# Capstone #1 Project - Inferential Statistics Report

Pawandeep Jandir

**Introduction:** A part of the Capstone project, this report sums up the inferential statistics steps we took to investigate the cleaned Combine dataset even further. As a reminder of the project, the goal of the analysis uses NFL Scouting Combine data of drills, tasks and more to predict which draft round a player might get drafted. This report has an accompanying jupyter notebook (with the same name) which shows the full code, plots, and calculations described in this document.

**Exploration:** In the previous portion of this project (the data story), we focused on the quarterbacks (QB) of the dataset. We had concluded that jupyter notebook with two jointplots, one of which plotted the draft round vs. broad (jump). This showed the correlation between the two variables. Since we want to predict the draft round variable, correlations between that and the other variables in the dataset is necessary. However, the draft round is a categorical variable, even if it ranges from -1 to 7 (no zero). This means the usual correlation calculation, the Pearson correlation coefficient, is not truly appropriate to use this case. Instead we can use the Kendall rank correlation coefficient to determine correlation. This test measures the ordinal association and is better suited for our purposes than the Pearson correlation coefficient.

As a statistical test, the (Kendall) correlation also gives us a p-value. The null hypothesis in this case is that there is no correlation between the two variables being tested. So a low p-value suggests the calculated correlation is significant. We can assume the usual statistical significance level of 5%. However, in order to really trust the result of this test we can perform a permutation test of the correlation. This permutation test randomly exchanges the draft round labels and re-calculates the correlation. This permuted correlation is calculated many times to develop a histogram (see Fig. 1). From this we can determine a permuted, or estimated, p-value by dividing the number of permuted correlations greater than the observed correlation by the total number of permutations done. This estimated p-value can then be compared to the observed p-value to determine which correlations are significant and which ones are not.

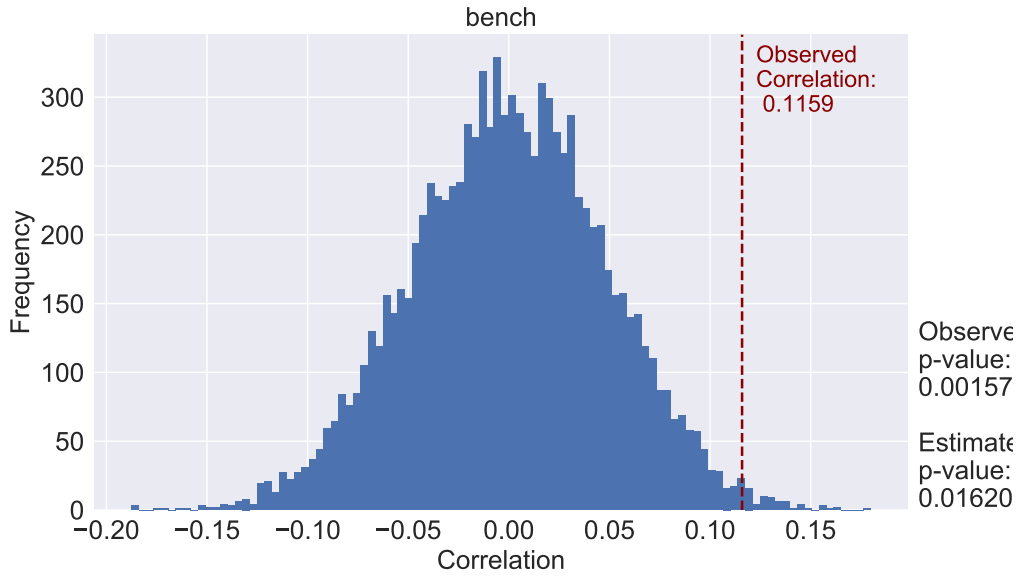


Figure 1: Permuted correlations.

After performing this test across the 12 variables of interest, only the height, bench and nflgrade columns show significant correlations to draft round. The bench correlation test is shown in Fig. 1. However, it must be noted that the highest measured correlation is about 0.12, which is quite weak. Regardless, even if they are weak correlations, they still seem to be real, significant ones.

