2023

AP® Statistics

Scoring Guidelines

Question 1: Focus on Exploring 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)	The histogram of dissolved oxygen concentration in Alaskan streams with water temperatures colder than 8°C is unimodal and skewed left with a median between 11 and 12 mg/l. The first quartile is in the bin from 10-11 mg/l and the third quartile is in the bin from 12-13 mg/l, so the IQR is approximately 2 mg/l. There do not appear to be any high outliers, but there are several potential low outliers because the values in the 2-3, 4-5, and 5-6 bins are all certainly more than 1.5 IQR below the first quartile.	**Essentially correct (E)** if the description of the distribution satisfies component 1 *AND* at least three of components 2-5: 1. Includes context of dissolved oxygen concentration 2. Shape: The distribution is skewed left 3. Center: The center of the distribution is between 11 mg/l and 12 mg/l 4. Spread: Refers to at least one measure of variation (i.e., the range is between 14 – 2 = 12 mg/l and 13 – 3 = 10 mg/l; all values are between 2 mg/l and 14 mg/l or the IQR is approximately 2 mg/l) 5. Unusual features: potential outliers or a gap between 3 mg/l and 4 mg/l **Partially correct (P)** if the response satisfies component 1 and two components out of components 2-5 *OR* if the response satisfies at least three out of components 2-5 but does not satisfy component 1. **Incorrect (I)** if the response does not meet the criteria for E or P.

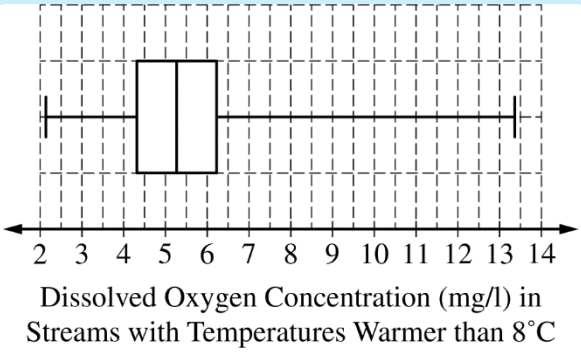
Additional Notes:

- **Context**
  - Component 1 can be satisfied with a reference to dissolved oxygen concentration, dissolved oxygen, amount of oxygen, or mg/l.

- **Shape:**
  - Component 2 cannot be satisfied if a response describes the histogram as normal or approximately normal.
- **Center:**
  - A response that addresses center using definitive language such as "the mean (median) of the distribution is 11.5" cannot satisfy component 3.
  - A response that addresses center using approximate language such as "the median of the distribution is approximately 11.5" must, for any single measure of center, specify a numeric value that is between 11 mg/l and 12 mg/l, inclusive, to satisfy component 3.
- **Spread:**
  - A response recognizing all values in the sample fall between 2 mg/l and 14 mg/l, satisfies component 4 only for these exact endpoints and need not appeal to a specific measure of spread such as the range.
  - A response that uses interval language must use it correctly. For example, "the observations range from 2 mg/l and 14 mg/l," satisfies component 4 because it correctly indicates that all observations are between 2 mg/l and 14 mg/l, inclusive. However, a statement such as "the range is between 2 mg/l and 14 mg/l," is incorrect because the range is a single number, i.e., $14 - 2 = 12$ mg/l.
  - A response that appeals to a specific measure of spread using approximate language, such as "the IQR is approximately 2," must specify a numeric value within the bounds appropriate to that measure of spread shown in the following table.

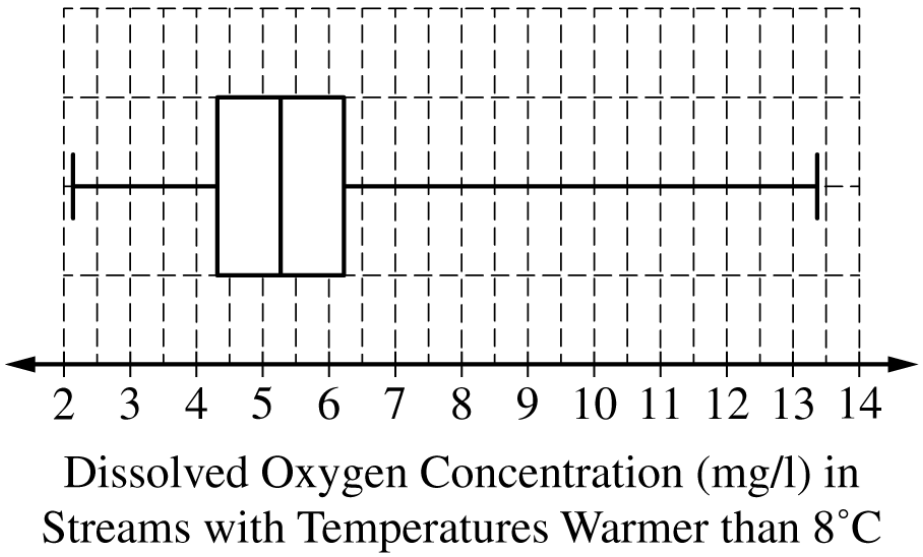
Statistic	Min (mg/l)	Max (mg/l)
Range	10	12
IQR	1.5	2.5
Standard Deviation	1.7	1.8

- **Unusual Features:**
  - Component 5 cannot be satisfied if a response indicates that the distribution has an unusual feature other than potential outliers or a gap.
  - Definitive language such as "there is an outlier" does not satisfy component 5.

