Midterm 1 Review

2/13/19

Coming up…

- **Lab Assignment 4** due tomorrow
- **Midterm 1** Monday 2/18

Feedback + Help
- *Problem Set 2 returned by Saturday morning.*
- *Weekend Office Hours: Saturday 2pm-4pm (Lavonne, Old Chem 203B)*

Midterm 1

- **Material Covered:**
  - Units 1-3.2
- **What to bring:**
  - Cheat sheet
    - 1 page (8.5" by 11")
    - Front/back ok
    - CAN be typed
  - Calculator (no phones)
- **Provided:**
  - Z-tables
- **Exam**
  - 4 written questions (like AEs)
  - 5 T/F
  - 10 MC

Midterm 1 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

Final Review Suggestions

- **Concept review:**
  - Lecture slides **(has material not in the videos/book)**
  - Video notes
  - 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)
    - Equations
    - 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.
    - Conditions
    - Common misconceptions (lecture notes)
    - What test to use under certain conditions

1

Unit 1 – Data Collection

Clicker question

Researcher would like to sample SAT scores from high schools in a state. Each high school has a diverse student body. They randomly select 12 high schools in the state and collect SAT scores from all students in the selected high schools. What type of sampling is this?

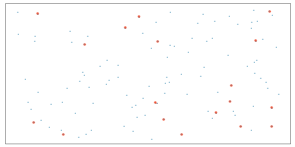
(a) Simple Random Sample
(b) Stratified Sampling
(c) Cluster Sampling
(d) Multistage Sampling

Clicker question

Researcher would like to sample SAT scores from high schools in a state. Each high school has a diverse student body. They randomly select 12 high schools in the state and collect SAT scores from all students in the selected high schools. What type of sampling is this?

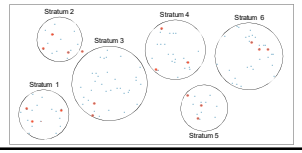
(a) Simple Random Sample
(b) Stratified Sampling
(c) Cluster Sampling
(d) Multistage Sampling

Simple random:
Drawing names from a hat

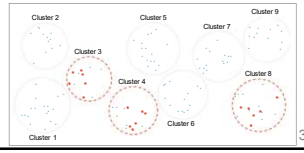


Cluster: heterogenous clusters
Sample all chosen clusters

Stratified: homogenous strata
Stratify to control for age group



Multistage: heterogeneous clusters
Random sample in chosen clusters



3

	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

Unit 1 – Data Visualization and Summary Statistics

Unit 1 - Visualizations

- What's an appropriate plot to use for each of the following types of data?
  1. **One numerical variable**
  2. **One categorical variable**
  3. **Two numerical variables**
  4. **One categorical variable and one numerical variable**
  5. **Two categorical variables**

Unit 1 - Visualizations

- 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, segmented barplot*

Unit 1 – Data Visualization and Summary Statistics

- One Numerical Variable

Unit 1 - Visualizations

- What four aspects of a numerical variable should we be prepared to discuss?

Unit 1 - Visualizations

- What four aspects of a numerical variable should we be prepared to discuss?
  - *Shape*: skewness, modality
  - *Center*: an estimate of a *typical* observation in the distribution (mean, median, mode, etc.)
    - Notation: μ: population mean, $\bar{x}$ sample mean
  - *Spread*: measure of variability in the distribution (standard deviation, IQR, range, etc.)
  - *Unusual observations*: observations that stand out from the rest of the data that may be suspected outliers

Clicker question

Which of the following sets of statistics are all robust to outliers?

(a) Variance
(b) Mean, Median
(c) Median, IQR, range
(d) Median, IQR
(e) IQR, Standard deviation

Clicker question

Which of the following sets of statistics are all **robust to outliers**?

~~(a) Variance~~
~~(b) Mean~~, Median
(c) Median, IQR, ~~range~~
(d) Median, IQR
(e) IQR, ~~Standard deviation~~

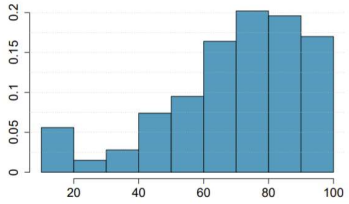
Robust statistics

- Mean and standard deviation are **easily affected by extreme observations** since the value of each data point contributes to their calculation.
- Median and IQR are **more robust to outliers.**
- Therefore we choose **median & IQR** (over mean&SD) when describing skewed distributions.
- We choose **mean & SD** when describing symmetric distributions, as they are more useful in using mathematical theory to make inferences.

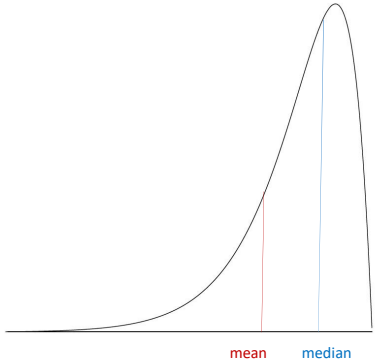
13

Clicker question

Which of the following is false?



