Chapter 1

Intermediate Statistical Investigations Test Bank

Question types:	FIB = Fill in the blank	Calc = Calculation
	Ma = Matching	MS = Multiple select
	MC = Multiple choice	TF = True-false

CHAPTER 1 TERMINAL LEARNING OUTCOMES

TLO1-1: Apply the six-step investigative process in the context of a well-designed experiment.

TLO1-2: Partitioning variation in the response variable into variation explained by the model and unexplained variation, and measuring and reporting the percentage of variation explained

TLO1-3: Assess the statistical significance of the difference between two groups on a quantitative response variable using both simulation and theory-based approaches

TLO1-4: Compare more than two treatments on a quantitative response using both simulation and theory-based approaches

TLO1-5: Apply Post-hoc analysis after significant *F*-test (pairwise differences, as well as confidence and prediction intervals for single means)

TLO1-6: Understand statistical power and how it is impacted by sample size, variability within groups, number of groups, and significance level

Section 1.1: Sources of Variation in an Experiment

LO1.1-1: Apply the six-step investigative process.

LO1.1-2: Distinguish experiments and observational studies.

LO1.1-3: Review basic study design principles such as inclusion criteria and random assignment.

LO1.1-4: Define terminology specific to an experimental study (e.g., treatments).

LO1.1-5: Produce a Sources of Variation diagram for an experiment.

Questions 1 through 3: A study published in *Psychological Science* in 2007 examined a possible link between mindset and health. The following is an excerpt from the abstract of the article: “84 female room attendants working in seven different hotels were measured on physiological health variables affected by exercise. Those in the informed condition were told that the work they do (cleaning hotel rooms) is good exercise and satisfies the Surgeon General's recommendations for an active lifestyle. Examples of how their work was exercise were provided. Subjects in the control group were not given this information.”

1. Identify the experimental units in this study.

   A. The eighty-four room attendants
   B. The seven different hotels
   C. The physiological health variables
   D. The two groups (informed and control)

   Ans: A; LO: 1.1-4; Difficulty: Easy; Type: MC

2. The researchers chose to include room attendants from seven different hotels (as opposed to using stricter inclusion criteria that would limit the study to room attendants at one particular hotel). Describe the consequences of this decision.

   Using broader inclusion criteria may ______ (increase/decrease) the amount of variation in the observed health outcomes. However, this decision also __________ (supports/limits) generalizability to a larger population of room attendants.

   Ans: increase, supports; LO: 1.1-3; Difficulty: Medium; Type: FIB

3. Room attendants were randomly assigned to either the informed condition or the control group. What is the most important reason for the random assignment?
   A. Random assignment ensures that the study is double-blind.
   B. Random assignment reduces the impact of outliers.
   C. Random assignment creates two groups of room attendants that are as similar as possible, which supports cause-and-effect conclusions.
   D. Random assignment makes it possible to generalize the results to the population.

   Ans: C; LO: 1.1-3; Difficulty: Medium; Type: MC

Questions 4 through 6: An online retailer is using an experiment to decide whether to modify their website. When visitors type in the web address or click a link to the site, they are randomly re-directed to one of two versions of the website: the version that has been in use for the last year (version A) or an updated version (version B). The retailer's goal is to maximize the amount of time (in minutes) visitors stay on the site.

4. Identify the experimental units and variables. *Note: One of the answer choices will not be used.*

   Experimental units:
   Explanatory variable:
   Response variable:

   A. Version of the website (A and B)
   B: Online retailers
   C: Visitors to the website
   D: Time spent on the website (in minutes)

   Ans: Experimental units: C, Explanatory variable: A, Response variable: D; LO: 1.1-4; Difficulty: Easy; Type: Ma

5. Consider two possible models for analyzing time spent on this retailer's website.

   Single-mean model:

   A. $\textit{Predicted time spent on site} = 12.33, \textit{SE of residuals} = 4.64$

   Separate-means model:

   B. $\textit{Predicted time spent on site} = \begin{cases} 9.9 \textit{ for Version A} \\ 14.7 \textit{ for Version B} \end{cases}, \ \textit{SE of residuals} = 3.98$

Does the version of the website appear to explain any of the variation in time spent on the site? *Note: If more than one of these justifications is appropriate, select multiple answers.*

A. Yes, because the mean time spent on the site is higher for Version B than for Version A.
B. Yes, because the SE of the residuals is smaller for the separate-means model than for the single-mean model.
C. No, because the mean time spent on the site is not the same for Version A and for Version B.
D. No, because the SE of the residuals is smaller for the separate-means model than for the single-mean model.

Ans: A, B; LO: 1.1-1; Difficulty: Medium; Type: MS

6. The researcher decides that the difference between Version A and Version B in this study is meaningful. Is it reasonable to generalize these results to all customers of this retailer?
   A. Yes, because visitors to the website were randomly assigned to either Version A or Version B.
   B. Yes, because the study's inclusion criteria would exclude potential subjects who are not customers.
   C. It depends whether visitors to the website knew about the research question being investigated. The study may not be double-blind.
   D. It depends who visited the website during the study period. The sample may not be representative.

Ans: D; LO: 1.1-1; Difficulty: Medium; Type: MC

Questions 7 through 8: Researchers at a university were interested in the effectiveness of a calculus workshop program for students who fail Calculus I and need to retake the course. As part of the study, students who were retaking Calculus I were allowed to enroll in a calculus workshop at their own discretion. At the end of the grading term, all students (even those with different instructors) took the same final exam. The researchers then compared the scores for those who enrolled in the workshop while re-taking calculus to those who re-took calculus without enrolling in the workshop.

7. Is this an experiment? Justify your answer.
   A. Yes, this is an experiment, because there was a treatment group who enrolled in the calculus workshop and a control group that did not.
   B. Yes, this is an experiment, because the study was double-blind (as long as the calculus teachers did not know which students enrolled in the workshop).
   C. No, this is an observational study, because it does not take place in a laboratory or other tightly controlled research environment.
   D. No, this is an observational study, because the choice of whether to participate in the workshop was made by the students not the researchers.

Ans: D; LO: 1.1-2; Difficulty: Easy; Type: MC

8. Which of the following are sources of unexplained variation in this study? Select all that apply.
    A. Whether or not students enrolled in the workshop
    B. Whether or not students had failed a calculus class in the past
    C. Student attendance in class (number of absences)
    D. Student motivation to study calculus
    E. Calculus instructor
    F. Difficulty of the final exam

    Ans: C, D, E; LO: 1.1-5; Difficulty: Easy; Type: MS

9. A study published in *Athletic Training* examined the effects of three different types of knee stabilizing braces on agility test speed. College football players from all different positions (running back, wide receiver, linebacker, lineman, etc.) were recruited to participate in the study. All players in the study had torn their ACL (anterior cruciate ligament) in the past, and needed to wear a knee brace to play football. Agility tests were administered in an outdoor football stadium, and the time to complete the test was recorded by a Lafayette photoelectric Cell and Light Time Unit (in seconds).

    Put the components of the study into the correct boxes in the Sources of Variation Diagram. *Note: Some boxes will include more than one answer.*

Observed variation in:	Sources of explained variation	Sources of unexplained variation
Inclusion criteria: Design:		

    A. College football players from all different positions
    B. Type of knee brace
    C. Players' current condition (health, mood, motivation, etc.)
    D. Time to complete agility test (in seconds)
    E. Measurement error
    F. History of torn ACL and need for a knee brace
    G. Details of the agility test
    H. Players' natural speed and agility
    I. Environmental factors (weather, wind, etc.)

    Ans: Observed variation in: D; Inclusion criteria: A, F; Design: G; Sources of explained variation: B; Sources of unexplained variation: C, E, H, I; LO: 1.1-5; Difficulty: Medium; Type: Ma

10. In a separate-means model, the standard error of the residuals can be thought of as the typical deviation of an observed response from:
    A. The residuals (prediction errors)
    B. The response predicted by the model (group mean)
    C. The overall mean of the response variable
    D. The overall mean of the explanatory variable

    Ans: B; LO: 1.1-1; Difficulty: Easy; Type: MC

11. A study published in the *Journal of Sports Science & Medicine* tested the effectiveness of the Power Balance © bracelet, which has been marketed as a way to improve balance, flexibility, strength, and power through the use of hologram technology. Subjects, who were all college athletes, completed tests of their athletic performance while wearing either a Power Balance © bracelet or a plain rubber placebo bracelet. The bracelets were covered with a wristband, so the athletes and those measuring their performance could not see which bracelet was being worn. Only the researcher who analyzed the data knew which measurements corresponded to the Power Balance © bracelet and which to the placebo bracelet. Classify this study.
    A. This study is not blinded.
    B. This is a single-blind study.
    C. This is a double-blind study.
    D. There is not enough information to classify this study.

Ans: C; LO: 1.1-4; Difficulty: Easy; Type: MC

12. Which of the following is an experiment? Select all that apply.
    A. Executives at a large department store chain selected 100 stores and randomly assigned 50 of them to reduce their hours, opening an hour later than before; hours for the other 50 stores were not changed. After six months, the executives compared revenue for the two groups of stores.
    B. A researcher recruited a group of American adults whose demographics were similar to the American population. The researcher measured each subject's forced expiratory volume, an indicator of lung function. Then each subject was asked whether or not they smoke cigarettes.
    C. A survey was administered to a large group of high school students. The survey asked whether the students were employed outside of school (in a paying job) and how much sleep they got the night before (in hours).

Ans: A; LO: 1.1-2; Difficulty: Medium; Type: MS

13. A university professor teaches two sections of introductory statistics: one section meets at 8:00 am and the other meets at 11:00 am. She wants to evaluate the effectiveness of a new method of teaching statistics compared to the standard method she has used in the past. She flips a coin to assign her sections to teaching methods and determines that she will use the standard method at 8:00 am and the new method at 11:00 am. At the end of the term, she compares the final exam scores for the two sections. Which of the following best describes the potential for confounding in this scenario?
    A. Different students have different levels of talent and motivation, so it is impossible to attribute differences in final exam scores to the teaching method.
    B. The two sections may not be exactly the same size, which would lead to inappropriate comparisons between the treatment and control groups.
    C. Students may find it difficult to pay attention at 8:00 am, which may negatively impact the exam scores of those taught with the standard method.
    D. Confounding is not a concern in this scenario, because the professor used random assignment as part of the study design.

Ans: C; LO: 1.1-1; Difficulty: Medium; Type: MC

14. True or False: The standard error of the residuals is a way to measure the amount of variation in the response variable that remains unexplained after applying the model.
    Ans: True; LO: 1.1-1; Difficulty: Easy; Type: TF

15. True or False: Because of the possibility of confounding, you should always avoid using causal language (action verbs like "affect" and "lead to") in your conclusions.

