

ISyE 2028 – Basic Statistical Methods - Fall 2015

Bonus Project: “Big” Data Analytics

Proposal (or Final Report)

NFL Draft Prospects:

A study on predicting top NFL drafts

Bhavna Choudhury

Georgia Institute of Technology

November 2015

Introduction

American Football plays a very large role in American media and culture. American football as a whole is the most popular sport in the United States. In the United States, professional football and college football are the most popular forms of the game, with the other major levels being high school and youth football. As of 2012, nearly 1.1 million high school athletes and 70,000 college athletes play the sport in the United States annually. The National Football League, the most popular American football league, has the highest average attendance of any sports league in the world; its championship game, the Super Bowl, ranks among the most-watched club sporting events in the world, and the league has an annual revenue of around \$10 billion.

Besides the Super Bowl and other high profile games, one of the most watched events of the sport is the National Football League (NFL) Draft. The draft is an event where professional football teams select some of the top college prospects to be a part of their team for the next season. However, the draft has a greater impact than only on the professional teams and the college prospects. In 2014, the NFL Draft was covered by the NFL Network, ESPN, and ESPN2 over the course of 3 days where 45.7 million people watched. With an event getting so much press and coverage, many fans look to predict the outcomes of the Draft, interested in who their team will draft and which select athletes will be the coveted first round draft picks. We hope to use some data about the top fifty NFL Draft prospects to get a better understanding of trends or patterns that can help make these predictions more precise and accurate. We would like to ask what the qualities of a player are that makes him a top prospect. And more importantly, what benchmarks makes an athlete worthy of being picked in the first round of the draft?

There exists various statistics that rank players with a heavy weight on different factors, including However, very few combine the statistics available and look at the performance of airports on a larger scale. As a result, a few questions remain: Is there a relationship between different kind of statistics on airport performance? If so, which factors may be influenced by other factors? Which are the most dependent on others and which are completely independent of others

Name	Concept	Variable Type	Possible Range
h2015	Height of NFL drafts (2015) in feet	Ordinal/Ratio	-(4, 9); -For this variable, the minimum possible height is 4 feet (more than midget status) and the maximum is a conservative estimation
hAll	Height of Hall of fame NFL players	Ordinal/Ratio	- (4, 9); -For this variable, the minimum possible height is 4 feet (more than midget status) and the maximum is a conservative estimation
weight	Weight of NFL drafts (2015) in pounds	Ordinal	-(100,400) -This is a conservative estimate for both the minimum and maximum values for weight - Lightweight, Midweight, Heavyweight classes ranked in order.
fortyT	Combined 40 yard dash time of NFL drafts (2015) in seconds	Ratio	-(0, 100) -The minimum 40 yard dash time cannot be less than zero seconds and the maximum is a conservative estimation

fortyA	Combined 40 yard dash time of Hall of Fame NFL players	Ratio	-(0, 100) -The minimum 40 yard dash time cannot be less than zero seconds and the maximum is a conservative estimation
threeCone	3- Cone drill timings in seconds of NFL drafts (2015)	Ratio	-(0, 100) -The minimum 3-cone drill time cannot be less than zero seconds and the maximum is a conservative estimation
Race	Ethnicity of the player	Categorical	-{Black, White, Other}

Table 1: Variable List

Data Distribution and one Sample Analysis:

threeCone is a ratio variable representing the three Cone drill timings of the NFL drafted players for 2015. The three-cone drill, or 3-cone drill is a test performed by American football athletes at the NFL Combine. It is primarily run to evaluate the agility, quickness and fluidity of movement of players by scouts. The sample size for this data set (n) = 50. The numerical values of the data and the conceptual meaning of the data's maximum, minimum, and mean (**the range and calculated average for this year's prospects three Cone drill timings**) indicates that this is a ratio variable. Since the data is measuring time, there are no units and the data less than 0 and it is assumed that the maximum is 100 seconds. The data set had to be cleaned since the data for a few of the players was missing.

Over the years, there have been many different types of drills tried out to test the performance of NFL players. Although the 40 yard dash time was originally considered to be the only important pre selection test there has been an increasing emphasis on agility, quickness and fluidity. Thus it is important to analyze the three cone drill.

The minimum of the cleaned data is 6.710 seconds and the maximum is 8.010 meaning that the average top 50 predicted drafted NFL player should be able to finish the drill in that range of time of [6.710, 8.010]. The first quartile of the data is 6.982 meaning that 25% of the data, a

quarter of the lie between 0 and 6.982. The median of the data is 7.16 seconds and the third quartile value is 7.5 seconds. The sample mean is 7.248 meaning that the top 50 NFL drafted prospects have an average run time of the aforementioned. The interquartile range is 0.565 and the sample standard deviation is 0.3672479 and the sample variance is 0.134871. Variance and median are the best measures of this data. Variance describes how the data varies greatly and is not precise and consistent. Median more accurately captures the middle data because the mean is greatly influenced by outliers even though the data set has been cleaned.

Numerical summary of threeCone:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
6.710	6.982	7.230	1206.000	7.695	9999.000

Numerical summary of good3Cone:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
6.710	6.945	7.160	7.248	7.510	8.010

Standard Deviation

0.3672479

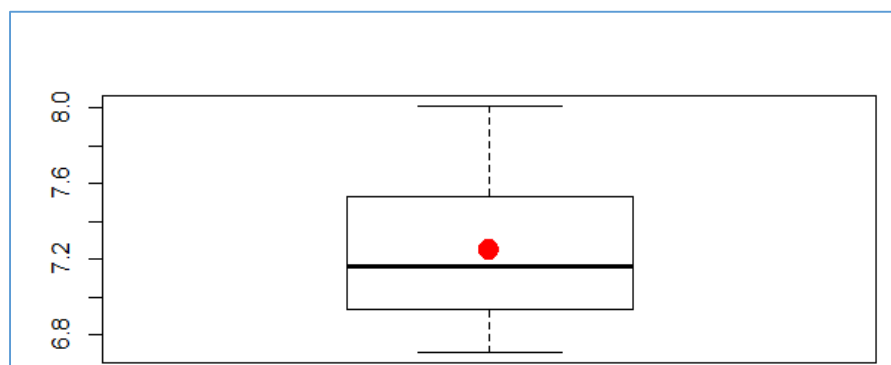


Fig 1. Boxplot of the cleaned data set for 3-Cone drill times with mean in red

$1.5 * IQR = 0.8475.$

$Q3 + 7.510 = 8.35.$

From the boxplot, one can see that the Inner Quartile range is 0.565 seconds so the box part is quite condensed. This means that the median and mean are relatively close. The red dot shows the mean and the middle bar shows the median. Small IQR means Q1 is near Q3 and the data in the box is concentrated around a small proximity, again indicating that the mean \sim median. Since the whisker reaches to the largest data point in this range it ends at 8.010. Likewise, on the lower end, the whisker reaches to the smallest data point 0. The box plot shows that if a normal distribution would be an appropriate model of this data with outliers on both ends of the box plot, both the right and left tails would be long to show a large variance. We decided to leave out the outliers on the higher end because they are only at most 14 more/less than the largest/smallest data point in the range within the whisker. After analyzing the boxplot, a QQ Plot was decided upon to see if the data did indeed model a normal distribution.

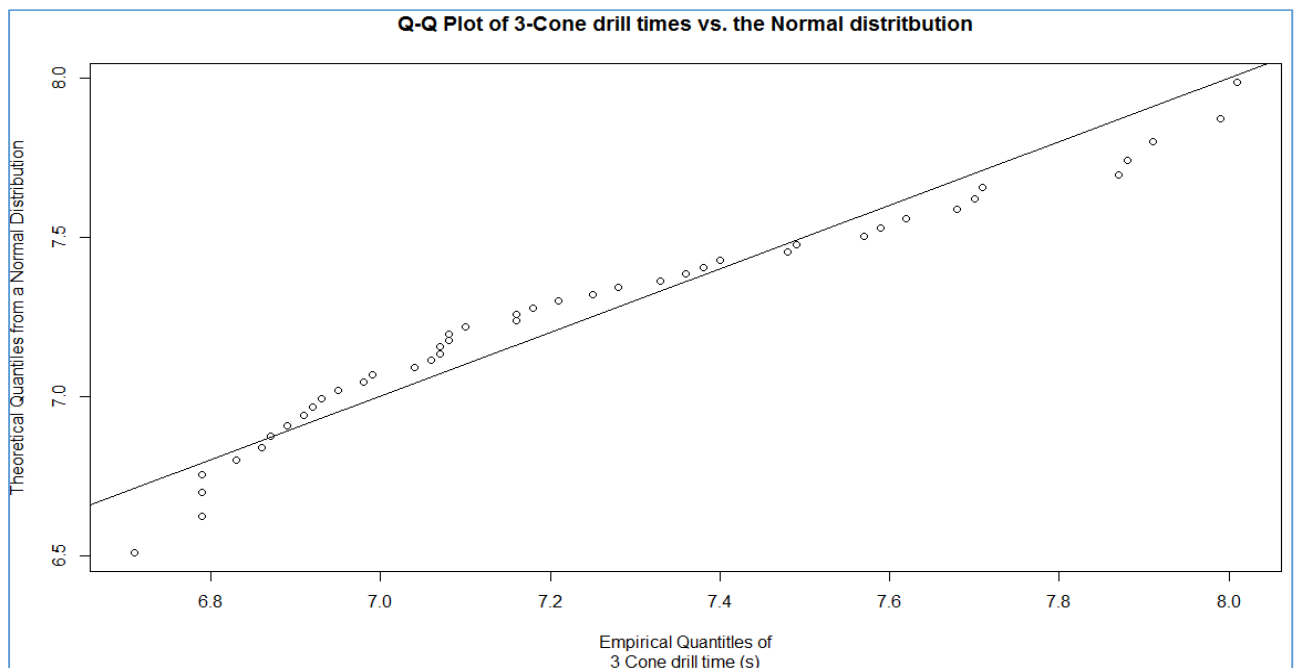


Fig 2. Q-Q plot of the average times in seconds of the 2015 NFL drafts against a normal curve

This Q-Q Plot, which compares theoretical(line) versus empirical(data) quantiles, shows that the data closely resembles a normal distribution. The empirical CDF of each point equals the theoretical CDF, corresponding to the specific values to compare. The very slight “S” shaped

curve indicates a larger variance. There is a lot of wiggle room for the quantile points of the data set and a majority of them lie off the line, but there are a few points that are close to it. Most points appear in the middle, meaning a high frequency and density. They are not tightly clustered towards the center so we can deduce that the confidence level should perhaps be a little smaller, let us assume that it is 90%. From this, the data looks like a normal distribution can approximately model a normal curve. The plot indicates that a normal distribution might be an appropriate model distribution for this data. We then proceed to lay our model distribution over the histogram to see how well it fits.

