MAS 250 Homework Assignment 1

Due: September 14 (Wednesday) 1 pm

Instruction: Turn in homework as a single (pdf) file. For the R parts, paste the relevant R command, outputs, and interpret the results as in the class notes.

1. Sarah would like to know the average sleep hours of all high school students in Korea. Therefore, she selected 1,000 students randomly and measured their sleep hours. Identify population, sample, parameter, and statistic.

2. Explain what is deficient in each of the following proposed experiments and explain how you would improve the experiment.

   (a) Two product promotion offers are to be compared. The first, which offers two items for $2, will be used in a store on Friday. The second, which offers three items for $3, will be used in the same store on Saturday.

   (b) A study compares two marketing campaigns to encourage individuals to eat more fruits and vegetables. The first campaign is launched in Florida at the same time that the second campaign is launched in Minnesota.

   (c) You want to evaluate the effectiveness of a new investment strategy. You try the strategy for one year and evaluate the performance of the strategy.

3. In each of the following situations, identify its sampling method.

   (a) A student organization has 55 members. A table of random numbers is used to select a sample of 5.

   (b) An online poll asks people who visit this site to choose their favorite television show.

   (c) Separate random samples of male and female first-year college students in an introductory psychology are selected to receive a one-week alternate instructional method.

4. Explain what is wrong with each of the following scenarios.

   (a) The population consists of all individuals selected in a simple random sample.

   (b) In a poll of an SRS of residents in a local community, respondents are asked to indicate the level of their concern about the dangers of dihydrogen monoxide, a substance that is a major component of acid rain and in its gaseous state can cause severe burns. (Hint: Ask a friend who is majoring in chemistry about this substance or search the Internet for information about it.)

   (c) Students in a class are asked to raise their hands if they have cheated on an exam one or more time within the past year.

5. Researchers examined the incidence of diabetes and height in adult men. They wonder if shorter men are at increased risk of diabetes. Among selected 100 adult men, they found that for short (less than 175 cm) men there was a 1.9% higher proportion that had diabetes versus tall ($>$ 175 cm) men.

   (a) What is the response variable?

   (b) What is the explanatory variable?

   (c) Is this an experiment or observational study?

   (d) Can researchers conclude that being short is the cause of diabetes?

   (e) What is necessary for researchers to generalize these results to a larger group? To what larger group would they be able to generalize the results?

6. (**R assignment.**) The dataset, **gallup.xlsx**, consists of responses to a telephone survey in Allegheny County, PA, collected in the '80s (source: Heinz School, Carnegie Mellon University). The variables in this dataset are arranged in columns. Each column is a variable, each row is an observation (an individual). The variables are:

location	5=Mon Valley, 6=Pittsburgh, 7=rest of count
age	age in years
race	1=white, 2=black, 3=other, 4=Hispanic
gend	1=male, 0=female
educ	1=4th grade or less, 2=grades 5-7, 3=grade 8, 4=high school incomplete (9-11) 5=high school graduate, 6=technical/trade/business school 7=college/university incomplete, 8=college/university graduate or more
emp	1=employed, 0=unemployed
wage	hourly wage if employed, 0 if unemployed
hours	hours worked per week
weeks	weeks worked per year if employed, 0 if unemployed
salary	annual salary if employed (equals wage×hours×weeks)
lsalary	(natural) log transformation of the variable salary
income	1=under \$10,000 per year, 2=10,000-14,999, 3=15,000-19,999 4=20,000-24,999, 5=25,000-29,999, 6=30,000-34,999 7=35,000-39,999, 8=40,000-49,999, 9=50,000+, 99=refused to answer
disloc	people who lost their jobs via mill closings: 1=have been dislocated, 0=not
train	1=participated in job training program, 0=not
monthu	number of months unemployed in previous year
rate	rating of community as place to live: 2=excellent, 1=good, 0=fair or poor

For each problem, paste the R code, edit the output, and answer the questions.

(a) Compare the distributions of the **salary** and the **lsalary** variables using histogram and boxplot.

(b) For the **lsalary** variable, report the mean, median, Q1, Q3, IQR, 35th percentile, minimum, and maximum.

(c) Compute the mean, median, standard deviation, minimum, maximum values of the **wage** for each **location**. Which location has the highest mean? Which location has the lowest variance?

(d) Draw a frequency table, pie chart and bar chart to show the frequency (or percentage) distribution of **educ**. What proportion of the observations is at least high school graduate?

(e) Create side-by-side boxplots to compare the mean of **lsalary** for different **race**. Compare the distribution shapes, centers, variations, and outliers.

(f) Create a two-way table to summarize the **rate** and **educ** together. Among the excellent rate group, what is the percentage of high school incomplete or high school graduate?

(g) Draw a plot that displays the relationship between **lsalary** (response) and **age** and calculate the correlation coefficient between the two variables. Interpret the relationship between the two variables using the plot and the correlation coefficient.

7. (Suggested: no submission) From the exercise problems in Chapter 1:

   3, 5, 6

8. (Suggested: no submission) From the exercise problems in Chapter 2:

   8, 10, 14, 19, 33, 34, 35