

BIOL 458 Biometry

LAB 3: Sampling, Sampling Distributions, and Point Estimation

Biologists and scientists and scholars in many other disciplines often wish to estimate the value of some attribute for one or more groups of subjects, and then use this information to ask questions about whether the value of the attribute differs from some standard specified in advance, or differs between groups. That part of statistics that deals with obtaining a measure of an attribute is called “estimation,” and the part that deals with using estimates to answer questions is called “hypothesis testing.”

In attempting to obtain an estimate of an attribute of some group of subjects, we must be careful to have clearly in mind the nature and extent of the group we are studying. For example, if we wish to determine the average abundance of fish in lakes that have not experienced a decrease in pH because of acid rain and compare this value to that obtained from lakes that have shown a decrease in pH because of acid rain, we must list all the lakes that fall into each of these groups. If our intent is to include all lakes in the world in our study groups of acidified and non-acidified lakes, listing them all would be very difficult because not all lakes have been monitored to determine if their pH has decreased because of acid rain. If our intent is to estimate fish abundance only for lakes east of the Mississippi River in the USA, then we may be able to categorize each lake concerning whether pH has decreased because of acid rain, since these lakes have been closely monitored. Assuming we are working with lakes east of the Mississippi River, we can define the **populations** whose attributes we wish to estimate by categorizing each lake as having experienced a decrease in pH or not. One population would be comprised of those lakes that experienced a decrease, and the other population would be comprised of those lakes that have not experienced a decrease. A **population** is the statistical name for the whole group of subjects whose attributes we wish to estimate, and about which we may want to test some statistical hypothesis. The use of the term **population** in this case should not be confused with the idea of a **biological population**, although sometimes when we estimate the attributes of a biological population the statistical and biological population we study are synonymous.

An important point to note here is that we have restricted the populations about which we wish to estimate attributes and test statistical hypotheses to only lakes east of the Mississippi River in the USA. Therefore, the estimates we obtain and any hypothesis we test using these estimates only applies to the populations we studied; lakes in the eastern USA. Because the abundance of fish in lakes in the western USA, Canada, or on other continents may be affected by other phenomena besides acid rain, our estimates for fish abundance in lakes in the eastern USA may be poor estimates for fish abundance in these other regions. We see here that one consequence of restricting the definition of our population to be lakes in the eastern USA is that the estimates we obtain and the hypothesis tests we perform are less general in their implications - they only apply to the eastern USA. While we may be tempted to claim that any results we get from our estimates and hypothesis tests may apply to other regions of the world, strictly speaking the statistical results only apply to the populations that were examined.

Even though we have restricted our population to be lakes east of the Mississippi in the USA, there are still hundreds of lakes, particularly if we include those created by humans. The cost of actually visiting each lake and estimating the abundance of fish say in number/m³ of water, or biomass/m³ of water is likely to be prohibitive both in terms of time and money. Therefore, we would like to estimate the average abundance of fish in acidified and non-acidified lakes in the eastern USA by examining a subset of the population. In other words, we would like to collect a **sample** of values of fish abundance from acidified and non-acidified lakes and use these as estimates of the average fish abundance for each of the whole **populations** of acidified and non-acidified lakes. The process of **sampling** then is a set of rules designed to insure that estimates of attributes obtained from examining a subset of the population of interest will be good estimates of the attribute in the whole population. The attribute in the underlying population that we wish to estimate using a sample of subjects drawn from that population is called a **population parameter**, and the attribute we calculate from our sample of subjects is called the **sample estimate** of that **population parameter**.

What rules must we use in selecting lakes to constitute our sample? We must develop a set of rules that insures that each acidified lake has an equal probability of being included in our sample of acidified lakes, and each non-acidified lake has an equal probability of being included in our sample of non-acidified lakes. One way to do this would be to name each lake and place the names of acidified lakes on equal sized pieces of paper in one hat, and do likewise for non-acidified lakes in another hat. We could then mix the pieces of paper thoroughly and draw names, without peeking in the hats. The names drawn from each hat would constitute our sample and we would visit these lakes to estimate fish abundance using a standardized procedure.

