Resource: Seeing Theory

Seeing Theory is a site developed by Brown University that includes many interactive tools that help to visualize statistical concepts. Throughout the course, we will be talking about many of these concepts. You may find some of the visualizations from Seeing Theory to be of interest!

The site is broken into six chapters, which includes:

1. Basic Probability
2. Compound Probability
3. Probability distributions
4. Frequentist Inference
5. Bayesian Inference
6. Regression Analysis

Note that the site is under constant development and improvement so the layout may change and chapters may be added or changed.

In each chapter, there are 3 parts. Some of the tools we will be touching upon throughout this course and in future courses include:

Chapter 3: Probability Distributions

Discrete and Continuous - This visualization gives an idea of what some of the common distributions look like:

- Select **discrete or continuous random variable**
- From the dropdown menu, select a **distribution** (these are some of the more common distributions that are used in statistics)
- The parameters that dictate the distribution will then be shown below the dropdown menu. They can be changed, and the visual on the right will alter based off those values.
- The visual on the right can be zoomed in or out to have a larger range of X and Y values
- The slider at the bottom shows the cumulative distribution up to a certain X point.

Central Limit Theorem - A distribution of our population can be set by varying the α/β values. These parameters are for a Beta (α, β) distribution that takes a variety of forms based on these 2 parameters.

- The sample size can be set, the larger it is the more bell-shaped (approximately normal) and narrower will be the resulting distribution for the sample means obtained from repeated samples taken from the population.
- The "**draws**" allows us to get multiple samples at a time
- The "**theoretical**" puts an overlay of what we would expect to see for the distribution of the sample mean.
- Sample away!

Chapter 4: Frequentist Inference

Confidence Interval - This visualization allows us to see a sample being drawn from a population and then a confidence interval is made to give a reasonable range of values for estimating our population mean.

- **Set a distribution** of the population, a sample size (n), and a confidence level ($1-\alpha$)
- Samples will be drawn from the population and a confidence interval for estimating the population mean will be made. If the CI has the true population mean in it then it will be colored green, if not then the color of the interval is red.
- Over time we would expect to see about ($1-\alpha$)% of the confidence intervals to contain the true population mean (to be green).

Chapter 6: Regression Analysis

Ordinary Least Squares - From this visualization we are shown that we can have data for which the relationships (via the scatter plots) look very different but the resulting regression lines are very similar

- Choose one of the quartlets and note how the regression line for each looks the same even though the relationships via the scatter plots of the data are drastically different

Correlation - The strength of a linear relationship can be summarized with the correlation. In the visualization a famous data set, the iris flower data set, is examined.

- **Select** 1 to 3 of the species of flower
- The correlation table below will fill with value depending on what species are selected
- Clicking on a certain box of the correlation table allows us to view the 2D scatter plot of the corresponding variables (ex: sepal length vs sepal width)

Analysis of Variance - also known as ANOVA, this visualization is essentially side by side boxplots where we can visually compare the centers (medians) and spreads (IQR and ranges) across groups, to see if the centers are similar or not. Note the column labeled "F" provides the test statistic value and the column labeled "p" provides the corresponding p-value which are used to assess if the population means are equal or is at least one different. You are able to see how these values change with the different data set visualizations

- **Choose** a data set to analyze
- **Compare** the centers in the visual on the right to certain F and p values of the table

Article: Jerzy Neyman on Population Inference

The article below is widely considered to be the seminal work by Jerzy Neyman that laid the foundation for making population inferences based on a single probability sample.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society, 97*(4), 558-625. doi:10.2307/2342192