Model Solution	Scoring
(b)  Dissolved Oxygen Concentration (mg/l) in Streams with Temperatures Warmer than 8˚C	**Essentially correct (E)** if the response satisfies four or five of the following five components: 1. Constructs the box beginning at Q1 2. Constructs the box ending at Q3 3. Locates the median within the box 4. Extends the lines on the boxplot to the minimum value 5. Extends the lines on the boxplot to the maximum value **Partially correct (P)** if the response includes a boxplot with only three of the five components. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- A response that shows a graph or plot other than a boxplot should be scored incorrect (I).
- To be correctly placed:
  - The minimum must be between the tick marks for 2 and 2.5
  - The first quartile must be between the tick marks for 4 and 4.5
  - The median must be between the tick marks for 5 and 5.5
  - The third quartile must be between the tick marks for 6 and 6.5
  - The maximum must be between the tick marks for 13 and 13.5



	Model Solution	Scoring
(c)	If the researchers' belief is correct, then streams with water temperature colder than 8°C are healthier for wildlife. The distribution of dissolved oxygen concentration for colder streams has a higher center because its median (between 11 mg/l and 12 mg/l) is larger than the median for warmer streams (5.43 mg/l). The shape of the distribution of dissolved oxygen concentration for colder streams is different from the shape of the distribution for warmer streams. The distribution of values of dissolved oxygen concentration for colder streams is skewed to the left but the distribution of values for warmer streams is skewed to the right. Both distributions have a similar spread because they both have similar IQR values — approximately 2 mg/l for the colder streams and 1.73 mg/l for the warmer streams.	**Essentially correct (E)** if the response satisfies component 1 *AND* at least two of components 2-4: 1. States colder streams are healthier for wildlife 2. Directly compares the centers of the two distributions 3. Indicates warmer streams are skewed right and colder streams are skewed left 4. Directly compares the spreads **Partially correct (P)** if the response satisfies component 1 *AND* one of components 2-4. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- While both distributions have potential outliers, it is not necessary to compare them for this part of the question.
- A response based on an incorrect part (a) or part (b) may be scored E or P if it justifies the choice consistent with their answers to part (a) or part (b).
- In order to satisfy component 4, a response may compare at least one of the values of the approximate range, IQR, or standard deviation with the value of the same statistic from the other distribution.

Scoring for Question 1	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 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: 1. Identifies the experimental units as the driveways 2. Identifies the treatments as concrete with fibers and concrete without fibers 3. 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: 1. 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 2. 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 3: Focus on Probability and Sampling Distributions — 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)	The random variable X is the dollar value of the cash prize in a bath fizzy. **(i)** The proportion of bath fizzies containing \$1 is equal to the $P(X = \$1)$ and $P(X = \$1)$ $= 1 - (0.2 + 0.05 + 0.05 + 0.01 + 0.01)$ $= 0.68$. **(ii)** The proportion of bath fizzies that contain at least \$10 is equal to the $P(X \geq \$10)$ and $P(X \geq \$10)$ $= 0.05 + 0.05 + 0.01 + 0.01$ $= 0.12$.	**Essentially correct (E)** if the response satisfies at least three of the following four components: 1. Correctly calculates the proportion of bath fizzies containing \$1 2. Provides correct supporting work for the value calculated in component 1 3. Correctly calculates the proportion of bath fizzies containing at least \$10 4. Provides correct supporting work for the value calculated in component 3 **Partially correct (P)** if the response satisfies only two of the four components. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- A response that provides a correct percentage, instead of a proportion, may satisfy components 1 and 3, such as 68 percent or 12 percent.
- An arithmetic or transcription error in a response can be ignored if correct work is shown.

	Model Solution	Scoring
(b)	Given a bath fizzy contains at least \$10, then the probability that it contains \$100 is $P(X = \$100 \mid X \geq \$10)$ $= \frac{0.01}{0.12}$ $\approx 0.0833.$	**Essentially correct (E)** if the response satisfies the following two components: 1. Correctly calculates the requested probability 2. Shows work consistent with their response to part (a-ii) **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 specific conditional probability statement is not required, but if correctly given should be considered a positive in holistic scoring.
- An arithmetic or transcription error in a response can be ignored if correct work is shown.

	Model Solution	Scoring
(c)	The expected value of the distribution of X is $E(X) = 1(0.68) + 5(0.2) + 10(0.05)$ $+ 20(0.05) + 50(0.01) + 100(0.01)$ $= \$4.68$. The expected value is the mean of the cash prizes that result from the long run of many, many trials of randomly selecting bath fizzies and determining the amount each contains.	**Essentially correct (E)** if the response satisfies both components 1 and 2 *AND* at least two of components 3–5: 1. States the correct expected value 2. Shows appropriate work to calculate the expected value 3. Interpretation includes the concept of repeating the selection process over a long period of time 4. Interpretation includes the concept of an average or mean 5. Interpretation includes the context of receiving a cash prize **Partially correct (P)** if the response does not meet the criteria for E but satisfies two or three of components 1–4. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- Supporting work for finding the expected value must include at least two of the terms in the equation to show the pattern, such as $1(0.68) + 5(0.2) + \ldots$.
- Calculator notation does not satisfy component 2, such as 1 – VAR STATS(L1, L2).
- Responses that satisfy only components 1 and 5 or only components 2 and 5 receive a score of I but should be considered a positive in holistic scoring.
- An arithmetic or transcription error in a response can be ignored if correct work is shown.

	Model Solution	Scoring
(d)	The expected value of the distribution of X in euros is $4.68(0.89) \approx 4.17$ euros.	**Essentially correct (E)** if the response correctly calculates the number of euros for the expected value showing work and including units. **Partially correct (P)** if the response correctly calculates the number of euros for the expected value but is missing either work or units. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- A response that does not have at least one decimal place in the final response (e.g., rounded to an integer) should be scored no more than P.
- The response can either use the expected value calculated in part (c) or first convert all the values in the probability distribution to euros to find the new expected value.
- An arithmetic or transcription error in a response can be ignored if correct work is shown.

