

Question 3: Focus on Sampling and Experimental Design**4 points****General Scoring Notes**

- Each part of the question (indicated by a letter) is initially scored by determining if it meets the criteria for essentially correct (E), partially correct (P), or incorrect (I). The response is then categorized based on the scores assigned to each letter part and awarded an integer score between 0 and 4 (see the table at the end of the question).
- The model solution represents an ideal response to each part of the question, and the scoring criteria identify the specific components of the model solution that are used to determine the score.

Model Solution	Scoring
(a) This is an observational study. The researchers had the car owners estimate their mileage. The car owners were not randomly assigned a car model, so no treatment was imposed.	Essentially correct (E) if the response satisfies the following three components: - Identifies an observational study - Provides a justification based on no treatment being imposed - Includes context Partially correct (P) if the response satisfies only two of the three components required for E. Incorrect (I) if the response does not meet the criteria for E or P.

Additional Notes:

- A response that states the study is an experiment receives a score of I for part (a).
- A response that states “No random assignment of treatment” may satisfy component 2.
- A response may satisfy component 3 with car models, drivers, cars, owners, or members. Mileage alone does not satisfy component 3.

Model Solution	Scoring
(b) Number the days in the experiment from 1 to 70. Using a random number generator, generate 35 unique integers from 1 to 70, inclusive. Assign the days with those 35 unique integers for James to drive the car with autopilot and assign the remaining 35 days for James to drive the car without autopilot. (Alternative solution) Using 70 equally sized slips of paper, label 35 “with autopilot” and 35 “without autopilot.” Mix the slips of paper in a bag. Each day for the 70 days, select a slip of paper (without replacement) to determine the driving method for that day.	Essentially correct (E) if the response satisfies the following three components: - Creates appropriate labels for the units/treatments - Describes how to correctly implement the random process so that every possible random assignment is equally likely - The response indicates a random process that results in 35 days assigned to using autopilot and 35 days assigned to not using autopilot Partially correct (P) if the response satisfies only two of the three components required for E. Incorrect (I) if the response does not meet the criteria for E or P.

Additional Notes:

- For responses that use slips of paper (or marbles or equivalent) to represent treatments
 - To satisfy component 1 some slips must be labeled or assigned to represent autopilot and some labeled or assigned to represent no autopilot (e.g., blue marbles represent autopilot and yellow marbles represent no autopilot).
 - To satisfy component 2 the slips of paper must be mixed/shuffled, and the response must clearly link the treatment selected to a day (e.g., each day James selects a slip to determine the driving method).
 - To satisfy component 3 the response must indicate that there are 35 slips (or marbles of a specific color) for each treatment and the response must indicate that slips of paper (marbles) are selected without replacement.
- For responses that use slips of paper labeled from 1 to 70 (or an equivalent interval)
 - To satisfy component 1 the days must be labeled 1 to 70.
 - To satisfy component 2 the slips of paper must be mixed/shuffled, and the response must clearly link the day number selected to autopilot or no autopilot.
 - To satisfy component 3 the response must indicate that slips of paper are selected without replacement and that 35 days are assigned to each treatment.
- For responses that use a random number generator with days labeled from 1 to 70 (or an equivalent interval)
 - To satisfy component 1 the days must be labeled 1 to 70.
 - To satisfy component 2 the response indicates that the random number generator selects numbers 1 to 70 inclusive, and the response clearly links the day number selected to autopilot or no autopilot.
 - To satisfy component 3 the response must indicate that numbers are selected without repeats and that 35 days are assigned to each treatment.

- For responses that flip a coin for each day (or roll a die and note odd/even, generate a random number 1 or 2, or equivalent)
 - To satisfy component 1 each outcome must be linked to a treatment (e.g., heads equals autopilot, tails equals no autopilot).
 - To satisfy component 2 the response must clearly link the outcome of the coin flip to a day (e.g., each day James flips a coin to determine the driving method). Note: If the response includes a stopping rule (e.g., when 35 days are assigned one treatment, the remaining days are assigned the other treatment), component 2 is not satisfied because this plan increases the probability that the last days will have the same treatment, which does not meet the equally likely random assignment requirement.
 - To satisfy component 3 the response must indicate that 35 days are assigned to each treatment using a stopping rule. If there is no stopping rule, component 3 is not satisfied.
- If a response uses a random number generator or slips of paper with the numbers 1 to 70 and does not initially number the days from 1 to 70, component 1 may be satisfied if the response indicates a link between the number selected and the day (e.g., if 3 is selected, James uses autopilot on the third day).
- Responses that do not use any random process should be scored 1. For example, “number the days from 1 to 70 and use autopilot on odd-numbered days.”
- Responses that use blocking do not satisfy component 2 because all possible random assignments are not equally likely.
- If the response describes two ways to perform the random assignment, assign the score for the weaker assignment process.

Model Solution	Scoring
(c) In order to generalize his findings to all Model D cars in his club, he would need to randomly select Model D cars from the club. He would then need to carry out a similar study using the Model D cars that were randomly sampled from the club.	Essentially correct (E) if the response satisfies the following three components: - The response indicates that more cars must be sampled - The response indicates that random sampling is required for generalization - The response is in context, which would include sampling from the population of interest (Model D cars from his club) Partially correct (P) if the response satisfies only two of the three components required for E. Incorrect (I) if the response does not meet the criteria for E or P.

Additional Notes:

- Component 3 may be satisfied by using “members” of the club or “owners” in the club rather than cars.
- Any discussion of experimental design beyond the sample selection (such as conditions or two-sample design) should be ignored.

Scoring for Question 3	Score
Complete Response Three parts essentially correct	4
Substantial Response Two parts essentially correct and one part partially correct	3
Developing Response Two parts essentially correct and no part partially correct <i>OR</i> One part essentially correct and one or two parts partially correct <i>OR</i> Three parts partially correct	2
Minimal Response One part essentially correct and no part partially correct <i>OR</i> No part essentially correct and two parts partially correct	1

Question 2: Focus on Sampling and Experimental Design**4 points****General Scoring Notes**

- Each part of the question (indicated by a letter) is initially scored by determining if it meets the criteria for essentially correct (E), partially correct (P), or incorrect (I). The response is then categorized based on the scores assigned to each letter part and awarded an integer score between 0 and 4 (see the table at the end of the question).
- The model solution represents an ideal response to each part of the question, and the scoring criteria identify the specific components of the model solution that are used to determine the score.

Model Solution	Scoring
(a) Experimental units: 60 driveways. Treatments: Concrete with fibers and concrete without fibers. Response variable: Rating of the severity of the cracks after one year, on a scale of 0 to 10.	Essentially correct (E) if the response satisfies the following three components: - Identifies the experimental units as the driveways - Identifies the treatments as concrete with fibers and concrete without fibers - Identifies the response variable as the rating of the severity of the cracks Partially correct (P) if the response satisfies only two of the three components. Incorrect (I) if the response does not meet the criteria for E or P.

Additional Notes:

- The number of experimental units is not necessary to satisfy component 1.
- The experimental units can be identified as “new homes needing driveways” to satisfy component 1.
- The phrases “on a scale of 0 to 10” and “after one year” are not required for component 3.
- Reasonable synonyms for “rating of the severity” such as “rating,” “severity,” or “on a scale of 0 to 10” are acceptable towards satisfying component 3.
- Identification of the response variable as a mean, or average, does not satisfy component 3.

Model Solution	Scoring
(b) Number the 60 driveways from 01 to 60. Using a random number generator, generate two-digit integers between 01 and 60. Ignore 00 and any number greater than 60 until 30 unique numbers are obtained. Assign the driveways with those 30 unique numbers to receive concrete with fibers and the remaining 30 driveways to receive concrete without fibers.	Essentially correct (E) if the response satisfies the following two components: - Describes how to correctly use a random number generator, or some other appropriate random process, to assign driveways that have been labeled 1 to 60 to concrete with fibers and concrete without fibers so that every possible random assignment is equally likely - The random process results in an equal number of driveways assigned to the concrete with fibers and the concrete without fibers Partially correct (P) if the response describes how to implement a random process that satisfies only one of the two components. Incorrect (I) if the response does not meet the criteria for E or P.

