**V506 Homework Exercise 1**

**100 total points**

Instructions: Complete the following homework exercise independently. Submit via Canvas. Responses to questions 1-9 should be typed into their own document. The SAS data set you create for question 10 should be saved as *LastName\_FirstName.sas7bdat*.

1. **Variables and their measurement.** For each of the below variables, please state whether the variables are considered quantitative or qualitative (categorical), discrete or continuous, and measured using an interval, nominal, or ordinal scale.

**Note:** There variables that can be considered continuous but are measured and collected in discrete units (such as age), yet they are treated as continuous variables in analysis. Please refer to these variables as continuous.

1. Occupation (plumber, teacher, secretary)
2. Occupational status (blue collar, white collar)
3. Social status (lower, middle, upper class)
4. Race
5. Statewide murder rate (number of murders per 1000 population)
6. County population size (number of people)
7. Population growth rate (in percentages)
8. Community size (rural, small town, large town, small city, large city)
9. Annual income (thousands of dollars per year)
10. Attitude toward affirmative action (favorable, neutral, unfavorable)
11. Lifetime number of sexual partners

**(1\*11 = 11 points)**

1. **Samples and Population.** **(1.16)**

A sociologist wants to estimate the average age at marriage for women in New England in the early eighteenth century. She finds within her state archives marriage records for a large Puritan village for the years 1700-1730. She then takes a sample of those records, noting the age of the bride for each. The average age in the sample is 24.1 years. Using a statistical method from Chapter 5, the sociologist estimates the average age of brides at marriage for the population to be between 23.5 and 24.7 years.

1. What part of this example is descriptive?
2. What part of this example is inferential?
3. To what population does the inference refer?

**(3\*3 = 9 points)**

1. **Statistics and Parameters. (1.2)**

In the 2014 gubernatorial election in California, a CBS News exit poll of 1824 voters stated that 60.5% voted for the Democratic candidate, Jerry Brown. Of all 7.3 million voters, 60.0% voted for Brown.

1. What was the population and what was the sample?
2. Identify a statistic and a parameter.

**(2\*4 = 8 points)**

1. **Description and Inference. (1.8)**

The Current Population Survey (CPS) is a monthly survey of households conducted by the U.S. Census Bureau. A CPS of 68,000 households in 2013 indicated that of those households, 9.6% of the whites, 27.2% of the blacks, 23.5% of the Hispanics, and 10.5% of the Asians had annual income below the poverty level.

1. Are these numbers statistics, or parameters? Explain.
2. A method from this text predicts that the percentage of *all* black households in the United States having income below the poverty level is at least 25% but no greater than 29%. What type of statistical method does this illustrate – descriptive or inferential? Why?

**(3+5 = 8 points)**

1. **Simple Random Sampling. (2.12)**

A local telephone directory has 400 pages with 130 names per page, a total of 52,000 names. Explain how you could choose a simple random sample of five names. Show how to select five random numbers to identify subjects for the sample.

**(6 points)**

1. **Observational versus Experimental Studies. (2.14)**

A study is planned about whether passive smoking (being exposed to secondhand cigarette smoke on a regular basis) lead to higher rates of lung cancer.

1. One possible study would take a sample of children, randomly select half of them for placement in an environment where they are passive smokers, place the other half in an environment where they are not exposed to smoke, and then 60 years later observe whether each person has developed lung cancer. Would this study be experimental or observational? Why?
2. For many reasons, including time and ethics, it is not possible to conduct the study in (a). Describe a way that *is* possible, and indicate whether it would be an experimental, or observational, study.

**(2\*5 = 10 points)**

1. **Sampling Bias. (2.16)**

Explain how the following had sampling bias, and explain what it means to call their samples “volunteer samples.”

1. The BBC in Britain requested viewers to call the network and indicate their favorite poem. Of more than 7500 callers, more than twice as many voted for Rudyard Kipling’s “If” than for any other poem. The BBC reported that this was the clear favorite.
2. A mail-in questionnaire published in *TV Guide* posed the question “Should the President have the Lin Item Veto to eliminate waste?” Of those who responded, 97% said yes. For the same question posed to a random sample, 71% said yes.

**(2\*4 = 8 points)**

1. **Cluster Sampling. (2.18)**

Refer to Problem #5 about selecting 5 of 52,000 names on 400 pages of a directory.

1. Select five numbers to identify subjects for a systematic random sample of five names from the directory.
2. Is cluster sampling applicable? How could it be carried out, and what would be the advantages and disadvantages?

**(2\*4 = 8 points)**

1. **Stratified Sampling (2.30)**

With a total sample size of 100, we want to compare Native Americans to other Americans on the percentage favoring legalized gambling. Why might it be useful to take a disproportional stratified random sample?

**(8 points)**

**Part II (24 points):** The following problems require you to use R to perform the indicated tasks. The data set is available in the “Data Sets” folder in the Canvas /Files/Data Sets section. You should copy the respective files to your workstation or a USB Flash Drive and then use R to import the data. You must submit a copy of the relevant pages of your R output when you turn in the exercise (histograms are relevant). Note that you should not include frequency distributions.

1. Import the “General Social Survey 2008” data set file in to R. Note that the file on Canvas is a Comma Separated Values file or .csv. Take a look at the dataset. You will notice that it contains dozens of variables and thousands of observations. You will only work with a few variables for this assignment, but we will look at this data again in the future. Below are relevant variable definitions, but you can find all variable descriptions in the General Social Survey Codebook.
2. Derive basic descriptive statistics including mean, median, mode, range, variance, and standard deviation for the following variables: **AGE**, **HRS1**, **TVHOURS**,and **EDUC**. (8 points)
3. Generate histograms for the following variables: **AGE**, **HAPMAR**, and **LIFE.** (6 points)
4. What does the histogram for **AGE** tell us? Is there anything distinctive about its shape? (2 point).
5. What do you notice about the histograms for **HAPMAR** and **LIFE**? (2 points)
6. Do people in the survey generally find their lives exciting, routine, or dull? (2 points)
7. How do people describe their marriages? (2 points)
8. Approximately what percent of individuals in the sample are older than 75 years? (2 points)

**Variables Definitions:**

**MARITAL:** Current marital status

1=married, 2=widowed, 3=divorced, 4=separated, 5=never married, 9=no answer

**EDUC:** Years of education

0=no formal education, 1-20=years of school (up to 8 years in university study), 98=don’t know, 99=no answer

**HAPMAR:** Taking things all together, how would you describe your marriage?

1=very happy, 2=pretty happy, 3=not too happy, 8=don’t know, 9=no answer, 0=not married

**LIFE:** Do you find your life exciting, pretty routine, or dull?

1=exciting, 2=routine, 3=dull, 8, 9, and 0= no answer or N/A

**AGE:** In years

**HRS1:** Number of hours worked the previous week (1-89 hours):

98=don’t know, 99=no answer, -1=not employed

**TVHOURS:** Total number of hours watched on a daily basis (0-24):

98=don’t know, 99=no answer, -1=not applicable