HW04

Stat 145

Due February 6, 2024 at 11:00 pm

Problem set policies. Please provide concise, clear answers for each question. Note that only writing the result of a calculation (e.g., "SD = 3.3") without explanation is not sufficient. For problems involving *R*, include the code in your solution, along with any plots. (Using R Markdown to write up your answers would be ideal.)

Please submit your homework set via MOM as a PDF, along with the R Markdown source file.

I encourage you to discuss problems with other students or me, but you must write your final answer in your own words. Solutions prepared "in committee" are not acceptable. If you do collaborate with classmates on a problem, please list your collaborators on your solution

Problem 1.

Since states with larger numbers of elderly residents would naturally have more nursing home residents, the number of nursing home residents in a state is often adjusted for the number of people 65 years or older (65+). That adjustment is usually given as the number of nursing home residents age 65+ per 1,000 members of the population age 65+. For example, a hypothetical state with 200 nursing home residents age 65+ and 50,000 people age 65+ would have the same adjusted number of residents as a state with 400 residents and a total age 65+ population of 100,000 – 4 residents per 1,000.

The data file found at the link

https://scofield.site/teaching/data/csv/harrington/nursingHome.csv

contains this adjusted number of residents for each state in the United States. The state names are saved under the variable name `state` and the adjusted number of residents under the variable name `resident`.[1]

a) Which state has the smallest number of nursing home residents per 1000 population 65 years of age and over? Which state has the largest number? Hint: use the R functions `min()` and `max()`, or perhaps `qdata()`. Alternatively, look directly at the data in *RStudio*.

b) What factors might influence the substantial amount of variability among different states? This question cannot be answered from the data; speculate using what you know about the demographics of the United States.

c) Construct a boxplot for the number of nursing home residents per 1,000 population.

d) Does the distribution depicted in the boxplot of part (c) symmetric or skewed? Are there any states that could be considered outliers?

e) Display the number of nursing home residents per 1000 population using a histogram. Do you find this graph to be more or less informative than the box plot? Explain your answer.

[1] The data originally appeared in Chapter 12 of *Case Studies in Biometry*, 1994, by Lange et al.

Problem 2.

The data file found at

https://scofield.site/teaching/data/csv/harrington/adolescentFertility.csv

contains data on the number of children born to women aged 15-19 from 189 countries around the world for the years 1997, 2000, 2002, 2005, and 2006.[2] The data are defined using a scaling similar to that used in the nursing home data. The values for the annual adolescent fertility rates represent the number of live births among women aged 15-19 per 1,000 women members of the population of that age.

For the years 2000-2006, the adolescent fertility rate for Iraq is coded `NA`, or missing. When calculating a mean or standard deviation in R for a variable `x` which has missing data, add `na.rm=TRUE` to the argument to perform the calculations without the missing observations: `mean( x, na.rm=TRUE); sd( x, na.rm=TRUE)`.

a) Calculate the mean, standard deviation, and five-number summary for the distribution of adolescent fertility in 2006 (`fert_2006`).

b) What is the 70^{th} percentile of the distribution? Write a sentence explaining the 70^{th} percentile in the context of this data.

c) Suppose those observations for Iraq between 2000 and 2006 could be added. Is it likely the addition of these numbers would have notable effect on the five-number summary?

d) Produce, using one command, side-by-side boxplots of the fertility rates for each of the five years in the dataset. What pattern do you see?

Problem 3.

A recently published analysis examined 10 studies that measured optimism and pessimism by asking participants about their level of agreement with statements like "In uncertain times, I usually expect the best," or "I rarely expect good things to happen to me". Optimistic people tend to expect that they will encounter favorable outcomes, whereas less optimistic people tend to expect that they will encounter unfavorable outcomes.[3]

These studies also measured other variables on participants, including factors related to heart disease. The analysis found that compared with pessimists, people with the most optimistic outlook had a 35% lower risk for cardiovascular events (e.g., heart attacks). The studies, on average, observed people over a 14-year period and compared the rate of cardiovascular events between those classified as optimists versus pessimists.

a) A popular newspaper reports on the analysis with the headline "Thinking Positively Improves Cardiovascular Health". Write a short response to the editor explaining clearly why the headline is potentially misleading. Be sure to use language accessible to a general audience without a statistics background. Limit your answer to at most five sentences.

b) Briefly describe a plausible study design that has the potential to demonstrate the effect of thinking positively on cardiovascular health.