(a) The box plot would have outliers only on the lower end.
(b) The median is between 70 and 80.
(c) More than 25% of the data is above 90.
(d) More than 50% of the data have positive Z scores.
(e) The mean is likely to be smaller than the median.



$$z = \frac{obs - mean}{standard\ deviation}$$

positive

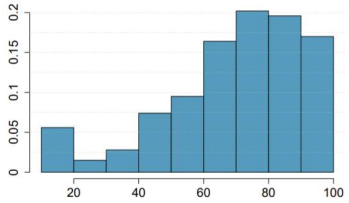
$$= \frac{(obs\ larger\ than\ mean - mean)}{standard\ deviation}$$

NEVER negative

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{n}}$$

Clicker question

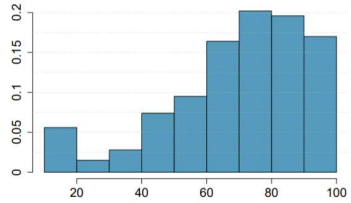
Which of the following is false?



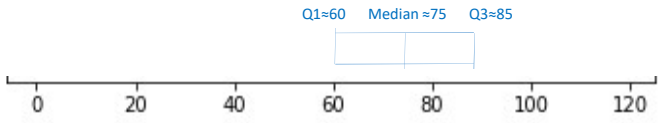
(a) The box plot would have outliers only on the lower end.
(b) The median is between 70 and 80.
(c) More than 25% of the data is above 90.
(d) ~~More than 50% of the data have positive Z scores.~~ (MEAN SMALLER THAN MEDIAN)
(e) ~~The mean is likely to be smaller than the median.~~ (LEFT SKEWED)

Clicker question

Which of the following is false?

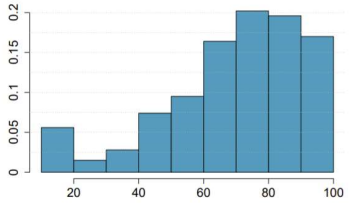


(a) The box plot would have outliers only on the lower end.
(b) The median is between 70 and 80.
(c) More than 25% of the data is above 90.
(d) More than 50% of the data have positive Z scores.
(e) The mean is likely to be smaller than the median.

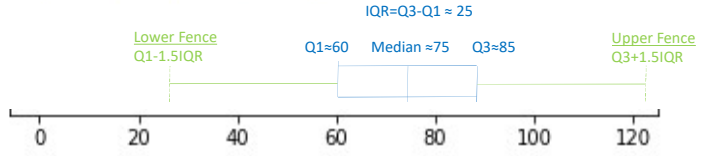


Clicker question

Which of the following is false?

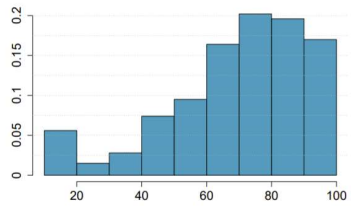


(a) The box plot would have outliers only on the lower end.
(b) The median is between 70 and 80.
(c) More than 25% of the data is above 90.
(d) More than 50% of the data have positive Z scores.
(e) The mean is likely to be smaller than the median.

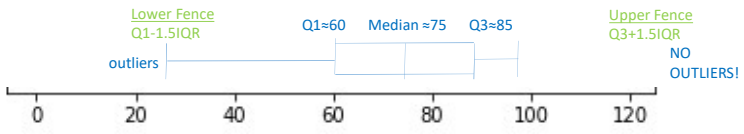


Clicker question

Which of the following is false?



(a) The box plot would have outliers only on the lower end.
(b) The median is between 70 and 80.
(c) More than 25% of the data is above 90.
(d) More than 50% of the data have positive Z scores.
(e) The mean is likely to be smaller than the median.



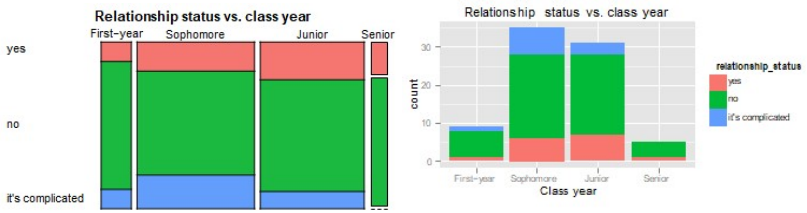
Unit 1 – Data Visualization and Summary Statistics

- Two Categorical Variables

Clicker question

Is a mosaic plot or a segmented bar plot better at determining if two categorical variables are independent/not-associated or dependent/assoc?

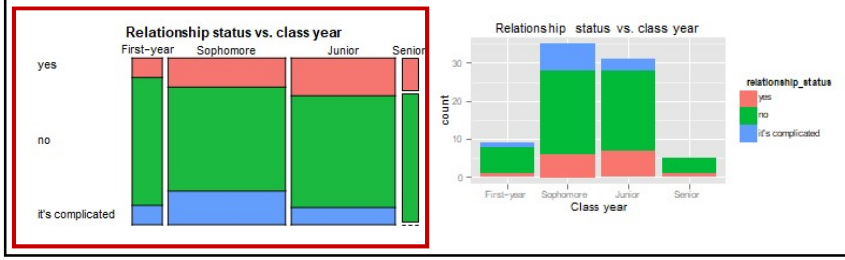
a) Mosaic plot
b) Segmented bar plot



Clicker question

Is a mosaic plot or a segmented bar plot better at determining if two categorical variables are independent/not-associated or dependent/assoc?

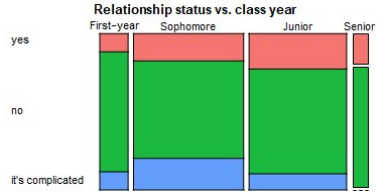
a) Mosaic plot
b) Segmented bar plot



Clicker question

What probabilities would you need to calculate to show that being in a relationship is associated with **class year**?

