

Moving from Populations to Samples

This assignment is intended to help you make a transition from "deductive reasoning" (deducing information about subsamples of the population from information about the entire population) to "inductive reasoning" (trying to learn about the population from information about a sample). In the first question we use old and new skills, to start making the transition from deductive to inductive reasoning. In the last part, we return to the WA state data set and SPSS with our skills in inductive reasoning.

Part I

In the entire population of water supplies in WA, the concentration of fluoride is known to be approximately normally distributed, with mean 3 ppm and standard deviation 1 ppm.

- a. What is the probability that a random WA water supply has a concentration of 2 ppm or higher? What is the chance that a random water supply contains 1 to 2 ppm?
- b. If you selected a second random WA water supply what is the chance that both water supplies have concentrations of 2 ppm or higher? What is the chance that both water supplies have between 1 and 2 ppm? Remember how to combine probabilities of random draws. Why are these answers different than those in part a?
- c. What is the probability that the average of the two water supplies concentrations is over 2 ppm? Between 1 and 2 ppm? What information about the population and the sample did you need to compute these probabilities? Why are these different than in b? [HINT: the sum of two normal random variables is also a normal random variable! And what do we know about sample means?]]
- d. If you took a random sample of 10 water supplies in the state, what would be the chance that the average concentration for the sample of 10 supplies would be over 2 ppm? Between 1 and 2 ppm? Why are these different than those from part c?
- e. If you knew that the distribution of the sample mean ($n=10$) of water supplies had a mean of 3ppm and a standard deviation of 1ppm, what could you say about the population distribution? If the sample means have the same mean and standard deviation, but were for samples of size 40 water supplies, what could you say? What can you say about the relationship between the population distribution and the sample mean distribution? How does it change with sample size? How does it change with population size?

Part II

1. Pulse rate is an important measure of the fitness of a person's cardiovascular system. The mean pulse rate for all U.S. adult males is approximately 72 heartbeats per minute. A random sample of 21 U.S. adult males who jog at least 15 miles per week had a mean pulse rate of 52.6 beats per minute and a standard deviation of 3.22 beats per minute.
 - 1a. Find a 95 percent confidence interval for the mean pulse rate of all U.S. adult males who jog at least 15 miles per week.
 - 1b. Interpret the interval found in part a. Does it appear that jogging at least 15 miles per week reduces the mean pulse rate for adult males?
 - 1c. What assumptions are required for the validity of the confidence interval and the inference, part b?
2. Past research has clearly indicated that the stress produced by today's lifestyles results in health problems for a sizeable proportion of society. An article in the *International Journal of Sports Psychology* (July-September 1992) evaluates the relationship between physical fitness

and stress. Employees of companies that participate in the Health Examination Program offered by Health Advancement Services (HAS) were classified into three fitness levels: poor, average, and good. Each person was tested for signs of stress. The results are reported for the three groups below:

Fitness Level	Sample Size	Proportion with Signs of Stress
Poor	242	.155
Average	212	.133
Good	95	.108

2a. Check to see whether each of the fitness level samples is large enough to use the large sample formula to construct a confidence interval for the population proportion of exhibiting signs of stress.

2b. Assuming each sample represents a random sample from its corresponding population calculate a 95% confidence interval for the proportion of people with signs of stress for each of the three fitness levels.

2c. Interpret each of the confidence intervals constructed in part b.

2d. How many employees from the "good condition" category would need to be sampled to estimate the true proportion for stress to within .01 with 95 percent confidence?

Part III

For this part refer back to assignment #1 and the WA State data set and SPSS information. This is also a chance to continue thinking about your policy report.

Requirements:

- Think of a hypothesis
- Create a bivariate table and construct confidence intervals for means in 2 or more categories
- Answer the discussion questions (***Just turn in answers, not the printout***)

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 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 are hypothesizing that there will be a different distribution for household income for households with different transportation modes. I'll use income as the outcome or "dependent variable" and mode of transport as the "explanatory variable". In this example, it would also be appropriate to switch this by looking at the proportion of people with variable transportation modes (outcome) for different income categories (explanatory variable).

B. Creating a bivariate table

Next we want to explore the relationship between the 2 variables selected. First let's look at the mean of the outcome variable separately for categories (or ranges) of the explanatory variable. If your explanatory variable is continuous or has many categories, you should transform it into a categorical.¹ For example, I transformed the transportation mode variable Q8P3 into a new categorical variable called "Ecocommute_Yes_1", with only 2 values (1=ecological methods of transport and 0=non-ecological methods of transport). How do you decide what the categories should be? It depends on how you define ecological commuting. I recoded as "1" for ecological commuting methods: van, bus, ferry, motorcycle, bicycle, walking, and work/study at home. I recoded as "0" for non-ecological commuting methods: car, truck, taxi, SUV. All other categories I recoded as system missing. You might have done this differently. 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?

Now look at the mean (or other statistics) of a continuous variable for various categories of a discrete variable. Make sure the table includes the count in each category.

In **SPSS** you can 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'd fill in income (HHINC) as my dependent variable and ecological 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?

What proportion of the population is likely to fall into the ecological transportation category? Construct a 99% confidence interval around the proportion of households that fall into one of your categories. Get the sample proportion by using the N for each category or the mean from an ANALYZE command.

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). 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.

Check your confidence interval for one mean by using

ANALYZE, COMPARE MEANS, ONE SAMPLE T TEST

Use "OPTIONS" to set CI level if you want something other than 95 percent.

C. Discussion Questions

1. What was your hypothesis about your 2 variables?
2. What did your confidence intervals for the proportion of households in one category tell you about the population of WA households?
3. Draw a picture of the confidence intervals for your category means, one picture for both categories. How do they look and why? Why do the confidence intervals differ by significance level? How would they differ by sample size?
4. Is the distribution of your "dependent" variable the same for each of your categories? What does this tell you about your hypothesis?
5. Write a short paragraph describing your findings to the client for your policy report.

ⁱ To transform a variable In SPSS:

Pull down the TRANSFORM menu, choose RECODE, then choose INTO DIFFERENT VARIABLES.

From the listed variables choose the variable you want to transform and click on the right arrow.

Type in the name of the new variable into 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 a the top circle to give one value or on one of the other circles to give or range, then type in the value in the blank next to the circle.]

Now move to the NEW VALUE box and type your new value into 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 in your new variable.

When you finished listing all the values than click on CONTINUE. In the re-code window, click on OK to finish the process. The new variable will be in the last column of the data spreadsheet.