What does this process of drawing names from hats achieve that makes us confident that each lake had an equal probability of being sampled? First, the process of drawing pieces of paper from the hats without peeking mimics the process of selecting lakes at **random**. If we insure that we sample our subjects at random from a clearly defined population, we have gone at least half way towards insuring that we place each subject at equal risk of being sampled. This is because random selection guards against any conscious or unconscious tendency to select lakes with low or high fish abundance. In other words, it guards against selecting a **biased** sample. Furthermore, if we sample subjects **independently** and **at random**, we will satisfy the rule that each subject has an equal probability or an equal risk of being sampled. The concept of **statistical independence** is crucial to obtaining good sample estimates, designing good experiments, and performing strong and accurate hypothesis tests. However, the concept of **statistical independence** is sometimes difficult to understand and apply to particular kinds of subjects.

Statistical independence means that selecting one subject to be included in your sample does not make it more or less likely that you will select any other particular subject. For example, to estimate fish abundance in lakes we could pick a particular lake to be our starting lake to begin the process of visiting and estimating fish abundance. We could then select not the next nearest lake, but the next one, and so on to be included in our sample. However, this process will not result in all lakes being at equal risk of being sampled. Lakes close to the first lake we selected will have a higher probability of being sampled than lakes farther away. Hence, this sampling protocol

does not satisfy the assumption that subjects were sampled independently, and therefore our sample estimates of the population parameter will not be a good.

So, making sure that we sample our subjects **independently** and **at random** insures that sample estimates of population parameters will be **unbiased** (not too large or too small) estimates of the underlying population parameters.

But, how is it that a sample of only, lets say for example 2 lakes, out of hundreds could provide a good estimate of fish abundance for this entire population of lakes?

For purposes of illustration let me define a hypothetical population of only 10 lakes with fish abundance expressed in number/m³.

Lakes (A - J)

| | | | |
|----------|-----|----------|-----|
| A | 1.2 | F | 1.9 |
| B | 0.7 | G | 0.9 |
| C | 2.0 | H | 0.4 |
| D | 1.1 | I | 1.3 |
| E | 1.0 | J | 0.7 |

Since these 10 lakes are our population, then we actually know the value of the population parameter, average fish abundance. In this example it is 1.12 fish/m³ of water. In real studies or experiments we do not know the value of the population parameter; we are sampling to obtain an estimate of it. In sampling programs we determine our sample size (usually abbreviated as n) in advance, and select one sample of that size to estimate our population parameter. Because we know the abundance of fish in all of the lakes in this hypothetical population we can actually look at all the samples of 2 lakes to see how well they estimate the underlying population parameter. For a group of 10 lakes taken 2 at a time, there are $(10!)/((10-2)!2!) = 45$ possible samples of $n = 2$ lakes.

Average Abundance of Fish from Hypothetical Population for all Samples of Size $n = 2$

| | | | | | | | | | |
|-----------|------|-----------|------|-----------|------|-----------|------|-----------|------|
| AB | 0.95 | BC | 1.35 | CE | 1.50 | DH | 0.75 | FH | 1.15 |
| AC | 1.60 | BD | 0.90 | CF | 1.95 | DI | 1.20 | FI | 1.60 |
| AD | 1.15 | BE | 0.85 | CG | 1.45 | DJ | 0.90 | FJ | 1.30 |
| AE | 1.10 | BF | 1.30 | CH | 1.20 | EF | 1.45 | GH | 0.65 |
| AF | 1.55 | BG | 0.80 | CI | 1.65 | EG | 0.95 | GI | 1.10 |
| AG | 1.05 | BH | 0.55 | CJ | 1.35 | EH | 0.70 | GJ | 0.80 |
| AH | 0.80 | BI | 1.00 | DE | 1.05 | EI | 1.15 | HI | 0.85 |
| AI | 1.25 | BJ | 0.70 | DF | 1.50 | EJ | 0.85 | HJ | 0.55 |
| AJ | 0.95 | CD | 1.55 | DG | 1.00 | FG | 1.40 | IJ | 1.00 |

Note that if we take the average of these 45 sample estimates we obtain the value 1.12 which is identical to the underlying population value! However, each of the individual sample estimates deviates from the true value of the population to some extent. We can calculate the error of each estimate as the difference between the true population parameter and the sample estimate. Greek letters usually represent population parameters. The underlying population mean is usually symbolized with the Greek letter, μ (mu). The sample estimate of the population mean is usually symbolized with a flat bar over the letter \bar{x} . Therefore, the error of each estimate is $\mu - \bar{x}$.

