Sampling Distribution for the slope of a regression line

March 23, 2022

Do students who sit in the front rows do better than students who sit farther away? A teacher randomly assigned 30 students to seats at the beginning of the semester and then recorded their exam scores at the end of the semester. Here are the results:

First Row Students	76	77	94	99	88	90
Second Row Students	83	85	74	79	77	79
Third Row Students	90	88	68	78	83	79
Fourth Row Students	94	72	101	70	63	76
Fifth Row Students	76	65	67	96	79	96

1. Is this an observational study or an experiment?

Solution: Experiment because treatment is being exposed.

2. How many variables are we measuring? Are they quantitative or categorical. What is the explanatory variable? What is the response variable?

Solution: We are measuring two quantitative variables. The explanatory variable is the row and the response variable is the score.

3. Using your calculator create a scatterplot of the data.

4. Find the least squares regression line.

Solution:

$$\hat{y} = 85.95 - 1.517x$$

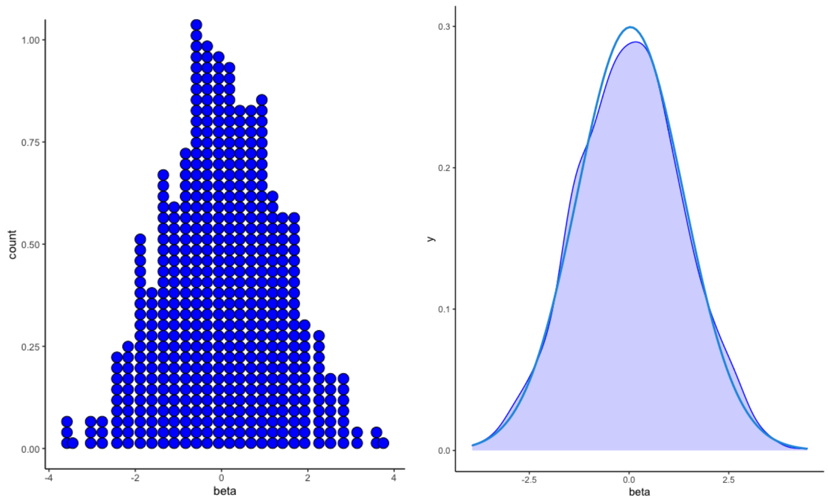
5. What is the slope of the LSRL? Interpret the slope in the context of the problem.

Solution: For each additional row, a student's predicted score goes down by 1.517 points.

Does the negative slope provide convincing evidence that sitting in a lower row causes higher exam scores, or is it plausible that the association is purely by chance because of random assignment?

In order to answer this question, we need to know more about "purely chance because of random assignment". If we assume that seat location has NO effect on exam score, then we could just randomly assign all 30 exam scores to each of the seat locations. We will do this by writing each of the 30 exam scores onto an index card, shuffle the index cards, and then randomly assign them to seat locations.

The sampling distribution for 100 random shuffles (dotplot) and 10000 shuffles are shown below:



The distribution for our random simulations is normal with a mean of $b = 0$ and a standard deviation of σ_b. We could use this sampling distribution to see where our data falls ($b = -1.517$).

It turns out that our sample slope for the regression line ($\hat{\beta}_1$ or b) is normally distributed with a mean of $\mu_b = \beta_1$ and a standard deviation of $\sigma_b = \dfrac{\sigma}{\sigma_x \sqrt{N}}$. This means that if the true value of $\beta = 0$, we get a distribution resembling the ones above.

6. Looking at the simulated plot and the result from our experiment, would you say that it looks like our slope is statistically significant?

7. You may have heard that your nose and ears grow through your whole life. While it is true that your nose and ears get bigger throughout life, it's not because they grow, but because of gravity. The cartilage in your nose and ears break down as we age and the "growth" people observe is the result of drooping. To quantify the expansion of ears over time, a random sample of 30 adults were selected. For each adult, their age (in years) was recorded and their ear height (cm) was measured. Below is the regression output. Is there evidence of a positive linear relationship between age and ear height? Assume the conditions for inference are met.

   (a) What is the estimate for β_0? Interpret this value.

   > **Solution:** $a = 2.8871$. For people that are 0 years old (newborns) we predict their ear height to be 2.8871 cm, on average.

   (b) What is our estimate for β_1? Interpret this value.

   > **Solution:** $b = 0.0021$. For each increase of 1 year in age, we predict the ear height to increase by 0.0021 cm.

   (c) What is the estimate for σ? Interpret this value.

   > **Solution:** $s = 0.3613$. The actual ear heights typically vary by about 0.03613 cm from the predicted ear height.

   (d) Find the standard error of the slope estimate. Interpret this value.

   > **Solution:** $SE_b = 0.0059$. The slope of the sample LSRL typically varies from the slope of the population regression line by about 0.0059 cm per year.