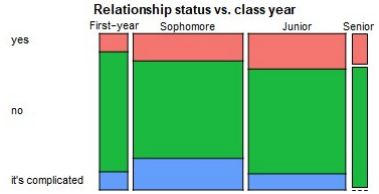
a) P(yes|first-year)≠ P(no|first-year) ≠ P(it's complicated|first-year)
b) P(yes|first-year)≠ P(yes|sophomore) ≠ P(yes|junior) ≠ P(yes|senior)
c) P(no|first-year)≠ P(no|sophomore) ≠ P(no|junior) ≠ P(no|senior)
d) Show (b) and (c)



Clicker question

What probabilities would you need to calculate to show that being in a relationship is associated with **class year**?

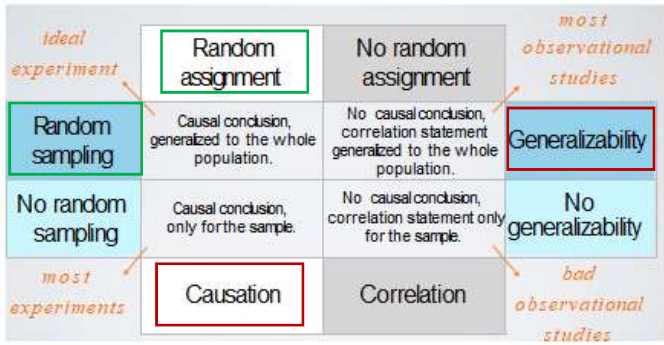
a) P(yes|first-year)≠ P(no|first-year) ≠ P(it's complicated|first-year)
b) P(yes|first-year)≠ P(yes|sophomore) ≠ P(yes|junior) ≠ P(yes|senior)
c) P(no|first-year)≠ P(no|sophomore) ≠ P(no|junior) ≠ P(no|senior)
d) Show (b) and (c)



Unit 1 – Making Inferences

Types of Inferences you Can Make and When You Can Make Them

Types of Inferences you Can Make and When You Can Make Them



Unit 1 – Making Inferences

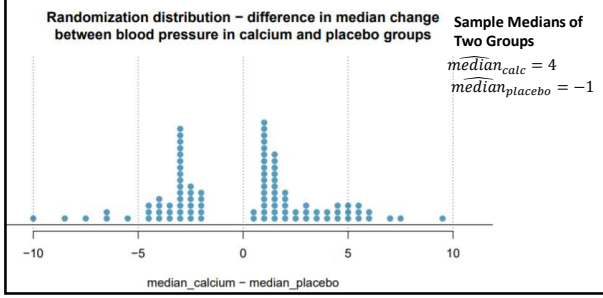
...with Randomization Testing

Calcium for treating blood pressure. Lyle et al. (1987) ran an experiment to study the effect of a calcium supplement on the blood pressure of African American males. A group of 10 men received a calcium supplement, and another group of 11 men received a placebo. The experiment lasted 12 weeks. Both before and after the 12-week period, each man had his systolic blood pressure measured while at rest. The changes in blood pressure are given in table below.

Calcium:	-5	-4	-3	-2	1	7	10	11	17	18	$Median_{calcium} = \frac{1+7}{2} = 4$
Placebo:	-11	-5	-3	-3	-1	-1	-1	2	3	5	12 $Median_{placebo} = -1$

Clicker question

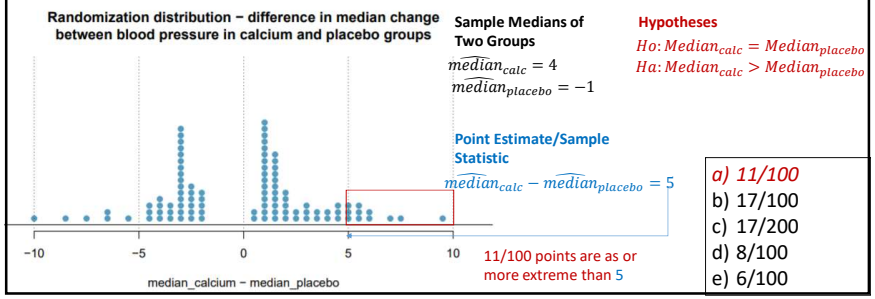
We would like to test if the median change in blood pressure of the calcium group is **greater** than that of the placebo group. Below is a randomization distribution created for this research question (100 simulations), what is the p-value for the randomization test?



a) 11/100
b) 17/100
c) 17/200
d) 8/100
e) 6/100

Clicker question

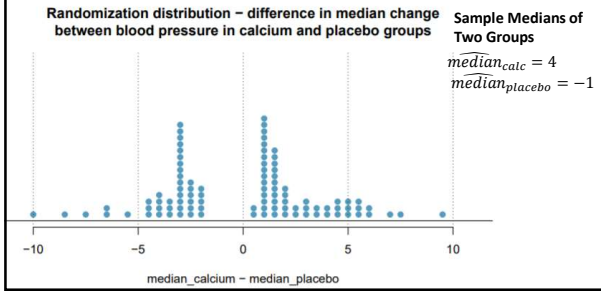
We would like to test if the median change in blood pressure of the calcium group is **greater** than that of the placebo group. Below is a randomization distribution created for this research question (100 simulations), what is the p-value for the randomization test?



