**V506 Homework Exercise 2**

**100 total points**

Instructions: Complete the following homework exercise independently. Submit via .doc or .docx format via Canvas.

**Part I:** Problems

1. **Basic Probability Rules.** A GSS estimates the probability than an American adult believes in heaven is 0.85. **(5 total points, 2.5 points for each part)**
   1. Estimate the probability that an American adult does not believe in heaven.
   2. Of those who believe in heaven, about 84% believe in hell. Estimate the probability a randomly chosen adult believes in both heaven and hell.
2. **Probability Distributions for Discrete and Continuous Variables.** Let *y* be the number of languages in which a person is fluent. According to Statistics Canada, for residents of Canada *y* has a probability distribution *P*(0) = 0.02, *P*(1) = 0.81, and *P*(2) = 0.17, with negligible probability for higher values of *y*. **(10 total points, 2.5 points for each part)**
   1. Is *y* a discrete, or a continuous variable? Why?
   2. Construct a table showing the probability distribution of *y*.
   3. Find the probability that a Canadian is *not* multilingual.
   4. Find the mean of this probability distribution.
3. **The Normal Probability Distribution***.* For a study in Aarhus University Hospital (Denmark), 5459 pregnant women who reported information on length of gestation until birth had mean = 281.9 days and standard deviation = 11.4 days. A baby is classified as premature if the gestation time is 258 days or less. **(5 total points, 2.5 points for each part)**
   1. If gestation times are normally distributed, what proportion of babies would be born prematurely?
   2. The actual proportion born prematurely during the period was 0.036. Based on this information, how would you expect the distribution of gestation time to differ from the normal distribution?
4. **Sampling Distribution of the Sample Mean.** The probability distribution associated with the outcome of rolling a balanced die has probability 1/6 attached to each integer, {1, 2, 3, 4, 5, 6}. Let (*y1*, *y2*) denote the outcomes for rolling the die twice. **(12.5 total points, 2.5 points for each part)**
   1. Enumerate the 36 possible (*y1*, *y2*) pairs (e.g., (2,1) represents a 2 followed by a 1).
   2. Treating the 36 pairs as equally likely, construct the sampling distribution for the sample mean y-bar of the two numbers rolled.
   3. Construct a histogram of the (i) probability distribution for each roll, (ii) sampling distribution of y-bar in (b). Describe their shapes.
   4. What are the means of the two distributions in (c)? Why are they the same?
   5. Explain why the sampling distribution of y-bar has relatively more probability near the middle than at the minimum and maximum values. (*Hint*: Note there are many more (*y1*, *y2*) pairs that have a sample mean near the middle than near the minimum or maximum.)
5. **Point and Interval Estimates.** One question on a recent General Social Survey asked, “Do you think that it should be government’s responsibility to reduce income differences between the rich and the poor?” Those answering *yes* included 90 out of the 142 subjects who called themselves *strong Democrats* in political party identification and 26 of the 102 who called themselves *strong Republicans*. **(5 total points, 2.5 points for each part)**
   1. Find the point estimate of the population proportion who would answer *yes* for each group.
   2. The 95% confidence interval for the population proportion of *yes* responses is (0.55, 0.71) for strong Democrats and (0.17, 0.34) for strong Republicans. Explain how to interpret the intervals.
6. **Confidence Interval for a Proportion.** When the GSS most recently asked whether human beings developed from earlier species of animals, 53.8% of 1095 respondents answered that this was probably or definitely not true. Find a 99% confidence interval for the corresponding population proportion, and indicate whether you can conclude that a majority of Americans felt this way. **(2.5 points)**
7. **Confidence Interval for a Mean.** In response to the GSS question in 2014 about the number of hours daily spent watching TV, the responses by the seven subjects who identified themselves as being raised Islamic were 1, 1, 1, 2, 2, 3, 6. **(5 total points, 2.5 points for each part)**
   1. Estimate the mean, standard deviation, and standard error.
   2. Construct a 95% confidence interval for the population mean, specifying its assumptions. Interpret.
8. **Choice of Sample Size.** A 2011 poll in Canada indicated that 41% of Canadians favored bringing back the death penalty for convicted murderers. (The United States is the only Western democracy that has it.). A report by Amnesty International on this and related polls (www.amnesty.ca) did not report the sample size but states, “Polls of this size are considered to be accurate within 2.5 percentage points 95% of the time.” If this is the case, about how large was the sample size? **(2.5 points)**
9. **The Five Parts of a Significance Test**. You want to know whether adults in your country think the ideal number of children is equal to 2, or higher or lower than that. **(7.5 total points, 2.5 points for each part)**
   1. Define notation and state the null and alternative hypotheses for studying this.
   2. For responses in a recent GSS to the question “What do you think is the ideal number of children to have?” software shows results:

Test of mu = 2.0 vs mu not = 2.0

Variable n Mean StDev SE Mean T P-value

Children 1302 2.490 0.850 0.0236 20.80 0.0000

Report the test statistic value, and show how it was obtained from other values reported in the table.

* 1. Explain what the *P*-value represents, and interpret its value.

