PHP2510

ASSIGNMENT #2 WEEKS 6-9

PURPOSE

Statistical inference is the process of drawing insights about the world from limited data. The purpose of this assignment is to provide more practice working with sampling distributions, performing hypothesis tests, and constructing confidence intervals. This assignment will help you understand the concepts behind these techniques, interpret results, and gain more mastery of using R for executing statistical analyses.

Topics Covered: Sampling Distributions, Hypothesis Testing, Confidence Intervals (parametric only)

INSTRUCTIONS

Collaboration between students is allowed, but students should submit their own, individual work. Answers will be a mix of code and short-form responses; please include all relevant code. A templated R Markdown file is available for use (optional, but preferred). To assist with grading, we ask that you submit .pdf file(s) that show that your code has run (via knitting a .rmd file or adding screenshots of a console)

Grading: Each question (or subquestion) is worth 3 points, awarded as follows:

- 3 points if fully correct, or with very minor flaws in reasoning/arithmetic
- 2 points for an attempt that contains some flaws but shows comprehension of most core concepts
- 1 point for an attempt that contains major flaws or misunderstanding of most core concepts
- 0 points for no attempt whatsoever

Logistics: Per the syllabus, the problem set will be accepted up to 3 days late, with a 5% penalty for each late day. If you have extenuating circumstances that prevent you from submitting the assignment on time, please contact me **before** the original due date to arrange an exception, which will be determined on a case-by-case basis.

Feedback: Homework grades, along with any feedback, will be available within one week of the due date for all students that submitted the problem set on time. Late submissions will be graded shortly thereafter.

PROBLEM SET

1. [3 points] Explain, in your own words, why sigma is considered a nuisance parameter when estimating a population mean.
2. [3 points] Explain, in your own words, why a confidence interval is not always symmetric. Give an example.
3. [3 points] Explain, in your own words, why reporting a confidence interval but no p-value is reasonable, but the other way (a p-value with no confidence interval) is not recommended. Please provide more insight than "a confidence interval contains more information"
4. [3 points] Give an example, using toy data that you create in R, that shows the benefit of a paired t-test in the presence of correlated data.

5. [15 points] A municipal water utility is subject to an environmental regulation that mandates the 90th percentile of daily arsenic concentration measurements must not exceed 5 µg/L. The public health inspector's office collects a daily arsenic concentration sample over a three month (92 day) period to determine compliance.
    a. What is the null hypothesis?
    b. What is the alternative hypothesis?
    c. What test statistic and decision rule maintains at least a 5% alpha, while avoiding any distributional assumptions about the arsenic concentration?
        i. *Hint: turn this into a test of proportions, checking how many samples "fail" our inspection criteria*
        ii. *Note: be careful with how R treats the inequality sign with qbinom() and pbinom()*
    d. What is the actual alpha? Interpret this.
    e. As an environmentally-conscious citizen, do you have any concerns with the approach?
6. [9 points] Choose any data set you'd like from this repository.
    a. Identify a reasonable research question that could be answered by performing a difference in group means test.
    b. Then execute your test end-to-end (with hypothesis, p-value, and confidence interval).
    c. Write a short memo, without technical jargon, to explain your findings. Separately, enumerate any assumptions you made.
7. [6 points] For whatever research question you chose in (6), perform some power calculations. Then explain your findings. For your power calculations, please choose: 3 different effect sizes (think small / medium / large) x 3 different sample sizes (think small / medium / large) - for 9 answers overall. You can use the data set to determine a reasonable standard deviation under the alternative (if necessary).
8. [6 points] Using the fast food data set at openintro.org, answer the following questions. You may use any reasonable definition of "healthy" and assume that the items in the dataset are a random sample from all possible items (that could exist from that restaurant).
    a. Which restaurant has the "least healthy" items, on average? Which restaurant has the "most healthy" items, on average?
    b. Formally test whether the above two restaurants are different (in terms of the healthiness of their items)? What do you conclude?
9. [6 points] Report a confidence interval for each of the 8 restaurants (using your same definition of "healthy") in (8), without using the t.test() function. Discuss the challenges you would encounter if I asked for you to formally test all the restaurants against each other for any pairwise differences (just words).
10. [3 points] If I amended (8) to instead ask about comparing the restaurants using just their single most unhealthy item, how would your response change? Discuss in words (no quantitative solution required), focusing on the assumptions relative to those in (8).