a) 11/100
b) 17/100
c) 17/200
d) 8/100
e) 6/100

Clicker question

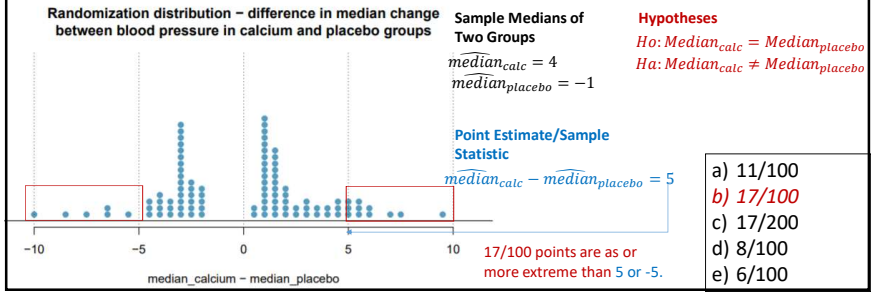
We would like to test if the median change in blood pressure of the calcium group is **different** than that of the placebo group. Below is a randomization distribution created for this research question (100 simulations), what is the p-value for the randomization test?



a) 11/100
b) 17/100
c) 17/200
d) 8/100
e) 6/100

Clicker question

We would like to test if the median change in blood pressure of the calcium group is **different** than that of the placebo group. Below is a randomization distribution created for this research question (100 simulations), what is the p-value randomization test?



a) 11/100
b) 17/100
c) 17/200
d) 8/100
e) 6/100

Unit 1 – Making Inferences

...what is the definition of a p-value?
…what conclusions can we make?

Definition

p-value = P(a sample statistic/data that is as "extreme" or "more extreme" than the one observed |null hypothesis is true)

Conclusions

p-value < significance level (commonly 0.05)

Reject the null hypothesis. There exists sufficient evidence to suggest the alternative hypothesis.

p-value ≥ significance level (commonly 0.05)

Fail to reject the null hypothesis. There does not exist sufficient evidence to suggest the alternative hypothesis.

Unit 2 – General Probability Properties

General Probability Rules

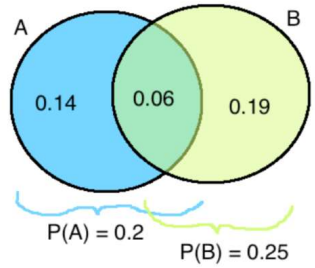
- **Disjoint (mutually exclusive) events** cannot happen at the same time
  - For disjoint A and B: P(A and B) = 0
- **Complementary events** A and $\bar{A}$ are disjoint events where $\bar{A}$ = "not A". (A or $\bar{A}$) represents all possibilities
  - For complementary A and $\bar{A}$: P(A) + P($\bar{A}$) = 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)

General Probability Rules

Clicker question

Which of the following is true?

(a) A and B are independent.
(b) P(A but not B) = 0.2
(c) P(A | B) = 0.06 / 0.14
(d) P(A or B) = 0.14 + 0.19 + 0.06
(e) P(neither A nor B) = 1 - 0.06

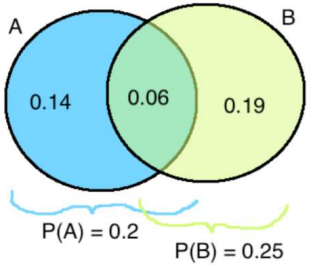


General Probability Rules

Clicker question

Which of the following is true?

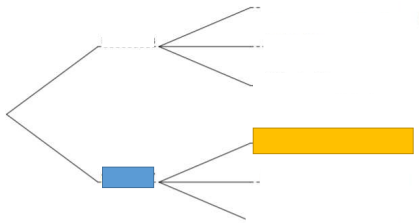
(a) A and B are independent.
(b) P(A but not B) = 0.2
(c) P(A | B) = 0.06 / 0.14
(d) P(A or B) = 0.14 + 0.19 + 0.06
(e) P(neither A nor B) = 1 - 0.06



Unit 2 – Binomial and Normal Probabilities

-How do we know which equations to use?

About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the posterior probability that they are identical?



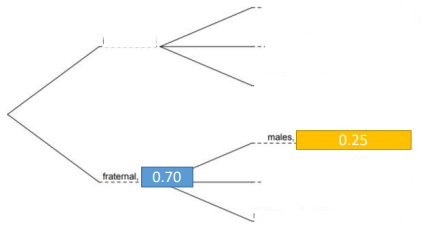
$$P(iden \mid f) = \frac{P(iden \ \& \ f)}{P(f)}$$

Clicker question

Which of the following *could* be the correct input for a probability tree?

a) 0.25 0.70
b) 0.70 0.25
c) 0.30 0.25
d) 0.50 0.30

About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the posterior probability that they are identical?



