realized that some students had already drank coffee, soda, and Redbull before arriving at the exam. What principle of experimental design could they employ to make the experiment better in light of this knowledge?

1) Randomize
2) Replicate
3) Control
4) Block

2

Let's make the experiment better…

The students realized that some students had already drank coffee, soda, and Redbull before arriving at the exam. What principle of experimental design could they employ to make the experiment better in light of this knowledge?

1) Randomize
2) Replicate
3) Control
4) Block

Control: Make the treatment/control groups as similar as possible.

- Have the student not ingest any caffeine before coming to the exam.
- Give the students in the "coffee group" exactly 2 cups of coffee at exactly the same time.
- Make sure everyone in the treatment group drinks it.

Issue # 6

2

Let's make the experiment better…

The students realized that Seniors tend to outperform Juniors in the SAT. What principle of experimental design could they employ to make the experiment better in light of this knowledge?

1) Randomize
2) Replicate
3) Control
4) Block

2

Let's make the experiment better…

The students realized that Seniors tend to outperform Juniors in the SAT. What principle of experimental design could they employ to make the experiment better in light of this knowledge?

1) Randomize
2) Replicate
3) Control
4) Block

Block: Minimize the effect of class year on the experiment by randomly assigning an equal amount to each group.

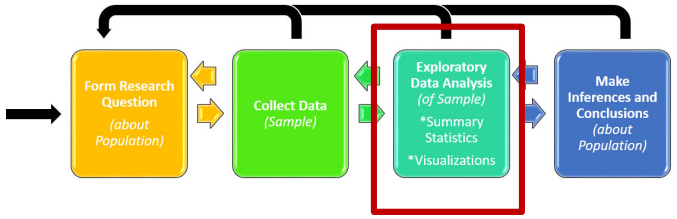
- Randomly assign half the **Seniors** and half the **Juniors** to the **treatment/coffee group.**
- Randomly assign half the **Seniors** and half the **Juniors** to the **control/no-coffee group.**

Issue # 7

2

Outline

Course Roadmap



What are good ways to visualize and summarize data?

Let's visualize the data.

Which plot would be appropriate to visualize the data from this experiment?

1) Scatter plot
2) Mosaic plot
3) Segmented frequency bar plot
4) Frequency bar plot
5) Side-by-side box plots

Let's visualize the data.

Which plot would be appropriate to visualize the data from this experiment?

1) Scatter plot
2) Mosaic plot
3) Segmented frequency bar plot
4) Frequency bar plot
5) ***Side-by-side box plots***

Numerical: SAT score
Categorical: coffee/no-coffee

What's an appropriate plot to use for each of the following types of data?

1. **One numerical variable** – *boxplot, histogram*
2. **One categorical variable** – *barplot, histogram*
3. **Two numerical variables** – *scatterplot*
4. **One categorical variable and one numerical variable** – *side-by-side boxplots*
5. **Two categorical variables** – *mosaic plot, frequency segmented barplot*

Clicker question

Which of the following is false?

(a) Box plots are useful for highlighting outliers, but we cannot determine skew based on a box plot.

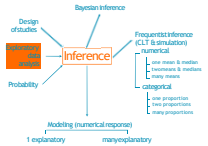
(b) Median and IQR are more robust statistics than mean and SD, respectively, since they are not affected by outliers or extreme skewness.

(c) When the response variable is extremely right skewed, it may be useful to apply a log transformation to obtain a more symmetric distribution, and model the logged data.

(d) Segmented bar plots are not good enough for evaluating the **relationship** between two categorical variables.

Clicker question

Which of the following is false?



(a) Box plots are useful for highlighting outliers, but we cannot determine skew based on a box plot.

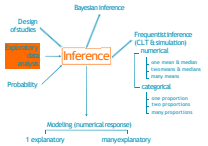
(b) Median and IQR are more robust statistics than mean and SD, respectively, since they are not affected by outliers or extreme skewness.

(c) When the response variable is extremely right skewed, it may be useful to apply a log transformation to obtain a more symmetric distribution, and model the logged data.

(d) Segmented bar plots are not good enough for evaluating the **relationship** between two categorical variables.

Clicker question

Which of the following is false?



(a) Box plots are useful for highlighting outliers, but we cannot determine skew based on a box plot. → We can determine skew with a boxplot! (BUT, we can't determine **modality** from a box plot)!

(b) Median and IQR are more robust statistics than mean and SD, respectively, since they are not affected by outliers or extreme skewness. → Why is this?

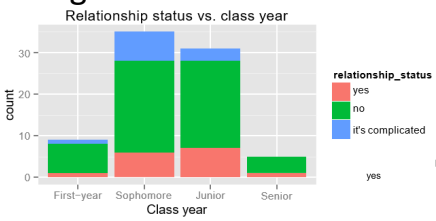
(c) When the response variable is extremely right skewed, it may be useful to apply a log transformation to obtain a more symmetric distribution, and model the logged data *(Slides 7.2)*

(d) Segmented bar plots are not good enough for evaluating the **relationship** between two categorical variables. → (See graph)

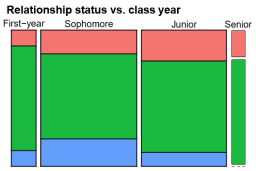
Clicker question

Which of the following is false?

Segmented Bar Plot



Mosaic Plot



(c)

(d) Segmented bar plots are not good enough for evaluating the **relationship** between two categorical variables. → (See graph)

Clicker question

A recent housing survey was conducted to determine the price of a typical home in Glendale, CA. Glendale is mostly middle-class, with one very expensive suburb. The mean price of a house was roughly $650,000. Which of the following statements is most likely to be true?

(a) Most houses in Glendale cost more than $650,000.

(b) Most houses in Glendale cost less than $650,000.

(c) There are about as many houses in Glendale that cost more than $650,000 than less than this amount.

(d) We need to know the standard deviation to answer this question

Clicker question

A recent housing survey was conducted to determine the price of a typical home in Glendale, CA. Glendale is mostly middle-class, with one very expensive suburb. The mean price of a house was roughly $650,000. Which of the following statements is most likely to be true?

(a) Most houses in Glendale cost more than $650,000.

(b) Most houses in Glendale cost less than $650,000.

(c) There are about as many houses in Glendale that cost more than $650,000 than less than this amount.

(d) We need to know the standard deviation to answer this question

- *Prices/income tend to be right skewed.*
- *Median < Mean in right skewed distributions.*
- *50% of data < Median < Mean = $650,000*

Outline

When finding the probability of an event, how do we decide which probability equations/techniques to use?

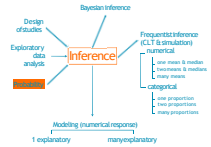
- General Probability Rules
- Bayes Equation
- Probability Tree
- Binomial Distribution Equations
- Normal Distribution Equations

General Probability Rules

- **Disjoint (mutually exclusive) events** cannot happen at the same time
  - For disjoint A and B: P(A and B) = 0
- **Complementary events** A and $\bar{A}$ are disjoint events where $\bar{A}$ = "not A". (A or $\bar{A}$) represents all possibilities
  - For complementary A and $\bar{A}$: P(A) + P($\bar{A}$) = 1
- If A and B are **independent events**, having information on A does not tell us anything about B (and vice versa)
  - If A and B are independent, ALL OF following are true:
    - P(A | B) = P(A)
    - P(B | A) = P(B)
    - P(A and B) = P(A) × P(B)
- If A and B are **dependent events**, having information on A DOES tell us things about B (and vice versa)
  - If A and B are dependent, ALL OF following are true:
    - P(A | B) ≠ P(A)
    - P(B | A) ≠ P(B)
    - P(A and B) ≠ P(A) × P(B)
- **General addition rule**: P(A or B) = P(A) + P(B) - P(A and B)
- **General multiplication rule**: P(A and B) =P(A | B) × P(B) = P(B | A) × P(A)
- **Bayes' theorem**: P(A | B) = P(A and B) / P(B)
- **"Splitting a Probability"**: P(B) = P(B and A) + P(B and $\bar{A}$) (where A and $\bar{A}$ are complements)

Clicker question

Which of the following is false?



(a) If A and B are independent, then having information on A does not tell us anything about B.

(b) If A and B are disjoint, then knowing that A occurs tells us that B cannot occur.

(c) Disjoint (mutually exclusive) events are always dependent since if one event occurs we know the other one cannot.

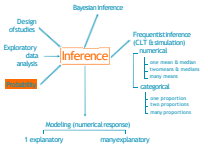
(d) If A and B are independent, then P(A and B) = P(A) + P(B).

(e) If A and B are not disjoint, then P(A or B) = P(A) + P(B) - P(A and B).

5

Clicker question

Which of the following is false?



(a) If A and B are independent, then having information on A does not tell us anything about B.

(b) If A and B are disjoint, then knowing that A occurs tells us that B cannot occur.

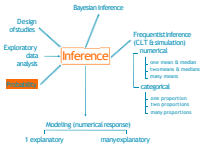
(c) Disjoint (mutually exclusive) events are always dependent since if one event occurs we know the other one cannot.

(d) If A and B are independent, then P(A and B) = P(A) + P(B).

(e) If A and B are not disjoint, then P(A or B) = P(A) + P(B) - P(A and B).

Clicker question

Which of the following is false?



(a) If A and B are independent, then having information on A does not tell us anything about B. → Independent means: P(A)=P(A|B)

(b) If A and B are disjoint, then knowing that A occurs tells us that B cannot occur. → Disjoint means P(A and B) = 0 and P(B|A)=0

(c) Disjoint (mutually exclusive) events are always dependent since if one event occurs we know the other one cannot. → Disjoint means P(A and B) = 0 and and P(B|A)=0

P(A) × P(B).

(d) If A and B are independent, then P(A and B) = ~~P(A) + P(B)~~.

~~(e) If A and B are not disjoint, then~~ P(A or B) = P(A) + P(B) - P(A and B).

General Addition Rule

Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the probability that at least two students from your high school (100 Seniors) get in this year (assume students at your school are independent)?

(a) Calculate exactly one binomial expression.

(b) Calculate multiple binomial expressions (and add/subtract/do other things with them).

(c) Approximate a binomial distribution with a normal distribution *(but don't check/show any conditions before you do this.)*

(d) Approximate a binomial distribution with a normal distribution *(but DO check/show conditions before you do this.)*

(e) NONE OF THE ABOVE (Use other probability equations that aren't specific to binomial and normal distributions).

Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the probability that at least two students from your high school (100 Seniors) get in this year (assume students at your school are independent)?

(a) Calculate exactly one binomial expression.

(b) Calculate multiple binomial expressions (and add/subtract/do other things with them).

(c) Approximate a binomial distribution with a normal distribution *(but don't check/show any conditions before you do this.)*

(d) Approximate a binomial distribution with a normal distribution *(but DO check/show conditions before you do this.)*

(e) NONE OF THE ABOVE (Use other probability equations that aren't specific to binomial and normal distributions).

Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the probability that at least two students from your high school (100 Seniors) get in this year (assume students at your school are independent)?

"Binomial event"

$P(\text{at least } k \text{ out of } n \text{ independent trials are a success})$

$P(\text{at least } 2 \text{ out of } 100 \text{ students at school get into Duke})$

Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the probability that at least two students from your high school (100 seniors) get in this year (assume students at your school are independent)?