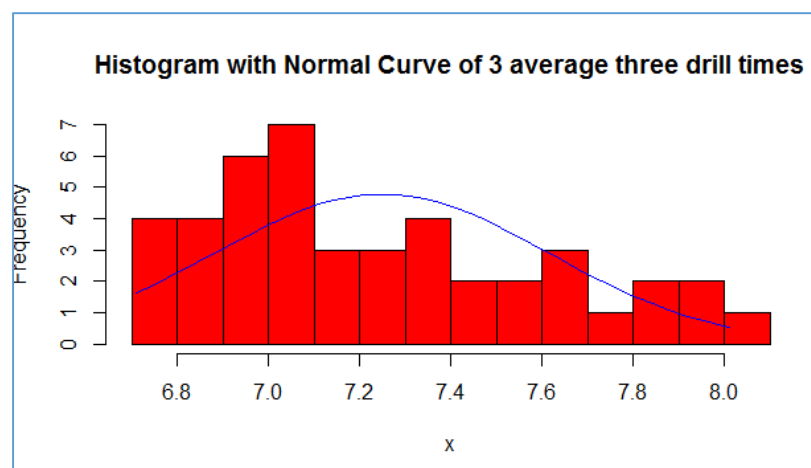


Fig 3. Histogram of the times of the cleaned data set for 3-Cone drill times

From the histogram we can see that the data set is right skewed and a normal curve fit to it and appears to be approximately normally distributed. The good amount of the data falls in the right side of the graph so is not perfectly symmetrical but does decrease towards the right edge. The tail on the left side is shorter than the one on the right. There are 12 breaks used in the histogram because it most accurately captures the data, although rule of thumb says there are 50 data points and $\sqrt{50} \sim 7$ to be used as bins. The area under the curve should be equal to the area in the bars which is 1. This means that the white gaps under the curve are compensated by the red parts of the bars that reach over the curve. The histogram's normal curve overlaid reflects a $N()$ with the sample mean and sample variance calculated as parameters. One estimated parameter of interest is the population mean, the average three Cone drill timings of all the top NFL prospects of 2015. This is because the sample mean can be calculated from the data, however, to know the average mean of the population, the population would have to be

measured. Another estimated parameter is the population variance, since likewise, there is no way to measure the variance of all football players. The sample variance is calculated, and so the population variance must be an estimated parameter. The data was chosen to be appropriately modeled by a normal distribution with the estimated parameters as the population mean of and a population variance of 700. The model with the sample data $N(7.248, 0.134)$ does not really fit the histogram well and the data lies close to the theoretical line drawn by the QQ Plot but completely on it, and so we conclude that the normal distribution is a loosely based approximate model. So a χ^2 model is probably a better fit for this data set.

This sample, of size 50, is not nearly as large as the total population of all the NFL drafted players over the past up until 2015 in the country.

Within the data set of all NFL drafted players, is the true mean time of a 3-cone drill test equal to 7.5 seconds?

Given an alpha of 0.10, construct a hypothesis test and confidence interval so that if all the past NFL players were taken into consideration, 90% of them would contain the true population mean time of the 3-cone drill test.

Our estimated true population mean of 7.5 seconds for the 4-cone drill test. Since we have one sample with an unknown population mean, we used a T-test, which has test statistic $T_0 = (\bar{X} - \mu_0) / (s / \sqrt{n})$. This test statistic has $n-1$ degrees of freedom, which would be 49 and $\alpha = 0.10$.

The questions asks us for a two sided hypothesis test on the estimated parameter and a two sided $100(1-\alpha)\%$ confidence interval:

$$100(1 - \alpha)\% \text{ Conf Int} = \bar{x} \pm Z_{\alpha/2} * S / \sqrt{n} = \text{point estimate} \pm \text{critical value} * \text{standard error}$$

One Sample t-test

$t = -4.5484$, $df = 43$, $p\text{-value} = 4.376e-05$

alternative hypothesis: true mean is not equal to 7.5

90 percent confidence interval:

7.155110 7.341254

sample estimates:

mean of x

7.248182

We should reject the null hypothesis, $\mu_0 = 7.5$, if it falls outside of the confidence interval, if $t_0 > t_{\alpha/2, n-1}$, or $t_0 < -t_{\alpha/2, n-1}$, or if the p-value is less than $\alpha=0.10$. Our p-value ~ 0.000045 which is less than 0.10, and so reject the null hypothesis that the true mean is equal to 7.5. This means that there is sufficient evidence to strongly claim the alternative hypothesis that the true time is not 7.5. Also, our $\mu_0 = 7.5$ does not lie within the range of the 90% confidence interval, (7.15, 7.34), which also leads to rejecting the null hypothesis. If we were to construct many of these intervals, only 90% percent of the intervals would contain the true population mean.

Since the hypothesis test resulted in a rejection of the true mean equaling 7.5 and we strongly concluded that the true mean time was not equal to 7.5 seconds, a second question is posed: If the true mean is not equal to 7.5, **is the true mean time for the 3-cone drill test equal to 7.3?**

This calls for a one sample, two sided t-test because we have an unknown population variance:

Assumptions: The average sample mean is 7.248 and the sample size $n = 44$. The data set is independent and identically distributed and is approximately normally distributed with unknown population variance.

Parameter of interest: The parameter of interest is the true mean time for the 3-cone drill test, μ .

Null hypothesis: $H_0 : \mu = 7.3$

Alternative hypothesis: $H_1 : \mu > 7.3$ or $\mu < 7.3$

We want to reject H_0 if the mean time does not equal 7.3 seconds.

Test statistic: The test statistic is $T_0 = (X - \mu_0) / (S / \sqrt{n})$

Reject H_0 if: Reject H_0 if the P-value is less than 0.10.

Computations: Because $X = 7.248$, $S = 0.367$, $\mu_0 = 7.3$ and $n = 44$:

One sample t-test:

data: x

$t = -0.9359$, $df = 43$, $p\text{-value} = 0.3545$

alternative hypothesis: true mean is not equal to 7.3

90 percent confidence interval:

7.155110 7.341254

sample estimates:

mean of x

7.248182

The p-value is greater than 0.10, and so we fail to reject the null hypothesis that the true mean time for the 3-cone drill test, μ is 7.3 seconds. This test is inconclusive. From these two hypothesis tests, we can see that the true population mean time for the 3-cone drill test may or may not be equal to 7.3.

By nature of the values of times as ratios, the data is not discrete. Since there is no theoretical model and the data is continuous, a Shapiro-Wilk Test was used to be most accurate. Just like the QQ Plot is a graphical test, the Shapiro-Wilk goodness of fit test is an analytical test for the normality of the data. A goodness of fit test is to test whether a hypothesized distribution fits the actual data or not.

Shapiro-Wilk Test

data: good3Cone

W = 0.93333 , p-value = 0.01539

W is the test statistic for the Shapiro-Wilk Test. Here, the null hypothesis is that the sample came from a normally distributed population. The p-value from the Shapiro-Wilk test is less than .10, indicating that the normal curve is not an appropriate model distribution for the good3Cone data. This is because we failed to reject the null hypothesis; there was insufficient evidence to strongly claim that the data is not from a normally distributed population. The conclusion that the normal distribution is not a well fit model for our data which verifies the conclusion found from the QQ Plot as well.

After choosing a distribution family and estimating parameters for g, we tested these estimated by graphing over a histogram and creating a QQ plot. The data for this variable is loosely close to being normally distributed, which points towards the Chi-squared distribution and our QQ plot verifies that using the sample mean and sample variance as parameters is not the

best way of finding the true mean. Also, as observed from the numerical summary, the variance is probably a better way of analyzing this variable as it is a ratio variable. The variance gives us a better understanding of this particular variable even though the mean ~ median since it gives us a frequency. Knowing that this data is well modeled by a normal distribution, we can make predictions based on variance of the times of 3-coned drill test of a NFL draft prospect.

Variable: "weight"

The weight class of NFL players has also been an important statistic in determining the ideal weight for an overall top prospect. In general, quarterbacks are lighter in weight however we are assuming that the sample selected is random. The performance of a football player is heavily dependent on the position played in and the weight. We are not considering position in relation to weight for the purpose of simplification in finding the true mean weight of a NFL draft prospect player.