The previous test focused on all of the QBs. What if we separate the QBs who were drafted from those who went undrafted? Are there differences in these types of players? For this test, we can measure the difference in means of each variable of interest. Then we can calculate a $z$-test and its associated p-value to get the statistical significance. The null hypothesis in this test is that the difference in means between the drafted and undrafted QBs is zero. A small p-value (i.e. $\alpha = 0.05$) implies that we can reject the null hypothesis. When performing this test, we determine that in most cases the p-value is smaller than 5%. There are only three variables with a p-value larger than 0.05: threecone, arms, and hands. Therefore, for these columns, we cannot reject the null hypothesis. However, in all others, we likely can. This means in general, there is a statistical difference between QBs that are drafted and those that are not drafted. An example variable of interest, the shuttle (run) is shown below.

```
   shuttle | z-test statistic :-3.125969
           | z-test p-value   : 0.001772
           | 95% Conf Int     :-0.088894 to -0.020380
-----------|
```

We can expand our focus beyond just QBs and include all players as well. By dividing players into their positiongroups, we can compare how the distribution of a variable of interest is different in any two positiongroups. We can again use a $z$-test to get the significance in the difference in means of the distributions. For example, we can compare how a defensive lineman's (DL) bench press distribution agrees with a linebacker's (LB) bench press distribution. A $z$-test and its p-value can help us determine whether there is a statistical difference or not. We can go further and loop through all combinations of

positiongroups for each variable of interest. The result is a heatmap of the p-values of the $z$-tests mapping each of these possibilities. In this type of plot, a low (light) value implies statistically signficant, while a high (dark) value implies no significance. The null hypothesis here is that the difference in means is zero. Remember each variable of interest generates a separate heatmap. The max color scale value is capped at 0.1 to ensure all dark colors land outside the $\alpha$ level we have set (at 5%). This heatmap is also symmetric by construction. An example of such a heatmap is given in Fig. 2.

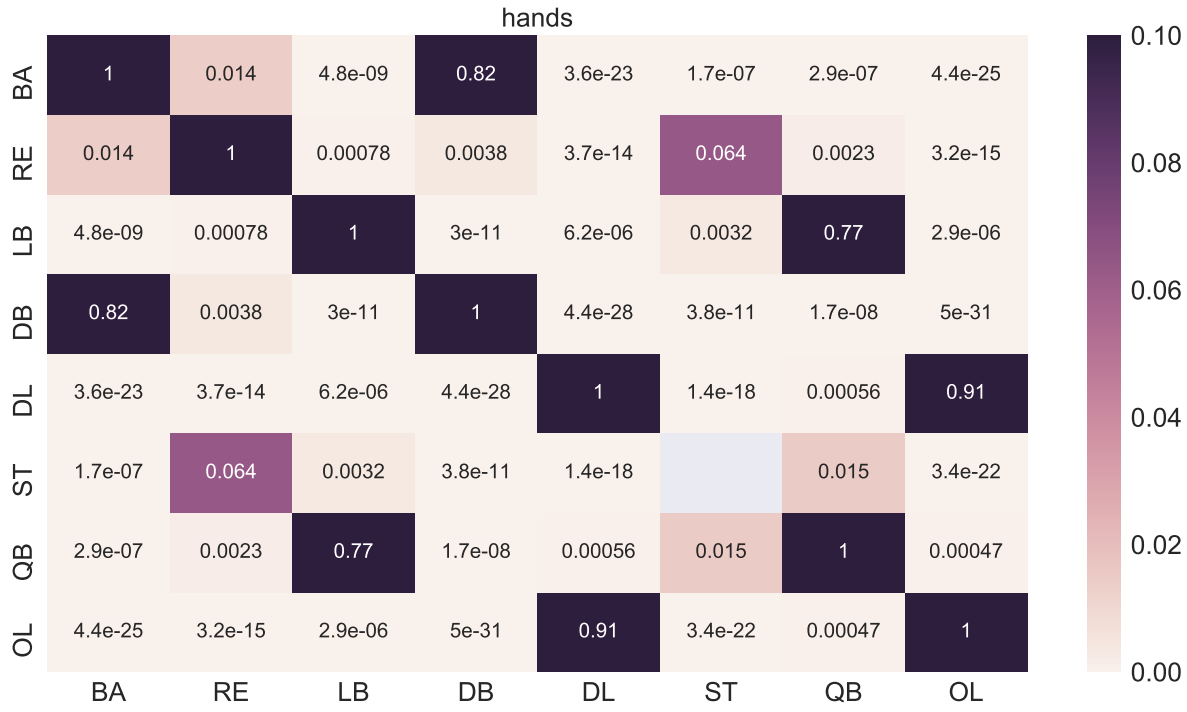


Figure 2: P-values.

From this example, we can see how pairs of positiongroups interact with each other for the hands variable. Offensive lineman (OL) and defensive lineman share a very similar mean as it has a p-value of 0.91. However the hands means are very different for OL and QBs as the p-value is 0.00047, well below our 5% significance threshold.

**Conclusion:** While some of the correlations are not as strong as we might have hoped, there are certainly indicators that the differences in the variables in the dataset are statistically different. This means there should be ways to leverage that information to make predictions about players.