"Binomial event"

$P(\text{at least } k \text{ out of } n \text{ independent trials are a success})$

$P(\text{at least } 2 \text{ out of } 100 \text{ students at school get into Duke})$

Binomial distribution is needed because:

a) n=100 trials (fixed)

b) each trial has probability 0.057

c) independent trials

d) only two possibilities (success=Duke, failure=No Duke)

Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the probability that at least two students from your high school get in this year (assume students at your school are independent)?

Calculate multiple binomial expressions (and add/subtract/do other things with them).

(a) $P(X \geq 2) = \binom{100}{2} 0.057^2 (1-0.057)^{100-2}$

(b) $P(X \geq 2) = \binom{100}{2} 0.057^2 (1-0.057)^{100-2} + \cdots + \binom{100}{100} 0.057^{100} (1-0.057)^{100-100}$ *(by hand)*

(c) $P(X \geq 2) = 1 - \left[\binom{100}{2} 0.057^2 (1-0.057)^{100-2}\right]$

(d) $P(X \geq 2) = 1 - \left[\binom{100}{0} 0.057^0 (1-0.057)^{100-0}\right]$

(e) $P(X \geq 2) = 1 - \left[\binom{100}{0} 0.057^0 (1-0.057)^{100-0} + \binom{100}{1} 0.057^1 (1-0.057)^{100-1}\right]$

Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the probability that at least two students from your high school get in this year (assume students at your school are independent)?

Calculate multiple binomial expressions (and add/subtract/do other things with them).

(a) $P(X \geq 2) = \binom{100}{2} 0.057^2 (1-0.057)^{100-2}$

(b) $P(X \geq 2) = \binom{100}{2} 0.057^2 (1-0.057)^{100-2} + \cdots + \binom{100}{100} 0.057^{100} (1-0.057)^{100-100}$ *(by hand)*

(c) $P(X \geq 2) = 1 - \left[\binom{100}{2} 0.057^2 (1-0.057)^{100-2}\right]$

(d) $P(X \geq 2) = 1 - \left[\binom{100}{0} 0.057^0 (1-0.057)^{100-0}\right]$

(e) $\mathbf{P(X \geq 2) = 1 - \left[\binom{100}{0} 0.057^0 (1-0.057)^{100-0} + \binom{100}{1} 0.057^1 (1-0.057)^{100-1}\right]}$

Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the probability that at least three students from your high school get in this year (assume students at your school are independent)?

Calculate multiple binomial expressions (and add/subtract/do other things with them).

$X = of\ students\ out\ of\ 100\ Seniors\ at\ your\ HS\ that\ get\ into\ Duke$

$X \sim Bin(n = 100, p = 0.57)$

Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the probability that at least three students from your high school get in this year (assume students at your school are independent)?

Calculate multiple binomial expressions (and add/subtract/do other things with them).

$X = of\ students\ out\ of\ 100\ Seniors\ at\ your\ HS\ that\ get\ into\ Duke$

$X \sim Bin(n = 100, p = 0.57)$

$P(at\ least\ 2\ out\ of\ 100\ students\ at\ school\ get\ into\ Duke) =$
$= \boldsymbol{P(X \geq 2)}$

Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the probability that at least three students from your high school get in this year (assume students at your school are independent)?

Calculate multiple binomial expressions (and add/subtract/do other things with them).

$X = of\ students\ out\ of\ 100\ Seniors\ at\ your\ HS\ that\ get\ into\ Duke$

$X \sim Bin(n = 100, p = 0.57)$

$P(at\ least\ 2\ out\ of\ 100\ students\ at\ school\ get\ into\ Duke) =$
$= \boldsymbol{P(X \geq 2)}$
$= \boldsymbol{P(X = 2) + P(X = 3) + \cdots P(X = 100)}$

Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the probability that at least three students from your high school get in this year (assume students at your school are independent)?

Calculate multiple binomial expressions (and add/subtract/do other things with them).

$X = of\ students\ out\ of\ 100\ Seniors\ at\ your\ HS\ that\ get\ into\ Duke$

$X \sim Bin(n = 100, p = 0.57)$

$P(at\ least\ 2\ out\ of\ 100\ students\ at\ school\ get\ into\ Duke) =$
$= \boldsymbol{P(X \geq 2)}$
$= \boldsymbol{P(X = 2) + P(X = 3) + \cdots P(X = 100)}$
$= \boldsymbol{1 - [P(X = 0) + P(X = 1)]}$

Binomial Equation: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the probability that at least three students from your high school get in this year (assume students at your school are independent)?

Calculate multiple binomial expressions (and add/subtract/do other things with them).

$X = of\ students\ out\ of\ 100\ Seniors\ at\ your\ HS\ that\ get\ into\ Duke$

$X \sim Bin(n = 100, p = 0.57)$

$P(at\ least\ 2\ out\ of\ 100\ students\ at\ school\ get\ into\ Duke) =$

$= \boldsymbol{P(X \geq 2)}$

$= \boldsymbol{P(X = 2) + P(X = 3) + \cdots P(X = 100)}$

$= \boldsymbol{1 - [P(X = 0) + P(X = 1)]}$

$= \boldsymbol{1 - \left[\binom{100}{0} 0.057^0 (1 - 0.057)^{100-0} + \binom{100}{1} 0.057^1 (1 - 0.057)^{100-1}\right]}$

Binomial Equation: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the probability that at least two students from your high school (100 Seniors) get in this year (assume students at your school are independent)?

Why not do it this way?

Approximate a binomial distribution with a normal distribution.

Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the probability that at least two students from your high school (100 Seniors) get in this year (assume students at your school are independent)?

Why not do it this way?

Approximate a binomial distribution with a normal distribution.

SF conditions don't hold:

- np=100(0.057)<10
- n(1-p)=100(1-0.057) ≥10

Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the shape of the sampling distribution of sample proportions $\hat{p}$ with sample size 100?

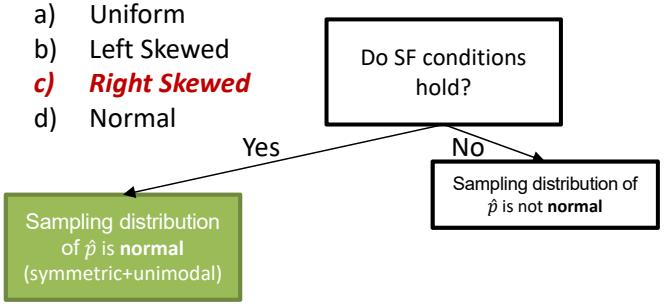
a) Uniform
b) Left Skewed
c) Right Skewed
d) Normal

Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the shape of the sampling distribution of sample proportions $\hat{p}$ with sample size 100?

a) Uniform
b) Left Skewed
c) Right Skewed
d) Normal

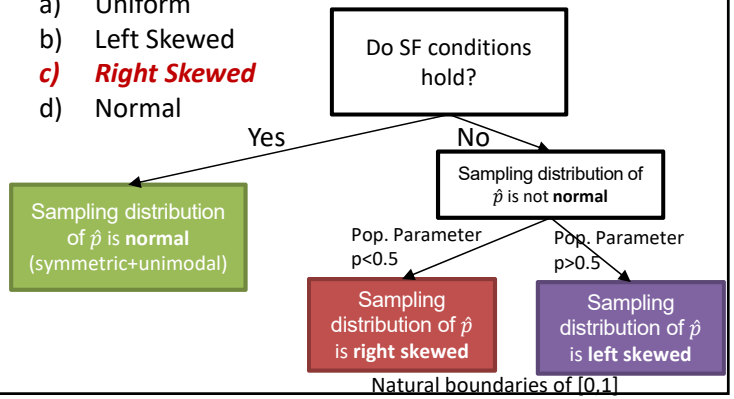


Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the shape of the sampling distribution of sample proportions $\hat{p}$ with sample size 100?

a) Uniform
b) Left Skewed
c) Right Skewed
d) Normal



Clicker question

The Duke acceptance rate for last year was 0.057. Is it unusual for 3 out of a random sample of 100 Seniors to get into Duke?

a) Yes
b) No
c) We cannot determine.

Clicker question

The Duke acceptance rate for last year was 0.057. Is it unusual for 3 out of a random sample of 100 Seniors to get into Duke?

a) Yes
b) No
c) We cannot determine

$$X \sim Bin(n = 100, p = .057)$$

$$\downarrow$$

$$X \sim N(\mu = np, \sigma = \sqrt{np(1-p)})$$

$$\downarrow$$

$$3 \text{ unusual if } 3 > \mu + 2\sigma \text{ or } 3 < \mu - 2\sigma$$

But why can we use this logic to show 3 is unusual/not unusual?

Clicker question

The Duke acceptance rate for last year was 0.057. Is it unusual for 3 out of a random sample of 100 Seniors to get into Duke?

a) Yes
b) No
c) We cannot determine because SF conditions don't hold!

$$X \sim Bin(n = 100, p = .057)$$

Only true when SF conditions hold

SF conditions don't hold:
- np=100(0.057)<10
- n(1-p)=100(1-0.057) ≥10

$$X \sim N(\mu = np, \sigma = \sqrt{np(1-p)})$$

Only true when X is normal

$$3 \; unusual \; if \; 3 > \mu + 2\sigma \text{ or } 3 < \mu - 2\sigma$$

Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the probability that at least 40 out of 200 randomly sampled Seniors got accepted

a) Yes
b) No
c) We cannot determine.

Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the probability that at least 40 out of 200 randomly sampled Seniors got accepted

a) Yes
b) No
c) We cannot determine.

$$3 < \mu - 2\sigma$$
$$3 < np - 2\sqrt{np(1-p)}$$
$$3 < 200(0.057) - 2\sqrt{200(0.057)(1-0.057)}$$

$$X \sim Bin(n = 200, p = .057)$$

Only true when SF conditions hold

SF conditions HOLD:
- np=200(0.057) ≥ 10
- n(1-p)=200(1-0.057) ≥10

$$X \sim N(\mu = np, \sigma = \sqrt{np(1-p)})$$

Only true when X is normal

$$3 \; unusual \; if \; 3 > \mu + 2\sigma \text{ or } 3 < \mu - 2\sigma$$

Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the probability that at least 40 out of 200 randomly sampled Seniors got accepted

(a) Calculate exactly one binomial expression.
(b) Calculate multiple binomial expressions (and add/subtract/do other things with them).
(c) Approximate a binomial distribution with a normal distribution *(but don't check/show any conditions before you do this.)*
(d) Approximate a binomial distribution with a normal distribution *(but DO check/show conditions before you do this.)*
(e) NONE OF THE ABOVE (Use other probability equations that aren't specific to binomial and normal distributions).

Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the probability that at least 40 out of 200 randomly sampled Seniors got accepted

(a) Calculate exactly one binomial expression.

(b) Calculate multiple binomial expressions (and add/subtract/do other things with them).

(c) Approximate a binomial distribution with a normal distribution *(but don't check/show any conditions before you do this.)*

(d) Approximate a binomial distribution with a normal distribution (but DO check/show conditions before you do this.)

(e) NONE OF THE ABOVE (Use other probability equations that aren't specific to binomial and normal distributions).

Clicker question

What would you need to do solve this problem?