    Ans: False; LO: 1.1-1; Difficulty: Easy Type: TF

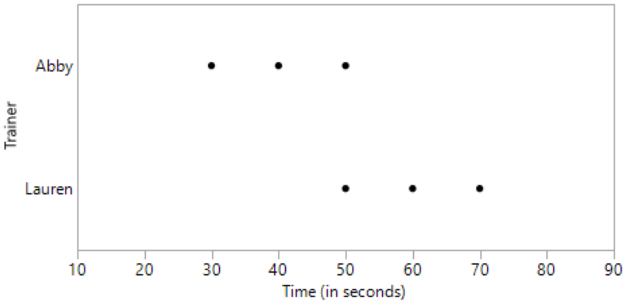
Section 1.2: Quantifying Sources of Variation

LO1.2-1: Partitioning variation in the response variable into variation explained by the model and unexplained variation.

LO1.2-2: Measuring percentage of variation explained.

LO1.2-3: Understanding effect size and practical significance.

Questions 1 through 3: Dog agility is a sport where trainers guide their dogs through an obstacle course as quickly as possible. Two trainers, Abby and Lauren, have dogs participating in agility competitions. Each dog completes the same agility course, and they measure the time it takes each dog to complete the course (in seconds). The times for Abby's three dogs were 30, 40, and 50. The times for Lauren's three dogs were 50, 60, and 70.



1. Calculate the Sum of Squares Total (SSTotal).

   Solution: $(30-50)^2+(40-50)^2+(50-50)^2+(50-50)^2+(60-50)^2+(70-50)^2=1000$
   Ans: 1000 ± 0; LO: 1.2-1; Difficulty: Medium; Type: Calc

2. Calculate the Sums of Squared Errors (SSError).

   Solution: $(30-40)^2+(40-40)^2+(50-40)^2+(50-60)^2+(60-60)^2+(70-60)^2=400$
   Ans: 400 ± 0; LO: 1.2-1; Difficulty: Medium; Type: Calc

3. Calculate the Sum of Squares for the Model (SSModel).

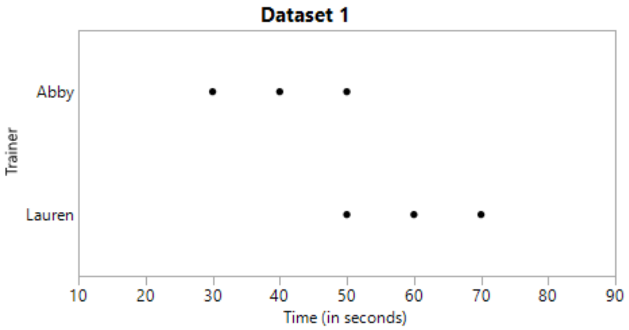
   Solution: $(40-50)^2+(40-50)^2+(40-50)^2+(60-50)^2+(60-50)^2+(60-50)^2=600$
   Ans: 600 ± 0; LO: 1.2-1; Difficulty: Medium; Type: Calc

Questions 4 and 5: Dog agility is a sport where trainers guide their dogs through an obstacle course as quickly as possible. Two trainers, Abby and Lauren, have dogs participating in agility competitions. Each dog completes the same agility course, and they measure the time it takes each dog to complete the course in seconds. Compare the sums of squares for two possible datasets that could occur in this context.

Dataset 1:

Times for Abby’s dogs: 30, 40, 50

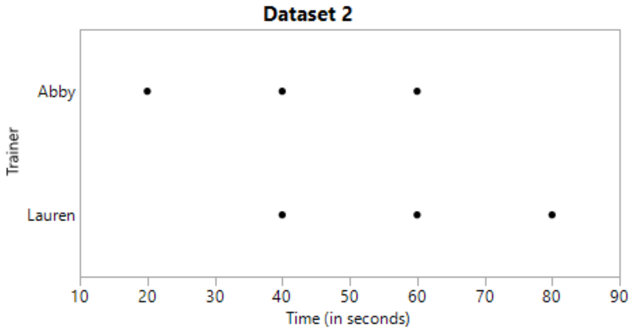
Times for Lauren’s dogs: 50, 60, 70



Dataset 2:

Times for Abby’s dogs: 20, 40, 60

Times for Lauren’s dogs: 40, 60, 80



4. SSTotal for Dataset 1 __________ (<, >, or =) SSTotal for Dataset 2

   SSModel for Dataset 1 __________ (<, >, or =) SSModel for Dataset 2

   SSError for Dataset 1 __________ (<, >, or =) SSError for Dataset 2

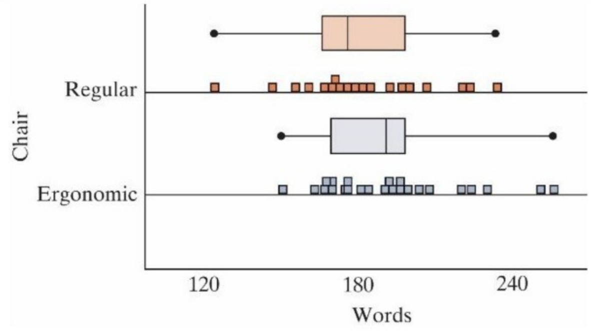
   Ans: <, =, <; LO: 1.2-1; Difficulty: Medium; Type: FIB

5. The R^2 value for Dataset 1 __________ (<, >, or =) the R^2 value for Dataset 2
   Ans: >; LO: 1.2-2; Difficulty: Medium; Type: FIB

Questions 6 and 7: An office designer claims that a new ergonomic desk chair makes typing at a computer terminal faster and easier. A client company plans to test it by asking 30 employees who do a lot of typing to take part in an experiment. They will randomly assign 15 employees to use the new ergonomic chair and 15 to use a regular chair. The 30 employees will then type a selected passage for 5 minutes, recording the total number of words that are typed correctly.

Consider two hypothetical data sets that could result from this experiment:

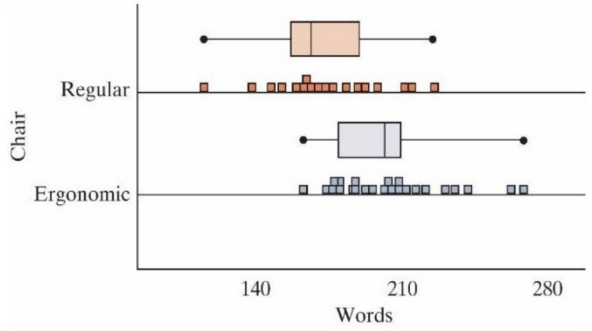
Dataset 1



Summary statistics:

```
            n   Mean    SD
Regular    30   180.73  25.43
Ergonomic  30   191.17  25.81
Residuals  60   0       25.62
```

Dataset 2



Summary statistics:

```
            n   Mean    SD
Regular    30   170.73  25.43
Ergonomic  30   201.17  25.81
Residuals  60   0       25.62
```

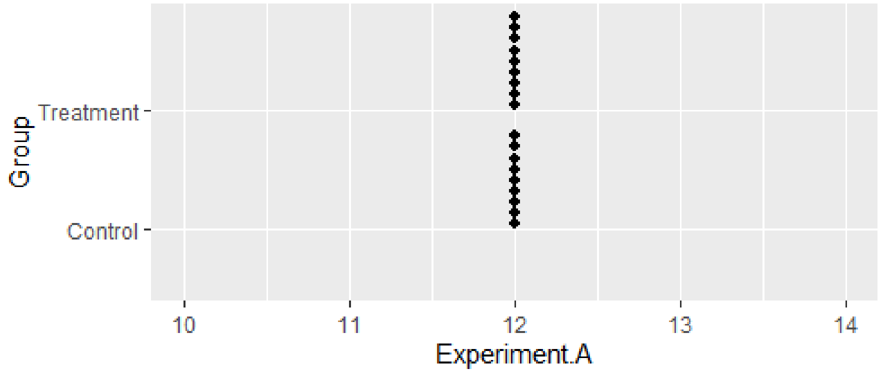
6. Which of the datasets would result in a larger R^2 value?
   - A. Dataset 1 would result in a larger R^2 value, because SSModel for Dataset 1 is larger than SSModel for Dataset 2.
   - B. Dataset 1 would result in a larger R^2 value, because SSError for Dataset 1 is smaller than SSError for Dataset 2.
   - C. Dataset 2 would result in a larger R^2 value, because SSModel for Dataset 2 is larger than SSModel for Dataset 1.
   - D. Dataset 2 would result in a larger R^2 value, because SSError for Dataset 2 is smaller than SSError for Dataset 1.

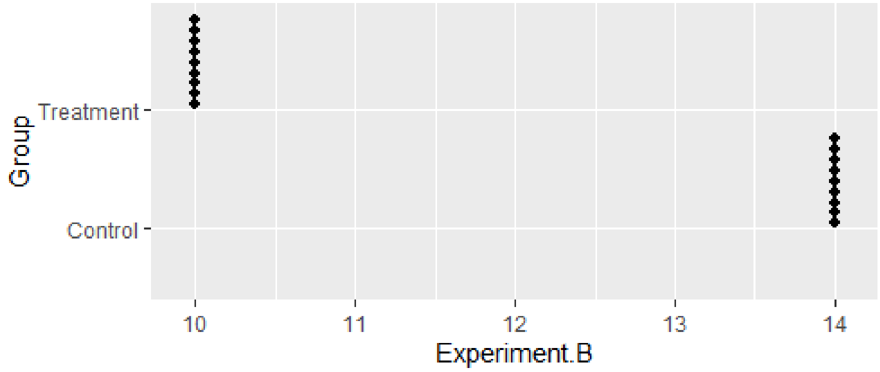
   Ans: C; LO: 1.2-2; Difficulty: Medium; Type: MC

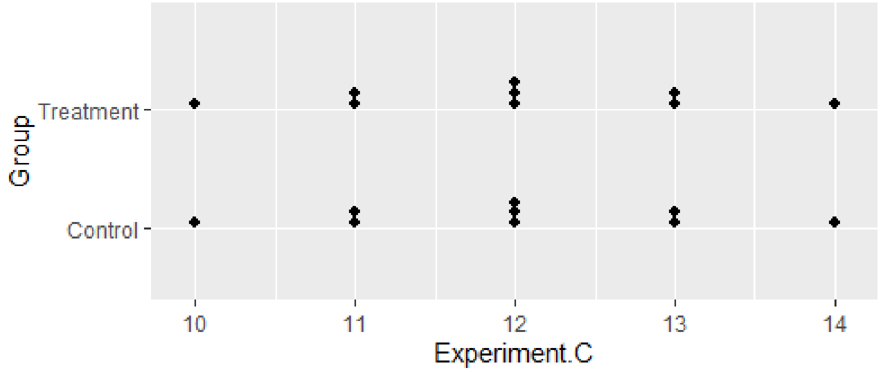
7. Compare the value of the *effects* for the two datasets.
   - A. The effects for Dataset 1 would be larger (in absolute value), because in Dataset 1 there is a smaller difference between the group means.
   - B. The effects for Dataset 2 would be larger (in absolute value), because in Dataset 2 there is a larger difference between the group means.
   - C. The effects for Dataset 1 would be the same size as the effects for Dataset 2, because the standard deviations are the same for both datasets.
   - D. The effects for Dataset 1 would be the same size as the effects for Dataset 2, because the sample sizes are the same for both datasets.