[2] Data from the CIA World Factbook

[3] Alan Rozanski, MD, et al. Association of optimism with cardiovascular events and all-cause mortality. *JAMA Network Open* 2019; 2(9):e1912200.

c) Suppose someone who is very optimistic reads about the analysis and concludes that the findings suggest he has a 35% lower risk for cardiovascular events than his friend who is extremely pessimistic. Explain why this is not necessarily the case.

Problem 4.

Suppose that you are interested in determining whether a relationship exists between the fluoride content in a public water supply and the dental caries experience of children using this water. The file

https://scofield.site/teaching/data/csv/harrington/water.csv

contains the data from a study examining 7,257 children in 21 cities from the Flanders region in Belgium.

The fluoride content of the public water supply in each city, measured in parts per million (ppm), is saved under the variable name `fluoride`; the number of dental caries per 100 children examined is saved under the name `caries`. The total dental caries number is obtained by summing the numbers of filled teeth, teeth with untreated dental caries, teeth requiring extraction, and missing teeth.[4]

a) Construct a scatterplot for these data, with `fluoride` as the x-variable and `caries` as the y-variable.

b) Do `fluoride` and `caries` appear to be positively or negatively associated? Explain your answer.

c) Later in the course, we will study methods for fitting a straight line to data.

   i. Add a straight line to the plot, and estimate its x-intercept and y-intercept. (*Hint*: Be sure to look at the limits on the axes...)

   ii. Based on the appearance of the plot, do you think that a straight line would be a reasonable way to represent these data? Explain your answer.

Problem 5.

This problem features data from the *famuss* (*Functional SNPs Associated with Muscle Size and Strength*) study discussed in Chapter 1. The study examined the possible genetic determinants of skeletal muscle size and strength, before and after training.

This problem uses the following variables from the FAMuSS data:

- `ndrm.ch`: the percent change in strength in a participant's non-dominant arm, from before training and after.
- `drm.ch`: the percent change in strength in a particpant's dominant arm.
- `actn3.r577x`: the genotype at residue *r577x* within the *ACTN3* gene.
- `race`: race of the participant, with values stored as text strings.

The `famuss` dataset is in the `oibiostat` package.

a) Make a table of the genotypes for the SNP `actn3.r577x`.

[4] These data appear in Table B21 in *Principles of Biostatistics*, 2nd ed. by Pagano and Gauvreau.

b) Construct a two-way table of `actn3.r577x` by race, with the genotypes in the columns of the table and races in the rows.

c) If you were to use numerical summaries to describe the `ndrm.ch` variable, would you prefer the mean and standard deviation or the five-number summary? Why?

d) Produce a graphical summary that shows the association between age and genotype at the SNP `actn3.r577x`. Describe what you see.

Problem 6.

Does smoking have the same association with cardiovascular disease in women as it does in men? Epidemiologists typically use data from observational studies to investigate possible causes of disease.

Aortic stenosis is a narrowing or stricture of the aorta that impedes blood flow to the body.[5]

The dataset contains three variables, for 215 study participants:

- `disease`: coded `Yes` if stenosis is present, `No` if it is absent.
- `smoke`: coded `Smoker` if the participant is a current or former smoker, `NonSmoker` if the participant has never smoked.
- `sex`: coded as either `Male` or `Female`

Use the data from the file at

https://scofield.site/teaching/data/csv/harrington/stenosis.csv

to answer the following questions.

a) Construct a two-way table for smoking status and disease presence. What percentage of the 215 participants were both smokers and had aortic stenosis? This percentage is one component of the *joint distribution* of smoking and stenosis; what are the other three numbers of the joint distribution?

b) Among the smokers, what proportion have aortic stenosis? This number is a component of the conditional distribution of stenosis for the two categories of smokers. What proportion of non-smokers have aortic stenosis?

c) Repeat part b) for males and females separately. To do this, first subset the data to create two datasets: one with only males, and one with only females. Include the tables in your solution. Are there any differences by sex in the proportion of smokers who suffer from aortic stenosis?

[5] The data appear in Table B20, *Principles of Biostatistics*, 2nd ed. by Pagano and Gauvreau.