The Duke acceptance rate for last year was 0.057. What is the probability that at least 40 out of 200 randomly sampled Seniors got accepted

$$X \sim Bin(n = 200, p = .057)$$

Only true when SF conditions hold

SF conditions HOLD:
- np=200(0.057) ≥ 10
- n(1-p)=200(1-0.057) ≥10

$$X \sim N(\mu = np, \sigma = \sqrt{np(1-p)})$$

Only true when X is normal

$$P(X \geq 40) = P\left(Z \geq \frac{40-u}{\sigma}\right) = P\left(Z \geq \frac{40-np}{\sqrt{np(1-p)}}\right)$$

$$= P\left(Z \geq \frac{40-200(0.057)}{\sqrt{200(0.057)(1-0.057)}}\right) < 0.0002$$

Clicker question

What would you need to do solve this problem?

The acceptance rate for College University (CU) last year was 20%. Assume the SAT scores for incoming CU freshmen was normally distributed (for Math+Verbal) with an average of 1200 and a standard deviation of 100. Assume the SAT scores for all students in the US was normally distributed (for Math+Verbal) with an average of 1060 and a standard deviation of 195. What's the probability of getting into CU with an SAT score of at least 1500?

(a) Bayes equation and a probability tree
(b) Bayes equation and the normal distribution
(c) Just use the normal distribution
(d) Multiple binomial distribution
(e) General probability equations

Clicker question

What would you need to do solve this problem?

The acceptance rate for College University (CU) last year was 20%. Assume the SAT scores for incoming CU freshmen was normally distributed (for Math+Verbal) with an average of 1200 and a standard deviation of 100. Assume the SAT scores for all students in the US was normally distributed (for Math+Verbal) with an average of 1060 and a standard deviation of 195. What's the probability of getting into CU with an SAT score of at least 1500?

(a) Bayes equation and a probability tree
(b) Bayes equation and the normal distribution
(c) Just use the normal distribution
(d) Multiple binomial distribution
(e) General probability equations

Clicker question

What would you need to do solve this problem?

The acceptance rate for College University (CU) last year was 20%. Assume the SAT scores for incoming CU freshmen was normally distributed (for Math+Verbal) with an average of 1200 and a standard deviation of 100. Assume the SAT scores for all students in the US was normally distributed (for Math+Verbal) with an average of 1060 and a standard deviation of 195. What's the probability of getting into CU with an SAT score of at least 1500?

$$P(get\ into\ CU|X > 1500) =$$

Clicker question

What would you need to do solve this problem?

The acceptance rate for College University (CU) last year was 20%. Assume the SAT scores for incoming CU freshmen was normally distributed (for Math+Verbal) with an average of 1200 and a standard deviation of 100. Assume the SAT scores for all students in the US was normally distributed (for Math+Verbal) with an average of 1060 and a standard deviation of 195. What's the probability of getting into CU with an SAT score of at least 1500?

$$P(get\ into\ CU|X > 1500) = \frac{P(X > 1500\ and\ get\ into\ CU)}{P(X > 1500)}$$

Bayes Equation

Clicker question

What would you need to do solve this problem?

The acceptance rate for College University (CU) last year was 20%. Assume the SAT scores for incoming CU freshmen was normally distributed (for Math+Verbal) with an average of 1200 and a standard deviation of 100. Assume the SAT scores for all students in the US was normally distributed (for Math+Verbal) with an average of 1060 and a standard deviation of 195. What's the probability of getting into CU with an SAT score of at least 1500?

$$P(get\ into\ CU|X > 1500) = \frac{P(X > 1500\ and\ get\ into\ CU)}{P(X > 1500)} = \frac{P(X > 1500|get\ into\ CU)P(get\ into\ CU)}{P(X > 1500)}$$

General Mult. Rule

Clicker question

What would you need to do solve this problem?

The acceptance rate for College University (CU) last year was 20%. Assume the SAT scores for incoming CU freshmen was normally distributed (for Math+Verbal) with an average of 1200 and a standard deviation of 100. Assume the SAT scores for all students in the US was normally distributed (for Math+Verbal) with an average of 1060 and a standard deviation of 195. What's the probability of getting into CU with an SAT score of at least 1500?

$$P(get\ into\ CU|X > 1500) = \frac{P(X > 1500\ and\ get\ into\ CU)}{P(X > 1500)} = \frac{P(X > 1500|get\ into\ CU)P(get\ into\ CU)}{P(X > 1500)} = \frac{P(Z > \frac{1500-1200}{100})(0.20)}{P(Z > \frac{1500-1060}{195})}$$

Clicker question

What would you need to do solve this problem?

The acceptance rate for College University (CU) last year was 20%. 10% of students that got into CU had a perfect high school GPA and 3% of students that did not get into CU had a perfect high school GPA. What's the probability of getting into CU with a perfect high school GPA?

(a) Bayes equation and a probability tree
(b) Bayes equation and the normal distribution
(c) Just use the normal distribution
(d) Multiple binomial distribution
(e) General probability equations

Clicker question

What would you need to do solve this problem?

The acceptance rate for College University (CU) last year was 20%. 10% of students that got into CU had a perfect high school GPA and 3% of students that did not get into CU had a perfect high school GPA. What's the probability of getting into CU with a perfect high school GPA?

(a) Bayes equation and a probability tree
(b) Bayes equation and the normal distribution
(c) Just use the normal distribution
(d) Multiple binomial distribution
(e) General probability equations

Clicker question

What would you need to do solve this problem?

The acceptance rate for College University (CU) last year was 20%. 10% of students that got into CU had a perfect high school GPA and 3% of students that did not get into CU had a perfect high school GPA. What's the probability of getting into CU with a perfect high school GPA?

Want:
$P(CU|perfect\ score)$

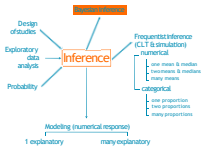
Know:
$P(CU) = 0.2$
$P(perfect\ score|CU) = 0.1$
$P(perfect\ score|not\ CU) = 0.03$

Probability Tree usually helpful when you know:

- Probability of what you want, without the "data".
- "Reverse given" probability of what you want.
- Probability of your "data", given all possibilties

Clicker question

Which of the following is false?

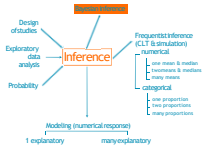


(a) Suppose you're evaluating 4 claims. If prior to data collection you don't have a preference for one claim over another, you should assign 0.25 as the prior probability to each claim.
(b) Posterior probability and the p-value are the equivalent.
(c) One advantage of Bayesian inference is that data can be integrated to the inferential scheme as they are collected.
(d) Suppose a patient tests positive for a disease that 2% of the population are known to have. A doctor wants to confirm the test result by retesting the patient. In the second test the prior probability for "having the disease" should be more than 2%.

7

Clicker question

Which of the following is false?



(a) Suppose you're evaluating 4 claims. If prior to data collection you don't have a preference for one claim over another, you should assign 0.25 as the prior probability to each claim.

(b) Posterior probability and the p-value are the equivalent.

(c) One advantage of Bayesian inference is that data can be integrated to the inferential scheme as they are collected.

(d) Suppose a patient tests positive for a disease that 2% of the population are known to have. A doctor wants to confirm the test result by retesting the patient. In the second test the prior probability for "having the disease" should be more than 2%.

Posterior = P(hypothesis | data), p-value ≈ P(data | hypothesis)

7

Frequentist Hypothesis Testing

Conducting an Analysis… where to start?

How many numerical and categorical variables are involved?
If categorical variables, how many levels do they have?
What visualization would be appropriate for visualizing this data?

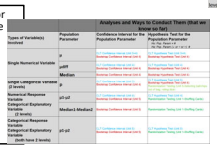
- Is it a combo of variables that **has one specified Population Parameter of Interest** where
  - We can create a **confidence interval** for the population parameter.
  - We can conduct a **hypothesis test** for the population parameter using:

$$Ho: Pop.Param = \#$$
$$Ha: Pop.Param\ (\neq\ or\ <\ or\ >)\ \#?$$

- Is it a combo of variables that **doesn't have one specified Population Parameter of Interest** and
  - We can't make a **confidence interval**.
  - Has a different **hypothesis test** set up?

Types of Variables	Analysis	Hypotheses
Numerical Response Variable Categorical Explanatory Variable (>2 levels)	**ANOVA**	Ho: $\mu_1=\mu_2=...=\mu_k$ Ha: at least one pair of groups has means that are different
Single Categorical Variable (>2 levels)	**Chi-Squared Goodness of Fit Test**	Ho: The data follows the specified distribution Ha: The data does not follow the specified distribution
Categorical Response Variable Categorical Explanatory Variable (at least one has >2 levels)	**Chi-Squared Independence Test**	Ho: The two variables are independent/not associated Ha: The two variables are dependent/associated

- If we are making a confidence interval or hypothesis test for this population parameter, when should we use each of the following **methods** to do this?
  - CLT methods
  - Randomization testing methods
  - Bootstrap methods



One Population Parameter of Interest

		Analyses and Ways to Conduct Them (that we know so far)	
Types of Variable(s) Involved	**Population Parameter**	**Confidence Interval for the Population Parameter**	**Hypothesis Test for the Population Parameter** *Ho: Pop. Param = #* *Ha: Pop. Param (≠ or < or >) #*
Single Numerical Variable	**μ**	CLT Confidence Interval (Unit 3+4) Bootstrap Confidence Interval (Unit 4)	CLT Hypothesis Test (Unit 3+4) Bootstrap Hypothesis Test (Unit 4)
	μdiff	CLT Confidence Interval (Unit 4) Bootstrap Confidence Interval (Unit 4)	CLT Hypothesis Test (Unit 4) Bootstrap Hypothesis Test (Unit 4)
	Median	Bootstrap Confidence Interval (Unit 4)	Bootstrap Hypothesis Test (Unit 4)
Single Categorical Variable (2 levels)	**p**	CLT Confidence Interval (Unit 5) Bootstrap Confidence Interval (Unit 4)	CLT Hypothesis Test (Unit 5) Bootstrap Hypothesis Test (Unit 4) Randomization Testing (Unit 5-Selecting ball/chips out of bag, rolling dice)
Numerical Response Variable Categorical Explanatory Variable (2 levels)	**μ1-μ2**	CLT Confidence Interval (Unit 4) Bootstrap Confidence Interval (Unit 5)	CLT Hypothesis Test (Unit 4) Randomization Testing (Unit 1-Shuffling Cards)
	Median1-Median2	Bootstrap Confidence Interval (Unit 5)	Randomization Testing (Unit 1-Shuffling Cards)
Categorical Response Variable Categorical Explanatory Variable (both have 2 levels)	**p1-p2**	CLT Confidence Interval (Unit 5) Bootstrap Confidence Interval (Unit 5)	CLT Hypothesis Test (Unit 5) Randomization Testing (Unit 1-Shuffling Cards)

When other certain conditions are met...

$\bar{x} \sim N(mean = \mu, standard\ dev./error = \frac{\sigma}{\sqrt{n}})$

and we can make CLT CIs and HTs for μ

What are these CLT conditions?

When other certain conditions are met...

$\bar{x}_1 - \bar{x}_2 \sim N(mean = \mu_1 - \mu_2, standard\ dev./error = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$

and we can make CLT CIs and HTs for μ1-μ2

When other certain conditions are met...

