Assignment 4: Moving towards inferential statistics
This assignment is intended to help you transition from "deductive reasoning" (deducing information about sub samples of the population from information about the entire population) to "inductive reasoning" (learning about the population from information about a sample). In Parts I and II we use old and new skills to make this transition. Then in Part III we return to SPSS and the WSPS data set to use our inductive reasoning skills.

Part I
The number of shirts donated to Goodwill each day is normally distributed with a mean of 3000 and a standard deviation of 1000 shirts.

1. What is the probability that Goodwill receives fewer than 2000 shirts two days in a row (assuming days are independent)?
   \[(x - \bar{x})/s = z\]
   \[(2000 - 3000)/1000 = z\]
   \[z = -1\]
   use z table
   \[.34+.5=.84\]
   \[P <= .84\]
   \[1-(Probability>2000 shirts)=1-.84=.26\]
   \[.26^2=.06\]

2. What is the probability that the average of two random samples will be less than 3000 shirts?
   \[se = s/sqrt(n)\]
   \[se = 1000/sqrt(2)\]
   \[se=706.1068\]
   The standard error here is the standard deviation of the two sample means.
   \[(3000 - 3000)/706.1068 = z\]
   \[z=0\]
   \[P(u<3000)=.5\]

3. What is the probability that the average of two random samples will be less than 2000 shirts?
   \[se = s/sqrt(n)\]
   \[se = 1000/sqrt(2)\]
   \[se=706.1068\]
   The standard error here is the standard deviation of the two sample means.
   \[(2000 - 3000)/706.1068 = z\]
   \[z=-1.41\]
   \[P(u<2000)=.08\]
Part II

1. You have recently conducted an analysis of Goodwill’s jobs program. You find that based on a random sample of 30 participants in Goodwill’s jobs program, the average age of individuals in the jobs program is 45, with a standard deviation of 6 years.

   a. Find a 95% confidence interval for the population mean. 
   Equation for standard error: 
      \( se = \frac{s}{\sqrt{n}} \) 
      \( se = \frac{6}{\sqrt{30}} \) 
      \( se = 1.10 \) 
   
      Equation for confidence interval: 
      \( \bar{x} \pm z * se \) 
      \( 45 \pm 1.96 * 1.10 \) 
      \( 42.84 \) to \( 47.15 \) 

   b. A Goodwill publication states that the mean age in the jobs program is between 43 and 47. Based on the mean collected in your survey, what level of confidence would this range represent? 
      \( \bar{x} \pm z \times se \), and because confidence intervals are symmetric, we know that, 
      \( 2 = z \times se \) 
      \( 2 = z \times 1.1 \) 
      \( 2/1.1 = z \) 
      \( z = 1.82 \) 
      Probability greater than mean = .4656 
      \( .4646 \times 2 = 93.1\% \) 

   c. Last year, a group of Fosters Business students conducted a survey for Goodwill, and found that the average age of individuals in the jobs program was 50. Given your analysis, what was the probability of the true population mean being greater than 50? What might explain this result, even if they sampled the same group of people, with no obvious sources of bias? Explain with 1 or 2 sentences. 
      \( 50 = 45 + z \times se \) 
      \( 50 = 45 + z \times 1.1 \) 
      \( 5 = z \times 1.1 \)
Given our data, this result is very unlikely. However, the normal distribution goes towards infinity in each direction, so it is entirely possible we got this result just by random chance.

2. In analyzing Goodwill’s policies, you notice that the pricing of coats at Goodwill is wildly inconsistent--perhaps an indication of poor standardization of resale pricing across outlets. You collect a random sample of 8 winter coat prices from Goodwill Retail Stores in the Puget Sound area. Your results are as follows.

