Extra practice problems

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

STAT218

## Test information

The test will consist of four questions with multiple parts, and one less structured question in which your task is to explore and summarize a dataset based on descriptive statistics. You’ll have 48 hours to work on the test, and will submit your answers via a fillable form exactly as in the homework assignment. You can use any class resources (notes, textbook, lecture slides, past homeworks) but are required to work alone. You’ll be provided with a Posit cloud project in which to carry out your analyses, and will be expected to upload your script as a supporting document when you fill out the submission form. The practice problems below provide a sample of questions that approximate the test problems (but are slightly shorter).

## Practice problems

1. Another version of the frog dataset from earlier also includes measurements of egg size and body size. Use this dataset to practice visualizing and describing distributions of numeric variables.
	1. Make histograms of each of the four numeric variables, with appropriate numbers of bins, and describe the shape and number of modes.
	2. For each variable, suggest an appropriate measure of center and measure spread and identify any measures that would *not* be appropriate.
	3. Make pairwise scatterplots of each of the four numeric variables and describe the association, if any. (*Hint: try pairs(frog) for a more efficient way to generate these plots.*)
	4. For linear associations in (c), compute and interpret the correlation.

# load and inspect data
load('data/frog.RData')
head(frog)

 site altitude clutch.size clutch.volume egg.size body.size
1 040 3,462 181.9701 177.8279 1.949845 3.630781
2 040 3,462 269.1535 257.0396 1.949845 3.630781
3 040 3,462 158.4893 151.3561 1.949845 3.715352
4 040 3,462 234.4229 223.8721 1.949845 3.801894
5 040 3,462 245.4709 234.4229 1.949845 3.890451
6 040 3,462 301.9952 288.4032 1.949845 3.890451

# part a: histograms of each numeric variable; describe shape and modes
par(mfrow = c(2, 2), mar = c(4, 4, 4, 1))
hist(frog$clutch.size)
hist(frog$clutch.volume)
hist(frog$egg.size)
hist(frog$body.size)



# part c: pairwise scatterplots of clutch volume, egg size, body size, clutch size
pairs(frog)



# part d: correlations for linear associations
cor(frog$clutch.size, frog$clutch.volume)

[1] 0.8077344

cor(frog$clutch.volume, frog$egg.size)

[1] 0.6462605

cor(frog$body.size, frog$clutch.volume, use = 'complete.obs')

[1] 0.6755435

cor(frog$body.size, frog$clutch.size, use = 'complete.obs')

[1] 0.6147564

1. The chick data data come from a study investigating the early growth of chicks on different diets. In the study, 47 chicks were randomly assigned one of four diets at birth and researchers measured body weight in grams daily. The data below show body weights at 18 days since birth for each chick. The question of interest is: which diet is best?
	1. Is this observational or experimental data? Explain your reasoning.
	2. Produce a visualization that compares body weight distributions by diet. For which diet have chicks grown the most? The least? Explain the statistic(s) or features of the distribution you used to make this determination.
	3. Based on your plot in (b), suggest a measure of center and measure of spread that would be appropriate for summarizing the data.
	4. Calculate the measures you suggested in (c) separately for each diet group.
	5. Assume that in the previous question you found that chicks on diet 3 grew the most, regardless of your actual answer. Can you conclude that diet 3 caused the fastest growth? Explain why or why not.

# load and inspect data
load('data/chick.RData')
head(chick)

# A tibble: 6 × 3
 chick.id weight diet
 <dbl> <dbl> <fct>
1 1 171 diet 1
2 2 187 diet 1
3 3 187 diet 1
4 4 154 diet 1
5 5 199 diet 1
6 6 160 diet 1

# part b: visualize body weights by diet
boxplot(weight ~ diet, data = chick)



# part c-d: determine and compute appropriate measures of spread and center
chick |>
 group\_by(diet) |>
 summarize(avg.weight = mean(weight),
 sd.weight = sd(weight))

# A tibble: 4 × 3
 diet avg.weight sd.weight
 <fct> <dbl> <dbl>
1 diet 1 159. 49.2
2 diet 2 188. 63.3
3 diet 3 233. 57.6
4 diet 4 203. 33.6

1. The gss dataset contains observations for 500 respondents in the General Social Survey on a small number of demographic categorical variables. Use this to practice tabular and graphical summaries for categorical variables.
	1. For each variable, determine whether the variable is nominal or ordinal.
	2. Make a contingency table of age bracket and whether participants have obtained a college degree.
	3. Visualize the relationship between age and having obtained a college degree.
	4. Does the proportion of respondents with a college degree differ by sex?
	5. By political party?
	6. By socioeconomic class?
	7. Make one additional comparison of your choice and interpret the result.

# load and inspect data
load('data/gss.RData')
head(gss)

# A tibble: 6 × 5
 age sex college.degree political.party class
 <fct> <fct> <fct> <fct> <fct>
1 (29,38] male degree ind middle class
2 (29,38] female no degree rep working class
3 [18,29] male degree ind working class
4 (38,50] male no degree ind working class
5 (29,38] male degree rep middle class
6 (29,38] female no degree rep middle class

# part b: contingency table of age and college degree
table(gss$college.degree, gss$age)

 [18,29] (29,38] (38,50] (50,87]
 degree 36 44 60 34
 no degree 99 74 64 89

# part c: visualize relationship between age and college degree
table(gss$college.degree, gss$age) |>
 proportions(margin = 2) |>
 barplot(legend = T)



# part d: does the proportion of respondents with a degree differ by sex?
table(gss$college.degree, gss$sex) |>
 proportions(margin = 2) |>
 barplot(legend = T)



# part e: by political party?
table(gss$college.degree, gss$political.party) |>
 proportions(margin = 2) |>
 barplot(legend = T)



# part f: by class?
table(gss$college.degree, gss$class) |>
 proportions(margin = 2) |>
 barplot(legend = T)



# part g: one additional comparison of your choosing

1. Long COVID is a multi-systemic and often debilitating condition that develops in at least 10% of patients following a COVID infection. The following is an excerpt of the abstract from a recent study seeking to identify symptoms and risk factors associated with long COVID and published in *Nature Medicine*[[1]](#footnote-42): *“We undertook a … study using a UK-based primary care database, Clinical Practice Research Datalink Aurum, to determine symptoms that are associated with confirmed SARS-CoV-2 infection beyond 12 weeks in non-hospitalized adults and the risk factors associated with developing persistent symptoms. We selected 486,149 adults with confirmed SARS-CoV-2 infection … Outcomes included 115 individual symptoms, as well as long COVID, defined as a composite outcome of 33 symptoms by the World Health Organization clinical case definition … Among the patients infected with SARS-CoV-2, risk factors for long COVID included female sex, belonging to an ethnic minority, socioeconomic deprivation, smoking, obesity and a wide range of comorbidities. The risk of developing long COVID was also found to be increased along a gradient of decreasing age.”*
	1. Identify the type of study.
	2. Identify the study population.
	3. Describe the sample.
	4. List the study outcomes of interest.
	5. Identify any non-outcome variables.
1. Subramanian *et al.* (2022). Symptoms and risk factors for long COVID in non-hospitalized adults. *Nature medicine*, *28*(8), 1706-1714. [↑](#footnote-ref-42)