Scoring for Question 3 Each essentially correct (E) part counts as 1 point, and each partially correct (P) part counts as ½ point.	**Score**
Complete Response	**4**
Substantial Response	**3**
Developing Response	**2**
Minimal Response	**1**
If a response is between two scores (for example, 2 ½ 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 4: Focus on Inference — 4 points

General Scoring Notes

- This question is scored in three sections. Each section is initially scored by determining if it meets the criteria for essentially correct (E), partially correct (P), or incorrect (I). The first section includes statements of the null and alternative hypotheses and identification of the appropriate hypothesis test. The second section includes verifying the conditions for the test identified in the first section and calculating the value of the test statistic and the corresponding *p*-value. The third section includes the conclusion for the test identified in the first section. The response is then categorized based on the scores assigned to each section 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 section 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
Section 1	Let μ_d represent the true mean difference (placebo minus omega-3) of irritability scores for all people with this medical condition. The null hypothesis is $H_0: \mu_d = 0$ and the alternative hypothesis is $H_a: \mu_d > 0$. The appropriate inference procedure is a matched pairs *t*-test for a mean difference.	**Essentially correct (E)** if the response satisfies the following four components: 1. Identifies a paired *t*-test for a population mean difference by name or by formula 2. States the hypotheses using a single mean (e.g., μ_d, μ) 3. States the correct equality for the null hypothesis (e.g., $H_0: \mu_d = 0$) AND states the correct direction for the alternative hypothesis (e.g., $H_a: \mu_d > 0$) 4. Provides sufficient context for the parameter, by including reference to the *population* mean difference *AND* the sampling units (people with the medical condition) *AND* the response variable (irritability score) **Partially correct (P)** if the response satisfies three of the four components. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- A response that states the null hypothesis as $H_0: \mu_d \leq 0$ may satisfy component 3.
- A response that states the name of the procedure as *t*-test, or one-sample *t*-test may satisfy component 1.
- To satisfy component 2, the hypotheses must be stated in terms of a mean. If a symbol other than μ or $\overline{X}$ is used to denote the mean, it must be clearly defined as a mean (but does not need to reflect the context of irritability score). It is acceptable to use μ_0 to denote the mean.

- A response that states the hypotheses in words (e.g., "the null hypothesis is that the mean difference is 0, and the alternative hypothesis is that the mean difference is greater than 0") may satisfy components 2 and 3. Neither context nor the concept of the *population* is required to satisfy component 2 or 3.
- A response that states the hypotheses in words [e.g., "the null hypothesis is that the mean difference in irritability score (placebo minus omega-3) for all people with this medical condition is equal to 0 and the alternative hypothesis is that the mean is greater than 0"] may satisfy components 2,3, and 4.
- The elements of component 4 do not have to be satisfied with the statement of the hypotheses. They may be satisfied by work presented anywhere in the response, most likely by the statement of the conclusion.
- If the statement of the hypotheses refers to population mean and the conclusion refers to sample mean (or vice versa), then the population aspect of component 4 is not satisfied.
- If the response clearly refers to the *sample* mean instead of the *population* mean using words or a symbol (e.g., $\bar{x}$), then component 4 is not satisfied unless the symbol used is defined as the *population* mean.
- A response may satisfy the population aspect of component 4 by doing the following:
  - referring to population in the statement of the conclusion of the inferential procedure.
  - using notation such as μ when defining the hypothesis statements.
- A response may satisfy the sampling units aspect of component 4 by referring to "patients" or "people with this medical condition" or a similar statement.
- If the response identifies the correct test by name, but also states an incorrect formula, then component 1 is not satisfied.
- If the response identifies the test by formula using a *z*-percentile instead of a *t*-percentile, then component 1 is not satisfied.

Confidence Interval Approach:

- If a one-sample *t*-interval for a population mean is identified correctly by name (e.g., "one-mean *t*-interval" or "one-sample *t*-interval" or "*t*-interval") or by formula, then component 1 is satisfied.
- If a response uses a one-sample *t*-interval for a population mean, then components 2 and 3 are satisfied if the response indicates that it is a confidence interval for the mean difference in irritability score (placebo minus omega-3).

Two-Sample Approach:

- A response that clearly defines the *population* difference in means with clearly defined parameters for the means of placebo and omega-3 *AND* the sampling units (people with the medical condition or people similar to those in the study) *AND* the response variable (irritability score) may earn credit for component 4.

	Model Solution	Scoring
Section 2	The independence condition for performing a paired *t*-test for a mean difference is satisfied because the data were obtained from a randomized experiment where the week in which the patient received the treatment was randomly assigned. The sampling distribution of the mean difference must be approximately normal. Although the sample size is less than 30 $(n = 19)$, this is satisfied because the boxplot for the sample differences shows an approximately symmetric distribution with no outliers. The value of the test statistic is: $t = \frac{\bar{x}_d - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{1.789 - 0}{\frac{2.485}{\sqrt{19}}} \approx 3.138$ Using 18 degrees of freedom, the corresponding *p*-value is $P(t > 3.138) \approx 0.0028$.	**Essentially correct (E)** if the response satisfies the following four components: 1. Checks the independence condition by referring to the random assignment 2. Indicates that the distribution of irritability scores for the differences (placebo minus omega-3) is not badly skewed supports the assumption that the sampling distribution of $\bar{x}_d$ is approximately normal 3. Correctly reports the value of the *t*-statistic consistent with the named test 4. Correctly reports the *p*-value, consistent with the stated alternative hypothesis and reported test statistic **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.

