Sampling Regression Lines Using an On-line Simulation

0. Go to the following web site: http://www.rossmanchance.com/applets/regcoeff/regcoeff.html

1. First you need to define the population of ordered pairs from which you will sample. Begin with a population for which the true regression model is $\mu_Y = 1 + 2x$, that is $\alpha = 1$ and $\beta = 2$, with $\sigma = 3$. Edit the values of *pop slope*, *pop intercept*, and *sigma* appropriately. We also have the opportunity to set the mean and standard deviation of the *x*-values. Let's begin with *x mean* = 10 and *x std* = 2. Now click on Set Population and you will see part of the population of ordered pairs from which you will be sampling.

2. You need to resize the window to see the entire population. Note that there are four boxes on the edges of the graph. Click on the rightmost box and change this value to 20; click on the uppermost box and change this value to 40. Press Enter for the changes to take effect. The yellow line superimposed on the scatter plot is the true regression line, $\mu_Y = 1 + 2x$.

3. Set *sample size* = 5 to investigate the sample regression lines that result from samples of 5 ordered pairs from this population. Each time you click on *Draw Samples*, you will see the five ordered pairs that comprise your sample, your sample regression line will appear on the scatter plot in red, and the equation of the sample regression line will be displayed below the graph. Click on Draw Samples at least 10 times. Record the sample slope value for each sample. What is the minimum value you obtained? What is the maximum value you obtained? Is the variability in your slopes larger than you would have expected? Why?

4. Continuing with the settings above, now set *num samples* to 100 and click on *Draw Samples*. After 100 lines have been computed based on samples of 5 points, the population will disappear and you will see the red sample lines on the graph with the population line. You also will see dotplots showing the 100 sample values for *a* and *b*. Describe the sampling distribution of your 100 sample slopes. Record the mean and standard deviation of your 100 sample slopes resulting from $n = 5$.

5. Does it appear that the sample slope *b* is an unbiased estimator of the population slope β? Explain why or why not.

6. What information does the standard deviation of the sample slopes provide? Be specific.

7. Investigate the effect of increasing the sample size on the sampling distribution of the sample slopes. Write a few sentences describing how you investigated the effect of increasing sample size and what you discovered.

8. In addition to sample size, there are two other quantities that affect the variability of sample slopes. One of these is sigma. What information does sigma provide? How do you think the variability of the sample slopes will change if we decrease sigma? How do you think the variability of the sample slopes will change if we increase sigma?

9. Investigate the effects of decreasing and increasing sigma on the sampling distribution of the sample slopes. Write a paragraph describing how you investigated the effect of sigma size and what you discovered.

10. In addition to sigma and sample size, the spread of the x-values in your sample will also impact the variability of the sample slopes. How do you think the variability of sample slopes will be affected if

there is little spread in the x-values? How do you think the variability of sample slopes will be affected if there is greater spread in the x-values?

11. You do not have direct control over the x-values in our sample, but changing the value of *x std* will result in changing the spread of x-values in the sample. At this point the value of *x std* is 2. Restore the value of sigma to 3, and change the value of *x std* to 1. Investigate the effect of decreasing the value of *x std* on the sampling distribution of the sample slopes. Then change the value of *x std* to 3 and investigate the effect of increasing the value of *x std* on the sampling distribution of the sample slopes. Write a paragraph to describe how changing the spread of *x*-values affects the sampling distribution of sample slopes.

12. Suppose a friend is designing an experiment to analyze the relationship between the amount of nutrients given to shrimp and their weight at maturity. Based on preliminary studies, you expect to see a linear relationship between amount of nutrient and weight. If you fit a regression line to the ordered pairs you obtain (using nutrient as the explanatory variable), what information will the slope provide?

13. If your friend plans to report a confidence interval for the true slope, and he wants this interval to be as small as possible with a high level of confidence, what advice would you offer him as he designs his experiment?