Additional Notes:

- Examples of alternative random assignment processes that satisfy both component 1 and component 2 include:
 - Put 60 equally sized slips of paper labeled from 1 to 60 into a hat, mix well, and randomly select 30 slips of paper out of the hat, sampling without replacement. Assign the driveways with the numbers on the 30 selected slips of paper to concrete with fibers and assign the remaining 30 driveways to concrete without fibers.
 - Put 30 white marbles and 30 black marbles in an urn and mix well. Randomly select one marble from the urn. If the marble is white, assign the driveway with label 1 to concrete with fibers, otherwise assign the driveway with label 1 to concrete without fibers. Randomly select a marble from the remaining 59 marbles in the urn. If the marble is white, assign the driveway with label 2 to concrete with fibers, otherwise assign the driveway with label 2 to concrete without fibers. Continue this process of randomly selecting marbles from the urn without replacement until 30 driveways are assigned to concrete with fibers. The other 30 driveways are assigned to concrete without fibers.
 - Spin a spinner with 60 equally sized sections numbered from 1 to 60, ignoring repeats, until 30 unique numbers are generated. Assign the driveways with those 30 numbers to concrete with fibers and assign the remaining 30 driveways to concrete without fibers.
- The following random assignment processes are examples that would satisfy component 2 but would not satisfy component 1.
 - An example that does not specify without replacement: Put 60 equally sized slips of paper labeled from 1 to 60 into a bowl, mix well, and randomly select one slip out of the bowl and record the number. Continue this process of randomly selecting slips until 30 numbers are recorded. Assign the driveways with the 30 recorded numbers to concrete with fibers and assign the remaining 30 driveways to concrete without fibers.

- An example of a stopping rule with equal probabilities: Put 30 white marbles and 30 black marbles into an urn and mix well. Randomly select one marble from the urn. If the marble is white, assign the driveway with label 1 to concrete with fibers, otherwise assign the driveway with label 1 to concrete without fibers. Place the selected marble back into the urn, mix well, and randomly select a marble from the urn. If the marble is white, assign the driveway with label 2 to concrete with fibers, otherwise assign the driveway with label 2 to concrete without fibers. Continue this process of randomly selecting marbles from the urn with replacement until 30 driveways are assigned to concrete with fibers. The other 30 driveways are assigned to concrete without fibers.
- An example where not all outcomes are equally likely: Toss 60 fair coins and record the number of heads. Continue tossing the 60 coins, and ignoring zero, until 30 unique numbers are recorded. Assign the driveways with the 30 recorded numbers to concrete with fibers and assign the remaining 30 driveways to concrete without fibers.
- If there is not some type of labeling system (numbering the driveways), then component 1 is not satisfied.
- If it is not clear that the random assignment process allows every possible random assignment of driveways to type of concrete to be equally likely, then component 1 is not satisfied.
- If the response does not clearly indicate that random numbers are selected without using repeats, then component 1 is not satisfied.
- A response that only assigns driveways to groups and does not indicate how the groups correspond to concrete with fibers and concrete without fibers does not satisfy component 1.
- If a response describes two separate random assignment processes in detail (e.g., how to randomly assign number labels to driveways and how to randomly assign driveways to concrete type), score the combined random assignment process according to the two components.
- If the response describes two ways to make the same random assignment (e.g., how to randomly assign driveways to treatments), assign the score for the weaker random assignment process.

Model Solution	Scoring
(c) The results were statistically significant, and because the driveways were randomly assigned to either the concrete with the fibers or the concrete without the fibers, there is evidence the treatment (type of concrete) caused the response (rating of severity of cracks).	Essentially correct (E) if the response satisfies the following two components: 1. The response indicates that random assignment enables the conclusion that type of concrete caused the rating of severity of the cracks 2. The explanation is in context of the problem Partially correct (P) if the response satisfies only component 1, OR the response indicates random assignment limits the effect of confounding variables in context. Incorrect (I) if the response does not meet the criteria for E or P.

Additional Notes:

- If the response also indicates an incorrect conclusion based on random assignment, such as generalization, representation of the population, reduction of bias, or reduction of variability, reduce the score from an E to a P or from a P to an I.
- Context includes either the treatment (concrete with fibers or concrete without fibers or “fibers”) or the severity rating.

Scoring for Question 2	Score
Complete Response Three parts essentially correct	4
Substantial Response Two parts essentially correct and one part partially correct	3
Developing Response Two parts essentially correct and no part partially correct OR One part essentially correct and one or two parts partially correct OR Three parts partially correct	2
Minimal Response One part essentially correct and no part partially correct OR No part essentially correct and two parts partially correct	1

Question 6: Investigative Task**4 points****General Scoring Notes**

- Each part of the question (indicated by a letter) is initially scored by determining if it meets the criteria for essentially correct (E), partially correct (P), or incorrect (I). The response is then categorized based on the scores assigned to each letter part and awarded an integer score between 0 and 4 (see the table at the end of the question).
- The model solution represents an ideal response to each part of the question, and the scoring criteria identify the specific components of the model solution that are used to determine the score.

	Model Solution	Scoring									
(a)	(i) The relative frequencies of successful and unsuccessful treatments in each clinic are presented in the following table: <table border="1"> <tr> <td></td><td>Clinic A</td><td>Clinic B</td></tr> <tr> <td>Unsuccessful Treatment</td><td>$\frac{51}{139} \approx 0.3669$</td><td>$\frac{33}{68} \approx 0.4853$</td></tr> <tr> <td>Successful Treatment</td><td>$\frac{88}{139} \approx 0.6331$</td><td>$\frac{35}{68} \approx 0.5147$</td></tr> </table> (ii) Clinic A appears to be more successful in treating allergy sufferers than Clinic B. Clinic A was successful for 63.3% of the allergy sufferers it treated, while Clinic B was successful for only 51.5% of the allergy sufferers it treated.		Clinic A	Clinic B	Unsuccessful Treatment	$\frac{51}{139} \approx 0.3669$	$\frac{33}{68} \approx 0.4853$	Successful Treatment	$\frac{88}{139} \approx 0.6331$	$\frac{35}{68} \approx 0.5147$	Essentially correct (E) if the response satisfies the following two components: - The response to part (a-i) provides correct numerical values for the relative frequencies in all four cells of the table - The response to part (a-ii) identifies the clinic that is more successful in treating allergy sufferers and includes a justification based on the relative frequencies reported in part (a-i) Partially correct (P) if the response satisfies only one of the two components OR if the response provides correct numerical values for the relative frequencies in only two or three of the cells of the table, and correctly identifies the clinic that has greater success, with or without providing adequate justification based on the values of the reported relative frequencies. Incorrect (I) if the response does not meet the criteria for E or P.
	Clinic A	Clinic B									
Unsuccessful Treatment	$\frac{51}{139} \approx 0.3669$	$\frac{33}{68} \approx 0.4853$									
Successful Treatment	$\frac{88}{139} \approx 0.6331$	$\frac{35}{68} \approx 0.5147$									

Additional Notes:

- Responses may be given as proportions or percentages.
- To satisfy component 1, numerical values for relative frequencies must be accurately reported to at least two decimal places, e.g., 0.36 or 0.37, 0.48 or 0.49, 0.63 or 0.64, 0.51 or 0.52.
- If a response in part (a-i) reports the relative proportions of the total (dividing the count in each of the four cells by 207) instead of the conditional relative frequencies, then component 2 may be satisfied in either of the following ways:
 - Concluding that Clinic A is more successful in treating allergies based on the larger proportion of successful cases (0.43 vs. 0.17)
 - Concluding that Clinic B is more successful in treating allergies based on the smaller proportion of unsuccessful cases (0.16 vs. 0.25).

- If a response reports incorrect values in part (a-i), then the justification in component 2 may use either the relative frequencies reported in the table in part (a-i) OR the correct conditional relative frequencies given in the model solution.
- To satisfy component 2, the response must make it clear that Clinic A is being compared to Clinic B, either by explicitly mentioning both clinics OR by reporting both relative frequencies that are being compared.

Model Solution	Scoring
(b) No. This is an observational study; it is not a randomized experiment. Cause and effect can only be established with a well-designed, randomized experiment. There may be other variables, besides where a patient was treated, that affect the success rates for treating allergies. For example, Clinic A may mostly treat mild allergy cases that are easy to treat successfully, while Clinic B may mostly treat severe allergy cases that are more difficult to treat successfully.	Essentially correct (E) if the response satisfies the following two components: - Indicates “no,” a causal inference is not justified - Provides at least one of the following explanations: - Notes that this is an observational study - Indicates that this is not a randomized experiment - Identifies a plausible confounding variable that could affect the success rates for treating allergies at the two clinics Partially correct (P) if the response satisfies component 2 but fails to indicate that researchers are not justified in making a causal inference OR if the response satisfies component 1 AND includes one of the following as justification: - A statement that association does not imply causation - A statement that there may be a confounding variable without specifying a reasonable confounding variable. Incorrect (I) if the response does not satisfy the criteria for E or P.

