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)?

2. What is the probability that the average of two random samples will be less than 3000 shirts?

3. What is the probability that the average of two random samples will be less than 2000 shirts?

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.

   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?

   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.

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.

   $24  $27  $35  $30  $29  $36  $36  $31
a. What is the sample mean and sample standard deviation?

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

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

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).

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.

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.

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?
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.\(^1\) Specifically, for this assignment, make sure your independent

---

\(^1\) 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,
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 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

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.
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
1. Clearly state your hypothesis connecting your two variables.
2. Why do you think this hypothesis is either interesting or important to test?
3. What did your confidence intervals for the proportion of households in one category tell you about the population of WA households?
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?
5. Write a short paragraph describing your findings to the client for your policy report.