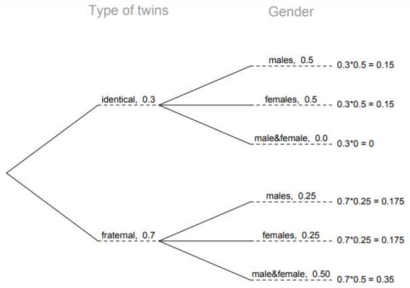
$$P(iden \mid f) = \frac{P(iden \ \& \ f)}{P(f)}$$

Clicker question

Which of the following *could* be the correct input for a probability tree?

a) 0.25 0.70
b) 0.70 0.25
c) 0.30 0.25
d) 0.50 0.30

About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the posterior probability that they are identical?



$$P(iden \mid f) = \frac{P(iden \ \& \ f)}{P(f)}$$

Clicker question

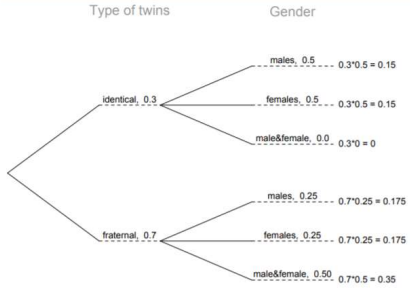
What is *P(iden|f)*?

a) $\frac{0.3}{0.5+0.25}$

b) $\frac{0.15}{0.15+0.175}$

c) $\frac{0.15}{0.175}$

About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the posterior probability that they are identical?



$$P(iden \mid f) = \frac{P(iden \ \& \ f)}{P(f)}$$

Clicker question

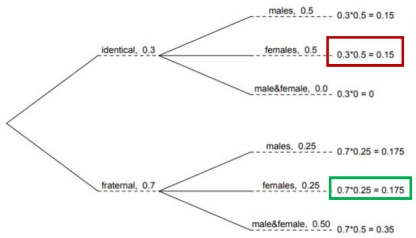
What is *P(iden|f)*?

a) $\frac{0.3}{0.5+0.25}$

b) $\frac{0.15}{0.15+0.175}$

c) $\frac{0.15}{0.175}$

About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the posterior probability that they are identical?



$$P(iden \mid f) = \frac{P(iden \ \& \ f)}{P(f)} = \frac{0.15}{0.15+0.175} = 0.46$$

Clicker question

What would you need to do solve this problem?

Suppose it is known that the probability a STA101 student had a horse growing up is 0.09. We collected a random sample of 100 STA101 students and asked them if they had a horse growing up. What's the probability that at least 3 students had a horse growing up?

(a) Calculate exactly one binomial expression.

(b) Calculate multiple binomial expressions (and/or 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?

Suppose it is known that the probability a STA101 student had a horse growing up is 0.09. We collected a random sample of 100 STA101 students and asked them if they had a horse growing up. What's the probability that at least 3 students had a horse growing up?

(a) Calculate exactly one binomial expression.

(b) Calculate multiple binomial expressions (and/or 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?

Suppose it is known that the probability a STA101 student had a horse growing up is 0.09. We collected a random sample of 100 STA101 students and asked them if they had a horse growing up. What's the probability that at least 3 students had a horse growing up?

Step 1: If X = # of students out of random sample of 100 STA101 student that had a horse, show that $X \sim Bin(n = 100, p = 0.09)$

Binomial distribution is appropriate because:

a) n=100 trials (and asking for the probability of exactly/at least/at most k=3 of these trials are a success=horse)

b) each trial has probability 0.09

c) independent trials

d) only two possibilities (horse/no horse)

Step 2: Calculate multiple binomial expressions (and/or add/subtract/do other things with them) to find P(X ≥ 3).

Clicker question

What would you need to do solve this problem?

Suppose it is known that the probability a STA101 student had a horse growing up is 0.09. We collected a random sample of 100 STA101 students and asked them if they had at least one horse growing up. What's the probability that at least 3 students had a horse growing up?

Step 2: Calculate multiple binomial expressions (and add/subtract/do other things with them) to find P(X ≥ 3).

(a) $P(X \geq 3) = \binom{100}{3} 0.09^3 (1-0.09)^{100-3}$

(b) $P(X \geq 3) = \binom{100}{3} 0.09^3 (1-0.09)^{100-3} + \binom{100}{4} 0.09^4 (1-0.09)^{100-4} + \cdots + \binom{100}{100} 0.09^{100} (1-0.09)^{100-100}$ *(because Dr. Ellison is mean)*

(c) $P(X \geq 3) = 1 - \left[\binom{100}{3} 0.09^3 (1-0.09)^{100-3}\right]$

(d) $P(X \geq 3) = 1 - \left[\binom{100}{0} 0.09^0 (1-0.09)^{100-0}\right]$

(e) $P(X \geq 3) = 1 - \left[\binom{100}{0} 0.09^0 (1-0.09)^{100-0} + \binom{100}{1} 0.09^1 (1-0.09)^{100-1} + \binom{100}{2} 0.09^2 (1-0.09)^{100-2}\right]$

Clicker question

What would you need to do solve this problem?

Suppose it is known that the probability a STA101 student had a horse growing up is 0.09. We collected a random sample of 100 STA101 students and asked them if they had at least one horse growing up. What's the probability that at least 3 students had a horse growing up?

Step 2: Calculate multiple binomial expressions (and add/subtract/do other things with them) to find P(X ≥ 3).

(a) $P(X \geq 3) = \binom{100}{3} 0.09^3 (1-0.09)^{100-3}$

(b) $P(X \geq 3) = \binom{100}{3} 0.09^3 (1-0.09)^{100-3} + \binom{100}{4} 0.09^4 (1-0.09)^{100-4} + \cdots + \binom{100}{100} 0.09^{100} (1-0.09)^{100-1}$ *(because Dr. Ellison is mean)*

(c) $P(X \geq 3) = 1 - \left[\binom{100}{3} 0.09^3 (1-0.09)^{100-3}\right]$

(d) $P(X \geq 3) = 1 - \left[\binom{100}{0} 0.09^0 (1-0.09)^{100-0}\right]$

(e) $\mathbf{P(X \geq 3) = 1 - \left[\binom{100}{0} 0.09^0 (1-0.09)^{100-0} + \binom{100}{1} 0.09^1 (1-0.09)^{100-1} + \binom{100}{2} 0.09^2 (1-0.09)^{100-2}\right]}$

Clicker question

What would you need to do solve this problem?