Additional Notes:

- A response that states that treatments (Clinic A, Clinic B) are not assigned (randomly assigned, imposed) to the patients satisfies component 2.
- A response that identifies a reasonable confounding variable does not need to explain why it is a confounding variable in order to satisfy component 2.
- The strength of the response in part (b) should be considered if holistic scoring is necessary.

Model Solution	Scoring
(c) (i) Clinic A: More successful in treating mild allergies. Clinic A successfully treated 75.0% of mild allergy sufferers, while Clinic A successfully treated only 28.6% of severe allergy sufferers. Clinic B: More successful in treating mild allergies. Clinic B successfully treated 91.7% of mild allergy sufferers, while Clinic B successfully treated only 42.9% of severe allergy sufferers. (ii) Clinic A: More likely to treat mild allergy sufferers than severe allergy sufferers. Of the 139 allergy sufferers treated at Clinic A, 104 (74.8%) suffered from mild allergies, while only 35 (25.2%) suffered from severe allergies. Clinic B: More likely to treat severe allergy sufferers than mild allergy sufferers. Of the 68 allergy sufferers treated at Clinic B, 56 (82.4%) suffered from severe allergies, while only 12 (17.6%) suffered from mild allergies.	Essentially correct (E) if the response satisfies the following four components: - The response to part (c-i) indicates that both Clinic A and Clinic B are more successful in treating mild allergies than treating severe allergies - The response to part (c-i) provides a justification based on correctly reported proportions (or percentages) of successfully treated allergy sufferers of each severity for each clinic - The response to part (c-ii) indicates that Clinic A is more likely to treat mild allergy sufferers, while Clinic B is more likely to treat severe allergy sufferers - The response to part (c-ii) provides a justification based on correctly reported frequencies or proportions (or percentages) of severe or mild allergy sufferers treated by each clinic Note: See the Q6 Correct Relative Frequencies tables at the end of this scoring guideline for the relative frequencies that should be used in each designated part. Partially correct (P) if the response satisfies two or three of the four components. Incorrect (I) if the response does not meet the criteria for E or P.

Additional Notes:

- A response to part (c-i) may satisfy component 2 in either of the following ways:
 - For each clinic, reports the correct proportion (or percentage) of severe allergy sufferers successfully treated AND the correct proportion (or percentage) of mild allergy sufferers successfully treated.
 - Makes a correct reference to the mosaic plot, comparing the ratio of shaded area to open area within the “Mild” rectangle to the ratio of shaded area to open area within the “Severe” rectangle for each clinic. Numerical values are not required.
- A response to part (c-ii) may satisfy component 4 in any of the following ways:
 - For each clinic, reports the correct proportion (or percentage) of treated allergy sufferers with severe allergies. (Clinic A: 0.252, Clinic B: 0.824)
 - For each clinic, reports the correct proportion (or percentage) of treated allergy sufferers with mild allergies. (Clinic A: 0.748, Clinic B: 0.176)

- Reports the correct proportion (or percentage) of treated allergy sufferers with severe allergies for one clinic AND reports the correct proportion (or percentage) of treated allergy sufferers with mild allergies for the other clinic.
- For each clinic, reports the correct frequency of treated allergy sufferers with mild allergies and the correct frequency of treated allergy sufferers with severe allergies. (Clinic A: 104 mild and 35 severe; Clinic B: 12 mild and 56 severe)
- Makes a correct reference to the mosaic plot, comparing the area or width of the “Mild” rectangle to the area or width of the “Severe” rectangle within each clinic. Numerical values are not required.

Model Solution	Scoring
(d) The more successful clinic identified in part (a-ii) is Clinic A, which is different from the physician's conclusion that Clinic B is better when taking allergy severity into account. This happens because for both clinics the success rate is much higher for mild allergy sufferers (75.0 percent versus 28.6 percent for Clinic A, and 91.7 percent versus 42.9 percent for Clinic B). Clinic A treats mostly mild allergy sufferers (74.8 percent of its patients) while Clinic B treats mostly severe allergy sufferers (82.4 percent of its patients). Therefore, when combining the results across allergy severity categories to obtain the table from part (a-i), the facts that Clinic A treats a larger proportion of mild allergy sufferers and mild allergy sufferers have a higher success rate make it appear as if Clinic A is better overall.	Essentially correct (E) if the response satisfies the following two components: 1. The response indicates that within each clinic, the proportion of successfully treated mild allergy sufferers is greater than the proportion of successfully treated severe allergy sufferers (0.75 vs. 0.286 for Clinic A and 0.917 vs. 0.429 for Clinic B) <i>OR</i> the response indicates that the overall proportion of successfully treated mild allergy sufferers (0.767) is much larger than the overall proportion of successfully treated severe allergy sufferers (0.374) 2. The response indicates that the allergy sufferers treated by Clinic A include a much greater proportion of mild cases than the allergy sufferers treated by Clinic B (sample proportions: 0.748 vs. 0.176) <i>OR</i> the response indicates that the allergy sufferers treated by Clinic B includes a much greater proportion of severe cases than the allergy patients treated by Clinic A (sample proportions: 0.824 vs 0.252) <i>OR</i> the response indicates that of the allergy sufferers treated at Clinic A, a higher proportion are mild (0.748) than severe (0.252), and of the allergy sufferers treated by Clinic B, a higher proportion are severe (0.824) than mild (0.176) Partially correct (P) if the response satisfies only one of the two components. Incorrect (I) if the response does not meet the criteria for E or P.

Additional Notes:

- A response that does not use the answer from (c-i) cannot satisfy component 1.
- A response that does not use the answer from (c-ii) cannot satisfy component 2.
- A response that compares successful treatments between clinics (0.75 vs. 0.917 or 0.286 vs. 0.429), does not satisfy component 1 or component 2.

- A response does not need to repeat the calculated values from part (c), but if it does, the numbers must be the same as those that were reported in part (c).
- A response that only compares the proportion of successful (mild and severe) treatments at Clinic A to the proportion of successful (mild and severe) treatments at Clinic B should be scored I.
- If the response in part (a-ii) concludes that Clinic B is more successful in treating allergies, then part (d) should be scored as follows:
 - **Essentially correct (E)** if the response includes the following four components:
 - (i) At least one of the two components listed for E above.
 - (ii) The response states that overall, Clinic B has a higher proportion of successfully treated allergy sufferers.
 - (iii) The response states that Clinic B has a higher proportion of successfully treated mild allergy sufferers.
 - (iv) The response states that Clinic B has a higher proportion of successfully treated severe allergy sufferers.
 - **Partially correct (P)** if the response satisfies component (i) AND two of the three remaining components (ii)–(iv) listed for E in this note.
- If the response in part (a-ii) concludes that Clinic B is more successful, the response can be scored no higher than 3.
- The strength of the response in part (d) should be considered if holistic scoring is necessary.

Q6 Correct Relative Frequencies

(a-i)

	Clinic A	Clinic B
Unsuccessful treatment	$\frac{51}{139} = 0.367$	$\frac{33}{68} = 0.485$
Successful treatment	$\frac{88}{139} = 0.633$	$\frac{35}{68} = 0.515$

(c-i)

	Clinic A	Clinic B
Successful treatment - Mild	$\frac{78}{104} = 0.75$	$\frac{11}{12} = 0.917$
Successful treatment - Severe	$\frac{10}{35} = 0.286$	$\frac{24}{56} = 0.429$

(c-ii)

	Clinic A	Clinic B
Mild cases treated	$\frac{104}{139} = 0.748$	$\frac{12}{68} = 0.176$
Severe cases treated	$\frac{35}{139} = 0.252$	$\frac{56}{68} = 0.824$

NOTE: To satisfy the components in part (c), the comparison needs to be made by severity WITHIN each clinic (a vertical comparison), not Clinic A to Clinic B.

Scoring for Question 6

Each essentially correct (E) part counts as 1 point, and each partially correct (P) part counts as $\frac{1}{2}$ point.

	Score
Complete Response	4
Substantial Response	3
Developing Response	2
Minimal Response	1

If a response is between two scores (for example, 2 $\frac{1}{2}$ points), use a holistic approach to decide whether to score up or down, depending on the strength of the response and quality of the communication.