The minimum of the data set is 186 pounds and the maximum is 339 pounds meaning that the average top predicted drafted NFL player should be in that range of [186,339] pounds. The first quartile of the data is 215.5 meaning that 25% of the data, a quarter of them, lie between 186 and 215.5 pounds. The median of the data is 246 pounds and the third quartile value is 305.8 pounds. The sample mean is 255.7 meaning that the top 50 NFL drafted prospects have an average weight of 255.7. The interquartile range is 90.25 and the sample standard deviation is 47.14 and the sample variance is 2222.17 which is extremely high for this data set of 50 players. Variance describes how the data varies greatly and is not precise and consistent. This leads us to hypothesize that perhaps the overall weight may not be directly co-related to the overall performance of an NFL without taking the position played into account. A large sample variance indicates that either the sample size is not big enough or that the data under observation is flawed or that we are asking the wrong question.

summary(weight)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
------	---------	--------	------	---------	------

```

186.0 215.5 246.0 255.7 305.8 339.0
sd(weight)
[1] 47.14702
> IQR(weight)
[1] 90.25

```

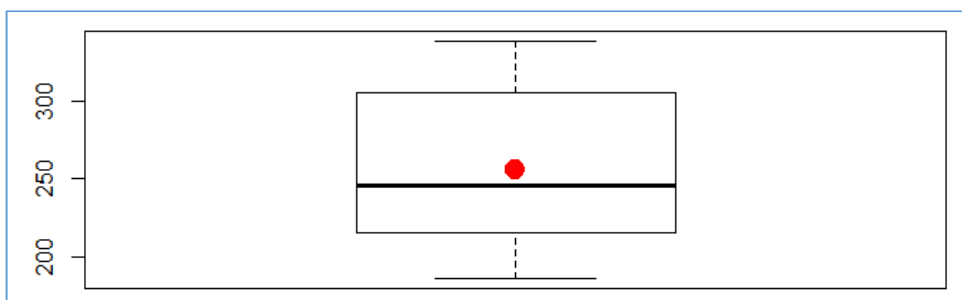


Figure 4. Boxplot of the weight of the top NFL draft prospects (2015)

From the boxplot, we can see that the median in the dark black line is very close to the mean but it is not equal. From this, it can be inferred that the distribution of the random variable is not perfectly normal but approximately normal. So we continue and do a barplot of the data to get a better understanding of the weights of NFL players.

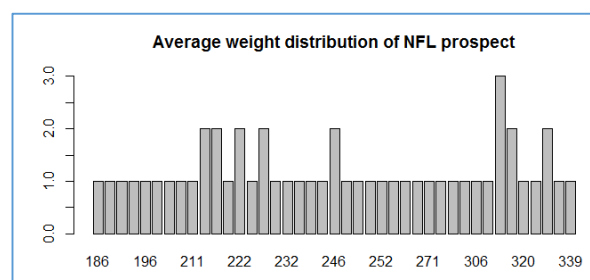


Fig 5. Bar plot of the average weight of the NFL draft prospects (2015)

From the bar plot, we get the weight class of each player represented by the bars and the density represented by the y-axis. Thus, we can conclude that the weight of the players is an ordinal variable.

Lastly, we conduct a Chi-square test with a null hypothesis suggesting that the data follows a uniform distribution model $U(186,339)$ to numerically evaluate the distribution model. In this chi square test, the sample size is 49, and the degree of freedom is 3 because there are 3 bins and we are estimating one parameter. (We collected a sample of 50, and only 49 data points are valid for this particular variable.) Noted since we are modeling a uniform distribution, we are not using any estimated parameters from our sample. We use a bin size of 3, one for each weight category: lightweight, midweight, heavyweight. The range of each bin is listed in Table 6, each with an expected count of 16.6667. Since we are comparing the observed and expected frequency, it is acceptable for the expected frequency to not be an integer.

Class Interval (X as rating)	Observed frequency (O_i)	Expected frequency (E_i)
$186 \leq x \leq 237$ (Lightweight)	21	16.6667
$237 \leq x \leq 288$ (Midweight)	14	16.6667
$288 \leq x \leq 339$ (Heavyweight)	14	16.6667

Fig 6. Table of the ordinal variable “weight”

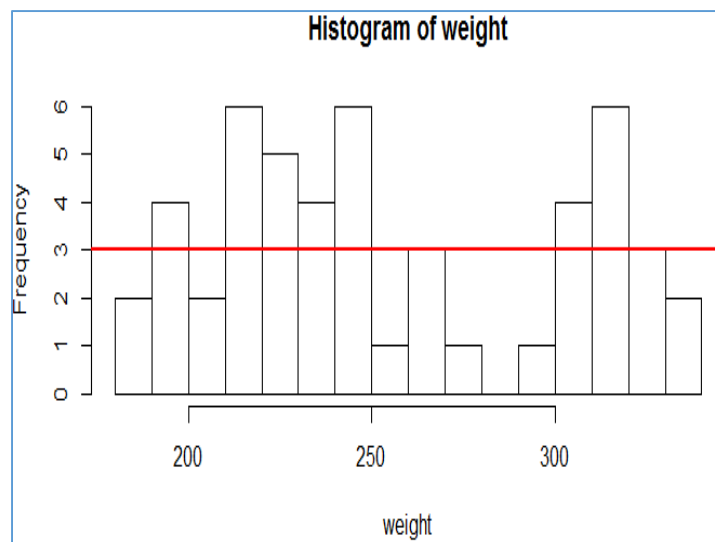


Fig 7. Histogram of weight and the uniform distribution in red

We juxtaposition the heights of the sample weights of the NFL draft prospects and the uniform distribution $U(186, 339)$ and it appears that the data is modeled well along the uniform

distribution. Since height has been considered an ordinal variable, it makes sense that the bins are of equal width. The test statistic comes back being 5.01, which is less than the critical value, 5.99 at $\alpha=0.05$ and 2 degrees of freedom. Therefore, our model is an adequate fit to the data since we are unable to reject the hypothesis that the weight classes of the NFL draft prospects follow a uniform distribution.

Categorizing players into weight classes is useful for organizing positions played. As discussed above, the numerical summary of this data does not give us much information because it was the wrong model to have been adopted. Thus, a barplot and histogram allows us to divide the data into 3 equal bins and we can deduce that the distribution of this ordinal variable is uniform.

Variable: “race”

The discrete random variable race is a categorical variable because there are only three categories but in no particular order: Black, White and Other. The question we are trying to ask here is: “Is the NFL biased in favor of African American players? If so, what percentage of the NFL consists of African American players and how does this compare to the NFL Hall of Fame players?”

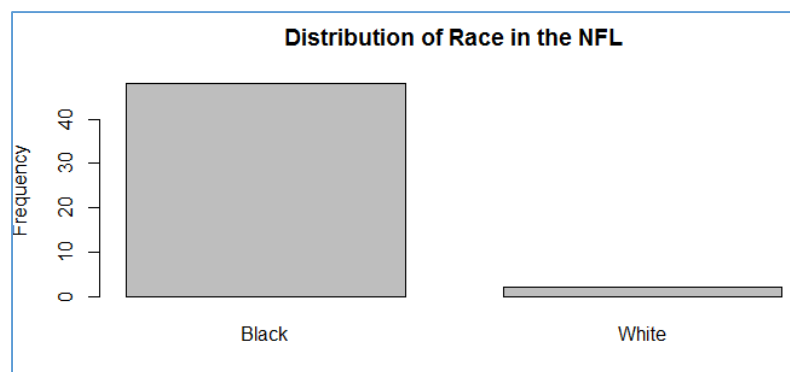


Fig 8. Racial distribution in the current draft of the NFL

It is very difficult to conduct a hypothesis test on a categorical variable. Although the current draft is binomial: having only two categories of “Black” and “White” football players, it is possible that players from other racial groups are drafted into the NFL. Thus we projected a bar plot to show all the categories instead of a Bernoulli plot. We are assuming that the sample selected is independent and identically distributed. We are also assuming that a binomial variable cannot be plotted on a histogram since we will not receive accurate information on categorical

variables through a histogram. From the bar plot however, it can be noted that the percentage of players in the sample that are Black is 85% whereas there are only 15% White players in this year's draft prospects. The statistics taken from the top 100 NFL players list by NFL.com report that 48% of the players were White, 48% were Black and 2% were others.

By comparing the two statistics, it is difficult to come to any conclusion about the performance of players based on race. Although the sample size is 50 and we can assume that it reflects a comparable measurement of racial representation, it is hard to predict the success rate of a player based on this variable.

Variable: h2015

The quantity the variable h2015 represents the height of an individual player in feet in the 2015 NFL Draft and is an ordinal variable. While height itself is a ratio variable, we are characterizing h2015 as an ordinal variable due to the fact that the way in which the height was measured is through categories of inches. With the heights being numerically binned into inches, we have decided to make h2015 an ordinal variable.