Additional Notes:

- Component 1 is not satisfied if the response indicates that the independence condition is met because the sample was selected at random from all patients with the medical condition.
- Component 1 is satisfied if the response indicates that the treatment was randomly assigned even if the terms “independent” or “independence” are not used.
- Discussion of degrees of freedom will be treated as extraneous in scoring for section 2.
- If the response incorrectly identifies the test as a *z*-test in section 1, the correct *z*-statistic of 3.138 satisfies component 3 and a *p*-value of 0.00085 satisfies component 4. However, if the test statistic is not reported, component 3 is not satisfied, but if the correct *p*-value (consistent with the alternative hypothesis) is reported then component 4 is satisfied.
- If the response does not identify a *z*-test in section 1 but reports the test statistic as $z = 3.138$ (instead of *t*) then component 3 is not satisfied; however, component 4 may still be satisfied if the correct *p*-value of 0.00085 is reported with a *z*-statistic of 3.138.
- If the response compares the value of the test statistic to a critical value instead of computing a *p*-value, then a comparison consistent with the stated alternative hypothesis, satisfies component 4.
- If a two-tailed alternative hypothesis is stated, then the *p*-value must be consistent with the stated alternative hypothesis to satisfy component 4.
- A response that reports the correct value for the *t*-statistic but contains errors in supporting work may still satisfy component 3.
- If the response satisfies component 4, any supporting work for the *p*-value may be treated as extraneous.

- If no test statistic is reported, and the *p*-value is consistent with the stated alternative hypothesis or equal to 0.0028 component 4 is satisfied.

Confidence Interval Approach:

- If the stated alternative hypothesis is correct or no alternative hypothesis is provided:
  - If either a one-sided 95 percent confidence interval for μ_d is correctly calculated as $(0.8004, \infty)$ or a two-sided 90 percent confidence interval for μ_d is correctly calculated as $(0.8004, 2.7776)$ then component 3 is satisfied.
  - If only the lower end of the confidence interval for μ_d is used to reach a conclusion, then component 4 is satisfied.
- If the stated alternative hypothesis is incorrect (two-sided or reversed direction), the confidence interval approach must be consistent with the stated alternative to satisfy components 3 and 4:
  - An interval consistent with the stated two-sided alternative will satisfy component 3. A two-sided 95 percent confidence interval for μ_d is $(0.591, 2.987)$. If the two-sided confidence interval is correctly interpreted based on whether zero is in the interval, then component 4 is satisfied.
  - An interval consistent with the incorrect reversed alternative will satisfy component 3. If either a two-sided 90 percent confidence interval for μ_d is correctly calculated as $(0.8004, 2.7776)$, or a one-sided 95 percent confidence interval for μ_d is correctly calculated as $(-\infty, 2.7776)$, then component 3 is satisfied. If the two-sided confidence interval is correctly interpreted based on whether zero is in the interval, then component 4 is satisfied; if only the upper end of the lower one-sided confidence interval is used to reach a conclusion, then component 4 is satisfied.
- If the difference calculated is μ_d, omega-3 minus placebo, the intervals will be the negative of those provided above.

Two-Sample Approach:

- A response that indicates that the distribution of irritability scores for each of the two samples, placebo *AND* omega-3, are not badly skewed may satisfy component 2.
- If the response correctly reports the value of the *t*-statistic as 2.256, then the response may satisfy component 3.
- Values for the *p*-value depend on how the degrees of freedom were determined. The following *p*-values all assume the correct alternative hypothesis and satisfy component 4:
  - The degrees of freedom for $t \approx 2.256$ on a two-sample solution computed with the Satterthwaite approximation used by many calculators is:

$$df = \frac{\left(\frac{s_{placebo}^2}{n_{placebo}} + \frac{s_{omega-3}^2}{n_{omega-3}}\right)^2}{\frac{1}{n_{placebo}-1}\left(\frac{s_{placebo}^2}{n_{placebo}}\right)^2 + \frac{1}{n_{omega-3}-1}\left(\frac{s_{omega-3}^2}{n_{omega-3}}\right)^2} = \frac{\left(\frac{(2.987)^2}{19} + \frac{(1.739)^2}{19}\right)^2}{\frac{1}{19-1}\left(\frac{(2.987)^2}{19}\right)^2 + \frac{1}{19-1}\left(\frac{(1.739)^2}{19}\right)^2}$$

$$\approx 28.94,$$

resulting in a *p*-value of 0.0159 or a *p*-value of 0.0318 if a two-sided alternative hypothesis was stated.

  - Using $t \approx 2.256$ and $(19 + 19 - 2) = 36$ degrees of freedom, the resulting *p*-value is 0.0151 or a *p*-value of 0.0302 if a two-sided alternative hypothesis was stated.

- Using $t \approx 2.256$ and $(19 - 1) = 18$ degrees of freedom, the resulting *p*-value is 0.0184 or a *p*-value of 0.0368 if a two-sided alternative hypothesis was stated.

- If no test statistic is reported, and the *p*-value is consistent with the stated alternative hypothesis or equal to 0.0159 or 0.0151 or 0.0184, component 4 is satisfied.

	Model Solution	Scoring
Section 3	Because the p-value ≈ 0.0028 is less than the significance level, $\alpha = 0.05$, the null hypothesis should be rejected. The data provide convincing statistical evidence that for patients similar to those in the study, the true mean difference (placebo minus omega-3) in irritability scores for people with this medical condition is greater than zero. This suggests the omega-3 fatty acids are helpful in reducing irritability scores in people with this medical condition.	**Essentially correct (E)** if the response satisfies the following two components: 1. Provides correct comparison of the p-value to alpha (p-value is less than/greater than alpha) *AND* provides a correct decision about the null and/or alternative hypothesis 2. States a conclusion in context, consistent with, and in terms of, the alternative hypothesis using non-deterministic language **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:

- In order to satisfy component 1, the response must clearly identify the number that is compared to alpha as a p-value.
- If the response provides an unreasonable p-value (that is clearly identified as the p-value) and correctly compares it to alpha, component 1 may be satisfied.
- To satisfy the p-value comparison in component 1, the response can compare the value of the test statistic to an appropriate critical value, e.g., $t_\alpha = 1.734,$ and $3.13 > 1.734$.
- An explicit decision is not required to satisfy component 1.
- If an explicit decision is stated and the conclusion is inconsistent with the decision, component 1 is not satisfied.
- The decision part of component 1 may be satisfied by implying the decision within the conclusion statement (sufficient evidence/insufficient evidence for the alternative hypothesis).
- To satisfy the context in component 2, the response must include the population parameter, sampling/experimental units, and the response variable.
- If the response omits hypotheses, assume the correct alternative hypothesis is provided when scoring component 2.
- If the response states incorrect hypotheses, component 2 may be satisfied by either stating a conclusion in terms of the stated alternative hypothesis or by answering the inference question.
- Examples of non-deterministic language in component 2 include "evidence to accept the alternative," "there is evidence for the alternative," "there is not sufficient evidence for the alternative."
- Examples of deterministic language in component 2 include "proves the null," "proves the alternative," "accepts the alternative," "there is not evidence for the alternative," and "no evidence for the alternative."
- If the comparison and decision are consistent with an incorrect p-value (or an incorrect value of the test statistic, or an incorrect confidence interval), the response may satisfy component 1.
- If components 1 and/or 2 are satisfied and the response provides an incorrect interpretation of the p-value, the score is lowered from E to P or P to I.
- In section 1, if component 4 is not satisfied only because of the omission of sampling units (patients), in section 3 component 2 may be satisfied by an appropriate conclusion that is only missing sampling units.

Confidence Interval Approach:

- Component 2 should be scored according to the rubric and component 1 should be scored with regard to a comparison of zero to the appropriate end of the reported confidence interval.
- If no alternative hypothesis is specified in the response, then assume the correct alternative hypothesis is provided when scoring component 2.
- If an incorrect two-sided alternative hypothesis is specified, then component 2 is satisfied if the justification is based on whether zero is included in the confidence interval.
- If the response includes an incorrect interpretation of the confidence interval, then the score for section 3 is lowered from E to P or from P to I.

Two-Sample Approach:

- Components should be scored according to the rubric using the reported *p*-value, or by comparing the *t*-statistic to an appropriate critical value.

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

Question 5: Multi-Focus — 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 section 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)	The scatterplot reveals a strong, positive, roughly linear association between the chest circumference and weight of tule elk. There are no points that seriously deviate from the straight-line pattern of the points in the plot.	**Essentially correct (E)** if the response provides a description that includes at least three of components 1-4 and component 5: 1. Direction of association (positive or increasing) 2. Strength of association (strong) 3. Form of association (linear or approximately linear) 4. Unusual features (no points with large discrepancies from the straight-line pattern exhibited by most of the points on the plot) 5. Context (association between chest circumference and weight of tule elk) **Partially correct (P)** if the response satisfies only one or two components out of components 1-4 and component 5 *OR* if the response satisfies at least two out of components 1-4 but does not satisfy component 5. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- To satisfy component 4 it is sufficient to simply indicate that there are no unusual features or no outliers.
- To satisfy component 5 it is minimally sufficient for the response to refer to the association or relationship between chest circumference or measurement and weight without explicitly mentioning tule elk.
- The quality of communication in part (a) should be considered if holistic scoring is required.

Model Solution	Scoring
(b) **(i)** The predicted weight of a male tule elk with a chest circumference of 145.9 cm is $-350.3 + 3.7455(145.9) \approx 196.17$ kg. **(ii)** The residual for a male tule elk with a chest circumference of 145.9 cm with an actual weight of 204.3 kg is $204.3 - 196.17 \approx 8.13$ kg.	**Essentially correct (E)** if the response satisfies the following two components: 1. Provides the correct value of 196.17 kg in part (b-i) with work shown 2. Provides the correct value of 8.13 kg in part (b-ii), or a value consistent with the predicted weight calculated for component 1, with work shown **Partially correct (P)** if the response satisfies only one of the two components required for an E *OR* the response gives both correct values with no work shown. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- An arithmetic or transcription error in a response can be ignored if correct work is shown.
- A response that includes an undefined or incorrect variable to identify work shown (e.g., x or $\hat{p}$) does not satisfy component 1.

	Model Solution	Scoring
(c)	The value of the slope of the least-squares regression line is 3.7455. This value indicates that the predicted weight of a tule elk increases by 3.7455 kilograms for each additional centimeter of chest circumference.	**Essentially correct (E)** if the response satisfies the following three components: 1. Identifies the value of the slope as 3.7455 2. Provides an interpretation that references an increase of a number of kilograms of weight for each one-centimeter increase in chest circumference 3. Indicates that the slope represents a change in a prediction using non-deterministic language such as "predicted," "estimated," "expected," or "average" **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 value of the slope, 3.7455, may be rounded to 3.746, 3.75, or 3.7, but not to 3 or 4, to satisfy the numerical requirement in component 1.
- A response that only contains 3.7455 in the interpretation satisfies component 1.
- A calculation of slope may satisfy component 1 provided that two points from the line are used in the calculation.
- Units of measurements must be correctly specified for both weight and length to satisfy component 2.
- It is not required to refer specifically to the "least-squares regression line."