Question 2: Focus on Collecting Data

4 points

General Scoring Notes

- Each part of the question (indicated by a letter) is initially scored by determining if it meets the criteria for essentially correct (E), partially correct (P), or incorrect (I). The response is then categorized based on the scores assigned to each letter part and awarded an integer score between 0 and 4 (see the table at the end of the question).
- The model solution represents an ideal response to each part of the question, and the scoring criteria identify the specific components of the model solution that are used to determine the score.

Model Solution

Scoring

- (a) Keeping daily journals could introduce response bias due to the self-reporting by subjects who may have a poor or incomplete memory of the amount of walking that was done. If most subjects who keep daily journals underreport the number of miles walked per day because they cannot remember all of their walking at the end of the day, then the estimate of mean daily miles walked for the target population will be biased too low. Wearing activity trackers would likely provide a more accurate record of daily miles walked by each subject in the study.

Essentially correct (E) if the response satisfies the following two components:

- Indicates that keeping a daily journal could result in a bias that would be avoided by using activity trackers AND provides a reasonable explanation
- Provides a description of a bias that refers to at least one of the following:
 - The use of a daily journal may result in a systematic/consistent underreporting, or systematic/consistent overreporting of daily miles walked
 - The use of a daily journal may result in a biased estimation (underestimation or overestimation) of a population parameter (e.g., mean daily miles walked for the members of the target population)

Partially correct (P) if the response satisfies only one of the two components.

Incorrect (I) if the response does not meet the criteria for E or P.

Additional Notes:

- A response does not need to specifically name a type of bias (e.g., response bias).
- The response may refer to the explanatory variable as “activity level.”
- The direction of the bias need not be specified in order to satisfy component 1.
- Examples of reasonable explanations for indicating that keeping a daily journal may result in a bias include:
 - “Because the subjects are self-reporting their daily miles walked.”
 - “Because the subjects may not accurately recall their daily miles walked.”
 - “Because the subjects may forget to complete an entry in their journal.”
- The direction of the bias must be specified in order to satisfy component 2.

- The response must indicate the underreporting or overreporting is systematic across the subjects (or there is a tendency to underreport or overreport) in order to satisfy component 2. Examples of responses that satisfy component 2 include:
 - “The subjects in the study may consistently underreport their daily miles walked.”
 - “Subjects are likely to underreport their daily miles walked.”
 - “Most subjects may overreport their daily miles walked.”
 - “The bias may result in an estimate of the mean daily miles walked by members of the target population that is lower than the target population mean.”
- A response that indicates the underreporting or overreporting for only some people does not satisfy component 2 (e.g., “Some people might record higher miles than they actually walk.”).

Model Solution	Scoring
(b) It is necessary to have a representative sample of subjects from the population in order to make an unbiased inference about the difference between the mean cholesterol levels for all adult members of the target population who walk fewer miles per day and the mean cholesterol levels for all adult members of the target population who walk more miles per day.	Essentially correct (E) if the response satisfies the following two components: - Provides an explanation that the use of a representative sample is necessary in order to make a valid generalization about the target population - Refers to estimation, or inference, for cholesterol levels in the target population OR an association between cholesterol level and amount of walking in the target population Partially correct (P) if the response satisfies only one of the two components. Incorrect (I) if the response does not meet the criteria for E or P.

Additional Notes:

- A response that discusses the accuracy or validity of a significance test does not satisfy component 1 unless the response makes it clear that the inference is being generalized to the target population.
- In order to satisfy component 2, the response need not state a specific population parameter(s).
- If a parameter is specified, it must be relevant to cholesterol level or the association between cholesterol level and amount of walking. Some examples include:
 - Individual population mean cholesterol level
 - One or more differences between population mean cholesterol levels
 - Individual population median cholesterol level
 - One or more differences between population median cholesterol levels
 - A population correlation between cholesterol level and amount of walking
 - A population regression model for cholesterol level and amount of walking

Model Solution	Scoring
(c) No, since the treatments (amounts of walking) were not randomly assigned to the subjects in the study, it would not be valid to claim that increased walking causes a decrease in average cholesterol levels for adults in the target population. The researchers would only be able to conclude that cholesterol level has a negative association with daily miles walked for adults in the target population. There may be one or more confounding variables that are the actual cause of the relationship. For example, people who walk more may be more concerned about maintaining a healthy diet and eat more foods that are low in cholesterol, while people who walk less may eat more foods that are high in cholesterol. Consequently, the association between cholesterol and daily miles walked could actually be caused by differences in diets and not differences in amount of walking.	Essentially correct (E) if the response satisfies the following two components: - Indicates that a causal inference cannot be made - Provides a valid explanation that is based on one of the following: - the lack of (random) assignment of treatments to subjects - being an observational study/not an experiment - the existence of a possible confounding variable that is associated with amount of walking and associated with cholesterol level Partially correct (P) if the response satisfies only component 1 AND provides a weak explanation. Incorrect (I) if the response does not meet the criteria for E or P.

Additional Notes:

- A response that provides an explanation that is based on the existence of a possible confounding variable may or may not identify a specific confounding variable. In either case, the response must indicate that the confounding variable has an association with amount of walking AND also indicate that the confounding variable has an association with cholesterol level in order to satisfy component 2. Examples of responses that satisfy component 2:
 - A response that identifies a reasonable confounding variable: “Diet could be a confounding variable. People who walk more may tend to eat more foods that are low in cholesterol, while people who walk less may tend to eat more foods high in cholesterol.”
 - A response that does not identify a confounding variable: “There could be a confounding variable that has an association with cholesterol level and also has an association with amount of walking.”
- If a response identifies a specific confounding variable, then any variable that is reasonable (e.g., diet, weight, body mass index, etc.) should be accepted in scoring component 2.
- In component 2, the following are examples of weak explanations:
 - The response indicates the existence of a confounding variable but does not indicate that the confounding variable has an association with amount of walking AND an association with cholesterol level.
 - The response communicates that an association between cholesterol level and amount of walking does not imply that there is a causal relationship between cholesterol level and amount of walking. However, a general statement, without context, that association does not imply causation should be scored incorrect (I).

- A response that only references specific elements of an experiment (e.g., placebo, control group, replication) aside from assignment of treatments to subjects should be scored incorrect (I).
- A response that states that a causal relationship can be concluded due to the statistically significant result and goes on to say that there may be a confounding variable that is associated with amount of walking and cholesterol level (e.g., diet) should be read as parallel solutions and scored incorrect (I).
- Responses in parts (a) or (b) cannot be carried down to part (c) to satisfy component 2 unless the response in part (c) refers to specific statements in part (a) or (b).

Scoring for Question 2	Score
Complete Response Three parts essentially correct	4
Substantial Response Two parts essentially correct and one part partially correct	3
Developing Response Two parts essentially correct and no part partially correct <i>OR</i> One part essentially correct and one or two parts partially correct <i>OR</i> Three parts partially correct	2
Minimal Response One part essentially correct and no part partially correct <i>OR</i> No part essentially correct and two parts partially correct	1

2018

AP[®] CollegeBoard

AP Statistics

Scoring Guidelines

2018 SCORING GUIDELINES

Question 1

Intent of Question

The primary goals of this question were to assess a student's ability to (1) identify various values in regression computer output; (2) interpret the intercept of a regression line in context; (3) interpret the coefficient of determination (r^2) in context; and (4) identify an outlier from a scatterplot.

Solution

Part (a):

The estimate of the intercept is 72.95. It is estimated that the average time to finish checkout if there are no other customers in line is 72.95 seconds.

Part (b):

The coefficient of determination is $r^2 = 73.33\%$. This value indicates that 73.33% of the variability in the times it takes customers to finish checkout, including time waiting in line, can be explained by knowing how many customers are in line in front of the selected customer.

Part (c):

The outlier is the point with $x = 3$ and y close to 0. This point is considered an outlier because the combination of x and y values differs from the pattern of the rest of the data. Specifically, the value of y (time to finish checkout) is much lower than would be expected when there are $x = 3$ customers in line in front of the selected customer, given the remaining data.

Scoring

Parts (a), (b), and (c) are scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

1. Correctly identifies 72.95 as the intercept.
2. Communicates the concept of a y -intercept in a context that includes both time and zero customers.
3. Indicates that the value of the intercept is a prediction by using language such as "predicted," "estimated," or "average" value of y .

Partially correct (P) if the response includes only two of the three components.

Incorrect (I) if the response includes at most one of the three components.

