

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 <i>population</i> 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 $\bar{X}$ 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:

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

$$\approx 28.94,$$

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) <i>AND</i> 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 <i>OR</i> One section essentially correct and one or two sections partially correct <i>OR</i> Three sections partially correct	2
Minimal Response One section essentially correct and no section partially correct <i>OR</i> No section essentially correct and two sections partially correct	1