Model Solution	Scoring
(d) **(i)** The degrees of freedom for the test of slope are $n - 2 = 30 - 2 = 28$. The t-table shows that for 28 degrees of freedom, the p-value for a one-sided test would be 0.001. Because this is a two-sided test, the p-value is $(2)(0.001) = 0.002$. **(ii)** Because the p-value $= 0.002$ is less than $\alpha = 0.05$, reject the null hypothesis. There is sufficient statistical evidence that the population slope for the linear regression of weight vs. chest circumference for male tule elk is different from 4.5 kg/cm.	**Essentially correct (E)** if the response satisfies the following three components: 1. Gives a correct p-value of 0.002, consistent with the two-sided alternative hypothesis 2. Provides correct comparison of the p-value to alpha (p-value is less than/greater than alpha) *AND* provides a correct decision about the null and/or alternative hypothesis 3. Provides a conclusion statement in context, consistent with, and in terms of, the alternative hypothesis using non-deterministic language **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 response need not make an explicit decision about the null hypothesis (reject H_o or fail to reject H_o).
- If the conclusion and justification are consistent with an incorrect p-value (or an incorrect value of the test statistic, or an incorrect confidence interval, or an incorrect critical value), the response may satisfy component 2.
- To satisfy the p-value comparison in component 2, the response can compare the value of the test statistic to the appropriate critical value (e.g., 2.048 or -2.048).
- If an explicit decision is stated and the conclusion statement is inconsistent with the decision, component 2 is not satisfied.
- The decision part of component 2 may be satisfied by implying the decision within the conclusion statement (sufficient evidence/insufficient evidence for the alternative hypothesis).
- If the response includes a statement that is equivalent to accepting the null hypothesis (e.g., "we conclude that the mean population slope is 4.5"), then component 2 is not satisfied.
- A student who doesn't give a p-value, but gives a correct conclusion based on a p-value $< \alpha$ with a consistent conclusion *OR* a p-value $> \alpha$ with a consistent conclusion may be scored P, e.g.,
  - If p were less than alpha, then reject H_0 and say there is sufficient evidence that the population slope for the linear regression of weight vs. chest circumference for male tule elk is different than 4.5 kg/cm.
  - If p were greater than alpha, then fail to reject H_0 and say there is not sufficient evidence that the population slope for the linear regression of weight vs. chest circumference for male tule elk is different than 4.5 kg/cm.
- If the response includes an incorrect interpretation of the p-value (e.g., "this is the probability that the null hypothesis is true"), then component 2 is not satisfied.
- The quality of communication in part (d) should be considered if holistic scoring is required.

Scoring for Question 5 Each essentially correct (E) part counts as 1 point, and each partially correct (P) part counts as ½ point.	**Score**
Complete Response	**4**
Substantial Response	**3**
Developing Response	**2**
Minimal Response	**1**
If a response is between two scores (for example, 2 ½ 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 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)	Let X represent the amount of gold applied to a necklace randomly selected from necklaces produced with this machine. The random variable X has an approximately normal distribution with mean 300 mg and standard deviation 5 mg. Then, $P(296 < X < 304)$ $= P(X < 304) - P(X \le 296)$ $= P\left(Z < \frac{304 - 300}{5}\right) - P\left(Z \le \frac{296 - 300}{5}\right)$ $= P(Z < 0.8) - P(Z \le -0.8)$ $\approx 0.7881 - 0.2119 \approx 0.5763.$	**Essentially correct (E)** if the response satisfies the following three components: 1. Indicates the use of a normal distribution with mean 300 and standard deviation 5 2. Specifies the correct event (boundary value and direction) or an event consistent with values reported in component 1 3. Provides the correct probability of 0.5763 or a probability consistent with components 1 and 2 **Partially correct (P)** if the response satisfies only two of the three components. **Incorrect (I)** if the response does not satisfy the criteria for E or P.

Additional Notes:

Component 1

- A response may satisfy component 1 by any of the following or a combination of the following:
  - Graphical: Displaying a graph of a normal density function with the appropriate scale on the horizontal axis showing the mean and the standard deviation for the distribution of the amount of gold applied to the necklaces.
  - Calculator function syntax: Labeling correct values of the mean and standard deviation in a "normalcdf" statement, such as:
    normalcdf(lower = 296, upper = 304, mean = 300, standard deviation = 5).
    Correct specification of the lower and upper bounds is not required to satisfy component 1.
  - Words: Using a statement such as "normal distribution with mean 300 and standard deviation 5."
  - Standard Notation: Using standard notation such as $N(300,5)$ or $N(300,(5)^2)$.
  - Z-score: Displaying the correct mean and standard deviation in a z-score calculation that includes "z," such as $z = \frac{304 - 300}{5}$.

Component 2

- A response may satisfy component 2 by any of the following or a combination of the following:
  - Graphical: Displaying a graph of a normal density function with the region of interest ($296 <$ amount of gold < 304 or $-0.8 < Z < 0.8$) clearly identified. The shaded area does not need to be proportional, but the boundaries should be on the proper side of the mean and the shading should be in the proper direction.
  - Calculator function syntax: Labeling the lower and upper bounds of the region of interest in a "normalcdf" statement, such as:
    - normalcdf(lower = 296, upper = 304, mean = 300, standard deviation = 5).
    - normalcdf(lower = −0.8, upper = 0.8, $\mu = 0$, $\sigma = 1$).
  - Words: Specifying the correct event in words with correct numerical value for the boundary value and correct direction, such as "the probability that the amount of gold applied to a randomly selected necklace is between 296 and 304 mg" or $P(296 <$ amount of gold $< 304)$.
  - Standard Notation: Using standard notation such as $P(296 < X < 304)$ or $P\left(\frac{296 - 300}{5} < Z < \frac{304 - 300}{5}\right)$ or $P(-0.8 < Z < 0.8)$.

General

- Minor errors in statistical notation may be ignored. However, this is considered poor communication if holistic scoring is required.
- If the only error in the response to part (a) is the reversal of the numerator for the *z*-score, $(300 - 296)$ or $(300 - 304)$, the response is scored P.
- An arithmetic or transcription error in a response can be ignored if correct work is shown.
- A response that satisfies components 1 and 2 by including more than calculator syntax (e.g., standard random variable notation and probability notation) is considered good communication and may be considered if holistic scoring is required.