   Ans: B; LO: 1.2-2; Difficulty: Medium; Type: MC

Questions 8 through 11: The graphs below display the outcomes of three different experiments to compare a treatment group with a control group.







8. For which of the experiments does SSModel = 0? Select one or more than one.
    A. Experiment A
    B. Experiment B
    C. Experiment C

   Ans: A, C; LO: 1.2-1; Difficulty: Easy; Type: MS

9. For which of the experiments does SSError = 0? Select one or more than one.
    A. Experiment A
    B. Experiment B
    C. Experiment C

   Ans: A, B; LO: 1.2-1; Difficulty: Easy; Type: MS

10. The R^2 value for Experiment B is _______ (0, 0.5, 1), because ________ (none, half, all) of the variability in outcomes is explained by the treatment group model.

    The R^2 value for Experiment C is ________ (0, 0.5, 1), because ________ (none, half, all) of the variability in outcomes is explained by the treatment group model.

    Ans: 1, all, 0, none; LO: 1.2-2; Difficulty: Medium; Type: FIB

11. The value of the *effect* for the treatment group in Experiment B is _____ (-4, -2, 0, 2, or 4).

    The value of the *effect* for the treatment group in Experiment C is _____ (-4, -2, 0, 2, or 4).

    Ans: -2, 0; LO: 1.2-3; Difficulty: Easy; Type: FIB

Questions 12 and 13: A statistics class conducted an experiment to investigate whether standing heart rates tend to be higher than sitting heart rates. Students were randomly assigned to either sit or stand, then the students measured their heart rates (in beats per minute). They used software to calculate the sums of squares and found that SSModel = 614.8 and SSTotal = 13232.1

12. Calculate SSError.

    Solution: $13232.1 - 614.8 = 12617.3$

    Ans: 12617.3 $\pm$ 0; LO: 1.2-1; Difficulty: Easy; Type: Calc

13. Calculate the R^2 value. *Give your answer as a proportion.*

    Solution: $614.8 / 13232.1 = 0.046$

    Ans: 0.046 $\pm$ 0.006; LO: 1.2-2; Difficulty: Easy; Type: Calc

14. An online retailer is using an experiment to decide whether to modify their website. When visitors type in the web address or click a link to the site, they are randomly re-directed to one of two versions of the website: the version that has been in use for the last year (version A) or an updated version (version B). The retailer's goal is to maximize the amount of time (in minutes) visitors stay on the site.

    Single-mean model:

    $$Predicted\ time\ spent\ on\ site = 12.3, SE\ of\ residuals = 4.64$$

    Separate-means model:

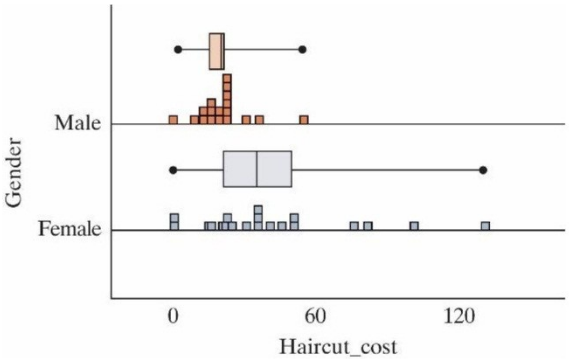
    $$Predicted\ time\ spent\ on\ site = \begin{cases} 9.9\ for\ Version\ A \\ 14.7\ for\ Version\ B \end{cases}, SE\ of\ residuals = 3.98$$

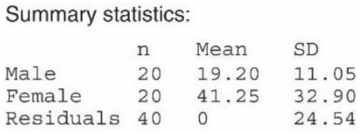
    Calculate the *effect* for Version A.

    Solution: $9.9 - 12.3 = -2.4$

Ans: -2.4 ± 0; LO: 1.2-3; Difficulty: Medium; Type: Calc

Questions 15 and 16: The following output displays the amount (in dollars) that a sample of male and female college students spent on their most recent haircuts.





Summary statistics:

```
          n    Mean    SD
Male      20   19.20   11.05
Female    20   41.25   32.90
Residuals 40   0       24.54
```

15. Calculate SSModel. *Note that the sample sizes are the same for the two groups.*

Solution: $$20\left(19.10-30.18\right)^2+20\left(41.25-30.18\right)^2=4906.23$$

Ans: 4906.23 ± 25 LO: 1.2-1; Difficulty: Medium; Type: Calc

16. Calculate the standard error of the residuals. *Note that the sample sizes are the same for the two groups.*

Solution: $$\sqrt{\frac{11.05^2+32.90^2}{2}}=24.54$$

Ans: 24.54 ± 0.1; LO: 1.2-1; Difficulty: Medium; Type: Calc

17. Match each sum of squares (SS) with its description.

SSModel: A. Measures the amount of variability in the response variable without accounting for groups

SSError: B. Measures the variability *within* groups, the variability unexplained by the model

SSTotal: C. Measures the variability *between* groups, the variability explained by the model

Ans: SSModel: C; SSError: B; SSTotal: A; LO: 1.2-1; Difficulty: Easy Type: Ma

18. In general, study results are considered *practically significant* when the R^2 value is ______ (large/small) and the effect sizes are ______ (large/small).

Ans: large, large; LO: 1.2-3; Difficulty: Easy; Type: FIB

19. What is the cutoff for determining whether study results are *practically significant*?
    A. Results are *practically significant* when the R^2 value is less than 0.05.
    B. Results are *practically significant* when the R^2 value is less than 0.5.
    C. Results are *practically significant* when the R^2 value is greater than 0.5.
    D. Results are *practically significant* when the R^2 value is greater than 0.95.
    E. There is no set cut-off for *practical significance*. It differs based on context.

    Ans: E; LO: 1.2-3; Difficulty: Easy; Type: MC

20. If a researcher were unhappy with his/her effect size or R^2 value, what steps could they take when planning a follow-up study?
    A. They could try to improve the model by adding new explanatory variables.
    B. They could try to reduce unexplained variation in the response through experimental controls or stricter inclusion criteria.
    C. Both of these strategies are reasonable.
    D. Neither of these strategies would impact the effect size or R^2 value.

    Ans: C; LO: 1.2-3; Difficulty: Medium; Type: MC

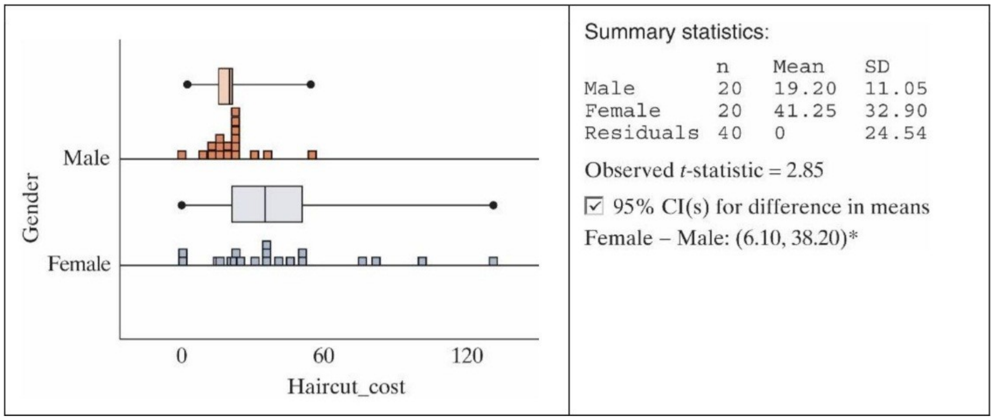
Section 1.3: Is the Variation Explained Statistically Significant

LO1.3-1: Carry out and evaluate a randomization test comparing two groups on a quantitative response variable.

LO1.3-2: Assess the statistical significance of a two-group comparison.

LO1.3-3: Apply two-sample t-procedures for tests of significance and confidence intervals.

Questions 1 through 3: The following output displays the amount (in dollars) that a sample of male and female college students spent on their most recent haircuts. You may assume that the sample is representative of a larger population.



```
Summary statistics:
            n    Mean    SD
Male        20   19.20   11.05
Female      20   41.25   32.90
Residuals   40   0       24.54
```

Observed t-statistic = 2.85

☑ 95% CI(s) for difference in means

Female – Male: (6.10, 38.20)*

1. Does this data provide strong evidence that female students spend more per haircut than male students, on average? Choose the appropriate statement of the null hypothesis.

   A. $H_0 : \mu_F - \mu_M = 0$

   B. $H_0 : \mu_F - \mu_M > 0$

   C. $H_0 : \overline{x}_F - \overline{x}_M = 0$

   D. $H_0 : \overline{x}_F - \overline{x}_M > 0$

   Ans: A; LO: 1.3-3; Difficulty: Medium; Type: MC

2. Interpret the 95% confidence interval.

   A. We are 95% confident that the sample mean for females is between \$6.10 and \$38.20 higher than the sample mean for males.

   B. We are 95% confident that the population mean for females is between \$6.10 and \$38.20 higher than the population mean for males.

   C. If we randomly select one male student and one female student from this sample, we are 95% confident that the haircut value for the female student will be higher.

   D. If we randomly select one male student and one female student from the population, we are 95% confident that the haircut value for the female student will be higher.