2018 SCORING GUIDELINES

Question 1 (continued)

Notes:

- Regression equations (such as $\hat{y} = 72.95 + 174.40x$) cannot be used to satisfy identification of the intercept in component 1, unless the intercept is explicitly labeled as such.
- A regression equation cannot be used to satisfy component 3.
- Incorrect regression equations are treated as extraneous and do not affect the scoring of any component.
- A response that interprets 72.95 as a slope does not satisfy components 1 or 2.

Part (b) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

1. Correctly identifies 73.33% as the coefficient of determination.
2. Provides a correct (possibly generic) interpretation of r^2 .
3. Interpretation includes context.

Partially correct (P) if the response satisfies only two of the three components;

OR

if the response satisfies the three components, but reverses the roles of number of customers in line and time to finish checkout in the interpretation.

Incorrect (I) if the response satisfies at most one of the three components.

Notes:

- In component 2 the correct interpretation of the coefficient of determination can take any of several equivalent forms, such as:
 - The percent variability in y that is attributed to the linear relationship between y and x or between x and y .
 - The proportion of the total variability in the dependent variable y that is explained by the independent variable x .
 - The proportion of variation in y that is accounted for by the linear model.
 - The proportionate reduction of total variation of the y values that is associated with the use of the independent variable x .
 - The proportionate reduction in the sum of the squares of vertical deviations obtained by using the least-squares line instead of the naïve prediction of $\bar{y}$.
- In component 2 common *incorrect* interpretations of the coefficient of determination include:
 - The percent variability in the *predicted* y values that is explained by the linear relationship between y and x .
 - The percent variability in the *data* that is explained by the linear relationship between y and x .
 - The percent variability that is explained by the linear relationship between y and x .
 - The percent variability in y that is *on average* explained by the linear relationship between y and x .
- For component 3 context must include mention of time or customers.

2018 SCORING GUIDELINES

Question 1 (continued)

Part (c) is scored as follows:

Essentially correct (E) if the response satisfies the following two components:

1. Correctly identifies the outlier.
2. Describes an unusual feature of the identified scatter plot point, relative to the remaining data points, that is sufficient to identify it as the outlier. Examples include:
 - The combination of x and y values is unusual compared to the other points.
 - The value of y is much lower than would be expected (or predicted), given the remaining data.
 - The residual for the point is unusually large relative to the other residuals.

Partially correct (P) if the response satisfies component 1 but does not satisfy component 2.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- In the absence of any point being circled on the graph, component 1 can still be satisfied by explicitly referring to the coordinates of the outlier. Valid coordinates for outlier identification must specify an x value of 3 and a y value that is strictly between 0 and 250.
- A response that does not make a comparison to the remaining data points, such as stating the outlier has a large residual or is nowhere near the regression line, does not satisfy component 2.
- A response that makes a comparison to the remaining data points based upon an unusual feature that is *insufficient* for outlier identification, such as stating the point is the only point with that particular y value, does not satisfy component 2.
- In the absence of explicit numerical calculation, a response that appeals to the influence that the outlier has on the regression coefficient estimates or on the sample correlation coefficient does not satisfy component 2.

2018 SCORING GUIDELINES

Question 1 (continued)

4 Complete Response

Three parts essentially correct

3 Substantial Response

Two parts essentially correct and one part partially correct

2 Developing Response

Two parts essentially correct and no parts partially correct

OR

One part essentially correct and one or two parts partially correct

OR

Three parts partially correct

1 Minimal Response

One part essentially correct

OR

No parts essentially correct and two parts partially correct

2018 SCORING GUIDELINES

Question 2

Intent of Question

The primary goals of this question were to assess a student's ability to (1) calculate the sample size when given the endpoints of a confidence interval for a proportion; (2) explain how bias could be present in a particular survey method; and (3) estimate a proportion from sample data collected using a method designed to decrease bias.

Solution

Part (a):

Using the standard formula for a confidence interval for one proportion, the interval (0.584 to 0.816) is found as follows. $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ where $\hat{p} = \frac{0.584 + 0.816}{2} = 0.7$, the margin of error is $0.816 - 0.7 = 0.116$, and $z^* = 1.96$.

Solving $1.96 \sqrt{\frac{0.7(1-0.7)}{n}} = 0.116$ yields $n = \frac{(1.96)^2(0.7)(1-0.7)}{(0.116)^2} \approx 59.95$. The sample size was 60.

Part (b):

Bias might have been introduced because students responded directly to the environmental science teacher. Because the students would know that an environmental science teacher cares about the environment, they might say yes when they actually don't recycle. This would result in a point estimate that is greater than the proportion of all students who would respond yes to the question.

Part (c):

(i) The expected number is $(300)\left(\frac{1}{2}\right) = 150$.

(ii) The point estimate is based on expecting 150 students to be required to say no and 150 students to truthfully answer the question. Of the 213 answers of no, we expect that $213 - 150 = 63$ were from students who truthfully answered the question. That means we expect that the remaining $150 - 63 = 87$ students truthfully answered the question and responded yes. So the point estimate for the proportion of all students at the high school who would respond yes to the question is $\frac{87}{150} = 0.58$.

2018 SCORING GUIDELINES

Question 2 (continued)

Scoring

Parts (a), (b), and (c) are scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is scored as follows:

Essentially correct (E) if the response satisfies the following five components:

1. Uses a standard error in the form $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ where $\hat{p}$ is between 0 and 1.
2. Shows evidence that $\hat{p} = 0.7$ was correctly used in the standard error.
3. Shows evidence that 0.116 was correctly used as the margin of error in the calculation.
4. Shows evidence that $z^* = 1.96$ was correctly used as the critical value in the calculation.
5. Includes a single, positive whole-number answer.

Partially correct (P) if the response satisfies only three or four of the five components.

Incorrect (I) if the response satisfies at most two of the five components.

Notes:

- Using an equation in the form $n = \frac{z^2 \hat{p}(1-\hat{p})}{\text{MOE}^2}$ satisfies component 1.
- A value of 0.21 in the numerator of the standard error implies that $\hat{p} = 0.7$ was correctly used in the standard error and satisfies component 2.
- An equation such as $0.816 = 0.7 + \text{MOE}$ implies that 0.116 was correctly used for the margin of error and satisfies component 3.
- Statements that suggest a whole-number answer is approximate (such as, “about 60” or “ ≈ 60 ”) satisfy component 5.
- Algebraic work between the set-up and final answer does not need to be shown to satisfy component 5.
- When calculating the values 0.7, 0.116, or 1.96, ignore minor arithmetic errors or transcription errors if they can be identified by the work shown.

2018 SCORING GUIDELINES

Question 2 (continued)

Part (b) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

1. Explains why the responses to the survey might differ from the truth about student recycling in this context (for example, the survey was not anonymous, the question was asked by an authority figure).
2. Explains how the responses to the survey might differ from the truth about student recycling (for example, “students might say yes when they actually don’t recycle,” “students lie and say yes,” “students don’t recycle but lie to the teacher”).
3. Describes the effect of the bias on the point estimate (or the proportion, percentage, number of yes responses in the sample) and doesn’t contradict the bias described.

Partially correct (P) if the response satisfies only two of the three components.

Incorrect (I) if the response satisfies at most one of the three components.

Notes:

- To satisfy component 1 the response must provide a reason that is based on a bias created by the teacher asking students in person. For example, a response that addresses the wording of the question, voluntary response, or sampling variability does not satisfy component 1.
- To satisfy component 2 the response needs to explicitly contrast what the students say with what they do.
- Evidence used to address component 3 cannot also be used to address component 2. For example, a response that says “Students might lie, producing an estimate that is too high” addresses the effect of the bias on the point estimate but should not be combined with the statement about students lying to infer that students do not actually recycle. However, a response that says “Students may lie and say yes, producing an estimate that is too high” satisfies both components 2 and 3.
- If the response is clearly about the population proportion and not about the point estimate, component 3 cannot be satisfied.
- Statements such as “the interval will be too high” do not satisfy component 3 because they don’t specifically address the point estimate.

2018 SCORING GUIDELINES

Question 2 (continued)

Part (c) is scored as follows:

Essentially correct (E) if the response gives an answer of 150 in (c-i) and gives an answer of 0.58 (or equivalent) in (c-ii).

Partially correct (P) if the response gives an answer of 150 in (c-i) and gives an answer of 0.42 (or equivalent) in (c-ii);

OR

if the response does not give an answer of 150 in (c-i) but gives an answer of 0.58 (or equivalent) with supporting work in (c-ii).