Suppose it is known that the probability a STA101 student had a horse growing up is 0.09. We collected a random sample of 100 STA101 students and asked them if they had at least one horse growing up. What's the probability that 3 students had a horse growing up?

(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?

Suppose it is known that the probability a STA101 student had a horse growing up is 0.09. We collected a random sample of 100 STA101 students and asked them if they had at least one horse growing up. What's the probability that 3 students had a horse growing up?

(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).

$$P(X = 3) = \binom{100}{3} 0.09^3 (1 - 0.09)^{100-3}$$

Clicker question

What would you need to do solve this problem?

Suppose it is known that the probability a STA101 student had a horse growing up is 0.09. We collected a random sample of 100 STA101 students and asked them if they had at least one horse growing up. What's the probability that the first 5 we ask had a horse, and then the last 95 we ask did not have a horse?

(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?

Suppose it is known that the probability a STA101 student had a horse growing up is 0.09. We collected a random sample of 100 STA101 students and asked them if they had at least one horse growing up. What's the probability that the first 5 we ask had a horse, and then the last 95 we ask did not have a horse?

(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?

Suppose it is known that the probability a STA101 student had a horse growing up is 0.09. We collected a random sample of 100 STA101 students and asked them if they had at least one horse growing up. What's the probability that the first 5 we ask had a horse, and then the last 95 we ask did not have a horse?

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

$$P(first\ 5\ had\ horse\ and\ last\ 95\ didn't)$$

$$= \underbrace{(0.09)(0.09)(0.09)(0.09)\ (0.09)}_{5}\underbrace{(0.91)(0.91)\ldots(0.91)}_{95}$$

$$= 0.09^5 0.91^{95}$$

Clicker question

What would you need to do solve this problem?

Suppose it is known that the probability a STA101 student had a horse growing up is 0.09 and the probability that they had a dog is 0.62 and the probability that they had a dog and a horse is 0.05. What is the probability a student had a dog or a horse growing up?

(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?

Suppose it is known that the probability a STA101 student had a horse growing up is 0.09 and the probability that they had a dog is 0.62 and the probability that they had a dog and a horse is 0.05. What is the probability a student had a dog or a horse growing up?

(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).

$$P(dog\ or\ horse) = (0.09) + (0.62) - (0.05)$$

Clicker question

What would you need to do solve this problem?

Suppose it is known that the probability a STA101 student had a horse growing up is 0.09. We collected a random sample of 200 STA101 students and asked them if they had at least one horse growing up. What's the probability that at least 50 students had a horse growing up?

(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?

Suppose it is known that the probability a STA101 student had a horse growing up is 0.09. We collected a random sample of 200 STA101 students and asked them if they had at least one horse growing up. What's the probability that at least 50 students had a horse growing up?

(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?

Suppose it is known that the probability a STA101 student had a horse growing up is 0.09. We collected a random sample of 200 STA101 students and asked them if they had at least one horse growing up. What's the probability that at least 50 students had a horse growing up?

Step 1: X = # of students out of random sample of 200 STA101 students that had a horse.
Show that $X \sim Bin(n = 200, p = 0.09)$

- ✓n=independent trials
- ✓Each trial is independent
- ✓Each trial is success (had a horse) or a failure (didn't have a horse)
- ✓Probability each trial is a success is 0.09

Step 2: Show that $X \sim N(\mu = np, \sigma = \sqrt{np(1-p)})$

SF Conditions Hold:

- ✓200(.09)≥10
- ✓200(1-.09)≥10

Step 3: $P(X \geq 50) = P\left(Z \geq \frac{50 - (np)}{\sqrt{np(1-p)}}\right) = P\left(Z \geq \frac{50 - (200 \cdot 0.09)}{\sqrt{200 \cdot 0.09 \cdot (1 - 0.09)}}\right) = P(Z \geq 7.9) \approx 0$

Use Z-tables

5. ***Cats on YouTube.*** If you randomly select a video on YouTube, the probability that it involves a cat is 0.11. Over the course of a week, you watch 100 videos on YouTube using an app that randomly selects videos (the random video picker).

How many cat videos would you need to see to suspect that the random video picker is biased towards **or** against cat videos? You can provide a range if need be.

(*Hint:* Think about what would be considered an expected number of videos.)

5. ***Cats on YouTube.*** If you randomly select a video on YouTube, the probability that it involves a cat is 0.11. Over the course of a week, you watch 100 videos on YouTube using an app that randomly selects videos (the random video picker).

How many cat videos would you need to see to suspect that the random video picker is biased towards **or** against cat videos? You can provide a range if need be.

(*Hint:* Think about what would be considered an expected number of videos.)

1. What does "biased toward/against" mean?
   - Biased towards cat videos = unusually hight # of cat videos.
   - Biased against cat videos = unusually low # of cat videos.

5. ***Cats on YouTube.*** If you randomly select a video on YouTube, the probability that it involves a cat is 0.11. Over the course of a week, you watch 100 videos on YouTube using an app that randomly selects videos (the random video picker).