   Ans: B; LO: 1.3-3; Difficulty: Medium; Type: MC

3. Based on the 95% confidence interval, we would expect the two-sided p-value to be _______ (>, <, or =) 0.05.

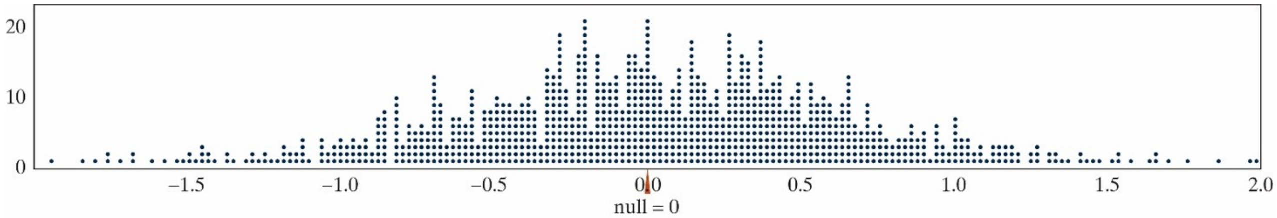
   Ans: <; LO: 1.1-2; Difficulty: Medium; Type: FIB

Questions 4 through 6: A study published in *Psychological Science* in 2007 examined a possible link between mindset and health. The following is an excerpt from the abstract of the article: "84 female room attendants working in seven different hotels were measured on physiological health variables affected by exercise. Those in the informed condition were told that the work they do (cleaning hotel rooms) is good exercise and satisfies the Surgeon General's recommendations for an active lifestyle. Examples of how their work was exercise were provided. Subjects in the control group were not given this information."

Over the course of four weeks, the informed group lost an average of 1.79 lbs and the uninformed group lost an average of 0.20 lbs. Are these results statistically significant? To decide, we can use a randomization test. The dotplot below shows 1000 simulated differences in mean weight

$\overline{y}_{informed} - \overline{y}_{uninformed}$

loss: .



4. Suppose we want to use the 3S Strategy to investigate whether being informed about how their work qualifies as exercise affects room attendants' weight loss. How would we design the simulation?
   A. Write the weight loss amounts on 84 cards. Shuffle and deal them into two groups to represent the informed and uninformed groups. Calculate the difference of means. Repeat.
   B. Write the group labels (informed or uninformed) on 84 cards. Shuffle and deal them into groups to represent weight loss. Calculate the difference of means. Repeat.
   C. Both of these designs are appropriate in this context.
   D. Neither of these designs is appropriate in this context.

   Ans: A; LO: 1.3-1; Difficulty: Medium; Type: MC

5. Why is the distribution of simulated statistics centered at 0?
   A. Because some of the room attendants in the sample gained weight and others lost weight, but the average is close to 0
   B. Because if we repeated this study again, some of the results would be positive and some of the results would be negative, but the average is close to 0
   C. Because the statistics were simulated under the assumption that being informed about how their work qualifies as exercise doesn't affect room attendants' weight loss
   D. Because the data in this study do not provide sufficient evidence to conclude that being informed about how their work qualifies as exercise affects room attendants' weight loss

   Ans: C; LO: 1.3-1; Difficulty: Medium; Type: MC

6. Estimate the p-value. *Include three decimal places in your answer.*

   Solution: Approximately 8 out of 1000 simulated differences were greater than or equal to 1.59, so we estimate that the p-value is 0.008

   Ans: 0.008 ± 0.004; LO: 1.3-1; Difficulty: Medium; Type: Calc

Questions 7 through 9: Researchers used an experiment to investigate whether cell phone use impairs drivers' reaction times. 64 students who volunteered to participate in the study were assigned to one of two driving conditions: cell phone use (n_1=32) or no distractions (n_2=32). The students then participated in a simulation of driving situations, pressing a brake button as soon as they saw a red light. A device recorded their reaction times (in milliseconds).

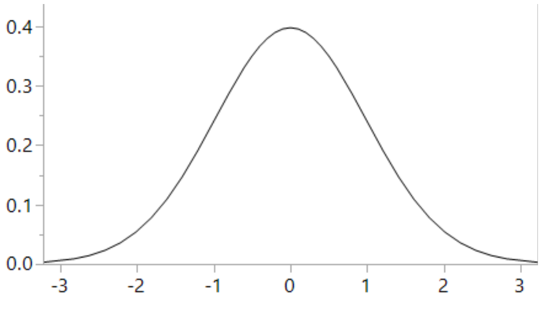
7. The standardized statistic for testing whether $\mu_{Cell} - \mu_{Control} = 0$ is $t = 2.72$. Interpret.
   A. The standard deviation of the reaction times is 2.72 milliseconds.
   B. The standard deviation of the reaction times is 2.72 milliseconds higher than we would have expected based on the null hypothesis.
   C. The sample mean for the cell phone group is 2.72 milliseconds above the sample mean for the control group.
   D. The sample mean for the cell phone group is 2.72 standard errors above the sample mean for the control group.

   Ans: D; LO: 1.3-3; Difficulty: Medium; Type: MC

8. You want to use a theory-based pooled *t*-test to assess the statistical significance of the difference between these two groups, so you would use a *t*-distribution with ____ degrees of freedom.

   Ans: 62 ± 0; LO: 1.1-3; Difficulty: Easy; Type: FIB

9. The graph below shows the appropriate *t*-distribution for assessing the statistical significance of the difference between these two groups. Which of the following statements includes a reasonable p-value and conclusion?

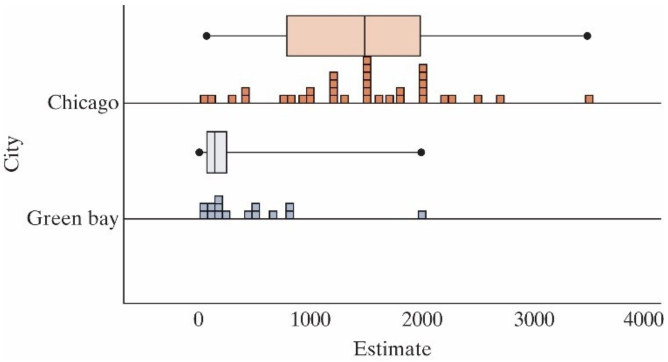


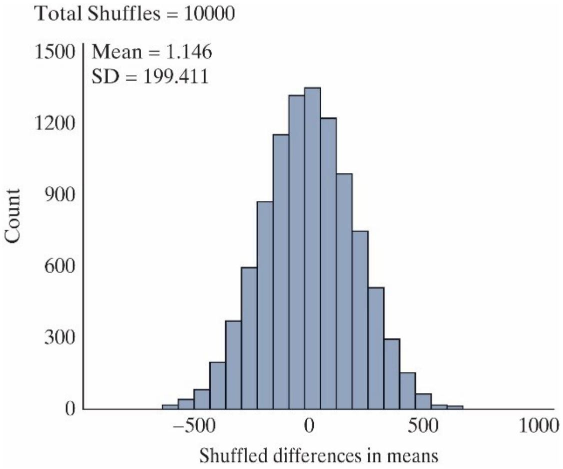
   A. The p-value = 0. This study provides only weak evidence to suggest that cell phone use impairs drivers' reaction times.
   B. The p-value = 0.0042. This study provides strong evidence to suggest that cell phone use impairs drivers' reaction times.
   C. The p-value = 0.2495. This study provides only weak evidence to suggest that cell phone use impairs drivers' reaction times.
   D. The p-value = 0.4856. This study provides strong evidence to suggest that cell

phone use impairs drivers' reaction times.

Ans: B; LO: 1.3-3; Difficulty: Hard; Type: MC

Questions 10 and 11: Anchoring is the common human tendency to rely too heavily, or "anchor", on one trait or piece of information when making decisions. A group of statistics students from California were asked to guess the population of Milwaukee, Wisconsin. Some of the students were randomly chosen to be told that the nearby city of Chicago, Illinois, has a population of about 3 million people, while the rest of the students were told that the nearby city of Green Bay, Wisconsin, has a population of about 100,000.





City	N	Mean	SD
Chicago	35	1357.34	802.21
Green Bay	34	271.38	370.96

10. Based on the information given above, is the effect of anchoring statistically significant in this context?
    A. Yes, because the difference of means observed in this sample would be unlikely to occur if anchoring really had no effect.
    B. Yes, because the sample means in this study are different and the sample sizes are both larger than 20.
    C. No, because the shuffled differences in means are centered at 0, so it is reasonable to conclude that anchoring has no effect.
    D. No, because shuffled differences in means generally fall between -500 and 500. That means the results of this experiment were an unlikely fluke.

    Ans: A; LO: 1.3-2; Difficulty: Medium; Type: MC

11. Are the validity conditions met for a theory-based pooled *t*-test?
    A. No, because the samples are not independent of each other.
    B. No, because the sample sizes are not the same.
    C. No, because the sample standard deviations for the two groups are very different.
    D. Yes. The only potential violation is the skewness in the sample distributions, but this is not a problem, because the sample sizes are both larger than 20.

    Ans: C; LO: 1.3-3; Difficulty: Easy; Type: MC

Questions 12 through 14: A psychology study (Rutchick, Slepian, and Ferris, 2010) investigated whether using a red pen causes people to assign lower scores than using a blue pen. A group of 128 students in an undergraduate psychology class were asked to grade 128 different eighth graders' essays on a scale of 0—100. Half of the students were randomly assigned a red pen while grading, and the other half were given blue. The results are given in the table below:

Pen Color	N	Mean Score	Standard Deviation
Red	64	76.20	12.29
Blue	64	80.00	9.36

12. State the alternative hypothesis.
    - A. $H_A : \mu_{Red} = \mu_{Blue}$
    - B. $H_A : \overline{x}_{Red} = \overline{x}_{Blue}$
    - C. $H_A : \mu_{Red} < \mu_{Blue}$
    - D. $H_A : \overline{x}_{Red} \neq \overline{x}_{Blue}$

    Ans: C; LO: 1.3-1; Difficulty: Medium; Type: MC

13. Is it appropriate to use a pooled *t*-test to compare these groups?
    - A. Yes, a pooled *t*-test is appropriate, because the group means are fairly similar.
    - B. Yes, a pooled *t*-test is appropriate, because the group standard deviations are fairly similar.
    - C. No, an unpooled two-sample *t*-test is more appropriate, because it is not reasonable to assume that the group means are equal in the population.
    - D. No, an unpooled two-sample *t*-test is more appropriate, because it is not reasonable to assume that the group standard deviations are equal in the population.

    Ans: B; LO: 1.3-3; Difficulty: Medium; Type: MC

14. Calculate the *t*-statistic. *Note that the sample sizes are equal.*

$$\sqrt{\frac{\left(12.29^2+9.36^2\right)}{2}}=10.924; t=\frac{80.00-76.20}{10.924*\sqrt{\frac{1}{64}+\frac{1}{64}}}=1.97$$

    Solution: SE of residuals =

    Ans: 1.97 ± 0.05; LO: 1.3-3; Difficulty: Medium; Type: Calc

15. In the context of a pooled *t*-test for comparing two groups, what does it mean to say that the study results are statistically significant? Select all that apply.
    - A. It means there is large difference between the means of the two groups.
    - B. It means there is small difference between the means of the two groups.
    - C. It means the observed sample difference in means would be unlikely to occur if there were really no difference between the two groups in the population.
    - D. It means the observed sample difference in means would be likely to occur if there were really no difference between the two groups in the population.