Error of Estimates for each random Sample of Size $n = 2$

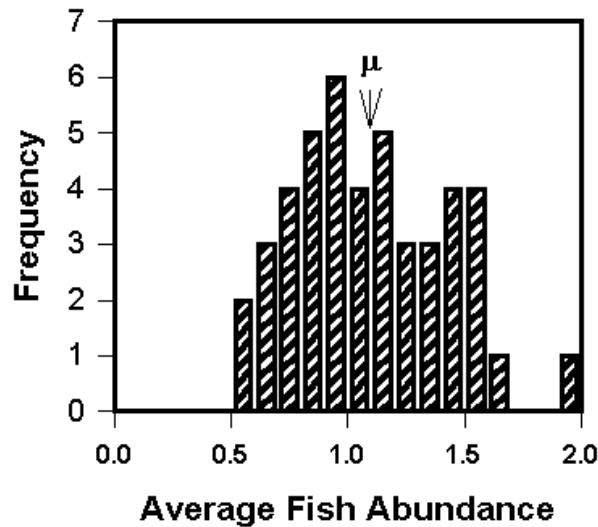
| | | | | | | | | | |
|-----------|-------|-----------|-------|-----------|-------|-----------|-------|-----------|-------|
| AB | 0.17 | BC | -0.23 | CE | -0.38 | DH | 0.37 | FH | -0.03 |
| AC | -0.48 | BD | 0.22 | CF | -0.83 | DI | -0.08 | FI | -0.48 |
| AD | -0.03 | BE | 0.27 | CG | -0.33 | DJ | 0.22 | FJ | -0.18 |
| AE | 0.02 | BF | -0.18 | CH | -0.08 | EF | -0.33 | GH | 0.47 |
| AF | -0.43 | BG | 0.32 | CI | -0.53 | EG | 0.17 | GI | 0.02 |
| AG | 0.07 | BH | 0.57 | CJ | -0.23 | EH | 0.42 | GJ | 0.32 |
| AH | 0.32 | BI | 0.12 | DE | 0.07 | EI | -0.03 | HI | 0.27 |
| AI | -0.13 | BJ | 0.42 | DF | -0.38 | EJ | 0.27 | HJ | 0.57 |
| AJ | 0.17 | CD | -0.43 | DG | 0.12 | FG | -0.28 | IJ | 0.12 |

What is the average deviation of the sample estimates from the underlying population mean, μ ? The average deviation is 0! This is because some of the estimates underestimate the population mean, and some overestimate the population mean. We could also calculate the mean square error (mse) as a measure of how far on average the sample estimates are of the population mean:

$$mse = \frac{\sum (\text{error - of - estimates})^2}{\text{number - of - estimates}} = \frac{4.712}{45} = 0.10471$$

The average deviation of the sample mean from the population mean is then the square root of the mse, which is 0.32359. This indicates that the average error of our estimates of approximately 29% of the population mean. If we had taken samples of size $n = 3$, $n = 4$, or some larger value then our error would have been smaller.

The graph below illustrates the relationship between the distribution of the values of the sample mean for samples of size $n = 2$, and the underlying population mean, μ . The mean of the distribution of the sample values is equal to the underlying population mean. As the size of the sample increases the spread of the values about the population mean becomes narrower, so that the average error associated with any single estimate become smaller.



To answer the question we asked previously, “...how is it that a sample of only 2 lakes out of hundreds could provide a good estimate of fish abundance for this entire population of lakes,” it is because for properly collected samples, the sample mean is an unbiased estimate of the population mean, and its deviation from the true population mean will decrease as the sample size increases. Strictly speaking this result only applies exactly when subjects are sampled independently and at random from a population that is **normally distributed**. However, even when populations are not normally distributed this result is approximately true, and the approximation improves as the sample size increases. The fact that the sample mean remains a good estimate of the population mean even when the population is not normally distributed means that statistical procedures are somewhat **robust** with regard to violation of the normality assumption.