$\hat{p} \sim N(mean = p, standard\ dev./error = \sqrt{\frac{p(1-p)}{n}})$

and we can make CLT CIs and HTs for p

When other certain conditions are met...

$\hat{p}_1 - \hat{p}_2 \sim N(mean = p_1 - p_2, standard\ dev./error = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}})$

and we can make CLT CIs and HTs for p1-p2

Central Limit Theorem

Confidence Interval for Population Parameter

$(point\ estimate) \pm (crit.value)SE$

When CLT conditions are met.

Types of Variable(s) Involved	Population Parameter	Point Estimate	Standard Error	Distribution: (1) To get Critical Values From (CI) (2) That the Test Statistic Follows (HT)
Single **Numerical** Variable	μ	$\bar{x}$	$\frac{\sigma}{\sqrt{n}}$	Z (if you know σ) T_{n-1} (if you don't know σ)
	μdiff	$\bar{x}_{diff}$	$\frac{\sigma_{diff}}{\sqrt{n}}$	Z (if you know σdiff) T_{n-1} (if you don't know σdiff)
Single **Categorical** Variable *(2 levels)*	p	$\hat{p}$	For Confidence Intervals: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ For Hypothesis Tests: $\sqrt{\frac{p_0(1-p_0)}{n}}$	z
Numerical Response Variable **Categorical** Explanatory Variable *(2 levels)*	μ1-μ2	$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	Z (if you know σ1 and σ2) $T_{min(n1-1, n2-1)}$ (if you don't know σ1 or σ2)
Categorical Response Variable **Categorical** Explanatory Variable *(both have 2 levels)*	p1-p2	$\hat{p}_1 - \hat{p}_2$	For Confidence Intervals: $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ For Hypothesis Tests with Ho:p1-p2=0: $\sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_1} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_2}}$	z

Central Limit Theorem

Hypothesis Testing for Population Parameter

$Ho: pop.param = null\ value$
$Ha: pop.param (\neq\ or\ >\ or\ <) null\ value$

$Test - Stat = \frac{(point\ estimate) - null\ value}{SE}$

When CLT conditions are met.

Types of Variable(s) Involved	Population Parameter	Point Estimate	Standard Error	Distribution: (1) To get Critical Values From (CI) (2) That the Test Statistic Follows (HT)
Single **Numerical** Variable	μ	$\bar{x}$	$\frac{\sigma}{\sqrt{n}}$	Z (if you know σ) T_{n-1} (if you don't know σ)
	μdiff	$\bar{x}_{diff}$	$\frac{\sigma_{diff}}{\sqrt{n}}$	Z (if you know σdiff) T_{n-1} (if you don't know σdiff)
Single **Categorical** Variable *(2 levels)*	p	$\hat{p}$	For Confidence Intervals: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ For Hypothesis Tests: $\sqrt{\frac{p_0(1-p_0)}{n}}$	z
Numerical Response Variable **Categorical** Explanatory Variable *(2 levels)*	μ1-μ2	$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	Z (if you know σ1 and σ2) $T_{min(n1-1, n2-1)}$ (if you don't know σ1 or σ2)
Categorical Response Variable **Categorical** Explanatory Variable *(both have 2 levels)*	p1-p2	$\hat{p}_1 - \hat{p}_2$	For Confidence Intervals: $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ For Hypothesis Tests with Ho:p1-p2=0: $\sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_1} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_2}}$	z

Analyses with a different Hypothesis Testing Structure and don't have Confidence Intervals

Types of Variables	Analysis	Hypotheses
Numerical Response Variable Categorical Explanatory Variable (>2 levels)	**ANOVA**	Ho: μ1=μ2=…=μk Ha: at least one pair of groups has means that are different
Single Categorical Variable (>2 levels)	**Chi-Squared Goodness of Fit Test**	Ho: The data follows the specified distribution. Ha: The data does not follow the specified distribution.
Categorical Response Variable Categorical Explanatory Variable (at least one has >2 levels)	**Chi-Squared Independence Test**	Ho: The two variables are independent/not associated. Ha: The two variables are dependent/associated.

Activity:

Test the hypothesis $H_0 : \mu =$ 10 vs. $H_A : \mu >$ 10 for the following 6 samples. Assume σ = 2.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	8:30am section	10:05am section	11:45am section
$p - value$			
$n = 5000$	1:25pm section	3:05pm section	Everyone
$p - value$			

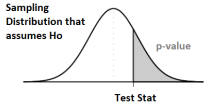
Activity:

Test the hypothesis $H_0 : \mu =$ 10 vs. $H_A : \mu >$ 10 for the following 6 samples. Assume σ = 2.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	8:30am section	10:05am section	11:45am section
$p - value$	0.45	0.39	0.29
$n = 5000$	1:25pm section	3:05pm section	Everyone
$p - value$	0.04	0.0002	≈ 0

- **Sample size ↑**
  - Test Statistic
  - P-value
- **Distance $\bar{x}$ is away from null value ↑**
  - Test Statistic
  - P-value

$$Test - Stat = \frac{\bar{x} - null\ val}{\sigma/\sqrt{n}}$$



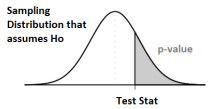
Activity:

Test the hypothesis $H_0 : \mu =$ 10 vs. $H_A : \mu >$ 10 for the following 6 samples. Assume σ = 2.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	8:30am section	10:05am section	11:45am section
$p - value$	0.45	0.39	0.29
$n = 5000$	1:25pm section	3:05pm section	Everyone
$p - value$	0.04	0.0002	≈ 0

- **Sample size ↑**
  - Test Statistic ↑
  - P-value ↓
- **Distance $\bar{x}$ is away from null value ↑**
  - Test Statistic ↑
  - P-value ↓

$$Test - Stat = \frac{\bar{x} - null\ val}{\sigma/\sqrt{n}}$$



Clicker question

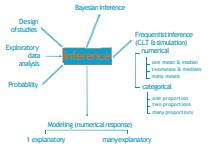
Which of the following is the best method for evaluating the if the distribution of a categorical variable follows a hypothesized distribution?

(a) chi-square test of independence
(b) chi-square test of goodness of fit
(c) anova
(d) linear regression
(e) t-test

Clicker question

Which of the following is the best method for evaluating the if the distribution of *one categorical variable follows a hypothesized distribution*?

(a) chi-square test of independence
(b) chi-square test of goodness of fit
(c) anova
(d) linear regression
(e) t-test

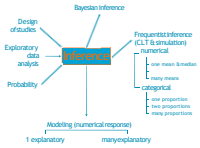


9

Clicker question

Which of the following is the best method for evaluating the relationship between a numerical and a categorical variable with >2 levels?

(a) z-test
(b) chi-square test of goodness of fit
(c) anova
(d) linear regression
(e) t-test

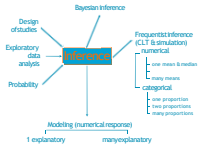


10

Clicker question

Which of the following is the best method for evaluating the relationship between a ***numerical and a categorical variable with >2 levels***?

(a) z-test
(b) chi-square test of goodness of fit
(c) anova
(d) linear regression
(e) t-test



10

Example - Breast Cancer & Age

It is theorized that an important risk factor for breast cancer is age at first birth. An international study was set up to test this hypothesis. Breast-cancer cases were identified among women in selected hospitals in the United States, Greece, Yugoslavia, Brazil, Taiwan, and Japan. Controls were chosen from women of comparable age who were in the hospital at the same time as the cases but who did not have breast cancer. All women were asked about their age at first birth.

The set of women with at least one birth was arbitrarily divided into two categories: (1) women whose age at first birth was less than or equal to 29 years and (2) women whose age at first birth was greater than of equal to 30 years. The following results were found among women with at least one birth: 683 of 3220 women with breast cancer (case women) and 1498 of 10,245 women without breast cancer (control women) had an age at first birth greater than or equal to 30. How can we assess whether this difference is significant or simply due to chance?

11

Example - Breast Cancer & Age

It is theorized that an important risk factor for breast cancer is age at first birth. An international study was set up to test this hypothesis. Breast-cancer cases were identified among women in selected hospitals in the United States, Greece, Yugoslavia, Brazil, Taiwan, and Japan. Controls were chosen from women of comparable age who were in the hospital at the same time as the cases but who did not have breast cancer. All women were asked about their age at first birth.

The set of women with at least one birth was arbitrarily divided into two categories: *(1) women whose age at first birth was less than or equal to 29 years and (2) women whose age at first birth was greater than of equal to 30 years*. The following results were found among women with at least one birth: 683 of 3220 women with breast cancer (case women) and 1498 of 10,245 women without breast cancer (control women) had an age at first birth greater than or equal to 30. How can we assess whether this difference is significant or simply due to chance?

Breast Cancer & Age - set-up

We are given 683 of 3220 women with breast cancer (case women) and 1498 of 10,245 women without breast cancer (control women) had an age at first birth greater than 30.

Is there a difference in the proportions of women who had their first child over 30 for those who do and do not have breast cancer.

		Breast Cancer (case)	No Breast Cancer (Controls)	Total
Age Having First Child	≤ 29	2537	8747	11284
	≥ 30	683	1498	2181
	Total	3220	10245	100465

Breast Cancer & Age - set-up

Analysis: Compare Two Proportions HT

- **cases**: 100465 women (hospital patients) with at least one child
- **variable(s)**: (1) breast cancer status - categorical, (2) age at first birth - categorical
- **parameter of interest**: $p_{case} - p_{ctrl}$
  - Note: p_{case} = $P(age \geq 30 \mid case)$ and p_{case} = $P(age \geq 30 \mid ctrl)$
- **test**: compare two population proportion of independent groups
- **hypotheses (two tailed)**:

$$Ho: p_{case} - p_{ctrl} = 0$$
$$Ha: p_{case} - p_{ctrl} \neq 0$$

Breast Cancer & Age - point estimate

Clicker question

Which of the following is the correct point estimate for this HT?

- **hypotheses (two tailed)**:

$$Ho: p_{case} - p_{ctrl} = 0$$
$$Ha: p_{case} - p_{ctrl} \neq 0$$

		BC (Case)	No BC (Controls)	Total
Age Having First Child	≤ 29	2537	8747	11284
	≥ 30	683	1498	2181
	Total	3220	10245	100465

(a) $\frac{683}{2181} - \frac{1498}{2181}$
(b) $\frac{683}{13465} - \frac{1498}{13465}$
(c) $\frac{2537}{11284} - \frac{683}{2181}$
(d) $\frac{683}{3220} - \frac{1498}{10245}$
(e) $\frac{683}{2181} - \frac{683}{3220}$

Breast Cancer & Age - point estimate

Clicker question

Which of the following is the correct point estimate for this HT?

- **hypotheses (two tailed):**
  $Ho: p_{case} - p_{ctrl} = 0$
  $Ha: p_{case} - p_{ctrl} \neq 0$

Success= Age Having First Child

	BC (Case)	No BC (Controls)	Total
≤ 29	2537	8747	11284
≥ 30	683	1498	2181
Total	3220	10245	100465

(a) $\frac{683}{2181} - \frac{1498}{2181}$
(b) $\frac{683}{13465} - \frac{1498}{13465}$
(c) $\frac{2537}{11284} - \frac{683}{2181}$
(d) $\frac{683}{3220} - \frac{1498}{10245}$
(e) $\frac{683}{2181} - \frac{683}{3220}$

14

Breast Cancer & Age - standard error

Clicker question

Which of the following is the correct standard error for this HT?