    Ans: C; LO: 1.3-3; Difficulty: Medium; Type: MS

16. Which of the following statistics reflect both the effect size *and* the sample size? In other words, which of the following statistics can be used to assess statistical significance?

Select all that apply.

A. R^2

B. Difference in means, $\overline{x}_1 - \overline{x}_2$

C. Standardized statistic, *t*

D. p-value

Ans: C, D; LO: 1.3-2; Difficulty: Medium; Type: MS

17. True or False: The p-value is the probability that the null hypothesis is true.

Ans: False; LO: 1.3-2; Difficulty: Easy; Type: TF

18. True or False: A small p-value indicates strong evidence against the null hypothesis.
Ans: True; LO: 1.3-2; Difficulty: Easy; Type: TF

19. True or False: A confidence interval is a range of plausible values for a parameter calculated using sample statistics.
Ans: True; LO: 1.3-3; Difficulty: Easy; Type: TF

Section 1.4: Comparing Several Groups

LO1.4-1: Compare more than two treatments using randomization tests.

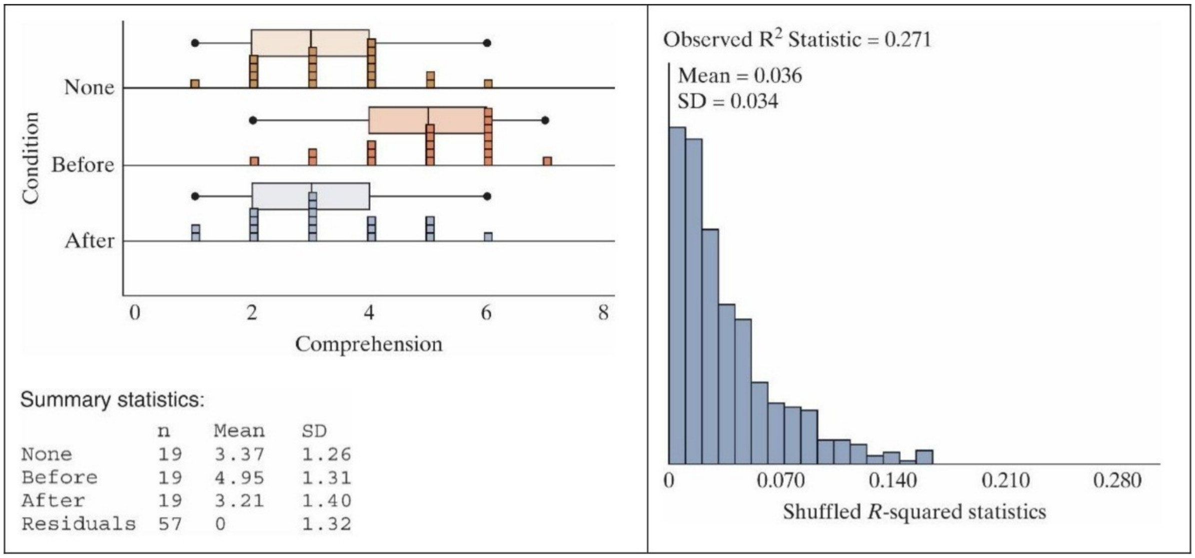
LO1.4-2: Calculate an *F*-statistic and use the *F*-distribution to find theory-based p-values.

LO1.4-3: Assess the validity of an *F*-test.

LO1.4-4: Complete an Analysis of Variance table.

Questions 1 through 4: Does seeing a picture have any effect on college students' understanding of ambiguous prose? 57 students were randomly assigned to three groups: 19 saw a picture before reading a difficult passage of text, 19 saw the picture after reading the passage, and 19 were shown no picture at all. The groups were then tested on their reading comprehension and assigned a quantitative score.

Does this data provide convincing evidence that seeing a picture has an effect on reading comprehension scores?



Summary statistics:

	n	Mean	SD
None	19	3.37	1.26
Before	19	4.95	1.31
After	19	3.21	1.40
Residuals	57	0	1.32

1. Which of the following is an appropriate statement of the null hypothesis? Select all that apply.

A. $H_0 : \mu_{None} = \mu_{Before} = \mu_{After}$

B. $H_0 : \mu_{After} < \mu_{None} < \mu_{Before}$

C. $H_0 :$ At least one μ differs from the others

D. $H_0 :$ There is an association between seeing a picture (before, after, or not at all) and reading comprehension scores.

E. $H_0:$ There is no association between seeing a picture (before, after, or not at all) and reading comprehension scores.

Ans: A, E; LO: 1.4-1; Difficulty: Medium; Type: MS

2. Which of the following is an appropriate statement of the alternative hypothesis? Select all that apply.
   A. $H_A: \mu_{None} = \mu_{Before} = \mu_{After}$
   B. $H_A: \mu_{After} < \mu_{None} < \mu_{Before}$
   C. $H_A:$ At least one μ differs from the others
   D. $H_A:$ There is an association between seeing a picture (before, after, or not at all) and reading comprehension scores.
   E. $H_A:$ There is no association between seeing a picture (before, after, or not at all) and reading comprehension scores.

   Ans: C, D; LO: 1.4-1; Difficulty: Medium; Type: MS

3. What does the graph of shuffled R-squared statistics represent?
   A. The values of R^2 for all 57 students who participated in this study
   B. The values of R^2 that would occur if the null hypothesis were really true
   C. The values of R^2 that would occur if the alternative hypothesis were really true
   D. The values of R^2 that would occur if we repeated this study many times in the real world

   Ans: C; LO: 1.4-1; Difficulty: Medium; Type: MC

4. Does this study provide strong evidence that seeing a picture affects reading comprehension?
   A. Yes, because an R^2 value of 0.271 does not appear in the graph of shuffled R-squared statistics, which suggests it would be unlikely to occur by chance alone.
   B. Yes, because the mean comprehension scores are different in each group, and all the validity conditions for the significance test are satisfied.
   C. No, because the graph of shuffled R-squared statistics is centered at 0.036, which suggests that very little variability is explained by the model.
   D. No, because an R^2 value of 0.271 does not appear in the graph of shuffled R-squared statistics, which suggests the data from this experiment was a fluke that occurred by chance alone.

   Ans: A; LO: 1.4-1; Difficulty: Hard; Type: MC

Questions 5 and 6: A study was carried out to investigate whether the type of message on the back of customer checks at a restaurant would affect tips (recorded as a percentage of the total bill). Sixty tables were selected to participate over a weekend at a restaurant in Philadelphia. Each table was randomly assigned to receive either (1) a picture of a happy face, (2) the words “Thank you!” written out, or (3) no message.

Does this study provide convincing evidence that the message written on the back of the check affects tip percentage? You can use a randomization test to decide.

5. How would you design the physical simulation?

   Write the tip percentages on cards. Shuffle and deal the cards into ____ groups.

   A. 2
   B. 3
   C. 20
   D. 60

   Ans: B; LO: 1.4-1; Difficulty: Medium; Type: MC

6. Which of the following statistics could be used to summarize each simulated sample? Select all that apply.

   A. Difference in means
   B. R^2
   C. *t*-statistic
   D. *F*-statistic

   Ans: B, D; LO: 1.4-1; Difficulty: Medium Type: MS

7. A food company was interested in how texture might affect the palatability of a particular food. They set up an experiment in which they looked at whether the "coarseness" of the final product (coarse or fine) affected the palatability scores given by 50 people. A partially filled in ANOVA table is given below. Calculate the *F*-statistic.

Source	df	SS	MS	*F*
Model				?
Error		6113		
Total		16722		

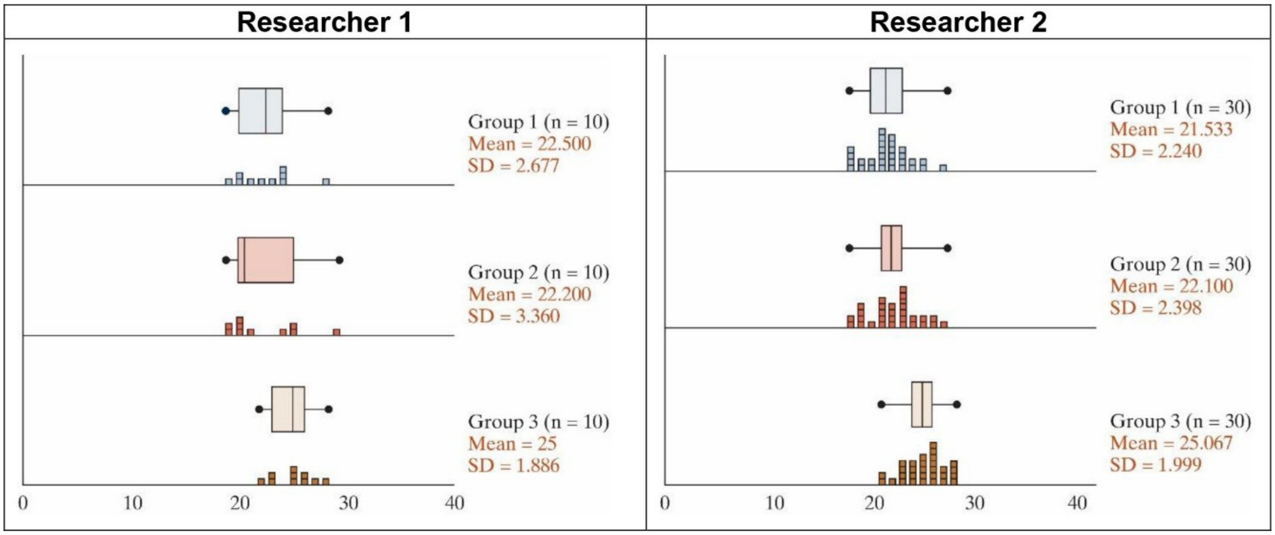
Solution:

$$SSModel = 16722 - 6113 = 10609; F = \frac{10609/(2-1)}{6113/(50-2)} = 83.30$$

Ans: 83.30 ± 0.5 LO: 1.4-4; Difficulty: Medium; Type: Calc

Questions 8 through 11: Multiple researchers have conducted studies to examine the time it takes for three different medications to register in a patient's blood system (in minutes). Each researcher wants to test whether the type of medication affects time.