Model Solution	Scoring
(b) **(i)** If the machine is working properly, and the sample mean amount of gold, $\bar{X}$, has a sampling distribution that follows a normal distribution with mean 300 mg and standard deviation $\frac{5}{\sqrt{2}} \approx 3.5355$ mg, then $P(\bar{X} > 303)$ $= P\left(Z > \frac{303 - 300}{\frac{5}{\sqrt{2}}}\right)$ $\approx P(Z > 0.8485)$ ≈ 0.198. **(ii)** Observing a sample mean amount of 303 mg would not provide convincing evidence that the population mean amount of gold being applied by the machine is something other than 300 mg because the probability of observing a sample mean that differs from 300 mg by 3 mg or more is large, around $0.198(2) = 0.396$.	**Essentially correct (E)** if the response satisfies components 1, 4, and 5 *AND* at least one of components 2 or 3: 1. In part (b-i), indicates the use of a normal (or an approximately normal) distribution and identifies the correct parameter values (mean 300 mg and standard deviation $\frac{5}{\sqrt{2}} \approx 3.5355$ mg) with work shown or a correct formula 2. In part (b-i), specifies the correct event (boundary value and direction) 3. In part (b-i), provides the correct probability of 0.198 or the probability consistent with components 1 and 2 4. In part (b-ii), indicates whether observing a sample mean of 303 mg would provide convincing evidence that the population mean is not 300 mg, consistent with the response to part (b-i) 5. In part (b-ii), provides justification based on the probability obtained in part (b-i) or a correctly computed probability in part (b-ii) **Partially correct (P)** if the response does not meet the requirements for E and satisfies at least three of the five components *OR* the response satisfies only components 4 and 5, consistent with the response to part (b-i) *OR* the response satisfies only component 1 and one of the other four components. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

Component 1

- A response may satisfy component 1 by any of the following or a combination of the following:
  - Graphical: Displaying a graph of a normal density function with the appropriate scale on the horizontal axis showing mean and standard deviation for the distribution of mean amount of gold.
  - Calculator function syntax: Labeling correct values of the mean and standard deviation in a "normalcdf" statement, such as "normalcdf" or "ncdf", e.g.,
    - normalcdf$\left(\text{lower} = 303, \text{upper} = \infty, \text{mean} = 300, \text{standard deviation} = \frac{5}{\sqrt{2}}\right)$.

    Correct specification of the upper and lower bounds is not required to satisfy component 1.
  - Words**:** Using a statement such as "approximately normal distribution with mean 300 and standard deviation $\frac{5}{\sqrt{2}}$."
  - Standard Notation: Using standard notation such as $N\left(300, \frac{5}{\sqrt{2}}\right)$ or $N\left(300, \left(\frac{5}{\sqrt{2}}\right)^2\right)$.
  - *Z*-score: Displaying the correct mean and standard deviation in a *z*-score calculation, which includes "*z*," such as $z = \frac{303 - 300}{\frac{5}{\sqrt{2}}}$.
  - To satisfy the supporting work criterion of component 1 the response must clearly indicate that the standard deviation for $\overline{X}$ is computed by dividing 5 by the square root of 2. This may be indicated with $\frac{5}{\sqrt{2}}$, words, or standard notation such as $\frac{\sigma}{\sqrt{n}}$.

Component 2

- A response may satisfy component 2 by any of the following or a combination of the following:
  - Graphical: Displaying a graph of a normal density function with the region of interest ($\overline{X} > 303$ or $Z > 0.8485$) clearly identified. The shaded area does not need to be proportional, but the boundary should be on the proper side of the mean and the shading should be in the proper direction.
  - Calculator function syntax: Labeling the lower and upper bounds of the region of interest in a "normalcdf" statement, such as:
    - normalcdf$\left(\text{lower} = 303, \text{upper} = \infty, \text{mean} = 300, \text{standard deviation} = \frac{5}{\sqrt{2}}\right)$.
    - normalcdf$(\text{lower} = 0.8485, \text{upper} = \infty, \mu = 0, \sigma = 1)$.
  - Words: Specifying the correct event in words with correct numerical value for the boundary and correct direction, such as the "probability that the mean amount of gold applied to two randomly selected necklaces is greater than 303 mg" or $P(\text{mean amount of gold} > 303)$.
  - Standard Notation: Using standard notation such as or $P\left(\overline{X} > 303\right)$ or $P\left(Z > \frac{303 - 300}{3.5355}\right)$ or $P(Z > 0.8485)$.

Component 3

- A response may satisfy component 3 if the reported probability is consistent with components 1 and 2. If a normal distribution with a mean of 300 mg and standard deviation 5 mg is used, the probability reported should be 0.274.

General

- Minor errors in statistical notation may be ignored. However, this is considered poor communication and may be considered if holistic scoring is required.
- If the only error in part (b) is the reversal of the numerator for the *z*-score $(300 - 303)$, the response is scored P.
- An arithmetic or transcription error in a response can be ignored if correct work is shown.

Alternate Solution:

- A response that calculates the probability that the ***total*** amount of gold on the two randomly selected necklaces is greater than 606 mg $(2 \times 303\,\text{mg})$ can be scored essentially correct if the response demonstrates:
  - The use of a normal (or approximately normal) distribution and the correct mean and standard deviation for the distribution of the total amount of gold applied to two randomly selected necklaces, (e.g., $N(600, 7.071)$, or mean $= 2(300) = 600$ mg and standard deviation $= \sqrt{2\left(5^2\right)} \approx 7.071$ mg, or $z = \dfrac{606 - 600}{7.071} \approx 0.8485$), satisfying component 1.
  - Specifies the correct event, including the correct boundary value of $x_{total} = 2(303) = 606$ or $z \approx 0.8485$, and the correct direction, satisfying component 2.
  - The probability is correctly computed using the mean and the standard deviation of the total amount of gold applied to two randomly selected necklaces. An arithmetic error can be ignored if correct work is shown and the result is between 0 and 0.5.
- While the probability of a difference in part (b-ii) is twice the probability found in part (b-i), it is not necessary for a response to double the probability in part (b-i). However, if the probability in part (b-i) is doubled, that should be considered if holistic scoring is required.