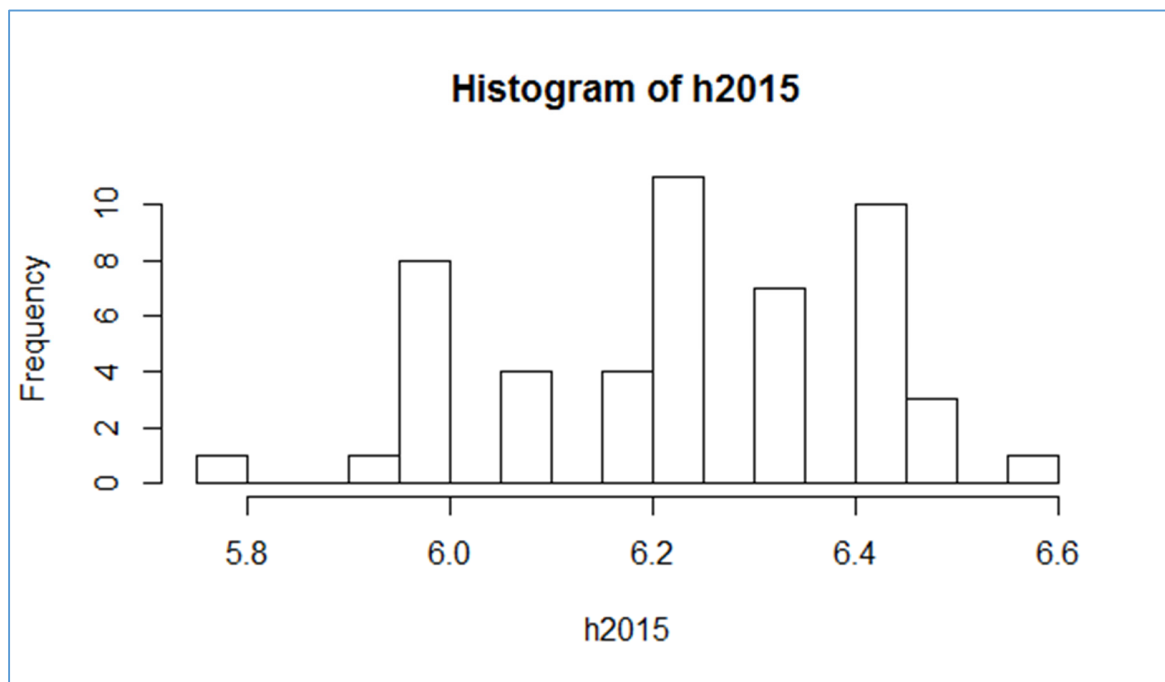


Fig 9. Histogram of the variable h2015, measuring the height classes

Figure 9 shows that there is no real good distribution for h2015, additionally, we have found only 10 different bins or values for our data, so we will treat our variable as an ordinal variable. With ordinal variables, the parameters are the probabilities of being in that category. To find the probability, we take the number of occurrences for each value and divide that by the total number of players. We find $P\{h2015=5.75\}=0.02$, $P\{h2015=5.92\}=0.02$, $P\{h2015=6.00\}=0.16$, $P\{h2015=6.08\}=0.08$, $P\{h2015=6.17\}=0.08$, $P\{h2015=6.25\}=0.22$, $P\{h2015=6.33\}=0.14$, $P\{h2015=6.42\}=0.20$, $P\{h2015=6.50\}=0.06$, $P\{h2015=6.58\}=0.02$. Again looking at Figure 9, one can see that there is not a very good distribution model similar to a normal curve, therefore we will also treat this as a discrete uniform model. Our histogram does not follow the general shape of any discrete distributions. However, in order to illustrate our understanding, we will model it as a discrete uniform. We expect that the analysis will reject the discrete uniform as an adequate model. In order to determine how appropriate this model is, we will perform a goodness of fit test with an alpha of 0.05.

The sample size of the heights of the players is 50 and we are assuming that the random variable is independent and identically distributed and follows a Uniform distribution. **Question: Is this model a good fit at a significance level of 0.05?**

Parameter of interest: distribution of population quantity

Null hypothesis (H_0): The distribution of population quantity is uniformly distributed

Alternative Hypothesis (H_a): The distribution of population quantity is not uniformly distributed

Test Statistic:

$$X_0^2 = \sum_{j=1}^k \frac{(O_j - E_j)^2}{E_j}$$

$X_0^2 \sim \chi^2$ with degree of freedom $k-p-1=10-1-1=8$

k = number of bins= 10

p = number of parameters estimated in model= 1

O_j = observed data count in bin j

E_j = expected count in bin j if good fit (H_0 is true)

$X_0^2=25.6$

25.6 > 15.51 (value found from Chi-Square Distribution table)

Because $X_0^2 > X_{0.05}^2$, we can reject H_0 in favor of H_a .

There is enough evidence to reject the claim that the uniform distribution is a good fit and support the claim that the uniform distribution is not a good fit.

Variable: hAll

The quantity the variable hAll represents the height of an individual player in feet in the NFL Hall of Fame and is an ordinal variable. While height itself is a ratio variable, we are characterizing hAll as an ordinal variable due to the fact that the way in which the height was measured is through categories of inches. With the heights being numerically binned into inches, we have decided to make hAll an ordinal variable.

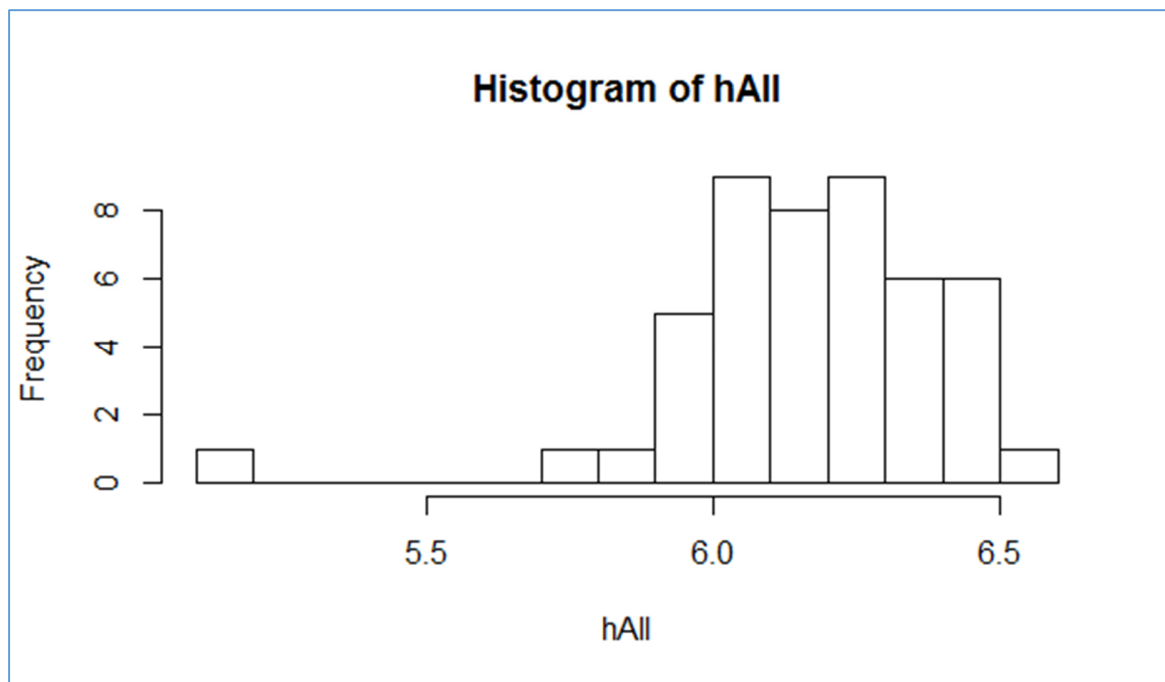


Fig 10. Histogram of heights of hall of fame NFL players

Figure 10 shows that there is potentially an uniform distribution for hAll, additionally, we have found only 10 different bins or values for our data, so we will treat our variable as an ordinal variable. With ordinal variables, the parameters are the probabilities of being in that category. To find the probability, we take the number of occurrences for each value and divide

that by the total number of players. We find $P\{h_{2015}=5.11\}=0.02$, $P\{h_{2015}=5.8\}=0.02$, $P\{h_{2015}=5.9\}=0.02$, $P\{h_{2015}=6.00\}=0.10$, $P\{h_{2015}=6.10\}=0.19$, $P\{h_{2015}=6.2\}=0.17$, $P\{h_{2015}=6.30\}=0.19$, $P\{h_{2015}=6.40\}=0.13$, $P\{h_{2015}=6.50\}=0.13$, $P\{h_{2015}=6.6\}=0.02$. Again looking at Figure 10, one can see that there is a somewhat uniform distribution model, therefore we will also treat this as a discrete uniform model. In order to determine how appropriate this model is, we will perform a goodness of fit test with an alpha of 0.05.

The sample size of the heights of the players is 48 and we are assuming that the random variable is independent and identically distributed and follows a Uniform distribution $\sim U(5.11, 6.60)$

Question: Is this model a good fit at a significance level of 0.05?

Parameter of interest: distribution of population quantity

Null hypothesis (H_0): The distribution of population quantity is uniformly distributed

Alternative Hypothesis (H_a) The distribution of population quantity is not uniformly distributed

Test Statistic:

$$X_0^2 = \sum_{j=1}^k \frac{(O_j - E_j)^2}{E_j}$$

$X_0^2 \sim \chi^2$ with degree of freedom $k-p-1=10-1-1=8$

k = number of bins= 10

p = number of parameters estimated in model= 1

O_j = observed data count in bin j

E_j = expected count in bin j if good fit (H_0 is true)

$X_0^2=22.08$

$22.08 > 15.51$ (value found from Chi-Square Distribution table)

Because $X_0^2 > X_{0.05}^2$, we can reject H_0 in favor of H_a .

There is enough evidence to reject the claim that the uniform distribution is a good fit and support the claim that the uniform distribution is not a good fit.

Variable: fortyT

The quantity the variable fortyT represents the 40 yard dash time of an individual player in seconds in the 2015 NFL Draft and is a ratio variable. Time can be measured as a ratio variable and thus we will treat fortyT as a ratio variable. We found the mean to be 4.75 seconds with a standard deviation of 0.315 seconds.

