Problem Set 1

Statistics 104

Due February 6, 2020 at 11:59 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*, be sure to include the code in your solution.*

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

We encourage you to discuss problems with other students (and, of course, with the course head and the TFs), 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.

For each of the following scenarios, discuss (in at most five sentences) the main issue(s) with respect to sampling or reporting bias.

a) A particular city has 14 architects who own their own firm. To select a survey sample, each architect was contacted via telephone by order of appearance in the telephone directory, then the first 8 that agreed to be interviewed formed the sample.

b) The September 1992 issue of *Prevention* magazine included a women's health survey; approximately 16,500 women responded to the survey. The May 1993 issue reported on the survey results, claiming that "92% of our readers rated their health as excellent, very good, or good".

c) Many scholars and policymakers are interested in estimating the prevalence of mental illness among the homeless population. In one study, the authors sampled homeless persons who received medical attention from a clinic that was part of the Health Care for the Homeless project, resulting in an estimated prevalence of 33%.[1] The authors maintain that selection bias is not a serious problem because the clinics are easily accessible to homeless people.

Problem 2.

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.[2]

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,

[1] This project is a federally funded program that brings general health and mental health services to homeless people.

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

observed people over a 14-year period and compared the rate of cardiovascular events between those classified as optimists versus pessimists.

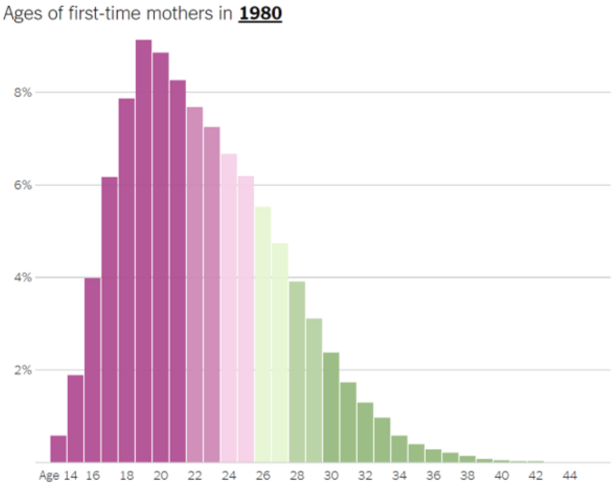
a) A popular newspaper reports on the analysis with the headline “Thinking Positively Improves Cardiovascular Health”. Write a short response to the editor explaining clearly why the headline is potentially misleading. Be sure to use language accessible to a general audience without a statistics background.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

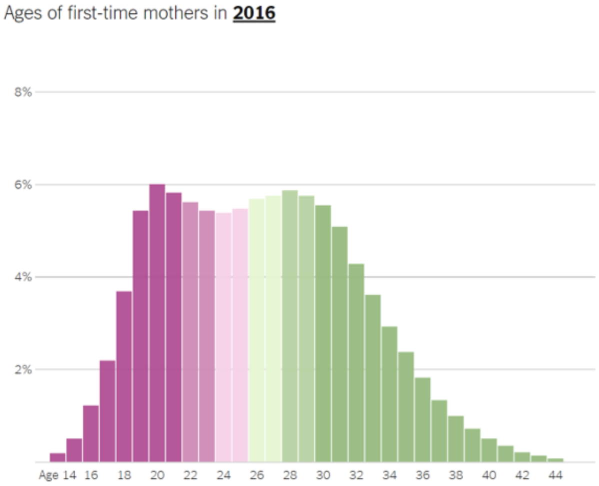
b) Briefly describe a plausible study design that has the potential to demonstrate the effect of thinking positively on cardiovascular health.

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 3.

The following graphs are based on data from the National Center for Health Certificates.





a) Describe what you see in the two graphs, with particular focus on the differences between the two distributions.

b) Economists are interested in the possible causes driving the shape of the age distribution in 2016.

   i. Discuss a possible reason behind the discrepancy between the 1980 distribution and the 2016 distribution; i.e., what is a potential factor driving the difference in the distributions?

   ii. Discuss a possible reason behind the shape of the age distribution in 2016.

Problem 4.

FiveThirtyEight is a data journalism site devoted to applying statistical analysis to a variety of current topics in politics, sports, science, economics, and culture. They recently published a series of articles on gun deaths in America, based on data collected from the Centers for Disease Control and Prevention (and other governmental agencies) on all gun deaths in the United States from 2012 - 2014.

The main dataset used for the analysis is available as the `gun_deaths` data. Each case represents a single gun death.

- `month`: Month of death, coded as a numerical variable taking values 1 - 12.
- `intent`: The underlying cause of death, either `Accidental`, `Homicide`, `Suicide`, or `Undetermined`. Deaths from legal intervention are coded as homicides.
- `police`: Coded `Yes` if the death is a result of legal intervention by police and `No` otherwise.
- `sex`: The sex of the individual who died, either `F` or `M`.
- `age`: Age of the individual who died, in years.
- `race`: Race of the individual who died, either `Asian/Pacific Islander`, `Black`, `Hispanic`, `Native American/Native Alaskan`, or `White`.
- `place`: Place of injury, either `Home`, `Residential institution`, `School/instiution`[sic], `Sports`, `Street`, `Trade/service area`, `Industrial/construction`, `Farm`, `Other specified`, or `Other unspecified`.
- `education`: Educational status of the individual who died, either `Less than HS`, `HS/GED`, `Some college`, or `BA+`.

a) Using numerical and graphical summaries, describe and compare the distribution of age (of death) between males and females.

b) Which underlying cause of death contributes the most toward gun deaths?

c) Identify the proportion of deaths classified as homicides that involved legal intervention by police.

d) During which season did gun deaths most often occur? Assume that spring is March - May, summer is June - August, fall is September - November, and winter is December - February.

e) Of the gun deaths in 2012 among individuals with at least a high school education, what proportion of those individuals were white males? Be sure to account for any missing values in the calculation.

f) Compare the proportion of gun deaths due to suicide versus homicide between race/ethnicity groups. In a few sentences, summarize the main findings.