How many cat videos would you need to see to suspect that the random video picker is biased towards **or** against cat videos? You can provide a range if need be.

(*Hint:* Think about what would be considered an expected number of videos.)

1. What does "biased toward/against" mean?
   - Biased towards cat videos = unusually high # of cat videos.
   - Biased against cat videos = unusually low # of cat videos.
2. Let X = number of cat videos you watch.
   IF $X \sim N(\mu = ?, \sigma = ?)$, then we can say:
   - $\mu + 2\sigma$ is a unusually high # of cat videos
   - $\mu - 2\sigma$ is a unusually low # of cat videos

5. ***Cats on YouTube.*** If you randomly select a video on YouTube, the probability that it involves a cat is 0.11. Over the course of a week, you watch 100 videos on YouTube using an app that randomly selects videos (the random video picker).

How many cat videos would you need to see to suspect that the random video picker is biased towards **or** against cat videos? You can provide a range if need be.

(*Hint:* Think about what would be considered an expected number of videos.)

1. What does "biased toward/against" mean?
   - Biased towards cat videos = unusually high # of cat videos.
   - Biased against cat videos = unusually low # of cat videos.
2. Let X = number of cat videos you watch.
   IF $X \sim N(\mu = ?, \sigma = ?)$, then we can say:
   - $\mu + 2\sigma$ is a unusually high # of cat videos
   - $\mu - 2\sigma$ is a unusually low # of cat videos

3. How can we check IF $X \sim N(\mu = ?, \sigma = ?)$ is true?

5. ***Cats on YouTube.*** If you randomly select a video on YouTube, the probability that it involves a cat is 0.11. Over the course of a week, you watch 100 videos on YouTube using an app that randomly selects videos (the random video picker).

How many cat videos would you need to see to suspect that the random video picker is biased towards **or** against cat videos? You can provide a range if need be.

(*Hint:* Think about what would be considered an expected number of videos.)

1. What does "biased toward/against" mean?
   - Biased towards cat videos = unusually high # of cat videos.
   - Biased against cat videos = unusually low # of cat videos.
2. Let X = number of cat videos you watch.
   IF $X \sim N(\mu = ?, \sigma = ?)$, then we can say:
   - $\mu + 2\sigma$ is a unusually high # of cat videos
   - $\mu - 2\sigma$ is a unusually low # of cat videos

3. How can we check IF $X \sim N(\mu = ?, \sigma = ?)$ is true?

4. IF:
   - $X \sim Bin(n = ?, p = ?)$ AND
   - $np \geq 10$ and $n(1-p) \geq 10$

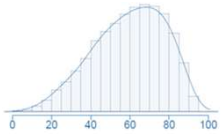
   THEN it is the case that
   $X \sim N(\mu = np, \sigma = \sqrt{np(1-p)})$.

See board!

Unit 3 – Properties of Sampling Distributions and the Central Limit Theorem

Clicker question

Four plots: Determine which plot (A, B, or C) is which.
(1) At top: distribution for a population ($\mu = 60, \sigma = 18$),
(2) a single random sample of 500 observations from this population,
(3) a distribution of 500 sample means from random samples with size 18,
(4) a distribution of 500 sample means from random samples with size 81.

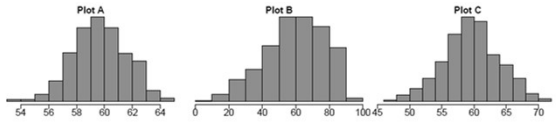


(a) (2) - B; (3) - A; (4) - C

(b) (2) - A; (3) - B; (4) - C

(c) (2) - C; (3) - A; (4) - D

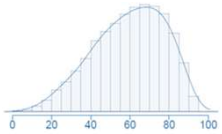
(d) (2) - B; (3) - C; (4) - A



17

Clicker question

Four plots: Determine which plot (A, B, or C) is which.
(1) At top: distribution for a population ($\mu = 60, \sigma = 18$),
(2) a single random sample of 500 observations from this population,
(3) a distribution of 500 sample means from random samples with size 18,
(4) a distribution of 500 sample means from random samples with size 81.

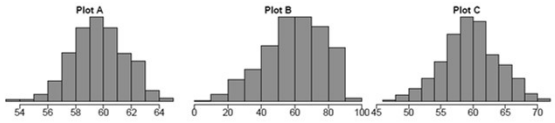


(a) (2) - B; (3) - A; (4) - C

(b) (2) - A; (3) - B; (4) - C

(c) (2) - C; (3) - A; (4) - D

(d) *(2) - B; (3) - C; (4) - A*



17

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

Would you expect most houses in Topanga to cost more or less than $1.3 million? Hint: What is most likely the shape of this distribution?

18

A random housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

Would you expect most houses in Topanga to cost more or less than $1.3 million? Hint: What is most likely the shape of this distribution?

Since the distribution is probably right skewed, the median would be less than the mean, and a majority of observations would be lower than the mean.

18

A random housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

Clicker question

Can we estimate the probability that a randomly chosen house in Topanga costs more than $1.4 million using the normal distribution?

(a) yes
(b) no

A random housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

Clicker question

Can we estimate the probability that a randomly chosen house in Topanga costs more than $1.4 million using the normal distribution?

(a) yes
(b) no

The POPULATION (of houses) distribution is NOT normal.
$X \sim N(\mu = \$1.3m, \sigma = \$0.3m)$

A random housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

Clicker question

Can we estimate the probability that the mean of 60 randomly chosen houses in Topanga is more than $1.4 million?