1. **Significance Test for a Mean.** By law, an industrial plant can discharge no more than 500 gallons of waste water per hour, on the average, into a neighboring lake. Based on other infractions they have noticed, an environmental action group believes this limit is being exceeded. Monitoring the plan is expensive, and a random sample of four hours is selected over a period of a week. Software reports **(7.5 total points, 2.5 points for each part)**

Variable No. Cases Mean StDev SE of Mean

WASTE 4 1000.0 400.0 200.0

* 1. Test whether the mean discharge equals 500 gallons per hour against the alternative that the limit is being exceeded. Find the *P*-value, and interpret.
  2. Explain why the test may be highly approximate or even invalid if the population distribution of discharge is far from normal.
  3. Explain how your one-sided analysis implicitly tests the broader null hypothesis that μ 500.

1. **Significance Test for a Proportion.** Same-sex marriage was legalized across Canada by the Civil Marriage Act enacted in 2005. Is this supported by a majority, or a minority, of the Canadian population? In an Ipsos Global poll conducted for *Reuter News* in May 2013 of 1000 Canadians that asked whether legalization should stand or be repealed, 63% supported legalization. Let π denote the population proportion of Canadian adults who support legalization. For testing *H0*: *μ* = 0.50 against *H­a*: *μ* 0.50 **(7.5 total points, 2.5 points for each part)**
   1. Find the standard error, and interpret.
   2. Find the test statistic, and interpret.
   3. Fine the *P*-value, and interpret in context.
2. **Decisions and Types of Errors in Tests.** Example 6.4 (page 148 in the 5th edition and page 151 in the 4th edition) tested a therapy for anorexia, using *Ho*: μ = 0 and *H­a*: μ 0 about the population mean weight change. **(7.5 total points, 2.5 points for each part)**
   1. In the words of that example, what is (i) Type I error? (ii) Type II error?
   2. The *P*-value was 0.018 (5th edition) or 0.017 (4th edition). If the decision for *α* = 0.05 were in error, what type of error is it?
   3. Suppose instead *α* = 0.01. What decision would you make? If it is in error, what type of error is it?
3. **Limitations of Significance Tests.** A report by the Collaborative on Academic Careers in Higher Education indicated that there is a notable gap between female and male academics in their confidence that tenure rules are clear, with men feeling more confident. The 4500 faculty members in the survey were asked to evaluate policies on a scale of 1 to 5 (very unclear to very clear). The mean response about the criteria for tenure was 3.51 for females and 3.55 for males, which was indicated to meet the test for statistical significance. Use this study to explain the distinction between statistical significance and practical significance. **(2.5 total points)**

**Part II:** The following problems require you to use R to perform the indicated tasks. The data sets are available under the “Data Sets” area in the Canvas Files section. You should copy the respective CSV files to your local workstation, Box account, or USB Flash Drive and then use R to import the data. You must submit relevant pages of your R code and output when you turn in the exercise.

1. Using the SCHOOLS.CSV data set (which contains data on a sample of school districts in Ohio),

1. Find the mean, median, mode, range, variance, and standard deviation for the variables: Number of Students (Students), the Percent of Families Receiving Welfare (Welfare), the Amount Spent per Pupil (Instruct), and Average Instructor Salary (Salary). **(2.5 pts)**
2. Create a new variable called “AmtPerPupil\_StdScore” which is equal to the Amount Spent per Pupil (Instruct) converted to standard scores (i.e., z-scores) and print out the original and new variable values. Then interpret the highest and lowest values for the Amt\_Per\_Pupil\_Std\_Score. **(2.5 pts)**
3. Create a plot where you overlay the normal curve on a histogram for average instructor salary (hint: to do this, you will need to use the curve and dnorm functions in R as demonstrated in the tutorial video). Be sure to add appropriate label axes and other elements necessary to enhance the readability of your graph. After constructing your plot, discuss any similarities or differences you observe between the observed distribution for instructor salary and the normal distribution **(2.5 pts)**.

2. Carry out the following tasks with the data from the Baltimore Longitudinal Study of Aging. This data set is named: BLSA.csv, which is a text file available from Canvas. The variables in this data set are, in order, subject number, sex, age, smoker, systolic blood pressure (SBP), diastolic blood pressure (DBP), height, and weight for a random sample of 720 adults. Ages are given in years, blood pressure readings are in millimeters of mercury, heights are given in centimeters, and weights are given in kilograms. Male gender is indicated with a value of M and female gender with a value of F. A smoker is given a value of Y and a non-smoker a value of N.

1. Construct a 95% confidence interval for the mean diastolic blood pressure of all males. Interpret the results. **(2.5 points)**
2. Construct a 99% confidence interval for the proportion of males that smoke. Interpret the results. **(2.5 points)**
3. Construct a 90% confidence interval for the mean systolic blood pressure of all females who are over 30 years old. Interpret the results. **(2.5 points)**
4. Test the claim, at the .05 level of significance, that males who are 50 and older have a mean systolic blood pressure that is greater than the mean systolic blood pressure for all men. Use the p-value method. **(2.5 points)**
5. Test the claim, at the .01 level, that females who have lower weight than the mean weight of all females also have a lower mean systolic blood pressure than the average female. Use the p-value method. **(2.5 points)**