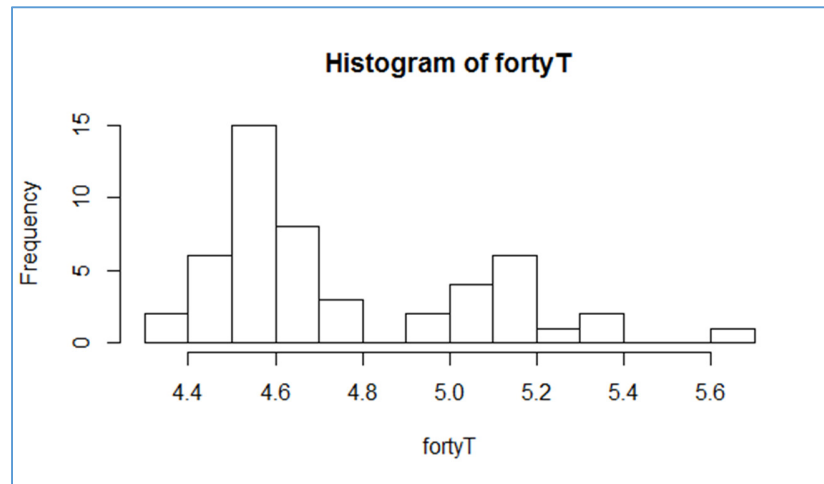


Fig 10. Histogram of the forty yard dash time (2015)

Fig 10. shows that there is actually a bi-modal normal distribution for fortyT. This seems appropriate because there are two types of players in football, linemen and “skilled” positions. The linemen tend to be slower than the other “skilled” positions, which explains the two normal curves. For our analysis, we will split the two curves and analyze them separately. While there may be some cases of “skilled” position players to be slower and linemen to be faster, we will still separate these curves at 4.80 seconds for the purpose of analysis. For the fast curve, we found the new mean to be 4.56 seconds and the standard deviation to be 0.109 seconds. Additionally, we found the median to be 4.57, which is very close to the mean, further justifying our Normal distribution model. Our parameters are the mean of 4.56 seconds, and variance of 0.0119 seconds².

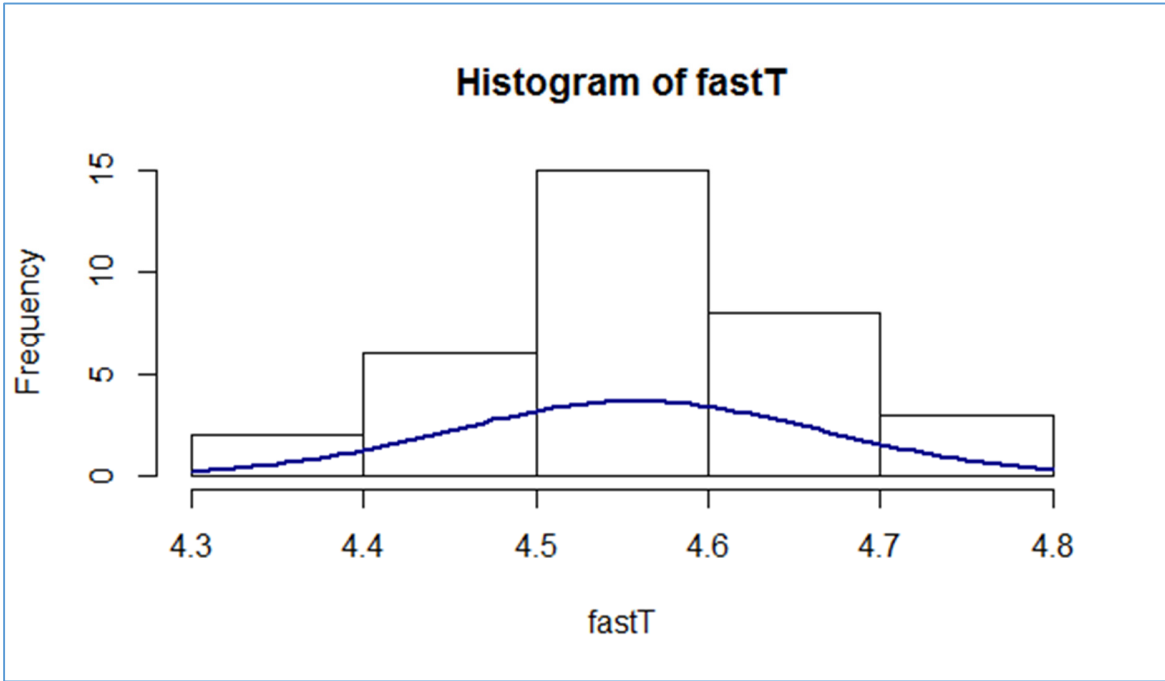


Fig 11. Histogram of Fast times with the normal curve

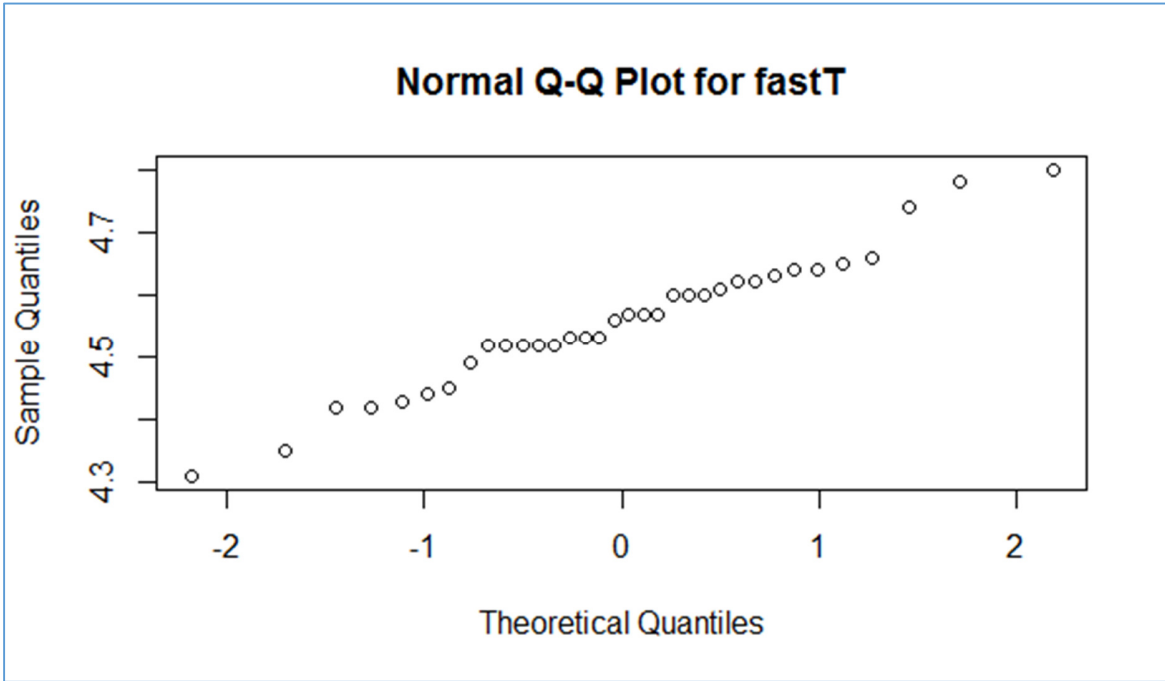


Fig 12. Normal Q-Q plot the variable fastT

The QQ plot of our variable indicates that our points are very close to the identity line, and therefore our theoretical and sample quantiles agree. To further justify our model, we will perform a goodness of fit test.

The sample size of the heights of the players is 50 and we are assuming that the random variable is independent and identically distributed and follows a Uniform distribution $\sim U(4.56, 0.0119)$

Question: Is this model a good fit at a significance level of 0.05?

Parameter of interest: distribution of population quantity

Null hypothesis (H_0): The distribution of population quantity is normally distributed

Alternative hypothesis (H_a): The distribution of population quantity is not normally distributed

Test Statistic:

$$W = \frac{\left(\sum_{i=1}^n a_i x_{(i)}\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

P value found: $0.646 > 0.05$

Because the p value of 0.646 is greater than 0.05, we fail reject H_0 in favor of H_a .

There is not enough evidence to reject the claim that the normal distribution is a good fit and support the claim that the uniform distribution is not a good fit. Therefore, we can conclude that the normal distribution is a good model.

For the slow curve, we found the new mean to be 5.15 seconds and the standard deviation to be 0.184 seconds. Additionally, we found the median to be 5.14, which is very close to the mean, further justifying our Normal distribution model. Our parameters are the mean of 5.15 seconds, and variance of 0.0339 seconds².