<table>
<thead>
<tr>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>$24</td>
</tr>
<tr>
<td>$27</td>
</tr>
<tr>
<td>$35</td>
</tr>
<tr>
<td>$30</td>
</tr>
<tr>
<td>$29</td>
</tr>
<tr>
<td>$36</td>
</tr>
<tr>
<td>$36</td>
</tr>
<tr>
<td>$31</td>
</tr>
</tbody>
</table>

a. What is the sample mean and sample standard deviation?

$$\bar{x} = \frac{(24 + 27 + 35 + 30 + 29 + 36 + 36 + 31)}{8}$$

$$\bar{x} = 31$$

$$s^2 = \frac{(24-31)^2 + (27-31)^2 + (35-31)^2 + (30-31)^2 + (29-31)^2 + (36-31)^2 + (36-31)^2 + (31-31)^2}{(8-1)}$$

$$s^2 = 19.43$$

$$s = \sqrt{19.43}$$

$$s = 4.41$$

b. Based on the data, construct a 95% confidence interval about the sample mean. Write a one sentence interpretation of the 95% confidence interval.

$$\bar{x} \pm t^* \left[ \frac{s}{\sqrt{n}} \right]$$

$$31 \pm 1.96 \left[ \frac{4.41}{\sqrt{8}} \right]$$

T distribution, $$\alpha/2 = .025$$ at 2.36

$$31 \pm 2.36 \times 1.56$$

$$27.315 \text{ to } 34.685$$

We are 95% confident that the average price of a coat at Goodwill is between $27.31 and $34.69.

c. Construct a 99% confidence interval. Compare this result to your one from b and
explain the difference.

\[ \bar{x} \pm t \times \left[ \frac{s}{\sqrt{n}} \right] \]

\[ 31 \pm t \times \left[ \frac{4.41}{\sqrt{8}} \right] \]

T distribution, \( \alpha/2 = .005 \) at 3.499

\[ 31 \pm 3.499 \times 1.56 \]

\[ 25.54 \text{ to } 36.45 \]

This interval is wider because it captures the true population mean with greater certainty.

3. The following is data on landowners within 2 miles of an unconventional gas development (UGD) site, broken down into subgroups of those who were paid for mineral rights and those who did not receive payment. The table provides response proportions for two questions: whether individuals support further UGD development and whether they believe more testing of impacts is necessary.

<table>
<thead>
<tr>
<th>Payment</th>
<th>Sample Size</th>
<th>Proportion supporting unconventional gas development (SUPPORT)</th>
<th>Is more testing needed on drilling impacts? (YES)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>72</td>
<td>.4</td>
<td>.6</td>
</tr>
<tr>
<td>Yes</td>
<td>103</td>
<td>.6</td>
<td>.6</td>
</tr>
</tbody>
</table>

a. Check to see whether each of the samples is large enough to use the large sample formula to construct a confidence interval for the population proportion of support for UGD development as well as for if more testing of drilling impacts is needed (HINT: use the rule involving an interval 3 standard errors wide).

\[ s = \sqrt{p \times (1-p)/n} \]

\[ s = \sqrt{(.4 \times (1-.4))/72} \]

\[ s = .06 \]

\[ 0 < .4 \text{ or } .6 \text{ plus or minus } 3 \times .06 < 1, \text{ so large sample okay} \]

b. Assuming each sample represents a random sample from its corresponding population, calculate a 95% confidence interval for the proportion of individuals in the paid and no payment group who believe further testing of impacts is needed. How did increasing the sample size affect your level of confidence regarding the sub-population? Explain in two sentences.

For payment group
\[
\text{se} = \sqrt{p(1-p)/n} \\
\text{se} = \sqrt{0.6 \times 0.4 / 72} \\
\text{se} = 0.058 \\
\bar{p} \pm z \times 0.058 \\
0.6 + 1.96 \times 0.058 = 0.71 \\
0.6 - 1.96 \times 0.058 = 0.49 \\
\text{Between 49 and 71\% of those not receiving payment support further testing of impacts.}
\]

For non payment group
\[
\text{se} = \sqrt{p(1-p)/n} \\
\text{se} = \sqrt{0.4 \times 0.6 / 103} \\
\text{se} = 0.048 \\
\bar{p} \pm z \times 0.048 \\
0.6 + 1.96 \times 0.048 = 0.69 \\
0.6 - 1.96 \times 0.048 = 0.51 \\
\text{Between 51 and 69\% of those receiving payment support further testing of impacts.}
\]

When the sample size increased, the standard deviation of the sample proportion decreased, which resulted in a decrease in the range the 95\% confidence interval about the mean. Thus, as sample size increases, we become more confident about our estimate--assuming the data follows the same distribution as we gather more data points.

c. Calculate 95\% confidence intervals for support of unconventional gas development in each subgroup. Do you think it is likely that the true population mean level of support in these two groups is the same? Why or why not. Explain in two sentences.