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- In part (c-i) the answer must be a single number. Responses such as “at least 150” or “147–153” are incorrect. However, responses such as “about 150” or “ ≈ 150 ” are acceptable.
- In part (c-ii) the proportion can be described verbally (e.g., “87 out of 150”).
- In part (c-ii) if the response clearly indicates that 0.58 (or 0.42) is the population proportion, lower the overall score in part (c) by one level (that is, from E to P, or from P to I). Using probability notation such as $P(\text{yes})$ does not clearly indicate a population proportion.
- In part (c-ii) if the response includes a point estimate of 0.58 or 0.42 but uses a confidence interval as the final answer, lower the overall score in part (c) by one level (that is, from E to P, or from P to I).
- If the answer is incorrect in part (c-i) and the answer in part (c-ii) uses numerator = 87 and denominator = 300 – answer to (c-i), the response should be scored P.

2018 SCORING GUIDELINES

Question 2 (continued)

4 Complete Response

Three parts essentially correct

3 Substantial Response

Two parts essentially correct and one part partially correct

2 Developing Response

Two parts essentially correct and no parts partially correct

OR

One part essentially correct and one or two parts partially correct

OR

Three parts partially correct

1 Minimal Response

One part essentially correct

OR

No parts essentially correct and one or two parts partially correct

2018 SCORING GUIDELINES

Question 3

Intent of Question

The primary goals of this question were to assess a student's ability to (1) compute a probability based on a weighted mixture of two populations; (2) compute a conditional probability; and (3) recognize a binomial random variable and compute the probability associated with it.

Solution**Part (a):**

Let L denote left-handed, M denote multiple birth, and S denote single birth.
The probability that a randomly selected child born in the region is left-handed is:

$$P(L) = P(M)P(L | M) + P(S)P(L | S) = (0.035)(0.22) + (0.965)(0.11) = 0.0077 + 0.10615 = 0.11385.$$

Part (b):

From part (a), $P(L) = 0.11385$. Therefore,

$$P(M | L) = \frac{P(L \text{ and } M)}{P(L)} = \frac{(0.035)(0.22)}{0.11385} = \frac{0.0077}{0.11385} \approx 0.0676.$$

Part (c):

Let X represent the number of children who are left-handed in a random sample of 20 children from the region. X has a binomial distribution with $n = 20$ and $p = 0.11385$ (found in part (a)). Using the binomial distribution,

$$\begin{aligned} P(X \geq 3) &= 1 - P(X \leq 2) \\ &= 1 - \binom{20}{0}(0.11385)^0(0.88615)^{20} - \binom{20}{1}(0.11385)^1(0.88615)^{19} - \binom{20}{2}(0.11385)^2(0.88615)^{18} \\ &\approx 1 - 0.598 \approx 0.402. \end{aligned}$$

2018 SCORING GUIDELINES

Question 3 (continued)

Scoring

Parts (a), (b), and (c) each scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is scored as follows:

Essentially correct (E) if the probability is computed correctly, *AND* work is shown that includes correct numerical values using a formula, end results from a tree diagram, or some other appropriate strategy.

Partially correct (P) if the response provides a reasonable strategy for finding the probability, such as a formula or tree diagram, but uses one or more inappropriate values;

OR
if the response gives the correct probability but not enough work is shown to determine how that probability was found.

Incorrect (I) if the response does not meet the criteria for E or P.

Note: A reasonable strategy needs to include summing the results of two multiplications.

Part (b) is scored as follows:

Essentially correct (E) if the probability is computed correctly, with work shown that includes appropriate numerical values for both the numerator and denominator.

Partially correct (P) if the response includes a numerator and denominator in calculating the conditional probability, with one appropriate term (numerator or denominator) and the other inappropriate.

Incorrect (I) if the response does not meet the criteria for E or P.

Note: Appropriate values include incorrectly calculated values from part (a).

2018 SCORING GUIDELINES

Question 3 (continued)

Part (c) is scored as follows:

Essentially correct (E) if the response satisfies the following five components:

1. Uses a calculation based on the binomial distribution to find the probability of the number of children in the sample who are left-handed.
2. Specifies appropriate values for n and p .
3. Uses correct endpoint value for the probability.
4. Uses correct direction to calculate the probability of at least three left-handed children.
5. Correctly calculates a binomial probability consistent with the previous work.

Partially correct (P) if the response satisfies component 1 and only two or three of the other four components;
OR

if components 2, 3, 4, and 5 are met, and the response does not explicitly indicate the binomial distribution is used by name or formula.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- “Appropriate” values include incorrectly calculated values from part (a) or a recalculated probability from part (b).
- An unlabeled numerical value in a calculator statement cannot be used to satisfy a component.
- A response which calculates $P(X \leq 3)$ satisfies component 3 but does not satisfy component 4.
- A normal approximation to the binomial is not appropriate because $np = 20 \times 0.11385 = 2.277 < 5$.
A response using the normal approximation can score at most P. To earn a score of P, the response must include all of the following:
 - a correct mean and standard deviation based on the binomial parameters
 - clear indication of boundary and direction with a z-score or diagram
 - the probability computed correctly

Notes for all parts:

- If the resulting probability or part of the calculation of the probability uses a value that is not between 0 and 1, inclusive, the score for that part is lowered by one level (that is, from E to P, or from P to I).
- An arithmetic or transcription error in a response can be ignored if correct work is shown. For example, $0.0077 + 0.10615 = 0.1385$.

2018 SCORING GUIDELINES

Question 3 (continued)

4 Complete Response

Three parts essentially correct

3 Substantial Response

Two parts essentially correct and one part partially correct

2 Developing Response

Two parts essentially correct and no parts partially correct

OR

One part essentially correct and one or two parts partially correct

OR

Three parts partially correct

1 Minimal Response

One part essentially correct

OR

No parts essentially correct and two parts partially correct

2018 SCORING GUIDELINES

Question 4

Intent of Question

The primary goals of this question were to assess a student's ability to (1) determine whether a cause-and-effect conclusion can be made based on how a study was conducted and (2) set up, perform, and interpret the results of a hypothesis test, in the context of the problem.

Solution

Part (a):

Yes, it would be reasonable to conclude that the new procedure causes a reduction in recovery time, for patients similar to those in the study. The patients in the study were randomly assigned to the two procedures, which reduces the chance that confounding variables will affect the results. Therefore the statistically significant reduction in mean recovery time can be attributed to the new procedure being superior to the standard procedure.

Part (b):

Step 1: State a correct pair of hypotheses.

Let μ_S represent the mean recovery time among all patients similar to those in the study if they were to receive the standard treatment.

Let μ_N represent the mean recovery time among all patients similar to those in the study if they were to receive the new treatment.

The hypotheses to be tested are $H_0 : \mu_S = \mu_N$ versus $H_a : \mu_S > \mu_N$.

Step 2: Identify a correct test procedure (by name or by formula) and check appropriate conditions.

The appropriate procedure is a two-sample t -test for a difference between means.

Because this is an experiment, the first condition is that subjects were randomly assigned to one treatment group or the other. In this case the condition is satisfied because we were told that the subjects were randomly assigned to either the standard or new procedure.

The second condition is that the recovery times of the two populations are normally distributed or the sample sizes are sufficiently large to presume that the distribution of the difference in the sample means is approximately normal. In this case the condition is met because the sample sizes of 110 and 100 are both sufficiently large.

2018 SCORING GUIDELINES

Question 4 (continued)

Step 3: Correct mechanics, including the value of the test statistic, degrees of freedom, and p -value (or rejection region).

$$\text{The test statistic is } t = \frac{\bar{x}_S - \bar{x}_N}{\sqrt{\frac{s_S^2}{n_S} + \frac{s_N^2}{n_N}}} = \frac{217 - 186}{\sqrt{\frac{34^2}{110} + \frac{29^2}{100}}} \approx 7.13.$$

The p -value is the area greater than 7.13 for a t -distribution with $df = 207.18$, which is essentially 0 (8.36×10^{-12}).

Step 4: State a correct conclusion in the context of the problem, using the result of the statistical test.

Because the p -value is very small, we have sufficient evidence to conclude that for patients similar to the ones in the study, those receiving the new procedure would have less recovery time, on average, than those receiving the standard procedure.