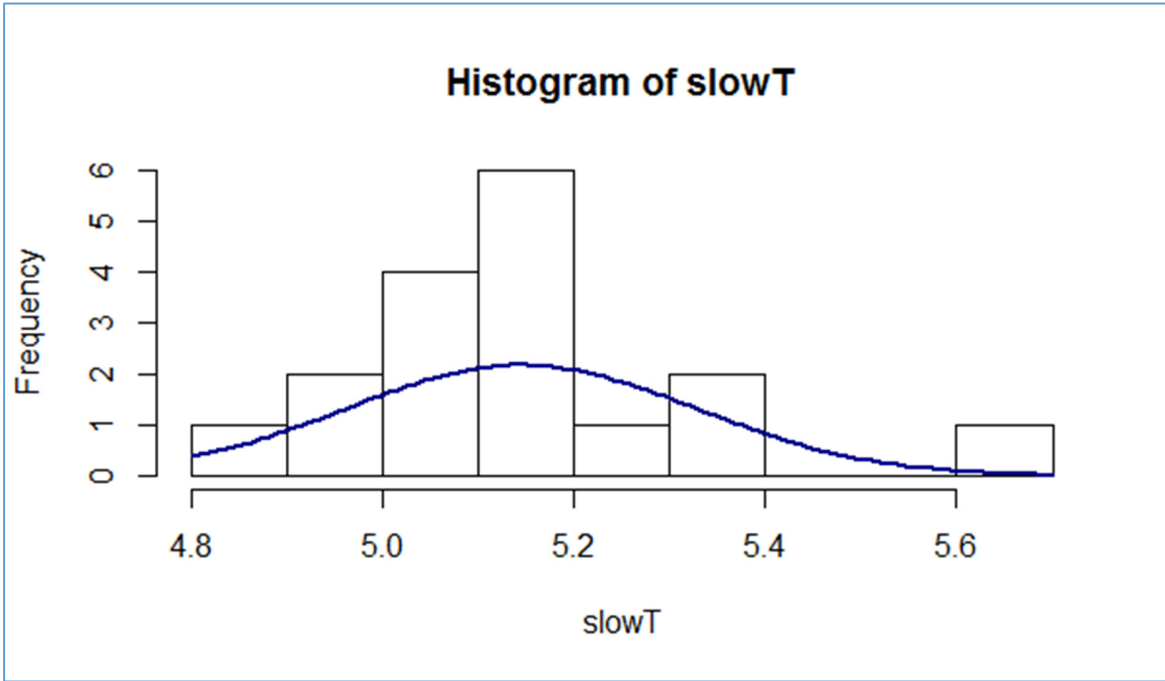


Fig 13. Histogram of the slow curve data against the normal curve

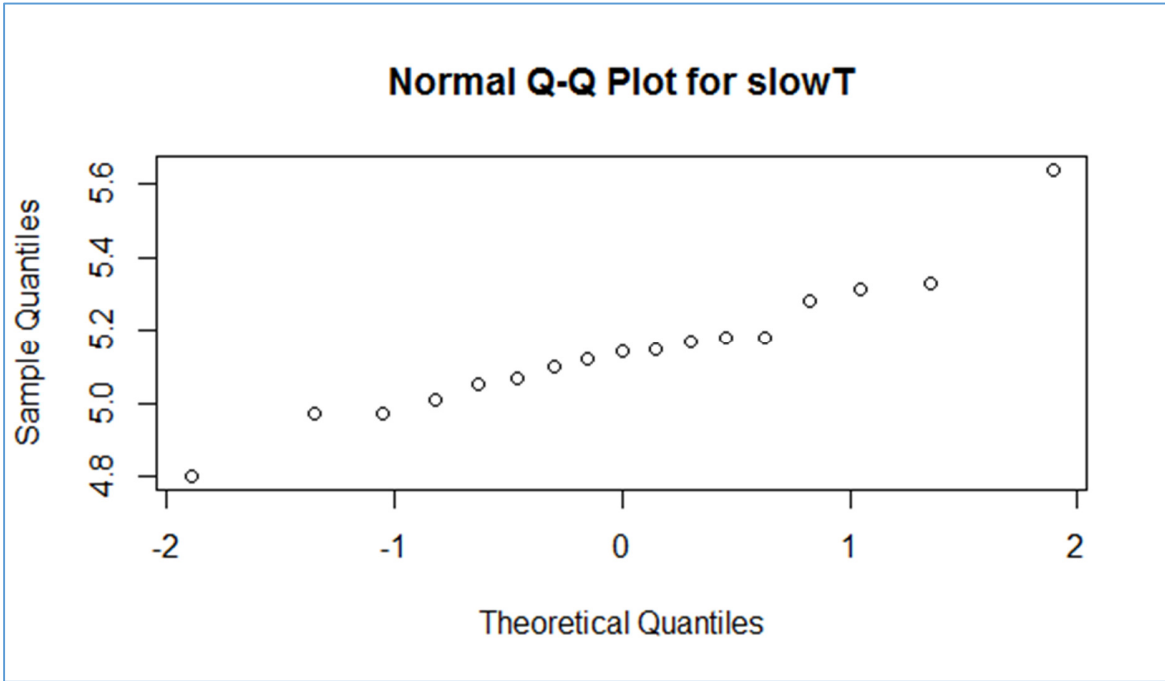


Fig 14. Normal Q-Q plot against the theoretical model for slowT

The QQ plot of our variable indicates that our points are very close to the identity line, and therefore our theoretical and sample quantiles agree. To further justify our model, we will perform a goodness of fit test.

The sample size of the heights of the players is 50 and we are assuming that the random variable is independent and identically distributed and follows a Normal distribution $\sim N(5.11, 0.0339)$

Question: Is this model a good fit at a significance level of 0.05?

Parameter of interest: distribution of population quantity

Null Hypothesis (H_0): The distribution of population quantity is normally distributed

Alternative Hypothesis (H_a): The distribution of population quantity is not normally distributed

Test Statistic:

$$W = \frac{\left(\sum_{i=1}^n a_i x_{(i)}\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

P value found: $0.2718 > 0.05$

Because the p value of 0.2718 is greater than 0.05, we fail reject H_0 in favor of H_a .

There is not enough evidence to reject the claim that the normal distribution is a good fit and support the claim that the uniform distribution is not a good fit. Therefore, we can conclude that the normal distribution is a good model.

Two sample analysis

Variable: fortyA

The quantity the variable fortyA represents the 40 yard dash time of an individual player in seconds in the NFL Hall of Fame and is a ratio variable. Time can be measured as a ratio variable and thus we will treat fortyA as a ratio variable. We found the mean to be 4.57 seconds with a standard deviation of 0.322 seconds.

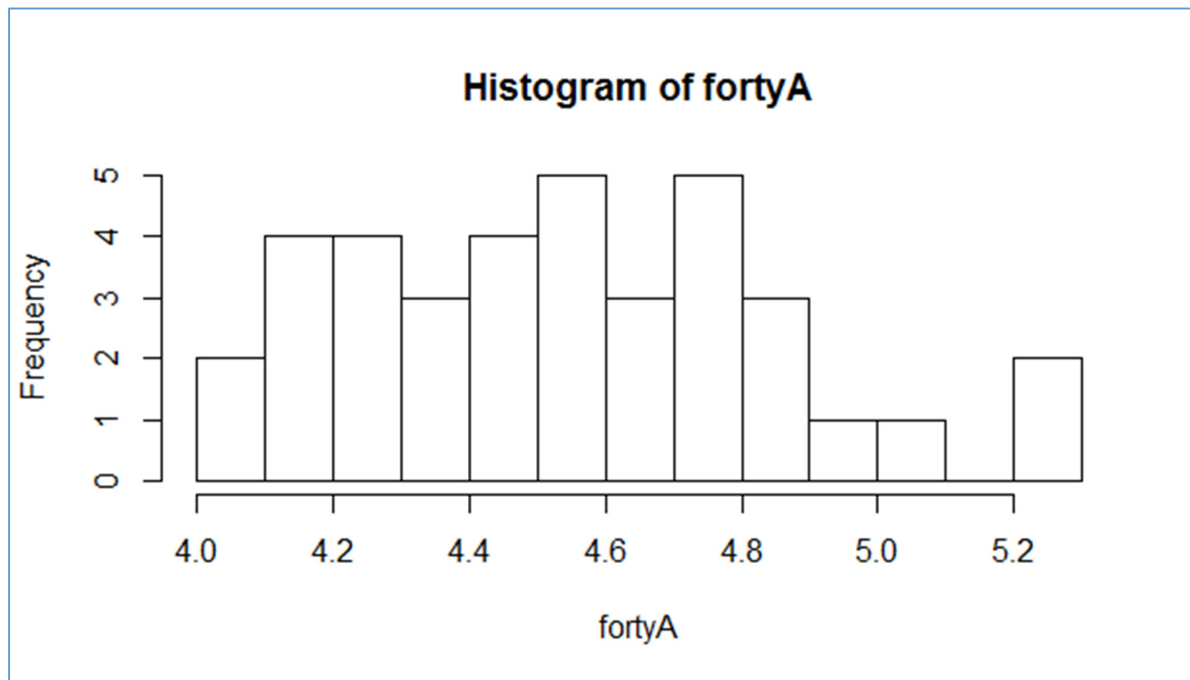


Fig 15. Histogram of forty yard dash times for hall of fame NFL players

Figure 15 shows that there is a uniform distribution for fortyA. This seems appropriate because the NFL Hall of Fame typically includes more “skilled” positions, rather than linemen. The similar 40 yard times would be expected, since many of the positions are the same. For our analysis, we will model this distribution as a uniform distribution. For the uniform distribution, the parameters are the minimum and maximum values, which are 4.00 and 5.30 respectively.

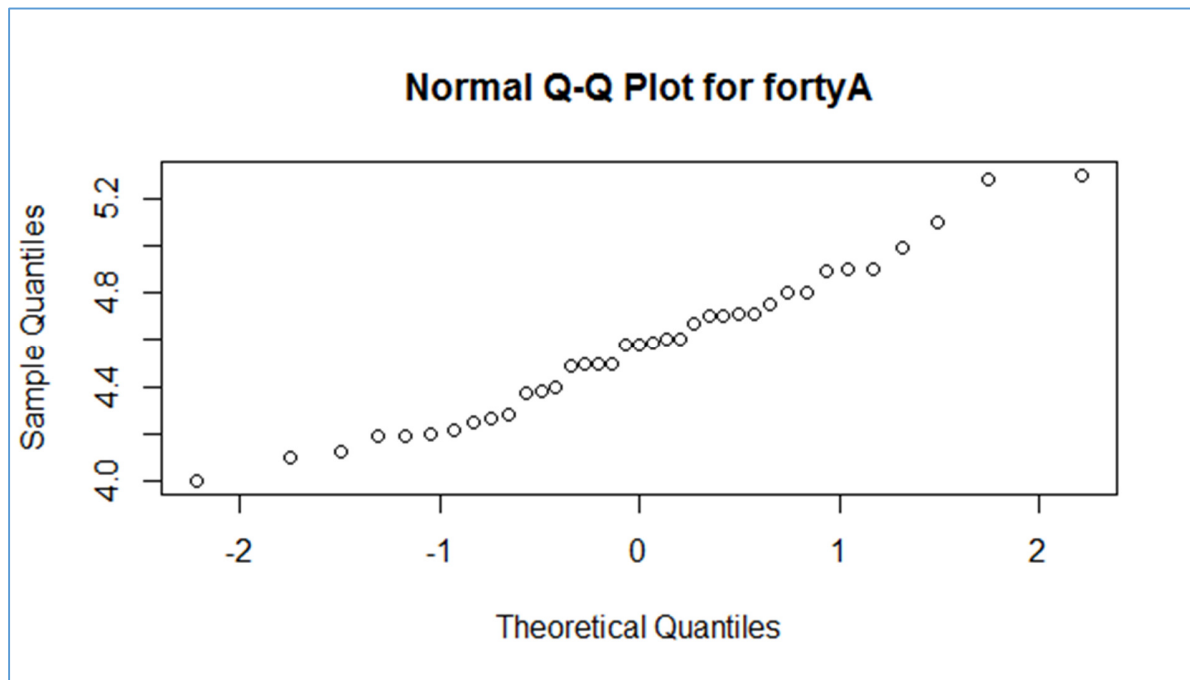


Figure 16: QQ Plot

The QQ plot of our variable indicates that our points are very close to the identity line, and therefore our theoretical and sample quantiles agree. To further justify our model, we will perform a goodness of fit test.