For payment group
\[
\text{se} = \sqrt{p(1-p)/n} \\
\text{se} = \sqrt{0.4 \times 0.6 / 72} \\
\text{se} = 0.058
\[ \bar{p} \pm z \cdot \overline{.058} \]
\[ .4 + 1.96 \cdot .058 = .51 \]
\[ .4 - 1.96 \cdot .058 = .29 \]

We are 95\% confident the proportion of those not receiving payments supporting development is between 29 and 51\%.

For non payment group
\[ se = \sqrt{p \cdot (1-p)}/\sqrt{n} \]
\[ se = \sqrt{.6 \cdot .4}/\sqrt{103} \]
\[ se = .048 \]
\[ \bar{p} \pm z \cdot \overline{.048} \]
\[ .6 + 1.96 \cdot .048 = .69 \]
\[ .6 - 1.96 \cdot .048 = .51 \]

We are 95\% confident the proportion of those not receiving payments supporting development is between 51 and 69\%.

It is unlikely the true mean of the support is the same, but there is at least as much as a 2.5\% chance they are the same. Given this, I would be at least 95\% confident they are not the same.

d. How many individuals who did not receive payment would need to be sampled to estimate the true proportion who support more testing of impacts to within .01?
\[ n = Z^2 \cdot (p \cdot q)/B^2 \]
\[ n = (1.96^2 \cdot (.6 \cdot .4))/.01^2 \]
\[ n = 9219.84 \]

Part III

For this part refer back to the WSPS data set and SPSS information from Assignment #1 and #2. This is also an opportunity to make progress on your policy report.

Requirements:
- Think of a hypothesis to test.
- Create a bivariate table and construct confidence intervals for the means of two categories.
- Answer discussion questions.

(Only turn in answers to the discussion questions, not your SPSS output.)
A. Thinking of a hypothesis
For this assignment think of a hypothesis involving 2 variables from the Washington state data set.

To illustrate, let’s hypothesize that people in higher income households are less likely to use eco-friendly commuting modes because they can afford the cost of personal transportation and are willing to pay to avoid the time cost of public transportation, walking, etc. Here we hypothesize that households with different distributions of household incomes will use different transportation modes. I will use mode of transport as the independent or explanatory variable and income as the dependent or outcome variable.

B. Creating a bivariate table
Next we want to explore the relationship between the 2 variables selected. First look at the mean of the outcome variable separately for categories (or ranges) of the explanatory variable. If your independent or explanatory variable is continuous or has many categories, you should transform it into a categorical variable. Specifically, for this assignment, make sure your independent variable is made up of two groups (ie Urban versus rural, low income versus all other incomes).

For example, I transformed the transportation mode variable Q8P3 into a new categorical variable called “Ecocommute_Yes_1”, with only 2 values (1=eco-friendly modes of transport, 0=non-eco-friendly modes of transport).

How do you decide what the categories should be? It depends on how you define your groups. Thus, you should be explicit in your analysis about how you decided to recode your variables and why you chose to do so. Are there logical breaking points in the distribution of the

1 To transform a variable In SPSS:
Go to the menu TRANSFORM >> RECODE >> INTO DIFFERENT VARIABLES.
From the listed variables, choose the variable you want to transform and click on the right arrow.
Type the name of the new variable in the OUTPUT name box on the upper right hand corner and click on CHANGE.
To tell SPSS which values to change, click on OLD AND NEW VALUES.
On the left is a box labeled OLD VALUE. Here you will give one of the old values or range of old values that you want to change. [Click on the top circle to give one value or on one of the other circles to give range. Then type the value in the blank next to the circle.]
Now move to the NEW VALUE box and type your new value in the blank.
Click on the ADD button to put this on the list. You can go through this process for each new value you want to create for your new variable. When you have finished listing all the values, click on CONTINUE.
In the recode window, click OK to finish the process. The new variable will be in the last column of the data spreadsheet and at the bottom of your variable list.
explanatory variable if it is quantitative? Check its distribution to see. Is there a theoretical reason for establishing a threshold?

I recoded the values for van, bus, ferry, motorcycle, bicycle, walking, and work/study at home all as “1.” I recoded the values for car, truck, taxi, and SUV all as “0.” All other categories I recoded as system missing. You may do this differently.