Scoring

This question is scored in three sections. Section 1 consists of part (a); section 2 consists of step 1, step 2, and the test statistic in step 3 in part (b); and section 3 consists of the p -value in step 3 and step 4 in part (b). Sections 1, 2, and 3 are each scored essentially correct (E), partially correct (P), or incorrect (I).

Section 1 is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

1. Correctly states that it is reasonable to make a causal conclusion.
2. Justifies the causal conclusion based on random assignment of patients to procedures (or procedures to patients);
OR
justifies the causal conclusion by stating that a randomized experiment was conducted.
3. Includes the context of the situation.

Partially correct (P) if the response satisfies component 1 AND provides WEAK justification of the causal conclusion by stating that there was random assignment or a randomized experiment was conducted, but with no context;

OR
by stating that an experiment was conducted or there was assignment (without the word "randomized") AND the response includes context of the situation;

OR
by stating that the study design reduces the chance of confounding variables or balances the effects of uncontrolled variables across both groups in context without explicitly referring to the random assignment.

Incorrect (I) if the response does not meet the criteria for E or P.

2018 SCORING GUIDELINES

Question 4 (continued)

Notes:

- If the response states that it is *not* reasonable to make a causal conclusion because the result could have been due to random chance *AND* explains that there is evidence for a causal conclusion based on random assignment of patients to procedures or by stating that a randomized experiment was conducted, then the response is scored E.
- If the response discusses aspects of an experiment other than random assignment (such as, control, replication, or large samples), then those aspects are considered extraneous and the response can be scored E unless those aspects are incorrect for this study (such as, blocking is a requirement, or the study used blocking, or the study used a placebo) in which case the score should be lowered one level (that is, from E to P, or from P to I).
- If the response correctly states in context that it is reasonable to make a causal conclusion but includes incorrect or contradictory justification (such as, random selection of patients), then the response is scored I.

Section 2 is scored as follows:

Essentially correct (E) if the response satisfies the following four components:

1. Parameters are defined correctly.
2. Hypotheses imply equality in the null and correct direction in the alternative.
3. Correct test is identified by name or formula.
4. Correct test statistic for a difference in means is calculated.

Partially correct (P) if the response satisfies only two or three of the four components.

Incorrect (I) if the response satisfies at most one of the four components.

Notes:

- If standard symbols are used for the parameters with appropriate group labels (such as, μ_S, μ_N), component 1 is satisfied.
- If the correct test is identified, but the response states an incorrect formula or uses incorrect notation in the formula, component 3 is not satisfied.
- A pooled two-sample t -test is acceptable for component 3, but the student must also state and comment on the plausibility of the equal population variances assumption.
- If the response identifies a z -test for equal means as the correct test identification, component 3 is not satisfied but component 4 could be satisfied.

Confidence Interval approach:

- If a single two-sample t -interval for the difference in means is used, components 3 and 4 can be satisfied. Component 3 is satisfied if the t -interval is correctly identified by name or formula. Component 4 is satisfied if the correct interval is calculated. If an alpha level is stated, then an appropriate adjustment to the confidence level must be made because the appropriate test is one-sided.
- If two one-sample t -intervals are used, while not a recommended approach, component 3 is not satisfied but component 4 could be satisfied. Component 4 is satisfied if both intervals are calculated correctly.

2018 SCORING GUIDELINES

Question 4 (continued)

Section 3 is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

1. Makes reference to an approximately correct p -value that is consistent with the test statistic and alternative hypothesis for a difference in means.
2. Correctly justifies the conclusion based on the size of the p -value or the test statistic.
3. Correctly states the conclusion in context.

Partially correct (P) if the response satisfies only two of the three components.

Incorrect (I) if the response does not meet the criteria for E or P or includes a justification not based on the inferential results.

Notes:

Component 1:

- Is satisfied if the response makes reference to a large test statistic without referring to a p -value.

Component 2:

- No alpha level is needed to provide justification of the conclusion based on the size of the p -value.
- Is satisfied if the response states the p -value without reference to size, but it is contiguous to the conclusion and clearly indicates a continuous train of thought.
- A correct interpretation of the p -value with a complete explanation that obtaining a test statistic at least this extreme is unlikely due to chance alone is considered justification based on the size of the p -value.
- If an incorrect interpretation of the p -value is given, the score is lowered one level (that is, from E to P, or from P to I).
- A decision about the null hypothesis (reject H_0 or fail to reject H_0) is not required, but if an incorrect decision is stated based on the given p -value then component 2 is not satisfied.
- If a rejection region approach is used, a reasonable critical value replaces the p -value.

Component 3:

- A correct conclusion must be related to the alternative hypothesis in order to satisfy component 3.
- The following responses do not satisfy component 3:
 - States or implies that the null hypothesis is *accepted*
 - States or implies that the alternative hypothesis has been *proven*
 - States the conclusion in past tense (unless the response did not satisfy a component of section 2 for the use of past tense)

2018 SCORING GUIDELINES

Question 4 (continued)

Confidence Interval approach:

- If a single two-sample t -interval for the difference in means is used:
 - Component 1 is satisfied if the response indicates that zero is either included or not included in the calculated interval.
 - Component 2 is satisfied if the response indicates that the bounds are either both above or both below zero (consistent with alternative hypothesis) and uses that as justification for the conclusion.
 - Component 3 is satisfied if the conclusion is stated in context.
- If two one-sample t -intervals are used (which is not recommended) the response is scored at most P if all three components are satisfied, otherwise scored I:
 - Component 1 is satisfied if the response states that the intervals do not overlap.
 - Component 2 is satisfied if the conclusion indicates that the confidence interval for the new procedure lies below the confidence interval for the standard procedure.
 - Component 3 is satisfied if the conclusion is stated in context.

Note: If the three sections of the response are scored as E, to earn a score of 4 as a complete response, both conditions in step 2 must be correctly stated and justified. Additional condition(s) inappropriate for a two-sample t -test must not be stated. Otherwise, the response earns a score of 3 a substantial response.

2018 SCORING GUIDELINES

Question 4 (continued)

4 Complete Response

Three sections essentially correct with conditions for inference

3 Substantial Response

Three sections essentially correct without conditions for inference

OR

Two sections essentially correct and one section partially correct

2 Developing Response

Two sections essentially correct and no sections partially correct

OR

One section essentially correct and one or two sections partially correct

OR

Three sections partially correct

1 Minimal Response

One section essentially correct

OR

No sections essentially correct and one or two sections partially correct

2018 SCORING GUIDELINES

Question 5

Intent of Question

The primary goals of this question were to assess a student's ability to (1) determine which of two histograms represents data with a larger median; (2) calculate the mean of a combined data set when the separate means and sample sizes are known; and (3) calculate the probability that an individual randomly chosen from a finite population will have a value within one standard deviation of the mean, when provided with values for the mean, standard deviation, and all members of the population.

Solution

Part (a):

The median teaching year for High School A is any value with 100 data values at or below it and 100 data values at or above it. The median teaching year for High School B is the 111th value in the ordered list of values. For High School A the median is in the interval that starts at 7 and ends just before 10, because there are only 94 data values below 7 and 106 data values of at least 7. Therefore the median cannot be less than 7. For High School B the median is in the interval that starts at 4 and ends just before 7 because there are more than half (113) of the data values less than 7. Therefore the median must be less than 7. So High School A must be the one with a median of 7, and High School B must be the one with a median of 6.

Another way to determine which school has the median of 7 is to notice that the distribution for High School B is highly skewed to the right, whereas the distribution for High School A is bimodal with a few possible outliers on the right. A distribution that is highly right-skewed is likely to have a substantially larger mean than median. The mean of both distributions is given as 8.2 years, so it makes sense that the highly right-skewed distribution (High School B) is the one with the bigger gap between the mean and median and, therefore, the one with the lower median of 6.

Part (b):

The mean for the original 200 teachers was given as 8.2 years, and the mean for the additional 18 teachers is 2.5 years. Therefore the mean for the combined data set is:

$$\frac{(200)(8.2) + (18)(2.5)}{200 + 18} = \frac{1,640 + 45}{218} \approx 7.73 \text{ years.}$$

Part (c):

The interval mean plus or minus 1 standard deviation on either side of the mean is 8.2 ± 7.2 , or from 1.0 year to 15.4 years. Because teaching year is recorded as an integer, the interval includes teaching years 1 to 15. The number of teachers in that interval can be found by adding the heights of the five bars in the histogram for the intervals from 1 to 16, which includes $79 + 34 + 28 + 29 + 19 = 189$. Therefore the probability is $\frac{189}{221} \approx 0.8552$.