Step 0:

Given population quantity of 37 people

Given potential model: Uniform(4.00, 5.30)

Question: Is this model a good fit at a significance level of 0.05?

Step 1:

Parameter of interest: distribution of population quantity

Step 2:

H_0 : The distribution of population quantity is uniformly distributed

Step 3:

H_a : The distribution of population quantity is not uniformly distributed

Step 4:

Test Statistic:

$$X_0^2 = \sum_{j=1}^k \frac{(O_j - E_j)^2}{E_j}$$

$X_0^2 \sim \chi^2$ with degree of freedom $k-p-1 = 12-1-1 = 10$

$k =$ number of bins $= 12$

$p =$ number of parameters estimated in model $= 1$

$O_j =$ observed data count in bin j

$E_j =$ expected count in bin j if good fit (H_0 is true)

$X_0^2 = 6.819$

$6.189 < 18.31$ (value found from Chi-Square Distribution table)

Because $X_0^2 < X_{0.05}^2$, we fail reject H_0 in favor of H_a .

There is not enough evidence to reject the claim that the uniform distribution is a good fit and support the claim that the uniform distribution is not a good fit. Therefore, we can conclude that the uniform model is a good fit for our data.

As part of our larger question, we looked to find how the 40 yard times of great players compare. We previously discussed that the types of players in the Hall of Fame are heavily represented by “skilled” positions, and thus we will complete a two sample hypothesis test using the “skilled” positions times from the NFL 2015 Draft prospects and the NFL Hall of Fame 40 yard dash times. We are looking to answer the question: does the idea that athletes are getting faster hold true? Moreover, is the mean 40 yard dash time for the “skilled” position NFL Draft prospects greater than the NFL Hall of Fame mean 40 yard dash time?

Step 0:

Let fastT be the 40 yard dash times for the NFL Draft “skilled” positions and $n_1 = 34$

Let fortyA be the 40 yard dash times for the NFL Hall of Fame players and $n_2 = 37$

Assume data sets are random samples and the population values for fastT and fortyA are normally distributed and that their respective variances are unknown and unequal

Step 1:

Parameter of interest: $\mu_1 - \mu_2$, the difference in the mean 40 yard dash times of the Draft prospects and Hall of Fame players.

Step 2:

Null hypothesis: $H_0: \mu_2 \leq \mu_1$

The average dash time of Hall of Fame players less than or equal to the average dash time of “skilled” position draft prospects.

Step 3:

Alternative Hypothesis: $H_A: \mu_2 > \mu_1$

The average dash time of Hall of Fame players is greater than the average dash time of “skilled” position draft prospects.

Step 4:

Given two samples, two different populations, and variances are known, I would perform a t test by hand.

$$S_1=0.1092 \quad S_2=0.3220 \quad \frac{s_2-s_1}{s_1} = 1.94 \rightarrow \text{big \% difference}$$

Not willing to assume $\sigma_1^2 = \sigma_2^2$

t test

$$t_0 = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_0)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad t_0 \sim t_v$$

Step 6:

The p value is 0.5912 which is greater than α which is 0.05, so we will fail to reject H_0 in favor of H_A .

Step 7:

There is insufficient evidence to fail to reject H_0 , the null hypothesis. We conclude that the mean 40 yard dash times for NFL draft prospects is greater than or equal to the mean 40 yard dash times NFL Hall of Fame players.

This conclusion is rather surprising, mostly due to the popular notion that athletes now are faster and perform better than athletes in the past. However, we are also looking to answer other questions. Additionally, are players “growing” over time as well? Specifically, is the mean height for the “skilled” position NFL Draft prospects greater than the NFL Hall of Fame mean height?

Step 0:

Let h_{2015} be the height for the NFL Draft “skilled” positions and $n_1 = 34$

Let h_{All} be the height for the NFL Hall of Fame players and $n_2 = 37$

Assume data sets are random samples and the population values for h_{2015} and h_{All} are normally distributed and that their respective variances are unknown and unequal

Step 1:

Parameter of interest: $\mu_1 - \mu_2$, the difference in the mean heights of the Draft prospects and Hall of Fame players.

Step 2:

Null hypothesis: $H_0: \mu_2 \leq \mu_1$

The average height of Hall of Fame players is less than or equal to the average height of “skilled” position draft prospects.

Step 3:

Alternative Hypothesis: $H_A: \mu_2 > \mu_1$

The average height of Hall of Fame players is greater than the average height of “skilled” position draft prospects.

Step 4:

Given two samples, two different populations, and variances are known, I would perform a t test by hand.

$$S_1 = 0.1821 \quad S_2 = 0.2445 \quad \frac{S_2 - S_1}{S_1} = .342 \rightarrow \text{big \% difference}$$

Not willing to assume $\sigma_1^2 = \sigma_2^2$

t test

$$t_0 = \frac{(\bar{X}_1 - \bar{X}_2) - (10)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad t_0 \sim t_v$$

Step 5:

The p value is 0.2689 which is greater than α which is 0.05, so we will fail to reject H_0 in favor of H_A .

Step 6:

There is insufficient evidence to fail to reject H_0 , the null hypothesis. We conclude that the mean height for NFL draft prospects may be greater than or equal to the mean height of NFL Hall of Fame players.

Conclusion

Many online channels, sports news channels, fantasy football leagues have investigated the dynamics of the game, but few have investigated the productivity or performance of football players and how changes in the certain variables (inputs) may have affected them. Few utilize any type of productivity measure or performance indicator. There is no standard measure that can determine what makes a 'good football player'. This paper aims to assess the performance of NFL draft prospects of 2015 and how it can be related to making the model of a great football player. In particular, we have tested hypothesis and found correlations and lack of correlations where might have previously been thought to be absent.

We previously discussed that the types of players in the Hall of Fame are heavily represented by "skilled" positions, and thus we will complete a two sample hypothesis test using the "skilled" positions times from the NFL 2015 Draft prospects and the NFL Hall of Fame 40 yard dash times. The various comparison tests: 2 sample analysis of this year's prospected drafts to the Hall of Fame NFL players have shown us the correlation between the 40 yard dash times of top class performers versus the present draft. It was found that the 40 yard dash time, although is a good means of recruiting for professional football is not an accurate description of performance on the field.

It was also found that although the NFL is dominated by African-Americans, the performances may not be a direct result of racial prowess and physical superiority. The co-relation between performance and listing in the Hall of Fame is thus undetected for race.

We analyzed that the linemen tend to be slower than the other "skilled" positions. While there may be some cases of "skilled" position players to be slower and linemen to be faster, we will still separate these curves at 4.80 seconds for the purpose of analysis.

With that being said, our data is very specified to this year's NFL drafts and does not take into consideration the population of drafts from previous years or does any tests based on positions

played. This can be an area for future studies so as to determine the specifics of what variables are co-related to for example, a good Quarterback. There are multiple questions that can be answered to make the game of American football a lot more interesting. The scope for expanding sports consumerism to live football recommendations: especially in the field of Fantasy Football is great and all the raw data available on the internet can lead to a new revolution in Sports Statistics if someone makes the effort to utilize them.

