Test 1

Study design, data types, and descriptive statistics [L1, L2, L3]

STAT218

2024-04-19

## Instructions

You have 48 hours from the release of this assignment to complete and submit your work. You may refer to all class materials, notes, and textbooks, but must complete this assignment on your own. By submitting your work, you are affirming that your work is your own and you have not consulted with anyone else in preparing your answers or generated your answers or analyses using AI. Failure to adhere to this expectation will be considered an act of academic dishonesty and result in loss of credit.

You will find a project with a mostly empty script in the class Posit cloud workspace; use this to complete your analyses where required. Note that not all parts require you to perform any calculations; some questions are purely qualitative. Use the prompts as your guide, not the script.

Once you have completed your analyses for the portions requiring use of statistical software, submit your work by filling out the [[test 1 form](https://forms.office.com/r/MSsz35jkyk)] (also posted on the course website). The form will automatically save your work, so you can return to it over the course of the 48-hour test window.

The form will stop accepting responses at the deadline, so make sure you submit by **5pm on Friday 4/19**. Lastly, keep in mind that you will be given the opportunity to revise problems that you miss the first time around to earn back credit.

## Question prompts

1. [L2, L3] The yrbss dataset contains measurements on a small collection of variables from 10,587 survey responses collected as part of the CDC’s Youth Risk Behavior Surveillance System (YRBSS) from 1991-2013. The objective of the survey program is to track behaviors with potential negative physical and mental health impacts among adolescents. In this problem you’ll explore the amount of sleep that respondents get on school nights and the number of days per week respondents are physically active.
   1. [L2] Read briefly about the YRBSS on the CDC website: <https://www.cdc.gov/healthyyouth/data/yrbs/overview.htm>. Based on the overview and the description above, are the data observational or experimental?
   2. [L1] Based on the overview (link in part (a)), identify the study population.
   3. [LX] Load the dataset and identify the type of each variable. What kind of variable is sleep.hours?
   4. [LX] Examine the frequency distributions of age and grade level. Do any grades or ages seem over-represented in the sample?
   5. [L3] Make a bar plot showing the frequency distribution of hours of sleep on school nights. Based on the summary, what is the typical amount of sleep respondents get on school nights?
   6. [L3] Make a stacked bar plot showing levels of sleep by grade. Do older students sleep more on school nights than younger students?
   7. [L3] Visualize the frequency distribution of the number of days per week that survey participants are physically active. Describe the distribution and interpret any patterns.
2. [L2, L3] Diet restriction and longevity. The longevity dataset contains data from a study in which 237 mice were randomly allocated to one of four diets at different levels of restriction: no restriction (NP), normal 85kCal diet before and after weaning (N/N85), normal diet before weaning and restricted 50kCal diet after weaning (N/R50), and normal diet before weaning and restricted 40kCal diet after weaning (N/R40). Researchers recorded the lifetime in months of each mouse in the study.
   1. [L2] Was this an experiment or observational study and why?
   2. [L3] Find the average lifetime of mice in each diet group.
   3. [L3] Find the standard deviation of lifetimes in each diet group.
   4. [L3] Make a plot comparing lifetimes by diet group.
   5. [L2] Write a short summary of the study results based on your work in (b)-(d). Indicate specifically whether there appears to be a relationship between dietary caloric intake and lifetime.
3. [L3] Brain and body size. The mammals dataset contains average body weights (kg) and average brain weights (g) for 62 common species of mammal, as well as log-transformed versions of those weights.
   1. [L3] Make a scatterplot of *log* brain weights against *log* body weights and describe the apparent relationship, if any.
   2. [L3] Compute and interpret the correlation between log brain weight and log body weight.
   3. [L3] Make a histogram of brain weights (not on the log scale) with an appropriate number of bins. Describe the distribution. Are there outliers?
   4. [L3] Based on (c), compute an appropriate measure of center and spread for brain weight.
   5. [L3] Which species of mammal has the largest average brain weight? Inspect the data directly using view(...) to answer this question.
   6. [L3] Which species of mammal has the smallest ratio? The largest? Inspect the data directly using view(...) to answer this question.
4. [L1, L2] The following is an excerpt from the abstract of the study that reported the results of the Moderna Covid vaccine phase three clinical trial[[1]](#footnote-23): *“Vaccines are needed to prevent coronavirus disease 2019 (Covid-19) and to protect persons who are at high risk for complications. The mRNA-1273 vaccine is a lipid nanoparticle–encapsulated mRNA-based vaccine that encodes the prefusion stabilized full-length spike protein of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus that causes Covid-19. This [study] was conducted at 99 centers across the United States … The trial enrolled 30,420 [adult volunteers with no known history of SARS-CoV-2 infection and no circumstances that put them at high risk of infection or severe Covid-19 or both,] who were randomly assigned in a 1:1 ratio to receive either vaccine or placebo (15,210 participants in each group) … Symptomatic Covid-19 illness was confirmed in 185 participants in the placebo group and in 11 participants in the mRNA-1273 group; vaccine efficacy was 94.1%.”*
   1. [L2] Is this an experiment or observational study? Explain.
   2. [L1] Identify the study population.
   3. [L1] Describe the study sample.
   4. [L1] What outcome(s) were measured in the study?
   5. [L3] The moderna dataset contains simulated observations according to the study description. Make a contingency table and use it to construct a table showing the proportions of volunteers infected and not infected in each group.
   6. [L3] Optional. Find the relative risk of illness in the vaccine group compared with the placebo group. Can you determine how efficacy is defined?
5. [L1, L2, L3] The temps dataset contains physical data collected on a number of individuals. Explore the dataset and write a brief summary of your descriptive analysis. While open-ended, your analysis should include descriptions of the variables and their statistical properties, and descriptions of relationships between the variables. Include at least one graphic related to your summary. Your summary need not be exhaustive – in fact, it is better to pick 1-2 interesting findings and report those, rather than describe everything you tried.

1. Baden, L. R., *et al*. (2021). Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *New England journal of medicine*, *384*(5), 403-416. [↑](#footnote-ref-23)