2018 SCORING GUIDELINES

Question 5 (continued)

Scoring

Parts (a), (b), and (c) are scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

1. States that the median is 6 for High School B and the median is 7 for High School A.
2. Provides a reasonable explanation of how the decision was made.
3. Provides the definition of the median or explicitly applies the definition of a median as a criterion in reaching their decision.

OR

Essentially correct (E) if the response satisfies the following three components:

1. States that the median is 6 for High School B and the median is 7 for High School A.
2. States that High School B shows a skewed distribution (or High School A shows a less skewed distribution).
3. Provides a reasonable explanation of how the more skewed distribution (High School B) would be the one with a larger separation between the mean and median.

Partially correct (P) if the response satisfies the first component and only one of the other two components required for E.

Incorrect (I) if the response does not meet the criteria for E or P.

Note: An incorrect statistical statement in the response will result in E being lowered to P, but not P being lowered to I. For example,

- If either distribution is described as left skewed, normal, or approximately normal;
- If the discussion would indicate a median different than 7 for High School A or a median different than 6 for High School B.

Part (b) is scored as follows:

Essentially correct (E) if the response satisfies the following two components:

1. The correct answer that the mean is 7.73.
2. Enough work to show that the answer was obtained as a weighted average of the two individual means.

Partially correct (P) if the response satisfies only one of the two components.

Incorrect (I) if the response does not satisfy the requirements for E or P.

2018 SCORING GUIDELINES

Question 5 (continued)

Part (c) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

1. Calculates that the appropriate interval is 1 to 15.4 or 1 to 15 teaching years.
2. Correctly sums the counts of data values in the numerator based on the intervals provided.
3. Computes the probability using 221 as the denominator.

Partially correct (P) if the response satisfies only two of the three components;

OR

if the response reports the correct probability (0.8552) without supporting work.

Incorrect (I) if the response satisfies at most one of the three components.

Notes:

- If the response attempts to use the Empirical Rule or normal distribution to provide the desired probability, the response is scored I.
- If an incorrect count is shown in component 2, for instance by including the interval from 16 to 19, then component 3 is satisfied if that incorrect count is divided by 221 to find the reported probability.
- It is acceptable if the count is slightly off because of difficulty reading the exact heights of the bars in the histogram.
- If only one of component 2 or component 3 is missing, but the correct probability (0.8552) is reported, the response can be scored E.
- If the response recognizes that all values in the histogram bins up to 16 fall within one standard deviation of the mean and reports the interval as 1 to 16, component 1 is satisfied.

2018 SCORING GUIDELINES

Question 5 (continued)

4 Complete Response

Three parts essentially correct

3 Substantial Response

Two parts essentially correct and one part partially correct

OR

Part (a) essentially correct and two parts partially correct

2 Developing Response

Two parts essentially correct and no parts partially correct

OR

Part (b) or part (c) essentially correct and one or two parts partially correct

OR

Three parts partially correct

1 Minimal Response

One part essentially correct

OR

No parts essentially correct and one or two parts partially correct

2018 SCORING GUIDELINES

Question 6

Intent of Question

The primary goals of this question were to assess a student's ability to (1) describe what constitutes a Type II error for a specific hypothesis test; (2) specify a rejection region in terms of values of the sample mean; (3) compute the power of a test for a specific value in the alternative hypothesis; (4) recognize the definition of power; and (5) understand the impact of increasing the sample size on the power of a test.

Solution

Part (a):

A Type II error occurs when the alternative hypothesis is true, but the null hypothesis is not rejected. In this situation a Type II error would happen if the mean systolic blood pressure of the population of employees is greater than 122 mmHg, but the null hypothesis that it is 122 mmHg is not rejected. In other words a Type II error would happen if the mean blood pressure for the population of employees is higher than the national average, but the test does not conclude that it is higher.

Part (b):

The test is one-sided and the standard deviation is known, so the null hypothesis will be rejected if the test

statistic $z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > 1.645$. With $\mu = 122$, $\sigma = 15$, and $n = 100$, we get $\frac{\bar{x} - 122}{1.5} > 1.645$. Therefore,

$$\bar{x} > 124.4675.$$

Part (c):

If the actual population mean is 125, with $\sigma = 15$ and $n = 100$, then the sampling distribution of $\bar{x}$ is

approximately normal with mean of 125 and standard deviation $\frac{15}{\sqrt{100}} = 1.5$. Therefore,

$$P(\bar{x} > 124.4675) = P\left(\frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} > \frac{124.4675 - 125}{1.5}\right) = P(z > -0.355) = 0.64.$$

Part (d):

The probability found in part (c) is called the power of the test.

Part (e):

If the sample size is increased from 100 to something larger, the probability of rejecting the null hypothesis when the population mean is 125 will be higher than it is for a sample of size 100. Intuitively, more data provide a higher probability of a correct conclusion. The technical explanation is that the rejection region will still be $z > 1.645$, but the sampling distributions of the sample mean will have a smaller standard deviation; therefore, the minimum value of $\bar{x}$ for which we would reject the null hypothesis would be lower and, in return, the probability the null hypothesis is rejected will increase.

2018 SCORING GUIDELINES

Question 6 (continued)

Scoring

This question is scored in three sections. Section 1 consists of part (a) and part (b), section 2 consists of part (c), and section 3 consists of part (d) and part (e). Sections 1, 2, and 3 are scored as essentially correct (E), partially correct (P), or incorrect (I).

Section 1 is scored as follows:

Essentially correct (E) if the response satisfies the following four components:

1. Part (a) includes in the description of a Type II error the fact that the alternative hypothesis is true, either generically or in context of the situation.
2. Part (a) includes in the description of a Type II error the fact that the null hypothesis is not rejected, either generically or in context of the situation.
3. Part (b) includes a correct z-score for the upper 5 percent tail and indicates the correct direction for the rejection region.
4. Part (b) includes $\mu_{\bar{x}} = 122$, $\sigma_{\bar{x}} = 1.5$, and the resulting $\bar{x}$ value.

Partially correct (P) if the response satisfies only two or three of the four components.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- If the response in part (a) does not include context, the number of components satisfied is reduced by one (that is, from four to three, or from three to two, and so on). Context includes a reference to units, to blood pressure, to employees, etc.
- If a response in part (a) is clearly referring to an individual's blood pressure as opposed to the mean blood pressure of all employees, neither components 1 nor 2 are satisfied.

Section 2 is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

1. Recognizes that the null hypothesis will be rejected when $\bar{x} \geq 124.4675$, as found in part (b).
2. Provides the correct sampling distribution for the sample mean when the true mean is 125, including correct values for the mean and standard deviation, either explicitly or by plugging them into the test statistic formula.
3. Provides evidence of using the normal curve and finds the correct probability value.

Partially correct (P) if the response satisfies only two of the three components.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- Components 1 and 3 can still be satisfied if errors made in finding the rejection region in part (b) are carried into part (c).
- A calculator statement that does not include labels for input values does not satisfy component 2 but may still satisfy components 1 and 3.

2018 SCORING GUIDELINES

Question 6 (continued)

Section 3 is scored as follows:

Essentially correct (E) if the response satisfies the following four components:

1. Part (d) specifies power as the name of the probability.
2. Part (e) correctly states that the probability would be greater.
3. Part (e) correctly implies that the standard deviation of the sampling distribution decreases, either explicitly or by substituting values into a formula.
4. Part (e) indicates the minimum value of $\bar{x}$ for which the null hypothesis is rejected decreases, either explicitly or by substituting values into a formula.

Note: Component 4 can still be satisfied if a response indicates that a maximum value of $\bar{x}$ for which the null hypothesis is rejected increases if this direction is consistent with answers in parts (b) and (c).

Partially correct (P) if the response satisfies only two or three of the four components.

Incorrect (I) if the response does not meet the criteria for E or P.

2018 SCORING GUIDELINES

Question 6 (continued)

4 Complete Response

Three sections essentially correct

3 Substantial Response

Two sections essentially correct and one section partially correct

2 Developing Response

Two sections essentially correct and no sections partially correct

OR

One section essentially correct and one or two sections partially correct

OR

Three sections partially correct

1 Minimal Response

One section essentially correct

OR

No sections essentially correct and two sections partially correct

OR

Sections 1 and 2 incorrect, and section 3 partially correct with exactly three of the four components satisfied

OR

Section 1 partially correct with exactly three of the four components satisfied, and sections 2 and 3 incorrect