	BC (Case)	No BC (Controls)	Total
≤ 29	2537	8747	11284
≥ 30	683	1498	2181
Total	3220	10245	100465
$\hat{p}$	0.212	0.146	0.162

(a) $\sqrt{\frac{0.212\times(1-0.212)}{3220}} + \sqrt{\frac{0.146\times(1-0.146)}{10245}}$
(b) $\sqrt{\frac{0.212\times(1-0.212)}{3220} + \frac{0.146\times(1-0.146)}{10245}}$
(c) $\sqrt{\frac{0.162\times(1-0.162)}{3220} + \frac{0.162\times(1-0.162)}{10245}}$
(d) $\sqrt{\frac{0.212\times(1-0.212)}{13465} + \frac{0.146\times(1-0.146)}{13465}}$
(e) $\sqrt{\frac{0.162\times(1-0.162)}{13465} + \frac{0.162\times(1-0.162)}{13465}}$

15

Breast Cancer & Age - standard error

Clicker question

Which of the following is the correct standard error for this HT?

	BC (Case)	No BC (Controls)	Total
≤ 29	2537	8747	11284
≥ 30	683	1498	2181
Total	3220	10245	100465
$\hat{p}$	0.212	0.146	0.162

(a) $\sqrt{\frac{0.212\times(1-0.212)}{3220}} + \sqrt{\frac{0.146\times(1-0.146)}{10245}}$
(b) $\sqrt{\frac{0.212\times(1-0.212)}{3220} + \frac{0.146\times(1-0.146)}{10245}}$
(c) $\sqrt{\frac{0.162\times(1-0.162)}{3220} + \frac{0.162\times(1-0.162)}{10245}} = 0.0074$
(d) $\sqrt{\frac{0.212\times(1-0.212)}{13465} + \frac{0.146\times(1-0.146)}{13465}}$
(e) $\sqrt{\frac{0.162\times(1-0.162)}{13465} + \frac{0.162\times(1-0.162)}{13465}}$

Must use: **Pooled Proportion = (#successes)/(total)**

*Would we calculate the standard error the same way if we were doing a confidence interval?

15

Breast Cancer & Age - standard error

Clicker question

Which of the following is the correct standard error for this HT?

	BC (Case)	No BC (Controls)	Total
≤ 29	2537	8747	11284
≥ 30	683	1498	2181
Total	3220	10245	100465
$\hat{p}$	0.212	0.146	0.162

(a) $\sqrt{\frac{0.212\times(1-0.212)}{3220}} + \sqrt{\frac{0.146\times(1-0.146)}{10245}}$
(b) $\sqrt{\frac{0.212\times(1-0.212)}{3220} + \frac{0.146\times(1-0.146)}{10245}}$
(c) $\sqrt{\frac{0.162\times(1-0.162)}{3220} + \frac{0.162\times(1-0.162)}{10245}} = 0.0074$
(d) $\sqrt{\frac{0.212\times(1-0.212)}{13465} + \frac{0.146\times(1-0.146)}{13465}}$
(e) $\sqrt{\frac{0.162\times(1-0.162)}{13465} + \frac{0.162\times(1-0.162)}{13465}}$

Must use: **Pooled Proportion = (#successes)/(total)**

*Would we calculate the standard error the same way if we were doing a confidence interval? **NO**

15

Breast Cancer & Age - test statistic & p-value

$$Z = \frac{\hat{p}_{case} - \hat{p}_{ctrl} - 0}{SE} = \frac{0.212 - 0.146}{0.0074} = 8.92$$

$$\text{p-value} = P(Z > 8.92) + P(Z < -8.92) \approx 0$$

Original Question: How can we assess whether this difference is significant or simply due to chance?

Conclusion: Reject the null hypothesis. There exists sufficient evidence to suggest that there is a statistically significant difference in the proportion of women who have their first child at age at 30 or older in those that do and do not have breast cancer.

Breast Cancer & Age - test statistic & p-value

$$Z = \frac{\hat{p}_{case} - \hat{p}_{ctrl} - 0}{SE} = \frac{0.212 - 0.146}{0.0074} = 8.92$$

$$\text{p-value} = P(Z > 8.92) + P(Z < -8.92) \approx 0$$

What else can we conclude with this same result? (Hint: Think about independence).

Breast Cancer & Age - test statistic & p-value

$$Z = \frac{\hat{p}_{case} - \hat{p}_{ctrl} - 0}{SE} = \frac{0.212 - 0.146}{0.0074} = 8.92$$

$$\text{p-value} = P(Z > 8.92) + P(Z < -8.92) \approx 0$$

What else can we conclude with this same result? (Hint: Think about independence).

Conclusion: Reject the null hypothesis. There exists sufficient evidence to suggest that age of first birth in women (≥30/<30) and getting breast cancer are dependent.

Breast Cancer & Age - confidence interval

- Confidence level: 98%

Breast Cancer & Age - confidence interval

- Confidence level: 98%
- Theoretical: Using a critical value based on the Z distr. (z^*):

point estimate ± ME

= point estimate ± z^*× SE

Breast Cancer & Age - confidence interval

- Confidence level: 98%
- Theoretical: Using a critical value based on the Z distr. (z^*):

point estimate ± ME

= point estimate ± z^*× SE

For a confidence interval,

$$SE = \sqrt{\frac{\hat{p}_{case}(1-\hat{p}_{case})}{n_{case}} + \frac{\hat{p}_{ctrl}(1-\hat{p}_{ctrl})}{n_{ctrl}}}$$

$$= \sqrt{\frac{0.212(1-0.212)}{3220} + \frac{0.146(1-0.146)}{10245}} = 0.008$$

$$(0.212 - 0.146) \pm 2.33 \times 0.008 \approx 0.066 \pm 0.0186$$

$$= (0.0474,\ 0.0846)$$

Breast Cancer & Age - confidence interval

Clicker question

The conclusions we make with the following hypothesis test at the α=0.03 significance level is equivalent to conclusions made with a confidence interval with:

- **hypotheses**:

$Ho: p_{case} - p_{ctrl} = 0$

$Ha: p_{case} - p_{ctrl} > 0$

a) Confidence level = 97%
b) Confidence level = 94%
c) Confidence level =98.5%

Breast Cancer & Age - confidence interval

Clicker question

The conclusions we make with the following hypothesis test at the α=0.03 significance level is equivalent to conclusions made with a confidence interval with:

- **hypotheses (two tailed)**:

$Ho: p_{case} - p_{ctrl} = 0$

$Ha: p_{case} - p_{ctrl} > 0$

a) Confidence level = 97%
b) Confidence level = 94%
c) Confidence level =98.5%

For <u>One-Sided Test</u>, Conclusions Agree When:

- Conf. Level = (1- 2α)

Breast Cancer & Age - confidence interval

For <u>One-Sided Test</u>, Conclusions Agree When:

- Conf. Level = (1- 2α)
- α = $\frac{(1-Conf.Level)}{2}$

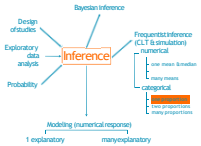
For <u>Two-Sided Test</u>, Conclusions Agree When:

- Conf. level = (1- α)
- α = (1 - Conf. Level)

17

Clicker question

n = 30 and $\hat{p}$= 0.6. Hypotheses: $H_0: p =$ 0.8; $H_A: p <$ 0.8. Which of the following is an appropriate method for calculating the p-value for this test?

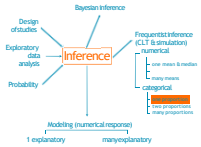


(a) CLT-based inference using the normal distribution
(b) simulation-based inference
(c) exact calculation using the binomial distribution

18

Clicker question

n = 30 and $\hat{p}$= 0.6. Hypotheses: $H_0: p =$ 0.8; $H_A: p <$ 0.8. Which of the following is an appropriate method for calculating the p-value for this test?



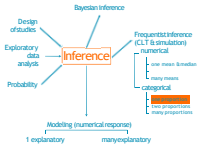
(a) CLT-based inference using the normal distribution
(b) simulation-based inference
(c) exact calculation using the binomial distribution

$$n(\ ) \geq 10?$$
$$n(1-(\ )) \geq 10?$$

18

Clicker question

n = 30 and $\hat{p}$= 0.6. Hypotheses: $H_0: p =$ 0.8; $H_A: p <$ 0.8. Which of the following is an appropriate method for calculating the p-value for this test?



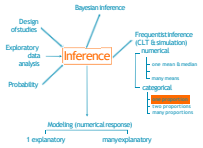
(a) CLT-based inference using the normal distribution
(b) simulation-based inference
(c) exact calculation using the binomial distribution

$$30(\ 0.8\ ) \geq 10?$$
$$\cancel{30(1-(\ 0.8))} \geq 10?$$

18

Clicker question

n = 30 and $\hat{p}$ = 0.6. Hypotheses: $H_0: p = 0.8$; $H_A: p < 0.8$. Which of the following is an appropriate method for calculating the p-value for this test?



(a) CLT-based inference using the normal distribution

(b) simulation-based inference

(c) exact calculation using the binomial distribution

$30(0.8) \geq 10?$

~~$30(1-(0.8)) \geq 10?$~~

SF Conditions not met. Can't use CLT methods.

18

Clicker question

A researcher has tried to set up the following randomization test for *p*. What were her hypotheses, observed sample proportion, and sample size?

Randomization Test for p
1. Place 100 chips in a bag, 20 red and 80 green.
2. Draw 10 chips from the bag without replacement.
3. Note the proportion of the 10 chips that were red and plot it in the randomization distribution.
4. Repeat (2) and (3) many times.
5. In the randomization distribution find the proportion of simulations where the sample proportion was 17% or less; or 23% or more.

a) Ho: p = 0.17; Ha: p < 0.17 $\hat{p} = 0.20$ n=100
b) Ho: p = 0.20; Ha: p ≠ 0.20 $\hat{p} = 0.17$ n=10
c) Ho: p = 0.17; Ha: p < 0.17 $\hat{p} = 0.20$ n=20
d) Ho: p = 0.20; Ha: p < 0.20 $\hat{p} = 0.17$ n=10
e) This is not a well-designed randomization test for a population proportion.

Clicker question

A researcher has tried to set up the following randomization test for *p*. What were her hypotheses, observed sample proportion, and sample size?

Randomization Test for p
1. Place 100 chips in a bag, 20 red and 80 green.
2. Draw 10 chips from the bag **without replacement**.
3. Note the proportion of the 10 chips that were red and plot it in the randomization distribution.
4. Repeat (2) and (3) many times.
5. In the randomization distribution find the proportion of simulations where the sample proportion was 17% or less; or 23% or more.

a) Ho: p = 0.17; Ha: p < 0.17 $\hat{p} = 0.20$ n=100
b) Ho: p = 0.20; Ha: p ≠ 0.20 $\hat{p} = 0.17$ n=10
c) Ho: p = 0.17; Ha: p < 0.17 $\hat{p} = 0.20$ n=20
d) Ho: p = 0.20; Ha: p < 0.20 $\hat{p} = 0.17$ n=10
e) This is not a well-designed randomization test for a population proportion…. Everytime you draw from the bag the proportion you're simulating will not be the same if you don't replace each time.