The **Central Limit Theorem** states that: If the sample size is sufficiently large, then the mean of a random sample from a population has a **sampling distribution** that is approximately normal, regardless of the shape of the underlying population distribution. As sample size increases, the sampling distribution will become closer to a normal distribution.

LAB 3 ASSIGNMENT

This week's lab exercise is simple, and should not take you a vast amount of time to complete now that you are becoming comfortable with the system. Thus use your time wisely; review the basics. The upcoming labs will involve more complicated statistical analyses and interpretation. Thus, now is the time to make sure by the end of this lab

that you are completely comfortable with managing **SPSS** files. I want you to be capable of making, saving, editing, and retrieving the **SPSS** output file (.spo) and the **SPSS** data file (.sav). You should know how to edit and print these files. You should also know how to import Text data files into **SPSS** and open **SPSS** data files, including using the **VARIABLE VIEW** tab in the **DATA Editor Window** to declare Missing Values. If you are having problems with these "basics," bring your list of questions to the lab.

The data file for lab 3 is called '[lab3.dat](#)'. It contains 1000 values that you will treat as the "population." While doing the lab, keep the concepts of SAMPLE versus POPULATION in mind. You will examine the differences in sample moments computed from different random samples taken from the same population and relate them to the size of the random samples. The objectives of this exercise are:

- 1) to demonstrate that sample moments such as the mean and standard deviation are random variables,
- 2) to demonstrate that like other random variables the sample mean and sample variance have distributions, and
- 3) to examine the relationship between the sample moments for random samples of size n and the moments of the underlying population for different sample sizes (different values of n).

a) Use the **SELECT CASES** command from the **DATA MENU** to sample 20 random sets (samples) of 5 numbers each. Chose the option under this command to **Select a Random Sample of Cases**. For each set (sample) calculate the mean and standard deviation. The **SELECT CASES** command should be performed with the data filtered rather than deleted. This will result in a filter variable being inserted in the next empty column in the Data Spreadsheet. The data points selected will have values of the filter variable equal to 1, and those not selected will have values equal to 0. Each time after you execute this sampling process calculate the mean and standard deviation using the command **DESCRIPTIVES** in the **ANALYZE MENU**. Calculate the mean and standard deviation for the variable name you used to name the data in column 1, not the filter variable. After obtaining the mean and standard deviation of a sample, click on the filter variable column in the **DATA EDITOR** window and delete the filter variable. You can repeat this process 20 times for samples of size $n = 5$ and 20 times for samples of size $n = 50$. Use the **SELECT CASES** command again to calculate the means and standard deviations for those observations that have filter variables = 1 for each sample individually. **Note:** Most of this lab work can be automated using the syntax window. Paste commands into the syntax window and repeatedly re-execute the set of commands in the **Syntax Window**. I will demonstrate the use of the **Syntax Window** during the lab period.

b) Repeat part a) only this time sample 20 sets of 50 numbers each.

c) Calculate the population mean, standard deviation, and variance. That is, calculate these parameters for the 1000 observations.

d) Plot a histogram of the means of samples with $n = 5$ superimposing a normal curve over the histogram. Do the same for the standard deviations from samples of $n = 5$. Calculate the mean and standard deviation of the means from the 20 random samples of $n = 5$. Do the same for the standard deviations of the 20 samples.

***Note: you are going to have to enter these values yourself.**

e) Plot a histogram of the means of samples with $n = 50$ superimposing a normal curve over the histogram. Do the same for the standard deviations from samples of $n = 50$. Calculate the mean and standard deviation of the means from the 20 random samples of $n = 50$. Do the same for the standard deviations of the 20 samples.

h) Interpret your results. What effect does increasing the sample size from 5 to 50 have on: **1)** the shape of the distribution of the sample means or sample standard deviations, **2)** the estimate of the mean and standard deviation of the distribution of sample means and standard deviations?

i) Turn in:

- 1) E-mail to me your SPSS data file with the 20 values of the mean and standard deviation for samples of size $n = 5$ and for $n = 50$. (You can make do this by using the **Save As** command under the **File Menu** to save your file as a Fixed ASCII file (.dat).)
- 2) The descriptive statistics (mean, standard deviation, and histograms) for each parameter (μ and s) for each sample size ($n = 5, 50$), your interpretation of the results (a paragraph), and include the notes section associated with one instance of each procedure you conduct.