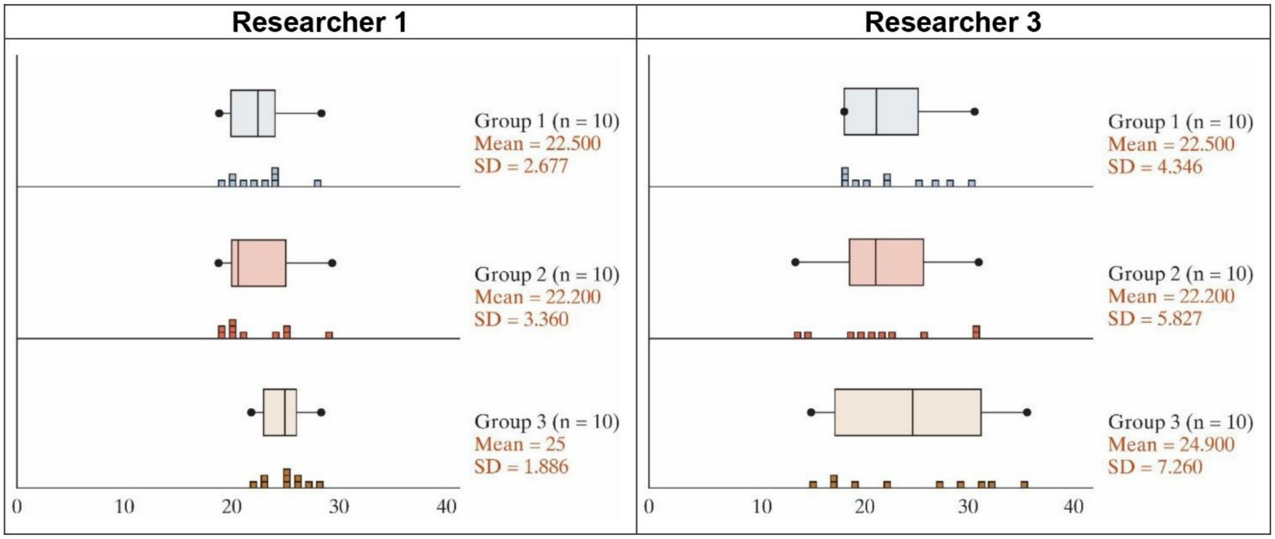
8. Which researcher would obtain a larger *F*-statistic based on their results?



A. Researcher 1 would obtain a larger *F*-statistic.
B. Researcher 2 would obtain a larger *F*-statistic.
C. Researchers 1 and 2 would obtain very similar *F*-statistics.
D. There is not enough information to determine which *F*-statistic would be larger.

Ans: B; LO: 1.4-2; Difficulty: Medium; Type: MC

9. Compare the results for Researcher 1 and Researcher 3.



Which researcher would obtain a larger *F*-statistic based on their results?

A. Researcher 1 would obtain a larger *F*-statistic.
B. Researcher 3 would obtain a larger *F*-statistic.
C. Researchers 1 and 3 would obtain very similar *F*-statistics.
D. There is not enough information to determine which *F*-statistic would be larger.

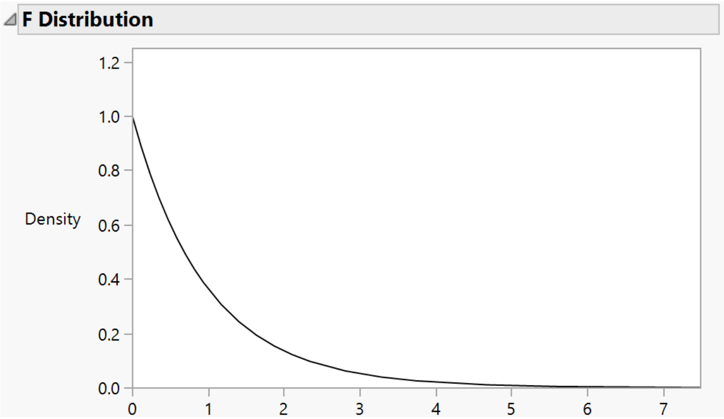
Ans: A; LO: 1.4-2; Difficulty: Medium; Type: MC

10. Researcher 4 had a total sample size of 90, with 30 patients per treatment group, and a standardized statistic of *F* = 12.72. She intends to use the *F*-distribution to find a theory-based p-value. Which *F*-distribution should she use?

She should use an *F*-distribution with Model df = _______ and Error df = ________.

Ans: 2 ± 0, 87 ± 0; LO: 1.4-2; Difficulty: Easy; Type: FIB

11. Researcher 4 had a total sample size of 90, with 30 patients per treatment group, and a standardized statistic of *F* = 12.72. The graph below shows the appropriate *F*-distribution for assessing the statistical significance of the association between type of medication and time. Which of the following statements includes a reasonable p-value and conclusion?



A. The p-value is close to 0. This study provides only weak evidence to suggest that type of medication has an effect on the time it takes for the medication to register.
B. The p-value is close to 1. This study provides only weak evidence to suggest that type of medication has an effect on the time it takes for the medication to register.
C. The p-value is close to 0. This study provides strong evidence to suggest that type of medication has an effect on the time it takes for the medication to register.
D. The p-value is close to 1. This study provides strong evidence to suggest that type of medication has an effect on the time it takes for the medication to register.

Ans: C; LO: 1.4-2; Difficulty: Hard; Type: MC

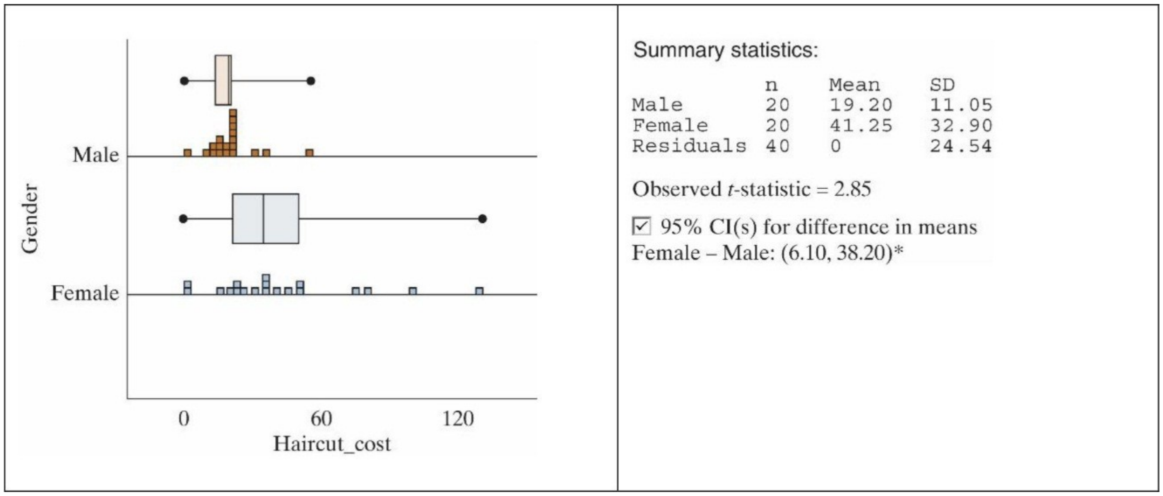
12. The table below summarizes the results of an experiment to compare yields (as measured by the dried weight of plants) obtained under a control and two different treatment conditions. Calculate SSModel.

	Sample Size	**Mean**	**SD**
Full sample	30	5.07	0.701
Treatment 1	10	5.03	0.583
Treatment 2	10	4.66	0.794
Treatment 3	10	5.53	0.443

Solution: $10(5.03-5.07)^2+10(4.66-5.07)^2+10(5.53-5.07)^2=3.81$

Ans: 3.81 ± 0.01; LO: 1.4-4; Difficulty: Medium; Type: Calc

13. The following output displays the amount (in dollars) that a sample of male and female college students spent on their most recent haircuts. You may assume that the sample is representative of a larger population. Calculate the *F*-statistic.



```
Summary statistics:

           n    Mean    SD
Male       20   19.20   11.05
Female     20   41.25   32.90
Residuals  40   0       24.54

Observed t-statistic = 2.85

☑ 95% CI(s) for difference in means
Female – Male: (6.10, 38.20)*
```

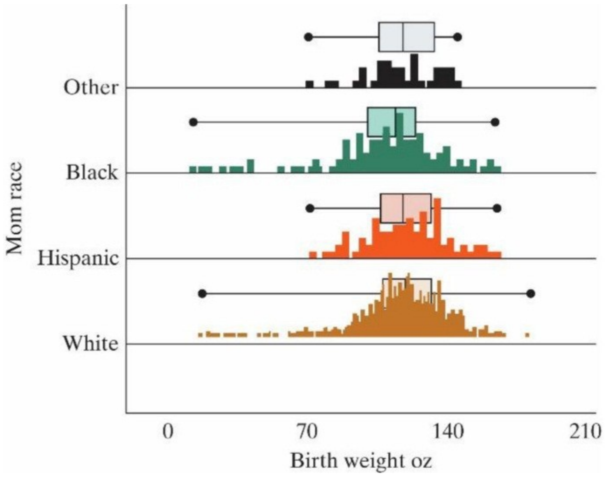
Solution: $2.85^2 = 8.12$

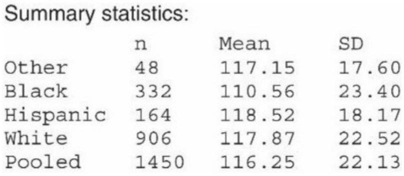
Ans: 8.12 ± 0.01; LO: 1.4-2; Difficulty: Easy; Type: Calc

14. A randomized experiment was conducted exploring the effectiveness of acupuncture in treating chronic lower back pain. Patients in the study were randomly assigned to one of three treatment groups: Verum acupuncture (traditional Chinese medicine), Sham acupuncture (placebo), and nonacupuncture therapy (drugs, physical therapy, etc.). After six months, each patient’s pain reduction was measured on a quantitative scale. Which inference procedure(s) could you use to test for an association between type of treatment and pain reduction?
    A. Two sample *t*-test (pooled or unpooled)
    B. ANOVA *F*-test
    C. Both of these tests are appropriate in this scenario.
    D. Neither of these tests is appropriate in this scenario.

Ans: B; LO: 1.4-2; Difficulty: Medium; Type: MC

15. A random sample of 1450 birth records was selected from the state of North Carolina in the year 2001. One question of interest is whether the distribution of birth weights (in ounces) differs based on the race/ethnicity of the mother (White, Black, Hispanic, or other).




Summary statistics:

```
          n      Mean     SD
Other     48     117.15   17.60
Black     332    110.56   23.40
Hispanic  164    118.52   18.17
White     906    117.87   22.52
Pooled    1450   116.25   22.13
```

Would an ANOVA *F*-test be valid for these data?

A. No, because the samples are not independent of each other.
B. No, because the sample sizes for the groups are not similar enough.
C. No, because distribution of weights is slightly skewed left for two of the groups.
D. Yes, because all of the validity conditions are met.