Clicker question

A researcher has tried to set up the following randomization test for *p*. What were her hypotheses, observed sample proportion, and sample size?

Randomization Test for p
1. Place 100 chips in a bag, 20 red and 80 green.
2. Draw 10 chips from the bag WITH REPLACEMENT.
3. Note the proportion of the 10 chips that were red and plot it in the randomization distribution.
4. Repeat (2) and (3) many times.
5. In the randomization distribution find the proportion of simulations where the sample proportion was 17% or less; or 23% or more.

a) Ho: p = 0.17; Ha: p < 0.17 $\hat{p} = 0.20$ n=100
b) Ho: p = 0.20; Ha: p ≠ 0.20 $\hat{p} = 0.17$ n=10
c) Ho: p = 0.17; Ha: p < 0.17 $\hat{p} = 0.20$ n=20
d) Ho: p = 0.20; Ha: p < 0.20 $\hat{p} = 0.17$ n=10
e) This is not a well-designed randomization test for a population proportion.

Clicker question

A researcher has tried to set up the following randomization test for *p*. What were her hypotheses, observed sample proportion, and sample size?

Randomization Test for p
1. Place 100 chips in a bag, 20 red and 80 green.
2. Draw 10 chips from the bag WITH REPLACEMENT.
3. Note the proportion of the 10 chips that were red and plot it in the randomization distribution.
4. Repeat (2) and (3) many times.
5. In the randomization distribution find the proportion of simulations where the sample proportion was 17% or less; or 23% or more.

a) Ho: p = 0.17; Ha: p < 0.17 $\hat{p} = 0.20$ n=100
b) Ho: p = 0.20; Ha: p ≠ 0.20 ***$\hat{p} = 0.17$*** ***n=10***
c) Ho: p = 0.17; Ha: p < 0.17 $\hat{p} = 0.20$ n=20
d) Ho: p = 0.20; Ha: p < 0.20 $\hat{p} = 0.17$ n=10
e) This is not a well-designed randomization test for a population proportion.

Clicker question

A researcher has tried to set up the following randomization test for *p*. What were her hypotheses, observed sample proportion, and sample size?

Randomization Test for p
1. Place 100 chips in a bag, 20 red and 80 green.
2. Draw 10 chips from the bag WITH REPLACEMENT.
3. Note the proportion of the 10 chips that were red and plot it in the randomization distribution.
4. Repeat (2) and (3) many times.
5. In the randomization distribution find the proportion of simulations where the sample proportion was 17% or less; or 23% or more.

0.20=20 red chips/100 total chips

Ho: p = 0.20; Ha: p ≠ 0.20 $\hat{p} = 0.17$ *n=10*

$p - value = P(\hat{p}\ at\ least\ extreme\ as\ one\ observed | Ho: p = 0.20)$

What the randomization distribution assumes/should be centered at.

Clicker question

A researcher has tried to set up the following randomization test for *p*. What were her hypotheses, observed sample proportion, and sample size?

Randomization Test for p
1. Place 100 chips in a bag, 20 red and 80 green.
2. Draw 10 chips from the bag WITH REPLACEMENT.
3. Note the proportion of the 10 chips that were red and plot it in the randomization distribution.
4. Repeat (2) and (3) many times.
5. In the randomization distribution find the proportion of simulations where the sample proportion was 17% or less; or 23% or more.

0.20=20 red chips/100 total chips

Ho: p = 0.20; Ha: p ≠ 0.20 $\hat{p} = 0.17$ *n=10*

$p - value = P(\hat{p}\ at\ least\ extreme\ as\ one\ observed | Ho: p = 0.20)$

What the randomization distribution assumes/should be centered at.

Clicker question

A researcher has tried to set up the following randomization test for *p*. What were her hypotheses, observed sample proportion, and sample size?

Randomization Test for p
1. Place 100 chips in a bag, 20 red and 80 green.
2. Draw 10 chips from the bag WITH REPLACEMENT.
3. Note the proportion of the 10 chips that were red and plot it in the randomization distribution.
4. Repeat (2) and (3) many times.
5. In the randomization distribution find the proportion of simulations where the sample proportion was 17% or less; or 23% or more.

Looking in both directions on the randomization distribution/sampling distribution for the p-value → 2-tailed test

0.20=20 red chips/100 total chips

Ho: p = 0.20; Ha: p ≠ 0.20 *n=10*

Sample proportion could have been $\hat{p} = 0.23$ *or* $\hat{p} = 0.17$

$p - value = P(\hat{p} \leq 0.17 or\ \hat{p} \geq 0.23 | Ho: p = 0.20)$

What the randomization distribution assumes/should be centered at.

Clicker question

We just conducted an ANOVA test comparing 4 means with a total sample size of n=100. The ANOVA hypothesis was found to be statistically significant at α=0.05.

Which of the following are the appropriate p-value and significance level to use for a post-hoc pairwise comparison test?

(a) $p-value = P\left(|t_{100-4}| > \frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{MSE}{n_1}+\frac{MSE}{n_2}}}\right), \quad \alpha^* = 0.05$

(b) $p-value = P\left(|t_{100-4-1}| > \frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{s_1^2}{100}+\frac{s_2^2}{100}}}\right), \quad \alpha^* = \frac{0.05}{\frac{4(4-1)}{2}}$

(c) $p-value = P\left(|z| > \frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{s_1^2}{100}+\frac{s_2^2}{100}}}\right), \quad \alpha^* = \frac{0.05}{\frac{4(4-1)}{2}}$

(d) $p-value = P\left(|t_{100-4}| > \frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}\right), \quad \alpha^* = \frac{0.05}{\frac{4(4-1)}{2}}$

(e) $p-value = P\left(|t_{100-4}| > \frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{MSE}{n_1}+\frac{MSE}{n_2}}}\right), \quad \alpha^* = \frac{0.05}{\frac{4(4-1)}{2}}$

Clicker question

We just conducted an ANOVA test comparing 4 means with a total sample size of n=100. The ANOVA hypothesis was found to be statistically significant at α=0.05.

Which of the following are the appropriate p-value and significance level to use for a post-hoc pairwise comparison test?

(a) $p-value = P\left(|t_{100-4}| > \frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{MSE}{n_1}+\frac{MSE}{n_2}}}\right), \quad \alpha^* = 0.05$

(b) $p-value = P\left(|t_{100-4-1}| > \frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{MSE}{n_1}+\frac{MSE}{n_2}}}\right), \quad \alpha^* = \frac{0.05}{\frac{4(4-1)}{2}}$

(c) $p-value = P\left(|z| > \frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{s_1^2}{100}+\frac{s_2^2}{100}}}\right), \quad \alpha^* = \frac{0.05}{\frac{4(4-1)}{2}}$

(d) $p-value = P\left(|t_{100-4}| > \frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}\right), \quad \alpha^* = \frac{0.05}{\frac{4(4-1)}{2}}$

(e) $p-value = P\left(|t_{100-4}| > \frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{MSE}{n_1}+\frac{MSE}{n_2}}}\right), \quad \alpha^* = \frac{0.05}{\frac{4(4-1)}{2}}$

Clicker question

Suppose it is known that the distribution of all SUV car sales prices is skewed to the right with mean $50,000 and standard deviation $10,000.

- We randomly select an SUV and note it's price of x=$25,000.
- We also randomly sample 100 SUVs from this population and note that their mean price is $\bar{x} = \$60{,}000$.

- **STATEMENT 1:** $P(X > \$20{,}000) = \mathrm{P}\left(\mathrm{Z} > \frac{\$25{,}000-\$50{,}000}{\$10{,}000}\right) = \mathrm{P}(\mathrm{Z} > -2.5) = 0.9938$
- **STATEMENT 2:** $P(\bar{x} > \$60{,}000) = \mathrm{P}\left(\mathrm{Z} > \frac{\$60{,}000-\$50{,}000}{\$10{,}000/\sqrt{100}}\right) = \mathrm{P}(\mathrm{Z} > 10) < 0.0002$

Which of the following are true statements?

(a) STATEMENT 1 is TRUE and STATEMENT 2 is FALSE
(b) STATEMENT 2 is TRUE and STATEMENT 1 is FALSE
(c) STATEMENT 1 and 2 are BOTH TRUE
(d) STATEMENT 1 and 2 are BOTH FALSE

Clicker question

Suppose it is known that the distribution of all SUV car sales prices is skewed to the right with mean $50,000 and standard deviation $10,000.

- We randomly select an SUV and note it's price of x=$25,000.
- We also randomly sample 100 SUVs from this population and note that their mean price is $\bar{x} = \$60{,}000$.

- **STATEMENT 1:** $P(X > \$20{,}000) = \mathrm{P}\left(\mathrm{Z} > \frac{\$25{,}000-\$50{,}000}{\$10{,}000}\right) = \mathrm{P}(\mathrm{Z} > -2.5) = 0.9938$
- **STATEMENT 2:** $P(\bar{x} > \$60{,}000) = \mathrm{P}\left(\mathrm{Z} > \frac{\$60{,}000-\$50{,}000}{\$10{,}000/\sqrt{100}}\right) = \mathrm{P}(\mathrm{Z} > 10) < 0.0002$

Which of the following are true statements?

(a) STATEMENT 1 is TRUE and STATEMENT 2 is FALSE
(b) STATEMENT 2 is TRUE and STATEMENT 1 is FALSE
(c) STATEMENT 1 and 2 are BOTH TRUE
(d) STATEMENT 1 and 2 are BOTH FALSE

Clicker question

Suppose it is known that the distribution of all SUV car sales prices is skewed to the right with mean $50,000 and standard deviation $10,000.

- We randomly select an SUV and note it's price of x=$25,000.
- We also randomly sample 100 SUVs from this population and note that their mean price is $\bar{x} = \$60{,}000$.

- **STATEMENT 1:** $P(X > \$20{,}000) = \mathrm{P}\left(\mathrm{Z} > \frac{\$25{,}000-\$50{,}000}{\$10{,}000}\right) = \mathrm{P}(\mathrm{Z} > -2.5) = 0.9938$
- **STATEMENT 2:** $P(\bar{x} > \$60{,}000) = \mathrm{P}\left(\mathrm{Z} > \frac{\$60{,}000-\$50{,}000}{\$10{,}000/\sqrt{100}}\right) = \mathrm{P}(\mathrm{Z} > 10) < 0.0002$

Which of the following are true statements?

Statement 1 not true:

- X is not normal.
- Cannot use z-score/z-tables to find probability as shown above.

Clicker question

Suppose it is known that the distribution of all SUV car sales prices is skewed to the right with mean $50,000 and standard deviation $10,000.

- We randomly select an SUV and note it's price of x=$25,000.
- We also randomly sample 100 SUVs from this population and note that their mean price is $\bar{x} = \$60{,}000$.

- **STATEMENT 1:** $P(X > \$20{,}000) = \mathrm{P}\left(\mathrm{Z} > \frac{\$25{,}000-\$50{,}000}{\$10{,}000}\right) = \mathrm{P}(\mathrm{Z} > -2.5) = 0.9938$
- **STATEMENT 2:** $P(\bar{x} > \$60{,}000) = \mathrm{P}\left(\mathrm{Z} > \frac{\$60{,}000-\$50{,}000}{\$10{,}000/\sqrt{100}}\right) = \mathrm{P}(\mathrm{Z} > 10) < 0.0002$

Which of the following are true statements?