Problem 5.

Employment statistics represent an important source of metrics for policymakers to use in gauging the overall health of the economy. In the United States, the government measures unemployment using the Current Population Survey (CPS). This survey collects demographic and employment information each month from about 60,000 occupied households.

To be eligible to participate in the survey, individuals must be 15 years of age or older and not in the Armed Forces. One person generally responds for all members of the household; this person is called the "reference person" and usually is the person who owns or rents the housing unit.

The `CPSData.csv` file contains data from the September 2016 survey. Descriptions of the relevant variables are as follows:

- `PeopleInHousehold`: number of people in the household, including the respondent
- `Region`: the census region the respondent lives in, either `Midwest`, `Northeast`, `South`, or `West`
- `State`: the state the respondent lives in, either one of the 50 states or the District of Columbia
- `Age`: age in years of the respondent, where `80` represents individuals ages 80-84, and `85` represents individuals ages 85 and higher
- `Married`: marital status of the respondent, either `Divorced`, `Married`, `Never Married`, `Separated`, or `Widowed`
- `Sex`: sex of the respondent, either `Female` or `Male`
- `Education`: highest educational level of the respondent, either `Associate degree`, `Bachelor's degree`, `Doctorate degree`, `High school`, `Master's degree`, `No high school diploma`, `Professional degree`, or `Some college, no degree`
- `Race`: race of the respondent, either `American Indian`, `Asian`, `Black`, `Multiracial`, `Pacific Islander`, `White`
- `Hispanic`: coded `1` if the respondent is of Hispanic ethnicity, and `0` otherwise
- `Citizenship`: citizenship status of the respondent, either `Citizen, Native`, `Citizen, Naturalized`, or `Non-Citizen`
- `EmploymentStatus`: employment status of the respondent, either `Disabled`, `Employed`, `Not in Labor Force`, `Retired`, or `Unemployed`
- `Industry`: industry of employment, available only if the respondent is employed

a) Explore the variables `Age`, `Sex`, and `Race`.

   i. Based on these three variables, write a short paragraph describing the basic demographics of the survey respondents. Reference numerical and graphical summaries as appropriate.

   ii. Do you notice anything odd about the distribution of `Age`? Point out what you think is unusual.

b) Describe the distribution of the number of people in a household, referencing appropriate numerical and graphical summaries.

c) Describe the distribution of citizenship status.

d) The CPS differentiates between race and ethnicity. For which races do 15% or more of respondents identify as ethnically Hispanic?

e) Create a graphical summary that shows the association between age and marital status. Describe what you see and comment on whether it is what you might expect intuitively.

Problem 6.

This problem uses stock data from Apple Corporation (AAPL) and Microsoft Corporation (MSFT). The following code uses commands from the quantmod package to fetch daily return data. The adjusted closing price can be thought of as the most accurate reflection of a stock's value at closing; the closing price factors in events that might affect the stock price after the market closes. The daily volume of a stock refers to how many shares were traded that day. The daily return quantifies how much value was gained/lost in a day.

```r
#load quantmod package
library(quantmod)

#load AAPL and MSFT data
getSymbols("AAPL", from = "2018-01-01", to = "2019-07-22")
getSymbols("MSFT", from = "2018-01-01", to = "2019-07-22")

#obtain adjusted closing prices
aapl.closing = Ad(AAPL)
msft.closing = Ad(MSFT)

#obtain daily volume
aapl.volume = Vo(AAPL)
msft.volume = Vo(MSFT)

#obtain daily returns
aapl.return = as.numeric(dailyReturn(Ad(AAPL)))
msft.return = as.numeric(dailyReturn(Ad(MSFT)))
```

a) Run the code in the template to plot Apple's and Microsoft's stock prices between 01 January 2018 and 22 July 2019. Describe what you see.

b) Compare the return from holding one share of AAPL for this time period versus one share of MSFT. The return over a period of time is the change in value divided by the original price.

c) Compute the standard deviation of daily returns for AAPL and MSFT over this time period. Based on standard deviation, which stock is less volatile? Volatility refers to the degree of fluctuation in a stock period over time.

d) Identify the highest and lowest adjusted closing prices of AAPL during this time period, as well as the dates on which they occurred. Do the same for MSFT.

e) Both MSFT and AAPL are in the S&P 500 Index, a stock index that measures the stock performance of 500 large publicly traded companies on the US market. It is a weighted index, such that larger companies contribute more to the index than smaller companies. The two largest components of the index are AAPL and MSFT. Thus, we might expect that the returns of these two companies are highly related.

   i. From a graphical summary, describe your impression of the relationship between daily return for AAPL and daily return for MSFT.

   ii. Calculate an appropriate numerical summary for the relationship observed in part i.

   iii. Interpret the value from part ii. in language accessible to an audience who has not taken a statistics course.