Ans: D; LO: 1.4-3; Difficulty: Medium; Type: MC

16. True or False: The validity conditions for the ANOVA *F*-test are the same as the validity conditions for the two-sample pooled *t*-test.

Ans: True; LO: 1.4-3; Difficulty: Easy; Type: TF

Section 1.5: Confidence and Prediction Intervals

LO1.5-1: Apply post-hoc analysis after significant *F*-test (pairwise differences).

LO1.5-2: Calculate and interpret confidence intervals on single means and differences in two means.

LO1.5-3: Calculate and interpret prediction intervals on quantitative variables.

LO1.5-4: Identify factors that impact widths of confidence intervals and prediction intervals.

Questions 1 through 3: Does seeing a picture have any effect on college students' understanding of ambiguous prose? 57 students were randomly assigned to three groups: 19 saw a picture before reading a difficult passage of text, 19 saw the picture after reading the passage, and 19 were shown no picture at all. The groups were then tested on their reading comprehension and assigned a quantitative score. The pairwise confidence intervals for the difference in mean comprehension scores are given below.

95% confidence interval for $\mu_{After} - \mu_{Before}$: (-2.60, -0.88)

95% confidence interval for $\mu_{After} - \mu_{None}$: (-1.02, 0.70)

95% confidence interval for $\mu_{Before} - \mu_{None}$: (0.72, 2.44)

1. Based on the confidence intervals, which conditions are significantly different from each other? Select all that apply.
   A. After is significantly different from Before.
   B. After is significantly different from None.
   C. Before is significantly different from None.
   Ans: A, C; LO: 1.5-1; Difficulty: Medium; Type: MS

2. Which of the following letters tables is consistent with the confidence intervals given above?

   A.

Group	Letters
Before	A
After	A
None	A

   B.

Group	Letters
Before	A
After	B
None	B

   C.

Group	Letters
Before	A
After	B
None	C

   D. The confidence intervals do not provide enough information to construct a letters table.

Ans: B; LO: 1.5-1; Difficulty: Medium; Type: MC

3. How could you use a confidence interval to decide whether the difference in comprehension scores for the Before group and the None group is important in a practical sense?
   A. Subtract the endpoints to find the width of the interval. If the interval is wide, then the true difference is practically important.
   B. Subtract the endpoints to find the width of the interval. If the interval is narrow, then the true difference is practically important.
   C. Look to see whether the interval includes 0. If the interval does not include 0, then the true difference is practically important.
   D. Look to see whether the endpoints of the interval are close to 0 or far from 0. If the endpoints are far from 0, then the true difference may be practically important.

   Ans: D; LO: 1.5-2; Difficulty: Medium; Type: MC

Questions 4 and 5: An experiment was conducted to compare yields (as measured by the dried weight of plants in kg) obtained under a control and two different treatment conditions. The pairwise confidence intervals for the difference in mean yields are given below.

95% confidence interval for $\mu_{Treatmen1} - \mu_{Treatment2}$: (-1.44, -0.29)

95% confidence interval for $\mu_{Treatment1} - \mu_{Control}$: (-0.94, 0.20)

95% confidence interval for $\mu_{Treatment2} - \mu_{Control}$: (-0.08, -1.07)

4. Based on the confidence intervals, which of the treatments are significantly different from each other?
   A. All three treatments are significantly different, because none of the confidence intervals have similar endpoints.
   B. All three treatments are significantly different, because none of the confidence intervals have midpoints that are equal to 0.
   C. None of the treatments are significantly different, because the widths of the confidence intervals are all very similar to each other.
   D. Treatment 1 is significantly different from Treatment 2, because the confidence interval for this comparison does not include 0.

   Ans: D; LO: 1.5-1; Difficulty: Hard; Type: MC

5. The grower's goal is to find a treatment that produces yields of at least 1.4 kg, on average. Which of the treatments meets this goal?
   A. None of the treatments meet the grower's goal.
   B. Only Treatment 2 meets the grower's goal.
   C. Both Treatment 1 and Treatment 2 meet the grower's goal.
   D. The pairwise confidence intervals for the difference in mean yields do not provide enough information to decide if the treatments meet the grower's goal.

   Ans: D; LO: 1.5-2; Difficulty: Hard; Type: MC

Question 6 through 8: In 2018, a sample of academic faculty were surveyed about their salaries (in US dollars). The results were classified according to academic rank: instructor, assistant professor, associate professor, and full professor. The table below shows 95% confidence intervals for the population mean of each rank.

Rank	Sample size	Group Mean	95% CI for μ_i
Instructor	75	63680	(54583, 72776)
Assistant	175	92029	(86073, 97984)
Associate	145	105133	(98591, 111676)
Full Professor	234	154509	(149359, 159659)

6. Which of the following statements is an appropriate interpretation based on the confidence intervals? *You may assume that this sample is representative of a larger population of academic faculty.*
    A. The average salaries of all four ranks are significantly different from each other, because none of the confidence intervals include 0.
    B. We are 95% confident that the population mean salary for instructors is between $54,583 and $72,776.
    C. Roughly 95% of instructors in the population earn between $54,583 and $72,776 per year.
    D. More than one of these statements is an appropriate interpretation of the confidence intervals.

    Ans: B; LO: 1.5-2; Difficulty: Medium; Type: MC

7. Why is the 95% confidence interval for μ_{full} narrower than the other intervals?
    A. The sample size for full professors is largest, and as sample size increases, the width of the confidence interval tends to decrease.
    B. The group mean for professors is largest, and as group mean increases, the width of the confidence interval tends to decrease.
    C. We can predict with a high level of certainty that full professors make more than other ranks, so the confidence interval provides a precise estimate.
    D. None of the justifications above are reasonable, so professors' salaries must be less variable (lower SD) compared to the other ranks.

    Ans: A; LO: 1.5-4; Difficulty: Medium; Type: MC

8. If we changed the confidence level from 95% to 99% (holding everything else constant), would the width of the confidence intervals change?
    A. The width of the confidence intervals would decrease.
    B. The width of the confidence intervals would increase.
    C. The width of some intervals would increase and the width of the other intervals would decrease.
    D. The width of all the confidence intervals would stay the same.

    Ans: B; LO: 1.5-4; Difficulty: Medium; Type: MC

9. An online retailer is using an experiment to decide whether to modify their website. When visitors type in the web address or click a link to the site, they are randomly re-directed to one of three versions of the website. The retailer's goal is to maximize the amount of time (in minutes) visitors stay on the site.

Version	Letters

1	A
2	AB
3	B

True or False: In this study, there is a statistically significant difference in mean time spent on the site for Version 1 and Version 2.

Ans: F; LO: 1.5-1; Difficulty: Easy; Type: TF

10. Body temperature measurements (in Fahrenheit) were taken from 65 healthy female volunteers aged 18 to 40 that were participating in vaccine trials. Based on this data, researchers calculated a 95% prediction interval: (96.90, 99.89). Interpret the interval. *You may assume that the sample is representative of a larger population.*

A. Roughly 95% of healthy females in this population would have body temperatures between 96.90 and 99.89.
B. We are 95% confident that the sample mean body temperature is between 96.90 and 99.89.
C. We are 95% confident that the population mean body temperature is between 96.90 and 99.89.
D. If we were to collect another sample of size 65, we are 95% confident that the sample mean body temperature would be between 96.90 and 99.89.

Ans: A; LO: 1.5-3; Difficulty: Medium; Type: MC

11. In a context with a quantitative response variable and a multi-level categorical explanatory variable, which of the following best describes the purpose of the ANOVA *F*-test?

A. The *F*-test helps us assess whether or not there is convincing evidence of an association between the variables.

B. The *F*-test helps us measure the strength of the association between the variables by indicating how much the groups differ in terms of the mean response.

C. The *F*-test helps us determine the direction of the association between the variables by indicating which group means are higher than others.

D. The *F*-test serves all three of the purposes listed above.

Ans: A; LO: 1.5-1; Difficulty: Medium; Type: MC

12. A researcher has conducted an experiment to study four different treatments, and they decide to analyze the data by comparing each treatment group mean to every other treatment group mean. This involves tests for six pairwise comparisons. If the researcher uses a significance level of $\alpha = 0.05$ for each test, then the probability of making at least one Type I error is ________ (>, <, =) 0.05.

Ans: >; LO: 1.5-1; Difficulty: Easy; Type: FIB

13. What should you do to protect against an inflated experiment-wise Type I error rate?

A. Conduct pairwise comparisons using *t*-procedures first and only conduct an *F*-test if the p-values for the *t*-tests are all large.
B. Conduct pairwise comparisons using *t*-procedures first and only conduct an *F*-test if the p-values for the *t*-tests are all small.
C. Conduct an *F*-test first and only conduct pairwise comparisons using *t*-procedures

if the p-value for the *F*-test is large.
D. Conduct an *F*-test first and only conduct pairwise comparisons using *t*-procedures if the p-value for the *F*-test is small.

Ans: D; LO: 1.5-1; Difficulty: Medium; Type: MC

14. How are prediction intervals different from confidence intervals?
A. Prediction intervals predict the population mean for a particular group, thus they are wider than confidence intervals.
B. Prediction intervals predict the population mean for a particular group, thus they are narrower than confidence intervals.
C. Prediction intervals predict the response of a new individual observation, thus they are wider than confidence intervals.
D. Prediction intervals predict the response of a new individual observation, thus they are narrower than confidence intervals.

Ans: C; LO: 1.5-3; Difficulty: Medium; Type: MC

15. The validity conditions for confidence intervals and prediction intervals on means require that the data distribution be reasonably bell-shaped and symmetric. Is this condition always important, even when the sample size is large?
A. This condition is not very important for confidence intervals or prediction intervals, as long as the sample size is large.
B. This condition is very important for prediction intervals. It is less of a concern for confidence intervals, as long as the sample size is large.
C. This condition is very important for confidence intervals. It is less of a concern for prediction intervals, as long as the sample size is large.
D. This condition is very important for both confidence intervals and prediction intervals, regardless of sample size.

Ans: B; LO: 1.5-4; Difficulty: Medium; Type: MC

16. Which of the following is the best way to reduce the width of a prediction interval?
A. Increase the confidence level
B. Increase the sample size
C. Reduce the unexplained variation within groups
D. Use a pooled estimate of the total variation

Ans: C; LO: 1.5-3; Difficulty: Medium; Type: MC

17. True or False: Suppose we test $\mathrm{H}_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ using an ANOVA *F*-test. This is preferable to using 6 pairwise *t*-tests, because testing all parameters at once controls the probability of Type II error.

Ans: False; LO: 1.5-1; Difficulty: Medium; Type: TF

Section 1.6: More Study Design Considerations

LO1.6-1: Understand statistical power and how it is impacted by sample size, variability within groups, number of groups, and significance level.