Now look at the mean (or other statistics) of a continuous variable (your response variable) for each category of your discrete independent variable. Make sure the table includes the count in each category. In SPSS, click on ANALYZE>>COMPARE MEANS >>MEANS. Then fill in your continuous variable as the DEPENDENT variable and your discrete variable as the INDEPENDENT variable. I would fill in income (hhinc) as my dependent variable and eco-commute mode as my independent variable. Click on OK to run it. You can also choose other statistics before you run it.

Does it look like the means are different for the 2 categories?

Construct confidence intervals around the point estimate of the mean for each category at the 5 and 10 percent levels of significance (the 95% and 90% confidence intervals). For now, do this by hand. Note that the standard deviation provided by SPSS for each category is your estimate of “s” for that category, so use it to get your standard error (SE). Also, your categories may differ in size, and the SE for the category will have to reflect that.

Next, we will check the confidence intervals. There are many ways of doing this in SPSS.

One way is to select DATA >> SPLIT FILE >>

Select your independent variable as the variable to split the data by, and check COMPARE GROUPS, and click OK.

Then, check your confidence interval for each mean by using ANALYZE >> COMPARE MEANS >> ONE SAMPLE T TEST

Use “OPTIONS” to set the CI level if you want something other than 95 percent.

When you are done with this, be sure to go back into DATA >> SPLIT FILE and reset the file splitting.

For your write up, it may be useful to know what proportion of the population is likely to fall into the each category. Construct a 95% confidence interval around the proportion of households that fall into one of your categories by applying the one sample t test procedure to your explanatory variable. You will not need to split your file to do this.

C. Discussion Questions NOTE: I HAVE INCLUDED SYNTAX SO YOU CAN REPLICATE THESE RESULTS. YOUR WRITEUPS SHOULD HAVE NO SYNTAX.
1. Clearly state your hypothesis connecting your two variables. 
   Individuals in Eastern Washington make less income than individuals in Western Washington.

2. Why do you think this hypothesis is either interesting or important to test? 
   Income between regions is potentially critical to test where social service provision may be needed.

3. What did your confidence intervals for the proportion of households in one category tell you about the population of WA households? 
   In Washington State, between 27 and 28% of residents live in Eastern Washington.

GET
   FILE='wapop10.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
FREQUENCIES VARIABLES=region
   /ORDER=ANALYSIS.

RECODE region (1=0) (2=0) (3=0) (9=0) (4=0) (5=S) (10=0) (ELSE=1)
   INTO eastw.
VARIABLE LABELS eastw 'eastw'.
EXECUTE.

T-TEST
   /TESTVAL=0
   /MISSING=ANALYSIS
   /VARIABLES=eastw
   /CRITERIA=CI(.95).

4. Draw a picture of the confidence intervals for your category means (one picture for both categories). How do they look and why? How would they differ by sample size? How would they differ if you changed the confidence level?

For each category, the confidence interval extends an equal distance on either side of the mean because the normal distribution is symmetric. The confidence interval...
is narrower for western washington because there are more observations in the sample. If either sample had an increased N with the same level of variance, the confidence interval would become narrower. If we changed the confidence interval to 99%, the intervals would become wider.

MEANS TABLES=wagehr09 BY eastw
/CELLS MEAN COUNT STDDEV.

SORT CASES  BY eastw.
SPLIT FILE LAYERED BY eastw.
T-TEST
/TESTVAL=0
/MISSING=ANALYSIS
/VARIABLES=wagehr09
/CRITERIA=CI(.95).

5. Write a short paragraph describing your findings to the client for your policy report.

Your office recently requested research on differences in household wage earnings between eastern and western Washington as part of your program to address social service gaps in the eastern part of the state. As part of the funding program, we need to know how incomes differ between eastern and western Washington. Using a random sample of Washington State Residents collected as part of the Washington State Population Survey in 2010, I have conducted an introductory analysis of the wage gap. Based on the survey, I estimate that between 27 and 28%* of Washington State Residents live in the eastern part of the state. This could be part of why social service provisions are so much stronger in western Washington. However, when analyzing hourly wage earnings, I find that the average wage of eastern Washington residents is only between $21.90 and $24.10 compared to $30.20 and $31.70 for residents of Eastern Washington. This analysis does not demonstrate the reason for this discrepancy, and I recommend future research as to why such a gap exists between the wages of the two groups.

*95% Confidence Intervals Reported