(a) yes
(b) no

A random housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

Clicker question

Can we estimate the probability that the mean of 60 randomly chosen houses in Topanga is more than $1.4 million?

(a) yes
(b) no

The sampling distribution IS normal....
$\bar{x} \sim N(\mu = \$1.3m, SE = \frac{\$0.3m}{\sqrt{60}})$

...because:

Independence

- Random sampling is used.
- n=60<10% of Topanga, CA homes.

Skewness/Sample size

- n>30

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than $1.4 million?

In order to calculate $P(\bar{X} > 1.4 \text{ mil})$ we need to first determine the distribution of $\bar{X}$. According to the CLT,

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than $1.4 million?

In order to calculate $P(\bar{X} > 1.4 \text{ mil})$ we need to first determine the distribution of $\bar{X}$. According to the CLT,

$$\bar{X} \sim N\left(mean = \quad , SE = \quad\right)$$

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than $1.4 million?

In order to calculate $P(\bar{X} > 1.4 \text{ mil})$ we need to first determine the distribution of $\bar{X}$. According to the CLT,

$$\bar{X} \sim N\left(mean = 1.3, SE = \frac{0.3}{\sqrt{60}} = 0.0387\right)$$

Remember…
If a random variable
□~N($mean_{□}$, $SD_{□}$), then
$\frac{□ - mean_{□}}{SD_{□}} \sim N(0,1)$

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than $1.4 million?

In order to calculate $P(\bar{X} > 1.4 \text{ mil})$ we need to first determine the distribution of $\bar{X}$. According to the CLT,

$$\bar{X} \sim N\left(mean = 1.3, SE = \frac{0.3}{\sqrt{60}} = 0.0387\right)$$

Remember…
If a random variable
□~N($mean_{□}$, $SD_{□}$), then
$\frac{□ - mean_{□}}{SD_{□}} \sim N(0,1)$

$$P(\bar{X} > 1.4) = P\left(Z > \frac{1.4 - 1.3}{0.0387}\right)$$

Z-table

$$= P(Z > 2.58) = 1 - P(Z \leq 2.58)$$

$$= 1 - 0.9951$$

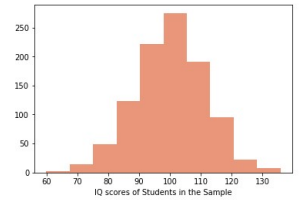
Unit 3 – Confidence Intervals and Margin of Error

A random sample of IQ scores 13 high school students was collected with a mean of 110 points and standard deviation 11 points. The sample distribution is plotted below.

Clicker question

Can we create a 98% confidence interval using Central Limit Theorem methods?

(a) yes
(b) no



A random sample of IQ scores 13 high school students was collected with a mean of 110 points and standard deviation 11 points. The sample distribution is plotted below.

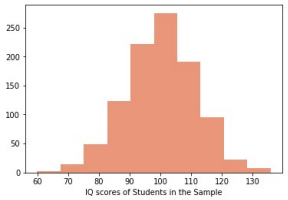
Clicker question

Can we create a 98% confidence interval using Central Limit Theorem methods?

(a) yes
(b) no

Independence
- ✓ Random sampling
- ✓ n<10% of all high school students

Skewness/Sample Size
- ~~n>30? OR~~
- ~~We know σ and population is normal~~

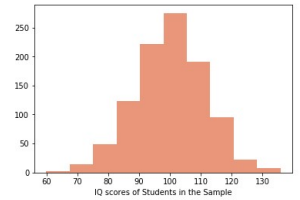


A random sample of IQ scores 13 high school students was collected with a mean of 110 points and standard deviation 11 points. The sample distribution is plotted below. Suppose we know that the population standard deviation is 15.

Clicker question

Can we create a 98% confidence interval using Central Limit Theorem methods?

(a) yes
(b) no



A random sample of IQ scores 13 high school students was collected with a mean of 110 points and standard deviation 11 points. The sample distribution is plotted below. Suppose we know that the population standard deviation is 15.

Clicker question

Can we create a 98% confidence interval using Central Limit Theorem methods?

(a) yes

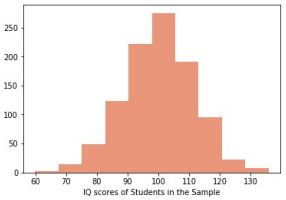
(b) no

Independence

- ✓ Random sampling
- ✓ n<10% of all high school students

Skewness/Sample Size

- ~~n>30? OR~~
- We know σ and population is normal



$$\bar{x} \pm z^*_{0.01}\frac{\sigma}{\sqrt{n}} = 110 \pm 2.33\frac{15}{\sqrt{13}}$$

Suppose we know that the population standard deviation of high school IQ scores is 15 and the scores are normally distributed.

What minimum sample size do we need to create 98% confidence interval using Central Limit Theorem methods with margin of error of at most 2?

Know:

$$ME = z^*_{0.01}\frac{\sigma}{\sqrt{n}}$$

$$ME \leq 2$$

Solve:

$$z^*_{0.01}\frac{\sigma}{\sqrt{n}} \leq 2$$

$$2.33\frac{15}{\sqrt{n}} \leq 2$$

$$2.33\frac{15}{2} \leq \sqrt{n}$$

$$17.45 \leq \sqrt{n}$$

$$304.5 \leq n$$

$$305 \leq n$$