LO1.6-2: Use statistical power analysis to plan the sample size of a study.

1. A Type II error occurs when researchers _________ (do/don't) find convincing evidence against the null hypothesis, when the null hypothesis is actually __________ (true/false).

   Ans: don't, false; LO: 1.6-1; Difficulty: Easy; Type: FIB

2. Suppose that you analyzed data from an experiment and obtained a large p-value. Which type of error is possible in this case?
   A. This result could be due to a Type I error.
   B. This result could be due to a Type II error.
   C. This result could be due to either a Type I error or a Type II error.
   D. As long as the validity conditions were met, the large p-value is not due to an error.

   Ans: B; LO: 1.6-1; Difficulty: Easy; Type: MC

3. The statistical power of a study is the probability that the researchers _______ (will/won't) find convincing evidence against the null hypothesis, when the null hypothesis is actually __________ (true/false).

   Ans: will, false; LO: 1.6-1; Difficulty: Easy; Type: FIB

4. Suppose researchers design a study such that the Type I error rate is 5% and the Type II error rate is 20%. Calculate the power of the study. *Include the % sign in your answer.*

   Solution: $100\% - 20\% = 80\%$
   Ans: 80% ± 0; LO: 1.6-1; Difficulty: Easy; Type: Calc

5. True or False: The aspects of a study that impact the strength of evidence (sample size, unexplained variation, number of groups, etc.) are the same ones that impact a study's power.

   Ans: True; LO: 1.6-1; Difficulty: Easy; Type: TF

6. How are the probabilities of Type I and Type II error affected by sample size? *Assume that variability within groups, number of groups, and significance level remain unchanged.*

   As the sample size increases, the probability of making a Type I error ________ (increases / decreases / stays the same), and the probability of making a Type II error ________ (increases / decreases / stays the same).

   Ans: stays the same, decreases; LO: 1.6-1; Difficulty: Medium; Type: FIB

7. How are the probabilities of Type I and Type II error affected by the significance level, α ? *Assume that sample size, variability within groups, and number of groups remain unchanged.*

   As the significance level increases, the probability of making a Type I error ________ (increases / decreases / stays the same), and the probability of making a Type II error ________ (increases / decreases / stays the same).

   Ans: increases, decreases; LO: 1.6-1; Difficulty: Medium; Type: FIB

8. True or False: When comparing groups with a quantitative response, using a smaller number of groups (fewer levels of the categorical variable) always increases the statistical power of the test.

   Ans: False; LO: 1.6-1; Difficulty: Medium; Type: TF

9. Suppose a researcher wants to design an experiment with a high level of statistical power. What should they do?

   Include a _________ (large/small) number of experimental units in the study.

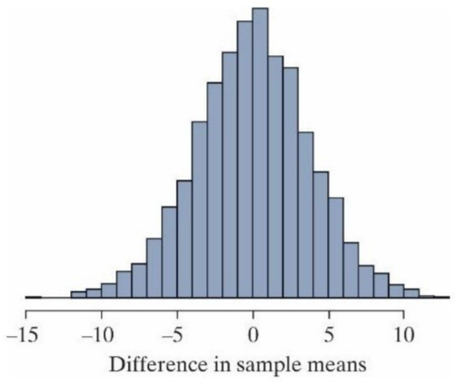
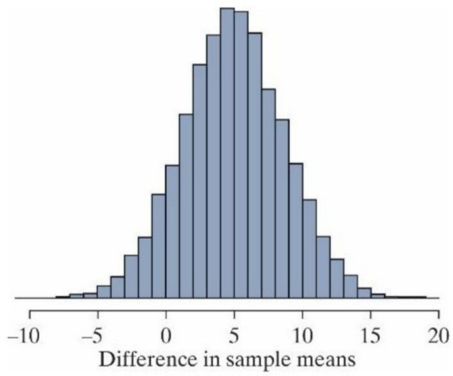
   Choose a number of groups that is as_________ (large/small) as possible without compromising the amount of variability explained.

   Take steps during study design and data collection to __________ (increase/decrease) the amount of variability within groups.

   Ans: large, small, decrease; LO: 1.6-2; Difficulty: Medium; Type: FIB

Questions 10 and 11: For a class project, a statistics student plans to conduct an experiment to investigate whether standing heart rates tend to be higher than sitting heart rates. They will randomly assign their participants to either sit or stand, then the participants' heart rates will be measured (in beats per minute).

The student considers a difference of 5 beats per minute to be practically important, so if the difference is 5 beats per minute or larger, they want to be able to detect it. The student expects the standard deviation of each group to be about 12 bpm and plans to use a significance level of $\alpha = 0.01$ and a sample size of 20 in each group.

Graph A: Simulated values of $\overline{x}_{Stand} - \overline{x}_{Sit}$ assuming $\mu_{stand} - \mu_{Sit} = 0$	**Graph B:** Simulated values of $\overline{x}_{Stand} - \overline{x}_{Sit}$ assuming $\mu_{stand} - \mu_{Sit} = 5$
	

10. Which of the graphs above would the student use to find the rejection region?
    - A. Graph A
    - B. Graph B
    - C. Either of these two graphs could be used to find the rejection region.
    - D. Neither of these two graphs could be used to find the rejection region.

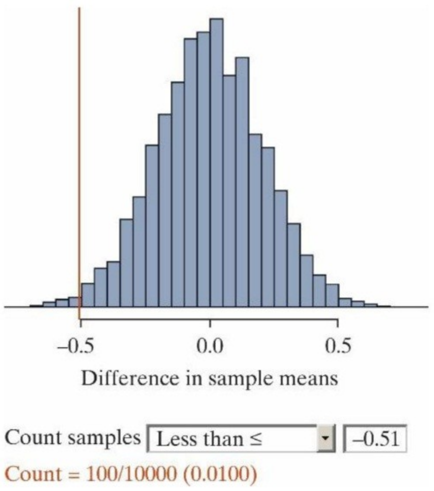
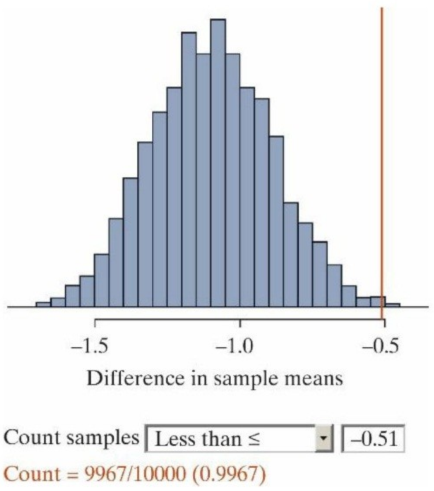
    Ans: A; LO: 1.6-2; Difficulty: Medium; Type: MC

11. The rejection region for this study is a difference of means $(\overline{x}_{Stand} - \overline{x}_{Sit})$ of 9.1 or higher. Which of the values below is closest to the power of the test, given that the difference in standing and sitting heart rates is really 5 beats per minute?
    - A. 1%
    - B. 15%
    - C. 50%
    - D. 90%

    Ans: B; LO: 1.6-2; Difficulty: Hard; Type: MC

Questions 12 and 13: Do older adults (ages 65+) have lower body temperatures than younger adults (ages 18-64), on average? Researchers decide to conduct a test of $H_0 : \mu_{older} = \mu_{younger}$ vs. $H_A : \mu_{older} < \mu_{younger}$ at the $\alpha = 0.01$ significance level with sample sizes of 25 in each group.

Suppose that the true mean body temperature for older adults is 97.5º F, the true mean body temperature for younger adults is 98.6º F, and both groups have a standard deviation of 0.75º F.

Graph 1: Simulated values of $\overline{x}_{Older} - \overline{x}_{Younger}$ assuming $\mu_{Older} - \mu_{Younger} = 0$	**Graph 2:** Simulated values of $\overline{x}_{Older} - \overline{x}_{Younger}$ assuming $\mu_{Older} - \mu_{Younger} = 97.5 - 98.6 = -1.1$
 Count samples Less than ≤ –0.51 Count = 100/10000 (0.0100)	 Count samples Less than ≤ –0.51 Count = 9967/10000 (0.9967)

12. Match each term to its representation in the graphs above.

Power:	A. The area to the left of the red line in Graph 1
Prob(Type I error):	B. The area to the left of the red line in Graph 2
Prob(Type II error):	C. The area to the right of the red line in Graph 2

Ans: Power: B, P(Type I): A, P(Type II): C; LO: 1.6-1; Difficulty: Medium; Type: Ma

13. The rejection region for this study is a difference of means ($\overline{x}_{Older} - \overline{x}_{Younger}$) of -0.51 or lower. Suppose the significance level was changed from $\alpha = 0.01$ to $\alpha = 0.05$. The boundary of the rejection region (the red line) would shift to the _____ (left/right), and the power would ______ (increase/decrease).

Ans: right, increase; LO: 1.6-1; Difficulty: Medium; Type: FIB

Question 14 and 15: *Olestra* was approved by the FDA for use in snack foods as a fat substitute in the 1990s. Because there were anecdotal reports of stomach (GI) problems associated with Olestra consumption, researchers planned to carry out an experiment to compare GI symptoms after consuming Olestra potato chips or regular potato chips.

The researchers consider a difference of proportions of 0.05 to be practically significant. That is, if 20% of people experience GI problems when eating Olestra and 15% experience GI problems while eating regular potato chips, they want to be able to detect the difference between the two. They use software to conduct a power analysis to decide how large the sample size needs to be in order for the study to have 80% power with a significance level of $\alpha = 0.05$ and a one-sided alternative hypothesis.

14. Statistical power is the probability of concluding that the risk of GI problems for those eating chips with Olestra is ___________ (higher than / the same as) the risk for those eating regular potato chips, given that the difference in proportions who experience GI problems is really equal to ____ (0 / 0.05).

    Ans: higher than, 0.05; LO: 1.6-2; Difficulty: Medium; Type: FIB

15. The power analysis suggests a sample size of at least 714. One of the researchers' colleagues is surprised that such a large sample size is necessary. Which of the following is the best explanation?
    A. The difference between 15% and 20% is small, and small effect sizes are more difficult to detect, so a large sample size is required.
    B. The significance level is fairly high. If the researchers changed the significance level from $\alpha = 0.05$ to $\alpha = 0.01,$ they wouldn't need such a large sample size.
    C. 80% power is an unusually high value that demands an unusually large sample. Lowering the power would make their plan more acceptable to funding agencies.
    D. One-sided tests always require larger samples. If they used a two-sided test, the necessary sample size would be roughly half as large.

    Ans: A; LO: 1.6-2; Difficulty: Medium; Type: MC