Statement 2 is true:

- *Because CLT conditions hold,* $\bar{x} \sim N\left(\mu = \$50{,}000, std.dev = \frac{\$10{,}000}{\sqrt{100}}\right)$
- *Thus we can use z-score/z-tables to find probability.*

CLT Conditions

Independence:

- Random sampling/assignment AND
- n<10% of population

Sample Size/Skew:

- n>30 OR population distribution is normal

Clicker question

Suppose it is known that the distribution of all SUV car sales prices is skewed to the right with mean $50,000 and standard deviation $10,000.

- We randomly select an SUV and note it's price of x=$25,000.
- We also randomly sample 100 SUVs from this population and note that their mean price is $\bar{x} = \$60{,}000$.

- **STATEMENT 1:** We can say that the price of our one randomly selected SUV (x=$25,000) is unusual, because x=$25,000 is lower than two standard deviations away from the population mean.
  - (ie $50,000-2($10,000)=$30,000)
- **STATEMENT 2:** We can say that the average price of our 100 randomly selected SUVs ($\bar{x} = \$60{,}000$) is unusual, because $60,000 is higher than two standard errors away from the population mean.
  - (ie ($\$50{,}000 + 2\frac{\$10{,}000}{\sqrt{100}} = \$52{,}000$)

Which of the following are true statements?

(a) STATEMENT 1 is TRUE and STATEMENT 2 is FALSE
(b) STATEMENT 2 is TRUE and STATEMENT 1 is FALSE
(c) STATEMENT 1 and 2 are BOTH TRUE
(d) STATEMENT 1 and 2 are BOTH FALSE

Clicker question

Suppose it is known that the distribution of all SUV car sales prices is skewed to the right with mean $50,000 and standard deviation $10,000.

- We randomly select an SUV and note it's price of x=$25,000.
- We also randomly sample 100 SUVs from this population and note that their mean price is $\bar{x} = \$60{,}000$.

- **STATEMENT 1:** We can say that the price of our one randomly selected SUV (x=$25,000) is unusual, because x=$25,000 is lower than two standard deviations away from the population mean.
  - (ie $50,000-2($10,000)=$30,000)
- **STATEMENT 2:** We can say that the average price of our 100 randomly selected SUVs ($\bar{x} = \$60{,}000$) is unusual, because $60,000 is higher than two standard errors away from the population mean.
  - (ie ($\$50{,}000 + 2\frac{\$10{,}000}{\sqrt{100}} = \$52{,}000$)

Which of the following are true statements?

(a) STATEMENT 1 is TRUE and STATEMENT 2 is FALSE
(b) STATEMENT 2 is TRUE and STATEMENT 1 is FALSE
(c) STATEMENT 1 and 2 are BOTH TRUE
(d) STATEMENT 1 and 2 are BOTH FALSE

Clicker question

Suppose it is known that the distribution of all SUV car sales prices is skewed to the right with mean $50,000 and standard deviation $10,000.

- We randomly select an SUV and note it's price of x=$25,000.
- We also randomly sample 100 SUVs from this population and note that their mean price is $60,000.

- **STATEMENT 1:** We can say that the price of our one randomly selected SUV (x=$25,000) is unusual, because x=$25,000 is lower than two standard deviations away from the population mean .
  - (ie $50,000-2($10,000)=$30,000)
- **STATEMENT 2:** We can say that the average price of our 100 randomly selected SUVs ($\bar{x} = \$60{,}000$) is unusual, because $60,000 is higher than two standard errors away from the population mean .
  - (ie ($50,000 + $2\frac{\$10,000}{\sqrt{100}} = \$52{,}000$)

Which of the following are true statements?

Rule: Unusual "observation" = "observations" that are 2 standard deviations away from the mean (or z-score >2 or <-2)

Clicker question

Suppose it is known that the distribution of all SUV car sales prices is skewed to the right with mean $50,000 and standard deviation $10,000.

- We randomly select an SUV and note it's price of x=$25,000.
- We also randomly sample 100 SUVs from this population and note that their mean price is $60,000.

- **STATEMENT 1:** We can say that the price of our one randomly selected SUV (x=$25,000) is unusual, because x=$25,000 is lower than two standard deviations away from the population mean .
  - (ie $50,000-2($10,000)=$30,000)
- **STATEMENT 2:** We can say that the average price of our 100 randomly selected SUVs ($\bar{x} = \$60{,}000$) is unusual, because $60,000 is higher than two standard errors away from the population mean .
  - (ie ($50,000 + $2\frac{\$10,000}{\sqrt{100}} = \$52{,}000$)

Which of the following are true statements?

Rule: Unusual "observation" = "observation" that are 2 standard deviations away from the mean (or z-score >2 or <-2)

→ Only works when the distribution of "observations" are normal.

Clicker question

Suppose it is known that the **distribution of SUV car sales prices is skewed to the right** with mean $50,000 and standard deviation $10,000.

- We randomly select an SUV and note it's price of x=$25,000.
- We also randomly sample 100 SUVs from this population and note that their mean price is $60,000.

- **STATEMENT 1:** We can say that the price of our one randomly selected SUV (x=$25,000) is unusual, because x=$25,000 is lower than two standard deviations away from the population mean .
  - (ie $50,000-2($10,000)=$30,000)
- **STATEMENT 2:** We can say that the average price of our 100 randomly selected SUVs ($\bar{x} = \$60{,}000$) is unusual, because $60,000 is higher than two standard errors away from the population mean .
  - (ie ($50,000 + $2\frac{\$10,000}{\sqrt{100}} = \$52{,}000$)

Which of the following are true statements?

Rule: Unusual "observation" = "observation" that are 2 standard deviations away from the mean (or z-score >2 or <-2)

→ Only works when the distribution of "observations" are normal.

Distribution of a single SUV sales price is not normal… can't use this rule.

Clicker question

Suppose it is known that the distribution of all SUV car sales prices is skewed to the right with mean $50,000 and standard deviation $10,000.

- We randomly select an SUV and note it's price of x=$25,000.
- We also randomly sample 100 SUVs from this population and note that their mean price is $60,000.

- **STATEMENT 1:** We can say that the price of our one randomly selected SUV (x=$25,000) is unusual, because x=$25,000 is lower than two standard deviations away from the population mean .
  - (ie $50,000-2($10,000)=$30,000)
- **STATEMENT 2:** We can say that the average price of our **n=100** randomly selected SUVs ($\bar{x} = \$60{,}000$) is unusual, because $60,000 is higher than two standard errors away from the population mean .
  - (ie ($50,000 + $2\frac{\$10,000}{\sqrt{100}} = \$52{,}000$)

Which of the following are true statements?

Rule: Unusual "observation" = "observation" that are 2 standard deviations away from the mean

→ Only works when the distribution of "observations" are normal.

Distribution of sample SUV means (sample size 100) IS normal because CLT conditions met… CAN use this rule!

Power Review

Are you here today?

a.) yes!

Clicker question

For a hypothesis test, all but one of these probabilities have been given a name in this class. Which one is it?

(a) P(reject Ho | Ho is true)

(b) P(fail to reject Ho | Ho is true)

(c) P(reject Ho | Ha is true)

(d) P(fail to reject Ho | Ha is true)

Clicker question

For a hypothesis test, all but one of these probabilities have been given a name in this class. Which one is it?

(a) P(reject Ho | Ho is true) = P(Type 1 Error) = α

(b) P(fail to reject Ho | Ho is true)

(c) P(reject Ho | Ha is true) = Power = 1- β

(d) P(fail to reject Ho | Ha is true) = P(Type 2 Error) = β = 1 - Power

4. Hypothesis tests are prone to decision errors

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.
  - **P(Type 1 Error) = α** = P(reject Ho|Ho is true)
- A *Type 2 Error* is failing to reject the null hypothesis when H_A is true.
  - **P(Type 2 Error) = β** = P(fail to reject Ho|Ha is true)
- *Power* is the probability of *correctly* rejecting H_0, and hence the complement of P(Type 2 Error).
  - **Power = 1 – β** = P(reject Ho|Ha is true)

To ↑ power/↓ β you can:
- ↑ n
- ↑ δ
- ↑ α

Regression and ANOVA Review

Clicker question

We are interested in how fat contributes to the amount of calories in a hotdog. Below is the ANOVA table for the multiple linear regression (where hotdog calories is the response variable). What percent of the variability in hotdog calories is explained by fat (or the SLR model)?

Regression ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Fat	1	10946	10946	319.34	1.78E-09
Residuals	11	377.1	34.3		
Total	12	11323.1			

(a) 10946

(b) $1-\left(\frac{377.1}{11323.1}\right)$

(c) $1-\left(\frac{377.1}{11323.1}\cdot\frac{12}{11}\right)$

(d) $\sqrt{1-\left(\frac{377.1}{11323.1}\right)}$

Clicker question

We are interested in how fat contributes to the amount of calories in a hotdog. Below is the ANOVA table for the multiple linear regression (where hotdog calories is the response variable. What percent of the variability in hotdog calories is explained by fat (or the SLR model)?

Regression ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Fat	1	10946	10946	319.34	1.78E-09
Residuals	11	377.1	34.3		
Total	12	11323.1			

(a) 10946

(b) $\mathbf{1-\left(\frac{377.1}{11323.1}\right)=\left(\frac{10946}{11323.1}\right)=R^2}$

(c) $1-\left(\frac{377.1}{11323.1}\cdot\frac{12}{11}\right)$

(d) $\sqrt{1-\left(\frac{377.1}{11323.1}\right)}$

Clicker question

We are interested in how fat contributes to the amount of calories in a hotdog. Below is the ANOVA table for the multiple linear regression (where hotdog calories is the response variable. What is the correlation between hotdog fat and calories?

Regression ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Fat	1	10946	10946	319.34	1.78E-09
Residuals	11	377.1	34.3		
Total	12	11323.1			

(a) 10946

(b) $1-\left(\frac{377.1}{11323.1}\right)$

(c) $1-\left(\frac{377.1}{11323.1}\cdot\frac{12}{11}\right)$

(d) $\sqrt{1-\left(\frac{377.1}{11323.1}\right)}$

Clicker question

We are interested in how fat contributes to the amount of calories in a hotdog. Below is the ANOVA table for the multiple linear regression (where hotdog calories is the response variable. What is the correlation between hotdog fat and calories?

Regression ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Fat	1	10946	10946	319.34	1.78E-09
Residuals	11	377.1	34.3		
Total	12	11323.1			

(a) 10946 = SSRegression

(b) $1-\left(\frac{377.1}{11323.1}\right)=R^2$

(c) $1-\left(\frac{377.1}{11323.1}\cdot\frac{12}{11}\right)=$ Adjusted R^2

(d) $\sqrt{1-\left(\frac{377.1}{11323.1}\right)}=\sqrt{R^2}$ *only works for Simple Linear Regression*

Clicker question

It is always the case that $\textbf{Adjusted } R^2 \leq R^2$.

What type of linear regression would give you $\textbf{Adjusted } R^2 = R^2$?

