1 Inference for Quantitative Data: Slopes

1.1 Sampling Distributions and Confidence Intervals for Slope

An "ideal" linear relationship can be described with a population regression line: $\mu_y = \alpha + \beta x$, where μ_y represents the mean value of the response variable y for any given value of the explanatory variable x.

α represents the population y-intercept and β represents the population slope.

An observed linear relationship can be described with a sample regression line: $\hat{y} = a + bx$.

If we took many LSRLs of the same size from the sample population, we can create a sampling distribution for our slope.

For a bivariate population with a given slope β, a standard deviation of residuals σ, and a standard deviation of x-values σ_x.

If you take all samples of size n and compute the slope of each of thoes samples, you get the sampling distribution:

- Shape: The distribution of sample slopes is approximately normal
- Mean: $\mu_D = \beta$
- Standard Deviation: $\sigma_b = \frac{\sigma}{\sigma_x \sqrt{n}}$, where σ is the standard deviation of the residuals, σ_x is the standard deviation for explanatory variable, and n is the sample size.

Once we develop a sampling distribution for our slope, we can begin to ask and answer our inference questions:

- Is there a linear relationship between x and y in the population, or could the pattern we see happen just by chance?
- In the population, how much will the predicted value of y change for each increase of 1 unit in x?

Conditions for Regression Inference

- Linear: x and y have a linear relationship. Check: make sure scatter plot can be described by a line.
- Independent: If sampling without replacement, check the 10% condition
- Normal Residuals: When x is fixed, y follows a normal distribution. Check - Make a histogram of the residuals and make sure it looks approximately Normal. If the graph has outliers or strong skewness, n should be larger than 30.
- Equal SD: Standard deviation of residuals doesn't vary with x. Check - Make a residual plot and check for a random pattern
- Random: Random sampling (SRS) or random assignment (experiment)

Together, this makes up the acronym, LINER, which can help you remember what conditions to check when creating a confidence interval or running a significance test.

A $C\%$ confidence interval is created to estimate the slope β of the population (true) regression line

$$b \pm t* \left(\frac{s}{s_x \sqrt{n-1}} \right)$$

with $df = n - 2$.

- $t*$ has $C\%$ of the area between $-t*$ and $t*$
- b is the point estimate (slope from our sample data)
- $SEb = \frac{s}{s_x \sqrt{n-1}}$ is the standard error of the slope

- s_x is the standard deviation of the x-values
- s is the standard deviation of the residuals

Interpretation: We are $C\%$ confident that the interval from (the interval) captures the true slope of the regression line between the x-variable and y-variable

1.2 Significance Test for a Slope

- The significance test for a slope is called a Linear Regression t-Test for Slope
- It can help us answer three different questions with the hypothesis.

Is the relationship between the explanatory and response variable negative?

- $H_0 : \beta = 0$
- $H_A : \beta < 0$

Is the relationship between the explanatory and response variable?

- $H_0 : \beta = 0$
- $H_A : \beta \neq 0$

Is the relationship between the explanatory and response variable positive?

- $H_0 : \beta = 0$
- $H_A : \beta > 0$

COnditions for Regression Inference: the same as above

The test statistic is as follows:

$$t = \frac{\text{statistic} - \text{parameter}}{\text{standard error}}$$

$df = 2$, and the p-value is calculated using your calculator, in the direction of the alternative hypothesis: $p - value =$ tcdf(Lower, Upper, df)

In your conclusion, you would state the results of your significance test (reject or fail to reject) and then interpret the findings in context

Note: Having a low p-value and finding evidence of the alternative hypothesis of some linear association does not mean the association is strong