Further Instructions for the Lab Exercise

Lab Exercise 3 requires you to use the **Select Cases** command from the **Data** menu to choose a random subset of 1000 values. Before doing so, you should go to the **Transform** Menu and choose the command **Random Number Generators**. From the sub-window of the **Random Number Generators** command, click on the box to "Set Active Generator." Leave the default selection of **SPSS 12 compatible** as the selected generator. Next, choose "Set Starting Point" and make sure that "Random" is the selected option. Then click "OK".

Now, you can proceed to the **Select Cases** command in the **Data** menu. From the Select cases sub-window, make sure you choose the variable name you are planning to select cases from by clicking on the name. Then click on the option "Random Sample of Cases." Next click the "sample" button just below. In the sub-window that appears click on the radio button next to "exactly," and fill in the appropriate numbers "5" or "50" cases from the first "1000" cases. Before clicking the "OK" button to executing this command, make sure that the radio button next to "Filter out unselected cases" has been selected.

After clicking OK, look at the data in the Data Entry Window. A one will be placed in the first empty column next to each case that has been selected, and a zero next to each case that has not been selected. This column of 0's and 1's is known as a filter variable. Also, a diagonal line will appear through the case number on the left side of the Data Entry Window for each case not selected. All subsequent procedures you run (except another Select Cases command) will use just the subset of cases selected and described by the current filter variable. One can delete the filter variable to have all cases ready to be used for a future command.

There is more than one way to obtain a histogram of a set of values. You could use the **Frequencies** Command from the **Descriptive Statistics** submenu of the **Analyze** menu, or the **Histogram** command from the **Graph** menu. Note that from the **Frequencies** command, one can also obtain descriptive statistics like the mean and standard deviation.

Given that you are asked to repeat the same set of steps many times in this lab, you can somewhat automate the process by using a **Syntax Window** in **SPSS**. When choosing all your options for **Select Cases**, just before you click "OK" to run the procedure, click "**Paste**." This will cause a new window, a **Syntax Window**, to open and all the **SPSS** syntax necessary to execute the **Select Cases** command you just set up will be copied to that **Syntax Window**. Now go to the **Frequencies** or **Descriptives** command in the **Descriptive Statistics** submenu of the **Analyze** menu. Chose the command you wish to use and the necessary options. Just before clicking "OK" to run the command, click "**Paste**" again. This will cause all the syntax required to execute the **Descriptives** or **Frequencies** command you choose to be pasted into the same **Syntax Window**. Now, from the **Syntax Window** in the **Run** menu choose the command **All**. This will cause both the **Select Cases** and the **Descriptives** or **Frequencies** command you pasted into the **Syntax Window** to be executed. New results should appear in the **Output Viewer Window** each time you Click - **Run** and then **All** in the **Syntax Window**. Hence, after 20 clicks you should have the data you need to proceed with your summary and analysis for either your sample size of 5 or 50. You can repeat this entire process to create a similar **Syntax Window** with the appropriate syntax for a sample size of 50. However, you could also just edit the syntax you pasted for samples of size 5. I think line 3 of the syntax you created just needs to be changed to a "50" to enable you to generate random samples of size 50 and obtain their means and standard deviations. Give it a try.

Finally, there is no easy way to enter your results back into **SPSS** other than manually typing in the mean and standard deviation for each random sample of size 5 and of size 50. So, create a new data file with column 1 the means of samples of size 5, column 2 the standard deviations of samples of size 5, column 3 the means of samples of size 50, and column 4 the standard deviations of samples of size 50. In addition to conducting your analysis of these data to complete the lab write-up, e-mail your data file (with the above described structure) to me.