Regression ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Fat	1	10946	10946	319.34	1.78E-09
Residuals	11	377.1	34.3		
Total	12	11323.1			

$$1-\left(\frac{377.1}{11323.1}\cdot\frac{12}{11}\right)=\text{Adjusted } R^2$$

$$1-\left(\frac{377.1}{11323.1}\right)=R^2$$

Clicker question

It is always the case that $\textbf{Adjusted } R^2 \leq R^2$.

What type of linear regression would give you $\textbf{Adjusted } R^2 = R^2$?

Regression ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
…	…	…	…	…	…
Residuals	n-k-1	SSresid	MSresid		
Total	n-1	SST			

$$1-\left(\frac{SSresid}{SST}\cdot\frac{n-1}{n-k-1}\right)=\text{Adjusted } R^2$$

$$1-\left(\frac{SSresid}{SST}\right)=R^2$$

Clicker question

It is <u>always</u> the case that **Adjusted $R^2 \leq R^2$.**

What type of linear regression would give you **Adjusted $R^2 = R^2$**?

Regression ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
…	…	…	…	…	…
Residuals	n-k-1	SSresid	MSresid		
Total	n-1	SST			

$$1 - \left(\frac{\text{SSresid}}{\text{SST}} \cdot \frac{n-1}{n-k-1}\right) = \text{Adjusted } R^2$$

$$1 - \left(\frac{\text{SSresid}}{\text{SST}}\right) = R^2$$

When k=0 (there are k=0 predictors in the equation… just an intercept.)

Clicker question

Which of the following is false?

(a) The SST and df_T are the same for ANOVA (which compares >2 means) and ANOVA for regression.

(b) The df_{error} is the same for ANOVA (which compares >2 means) and ANOVA for regression.

(c) $\sqrt{\frac{SST}{df_T}} = s_y$, where *y* is the response variable.

(d) $SS_{residuals}$ $(or\ SS_E)$ is equal to the sum of the squares of the residuals of the model.

(e) SSG for ANOVA (which compares >2 means), represents the total variability of y between groups.

Clicker question

Which of the following is false?

(a) The SST and df_T are the same for ANOVA (which compares >2 means) and ANOVA for regression.

(b) The df_{error} is the same for ANOVA (which compares >2 means) and ANOVA for regression.

(c) $\sqrt{\frac{SST}{df_T}} = s_y$, where *y* is the response variable.

(d) $SS_{residuals}$ $(or\ SS_E)$ is equal to the sum of the squares of the residuals of the model.

(e) SSG for ANOVA (which compares >2 means), represents the total variability of y between groups.

ANOVA for Comparing >2 Means

	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
Between groups	k-1	$SSG = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2$	MSG = SSG/(k-1)	MSG/MSE	P(F > MSG/MSE)
Error Within Groups	n-k	$SSE = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2$	MSE = SSE/(n-k)		
Total	n-1	$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2$			

<u>k: # of groups</u>; n: # of obs.

ANOVA for Linear Regression

	DF	Sum Sq	Mean Sq	F Value	Pr(>F)
Explanatory Variable 1	# of slopes for this expl. var. in model	◯	◯	◯	◯
Explanatory Variable 2	# of slopes for this expl. var. in model	◯	◯	◯	◯
...	...	...	...	...	...
Explanatory Variable w	# of slopes for this expl. var. in model	◯	◯	◯	◯
Error Residuals	n-k-1	$SS_{Res}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$MS_{Res}=SS_{Res}/Df_{Res}$		
Total	n-1	$SS_{Tot}=\sum_{i=1}^{n}(y_i-\bar{y})^2$			

k: # of slopes being estimate; n: # of obs.

Clicker question

Which of the following is false?

(a) $MSG=\frac{SSG}{df_G}$ for ANOVA (which compares >2 means), can be thought of as the average variability between groups.

(b) $MSE=\frac{SSE}{df_E}$ for ANOVA (which compares >2 means), can be thought of as the average variability within groups.

(c) A high F statistic for ANOVA (which compares >2 means), shows there is sufficient evidence to suggest the means are all different.

(d) For a simple linear regression, $1-\frac{SS_{Regression}}{SST}$ represents the percent of response variable variability that is NOT explained by the explanatory variable.

(e) In an ANOVA (which compares >2 means and the groups = levels of a categorical explanatory variable) $1-\frac{SS_{Group}}{SST}$ represents the percent of response variable variability that is NOT explained by the explanatory variable.

Clicker question

Which of the following is false?

(a) $MSG=\frac{SSG}{df_G}$ for ANOVA (which compares >2 means), can be thought of as the average variability between groups.

(b) $MSE=\frac{SSE}{df_E}$ for ANOVA (which compares >2 means), can be thought of as the average variability within groups.

(c) A high F statistic for ANOVA (which compares >2 means), shows there is sufficient evidence to suggest the means are all different.

(d) For a simple linear regression, $1-\frac{SS_{Regression}}{SST}$ represents the percent of response variable variability that is NOT explained by the explanatory variable.

(e) In an ANOVA (which compares >2 means and the groups = levels of a categorical explanatory variable) $1-\frac{SS_{Group}}{SST}$ represents the percent of response variable variability that is NOT explained by the explanatory variable.

Clicker question

Which of the following is true?

(a) A low R means that the two variables have no strong relationship or association.

(b) A high R means that a simple linear regression with variables will be a good fit (one variable as response, the other explanatory).

(c) The correlation between housing prices (in $US) and time will be different if you convert the currency to New Zealand dollars.

(d) An influence point is more likely to decrease R and change the regression slope, than a leverage point.

Clicker question

Which of the following is true?

(a) A low R means that the two variables have no strong relationship or association. →They could have a strong nonlinear relationship. R only measures strength of *linear relationships.*

(b) A high R means that a simple linear regression with variables will be a good fit (one variable as response, the other explanatory). →*A high R doesn't mean the regression conditions will hold. You could still have a nonlinear relationship or non-constant variance for instance.*

(c) The correlation between housing prices (in $US) and time will be different if you convert the currency to New Zealand dollars. →*R is unitless… not affected by unit conversions.*

(d) An influence point is more likely to decrease R and change the regression slope, than a leverage point. →*An influence point will fall further away from the* ***regression line****, causing the R to decrease, and the slope of the line to change.*

Clicker question

Would the result from the first step of backwards elimination agree if you used the p-value method vs. adjusted R^2 method?

Full Model: Hotdog Calories ~ **Fat +SatFat**

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   28.897      5.904   4.895 0.000628 ***
Fat           12.366      2.368   5.222 0.000389 ***
SatFat        -7.666      5.389  -1.422 0.185323
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.6 on 10 degrees of freedom
Multiple R-squared:  0.9723,    Adjusted R-squared:  0.9668
F-statistic: 175.5 on 2 and 10 DF,  p-value: 1.629e-08
```

Model that Removed **Fat**: Adjusted R^2 =0.9637
Model that Removed **SatFat**: Adjusted R^2 = 0.8837

(a) Yes
(b) No

Clicker question

Would the result from the first step of backwards elimination agree if you used the p-value method vs. adjusted R^2 method?

Full Model: Hotdog Calories ~ **Fat +SatFat**

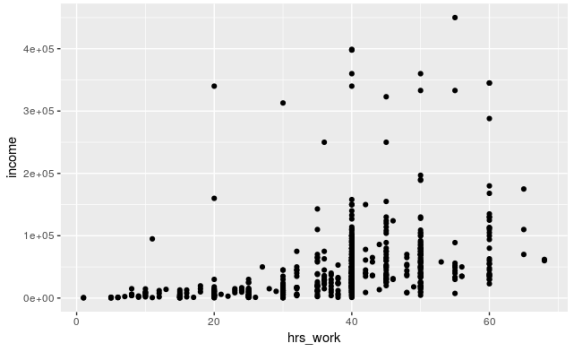
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   28.897      5.904   4.895 0.000628 ***
Fat           12.366      2.368   5.222 0.000389 ***
SatFat        -7.666      5.389  -1.422 0.185323
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.6 on 10 degrees of freedom
Multiple R-squared:  0.9723,    Adjusted R-squared:  0.9668
F-statistic: 175.5 on 2 and 10 DF,  p-value: 1.629e-08
```

Model that Removed **Fat**: Adjusted R^2 =0.9637
Model that Removed **SatFat**: Adjusted R^2 = 0.8837

(a) Yes
(b) No – Full model best using Adjusted R^2, Model with just Fat best using p-value method.

We are interested in predicting the annual income of a person that works 80 hours a week. What are THREE REASONS we SHOULD NOT fit a simple linear regression model and make this prediction with the data that we have (shown below)?



We are interested in predicting the annual income of a person that works 80 hours a week. What are THREE REASONS we SHOULD NOT fit a simple linear regression model and make this prediction with the data that we have (shown below)?

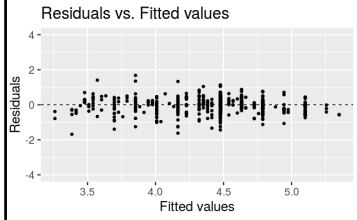
1. **Condition Violation:** The relationship is nonlinear.
2. **Condition Violation:** The variance of the residuals (when fitting a linear model) would increase as hours worked increases…. so non-homoscedastic residuals.
3. **Extrapolation:** 80 hrs/week working is outside the range of data *(highest our data goes is 65hrs/week)*. So our prediction would be an unreliable extrapolation.

Clicker question

We are interested in predicting the annual income of a person that works 45 hours a week with the same data set. The plots below are the the residuals vs. fitted values plots for two separate simple linear regression models. Which model would produce a smaller prediction interval?

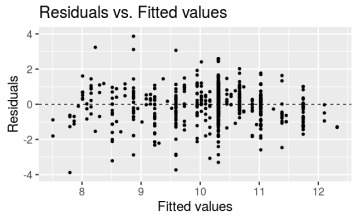
Model 1

$log_e(income) = b_0 + b_1(hrswork)$



Model 2

$log_{10}(income) = b_0 + b_1(hrswork)$



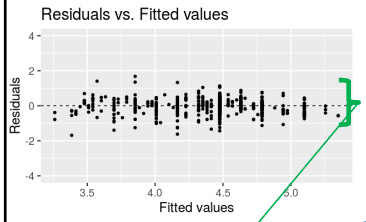
(a) Model 1
(b) Model 2

Clicker question

We are interested in predicting the annual income of a person that works 45 hours a week with the same data set. The plots below are the the residuals vs. fitted values plots for two separate simple linear regression models. Which model would produce a smaller prediction interval?

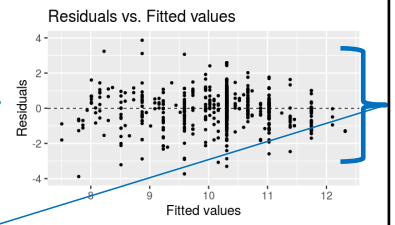
Model 1

$log_e(income) = b_0 + b_1(hrswork)$



Model 2

$log_{10}(income) = b_0 + b_1(hrswork)$



(a) Model 1
(b) Model 2

$$\hat{y} \pm t^{\star}_{n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x^{\star} - \bar{x})^2}{(n-1)s_x^2}}$$

Model 1 has smaller residual standard deviation, so will have a smaller prediction interval.