Model Solution	Scoring
(c) **(i)** The sampling distribution of the sample range for random samples of size $n = 2$ from a normal distribution with standard deviation $\sigma = 5$ is skewed to the right. Almost all values of the simulated ranges are between 0 mg and about 25 mg and the center of the distribution is about 6 mg. **(ii)** As the value of the population standard deviation increases, the variation (spread) in the distribution of the sample range increases and the mean of the distribution of the sample range also increases.	**Essentially correct (E)** if the response satisfies four or five of the following five components: 1. In part (c-i), describes the shape as positively skewed (or skewed to the right) 2. In part (c-i), describes the center of the distribution as about 6 mg 3. In part (c-i), describes the spread of the simulated sample ranges as having most values between 0 mg and 30 mg 4. In part (c-ii), indicates that the mean (or center) of the distribution of the sample range increases as the population standard deviation increases 5. In part (c-ii), indicates that the variation (or spread) of the distribution of the sample range increases as the population standard deviation increases **Partially correct (P)** if the response satisfies three of the five components *OR* the response satisfies only components 4 and 5 *OR* the response only satisfies components 2 and 4 *OR* the response satisfies only components 3 and 5. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- The centrality component of the response to part (c-i) may be satisfied with any reasonable description of center, such as a comment on the approximate value of mean or median of the distribution or indication that the center is somewhere between 4 and 10 mg.
- The spread component of the response to part (c-i) may be satisfied with a reasonable statement about the inter-quartile range. Any value between 5 and 7, inclusive, is considered reasonable.
- Although the actual range of the sample distribution of the sample range is infinite, a comment that the range is any value between 20 and 30 may be scored as satisfying the spread component of the response to part (c-i).
- The response to part (c-ii) need not comment on the skewness of the distributions. Any comment on the skewness of the distributions should be considered as extraneous in scoring the response.

Model Solution	Scoring
(d) **(i)** No, a sample range of 10 mg is not unusual if the machine is working properly with a standard deviation of 5 mg. Although observing a sample range around 10 mg or greater is much more likely if the population standard deviation is 8 mg or 12 mg than when the population standard deviation is 5 mg, the graph of the sampling distribution of the sample range for samples of size 2 from a normal distribution with $\sigma = 5$ mg indicates that 10 mg is not an unusual value for the range when $\sigma = 5$ mg. There is about a 20% chance that a random sample of two necklaces would yield a range of 10 mg or more when the machine is working properly. **(ii)** No, Cleo's sample mean of 303 mg and range of 10 mg do not indicate that the machine is not working properly. As noted in part (b-i), the probability that the sample mean would be equal to or greater than 303 mg when the machine is working properly is almost 20% so having a sample mean of 303 mg is not unusual. Furthermore, it is less than one standard deviation, $\frac{5}{\sqrt{2}} = 3.5355$ mg, away from 300 mg. As indicated in part (d-i), the probability of a range of 10 mg or greater when the population standard deviation is 5 mg is also about 20%, so not unusual. There is not statistically significant evidence to show the machine is not working properly.	**Essentially correct (E)** if the response satisfies four or five of the following five components: 1. In part (d-i), indicates a sample range of 10 mg is not unusual 2. In part (d-i), justifies the response to component 1 by indicating that the graph of the sampling distribution of the sample range when $\sigma = 5$ mg shows that values of 10 mg or greater occur often 3. In part (d-ii), indicates there is not convincing evidence that the machine is not working properly 4. In part (d-ii), justifies the response by indicating that a sample mean of 303 mg would not be unusual when the machine is working properly with one of the following • 303 mg is less than one standard deviation away from 300 mg • The probability that the mean is greater than or equal to 303 mg was shown in (b) to have probability of 0.198, which is not unusual • Some other reasonable justification 5. In part (d-ii), justifies the response by indicating that a sample range of 10 mg is not unusual when the machine is working properly as discussed in part (d-i) **Partially correct (P)** if the response does not meet the requirements for E and satisfies at least three of the five components *OR* the response satisfies components 1 and 2 *OR* the response satisfies component 3 and either component 4 or 5. **Incorrect (I)** if the response does not meet the criteria for E or P.

Additional Notes:

- An argument that a sample range of 10 mg or more is more likely to occur (or closer to the center of the sampling distribution of the range) when $\sigma = 8$ mg or $\sigma = 12$ mg is not required for component 2, but by itself does not satisfy component 2. Inclusion of this argument should be considered if holistic scoring is required.

- Other reasonable justifications to satisfy component 4 include arguments based on a *z*-value close to zero, a one-sample *z*-test or one-sample *t*-test for a population mean, or a one sample *z*-interval or *t*-interval for a population mean.

Alternate Solutions:

- A response that uses the intersection (or union) of the event 'the sample mean is greater than or equal to 303 mg' and the event 'the sample range is greater than or equal to 10 mg' can satisfy components 3, 4, and 5.
  - Components 3, 4, and 5 are satisfied if the response shows, for example:
    - Let $P(M) = P(\text{mean} \geq 303) = 0.198$ (from part b-ii).
    - Let $P(R) = P(\text{range} \geq 10 \text{ if } \sigma = 5) \approx 0.2$ (from part d-i).
    - For Intersection**:** $P(M \text{ and } R) = P(M)P(R) = (0.198)(0.2) = 0.0396,$ because this probability is small, this is evidence that the machine is not working properly.
    - For Union**:** $P(M \text{ or } R) = P(M) + P(R) - P(M)P(R) = 0.198 + 0.2 - (0.198)(0.2) = 0.3584,$ because this probability is large, this is not evidence that the machine is not working properly.

General Notes:

- Context (amount of gold in necklaces) is not required in parts (a), (b), (c), or (d) of the response. However, including context is considered good communication and may be considered if holistic scoring is required.

Scoring for Question 6

Each essentially correct (E) part counts as 1 point, and each partially correct (P) part counts as ½ point.

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

If a response is between two scores (for example, 2 ½ 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.