References

1. *NFL Statistics: The Top 100: NFL's Greatest Players* (2014), Available from:
<http://top100.nfl.com/all-time-100>
2. *NFLDraftScout.com, distributed by The Sports Xchange: NFL Draft Prospects* (2015),
Available from: <http://www.cbssports.com/nfl/draft/prospectrankings>
3. Independent Wikipedia searches for individual player statistics


Raw Data:

Name	State/College	Height	Weight	In State	40 Yard Time	3 cone drill	Position	Ethnicity
Leonard Williams	California	6.42	302	0	4.97	7.59	2	Black
Jameis Winston	Florida	6.33	231	0	4.97	7.16	1	Black
Marcus Mariota	Oregon	6.33	222	0	4.52	6.87	1	Black
Dante Fowler Jr.	Florida	6.25	261	1	4.6	7.4	3	Black
Vic Beasley	South Carolina	6.25	246	0	4.53	6.91	3	Black
Amari Cooper	Alabama	6.08	211	0	4.42	6.71	WR	Black
Kevin White	West Virginia	6.25	215	0	4.35	6.92	WR	Black
Randy Gregory	Nebraska	6.42	235	0	4.64	6.79	3	Black
Danny Shelton	Washington	6.17	339	1	5.64	7.99	2	Black
Brandon Scherff	Iowa	6.42	319	1	5.05	7.18	2	White
Shane Ray	Missouri	6.25	245	1	4.65	7.71	DL	Black
DeVante Parker	Kentucky	6.25	209	1	4.45	9.999	WR	Black
Trae Waynes	Michigan	6.00	186	0	4.31	7.06	CB	Black
Andrus Peat	Michigan	6.58	313	0	5.18	8.01	OL	Black
Alvin Dupree	Kentucky	6.33	269	0	4.56	7.49	3	Black
La'el Collins	Louisiana	6.33	305	1	5.12	7.7	OL	Black
Malcolm Brown	Texas	6.17	319	1	4.62	6.86	RB	Black
T.J. Clemmings	Philadelphia	6.42	309	0	5.14	7.68	OL	Black
Landon Collins	Alabama	6.00	228	0	4.53	7.38	Specialist	Black
Todd Gurley	Georgia	6.08	222	0	4.52	9.999	RB	Black
Ereck Flowers	Florida	6.50	329	1	5.31	9.999	OL	Black
Jaelen Strong	Arizona	6.17	217	0	4.44	7.33	WR	Black
Melvin Gordon	Wisconsin	6.08	215	1	4.52	7.04	RB	Black
Dorial Green- Beckham	Missouri	6.42	237	0	4.49	6.89	WR	Black
Owamagbe								
Odighizuwa	California	6.25	267	0	4.62	7.36	2	Black
Cameron Erving	Florida	6.42	313	0	5.15	7.48	OL	Black
Kevin Johnson	Tennessee	6.00	188	0	4.52	6.79	CB	Black
Maxx Williams	Minnesota	6.33	249	1	4.78	9.999	TE	White
Marcus Peters	Washington	6.00	197	0	4.53	7.08	CB	Black

Appendix A: Consulting Log

Start Date	Start Time	Location	Participants	Goals of Sessi	End Date	End Time	Goals Completed	Efficien	Total Time Spent	Billable Hours	Rate	Total
2/1/2015	3:00 PM	CULC	Polly, Rebecca	Module 1	2/1/2015	6:30 PM	Outlined Module 1	0.8	1.5	1.2	\$50	60
2/6/2015	4:00 PM	CULC	Polly, Rebecca	Module 1	2/6/2015	9:30 PM	Data research	0.9	1.5	1.35	\$50	67.5
2/15/2015	8:00 PM	Library	Polly, Rebecca	Module 1	2/15/2015	10:00 PM	Wrote meat of the outline	0.8	2	1.6	\$50	\$80
2/18/2015	8:30 PM	Library	Polly, Rebecca	Module 1	2/18/2015	10:00 PM	Completed Module 1	0.7	1.5	1.05	\$50	\$52.50
2/21/2015	3:00 PM	CULC	Polly, Rebecca	Homework	2/21/2015	4:30 PM	Homework	0.7	1.5	1.05	\$50	52.5
3/16/2015	5:00 PM	Home	Polly, Rebecca	New Module	3/16/2015	9:00 PM	Via messagers: Commea	0.7	1.5	1.05	\$50	52.5
3/17/2015	8:00 PM	Home	Polly, Rebecca	New Module	3/17/2015	8:30 PM	Remade module 1	0.6	1.5	0.9	\$50	45
3/22/2015	8:00 PM	CULC	Polly, Rebecca	Module 1 and	3/22/2015	7:00 PM	New module 1 and Module 2	0.85	4	3.4	\$50	170
3/23/2015	4:00 PM	CULC	Polly, Rebecca	Module 2/HV	3/23/2015	7:00 PM	Module 2 and coding	0.85	3	2.55	\$50	127.5
3/25/2015	8:00 PM	CULC	Polly, Rebecca	Module 2/HV	3/25/2015	10:00 PM	Completed Module 2/hw 5	0.85	2	1.7	\$50	85
3/27/2015	4:00 PM	Home	Polly, Rebecca	HW 6	3/27/2015	5:30 PM	Worked on some of the hw	0.7	1.5	1.05	\$50	52.5
3/27/2015	2:00 PM	Home	Polly, Rebecca	HW 6	3/27/2015	3:30 PM	Worked on some of the hw	0.85	3	2.55	\$50	127.5
3/29/2015	8:30 PM	Library	Polly, Rebecca	HW 6	3/29/2015	10:00 PM	Completed HW 6	0.7	1.5	1.05	\$50	52.5
4/5/2015	4:30 PM	Library	Polly, Rebecca	HW 7	4/5/2015	6:00 PM	Worked on some of the hw	0.8	1.5	1.2	\$50	60
4/6/2015	9:00 PM	CULC	Polly, Rebecca	HW 7	4/6/2015	11:00 PM	Completed HW 7	0.85	2	1.7	\$50	85
4/11/2015	3:00 PM	CULC	Polly, Rebecca	Final Project	4/20/2015	4:30 PM	Moving slowly on Module 4	0.7	1.5	1.05	\$50	52.5
4/12/2015	7:12 AM	Home	Polly, Rebecca	Final Project	4/20/2015	10:00 AM	Completed HW 8	0.85	3	2.55	\$50	127.5
4/13/2015	2:30 PM	Home	Polly, Rebecca	Final Project	4/20/2015	4:00 PM	Moving slowly on Module 5	0.7	1.5	1.05	\$50	52.5
4/14/2015	11:30 AM	CULC	Polly, Rebecca	Final Project	4/20/2015	1:30 PM	Moving slowly on Module 5	0.8	2	1.6	\$50	80
4/15/2015	11:30 AM	Library	Polly, Rebecca	Final Project	4/20/2015	1:15 AM	Completed Module 5	0.85	1.5	1.275	\$50	63.75
4/16/2015	11:30 AM	CULC	Polly, Rebecca	Final Project	4/20/2015	1:15 AM	Started Compiling Project	0.7	1.5	1.05	\$50	52.5
4/17/2015	7:00 PM	Home	Polly, Rebecca	Final Project	4/20/2015	11:00 PM	Project moving slowly but sur	0.85	4	3.4	\$50	170
4/18/2015	3:00 PM	Home	Polly, Rebecca	Final Project	4/20/2015	6:00 PM	Project moving slowly but sur	0.85	3	2.55	\$50	127.5
4/19/2015	4:00 PM	CULC	Polly, Rebecca	Final Project	4/20/2015	5:30 PM	Project moving slowly but sur	0.7	1.5	1.05	\$50	52.5
4/20/2015	8:00 AM	CULC	Polly, Rebecca	Final Project	4/20/2015	2:00 PM	Completed Project	0.7	1.5	1.05	\$50	52.5
									Total Billable Ho	40.025	Final Bill:	2001.25

Appendix B: Billing Invoice



Georgia Tech COC:101
Ph. +1(678)-938-8755

Greetings! Please find attached your invoice for service number 299.

Start Date	Start Time	Location	Participants	Goals of Sessi	End Date	End Time	Goals Completed	Efficien	Total Time Spent	Billable Hours	Rate	Total
2/1/2015	3:00 PM	CULC	Polly, Rebecca	Module 1	2/1/2015	6:30 PM	Outlined Module 1	0.8	1.5	1.2	\$50	60
2/6/2015	4:00 PM	CULC	Polly, Rebecca	Module 1	2/6/2015	9:30 PM	Data research	0.9	1.5	1.35	\$50	67.5
2/15/2015	8:00 PM	Library	Polly, Rebecca	Module 1	2/15/2015	10:00 PM	Wrote meat of the outline	0.8	2	1.6	\$50	\$80
2/18/2015	8:30 PM	Library	Polly, Rebecca	Module 1	2/18/2015	10:00 PM	Completed Module 1	0.7	1.5	1.05	\$50	\$52.50
2/21/2015	3:00 PM	CULC	Polly, Rebecca	Homework	2/21/2015	4:30 PM	Homework	0.7	1.5	1.05	\$50	52.5
3/16/2015	5:00 PM	Home	Polly, Rebecca	New Module	3/16/2015	9:00 PM	Via messagers: Commea	0.7	1.5	1.05	\$50	52.5
3/17/2015	8:00 PM	Home	Polly, Rebecca	New Module	3/17/2015	8:30 PM	Remade module 1	0.6	1.5	0.9	\$50	45
3/22/2015	8:00 PM	CULC	Polly, Rebecca	Module 1 anc	3/22/2015	7:00 PM	New module 1 and Module 2	0.85	4	3.4	\$50	170
3/23/2015	4:00 PM	CULC	Polly, Rebecca	Module 2/HW	3/23/2015	7:00 PM	Module 2 and coding	0.85	3	2.55	\$50	127.5
3/25/2015	8:00 PM	CULC	Polly, Rebecca	Module 2/HW	3/25/2015	10:00 PM	Completed Module 2/hw 5	0.85	2	1.7	\$50	85
3/27/2015	4:00 PM	Home	Polly, Rebecca	HW 6	3/27/2015	5:30 PM	Worked on some of the hw	0.7	1.5	1.05	\$50	52.5
3/27/2015	2:00 PM	Home	Polly, Rebecca	HW 6	3/27/2015	3:30 PM	Worked on some of the hw	0.85	3	2.55	\$50	127.5
3/29/2015	8:30 PM	Library	Polly, Rebecca	HW 6	3/29/2015	10:00 PM	Completed HW 6	0.7	1.5	1.05	\$50	52.5
4/5/2015	4:30 PM	Library	Polly, Rebecca	HW 7	4/5/2015	6:00 PM	Worked on some of the hw	0.8	1.5	1.2	\$50	60
4/6/2015	9:00 PM	CULC	Polly, Rebecca	HW 7	4/6/2015	11:00 PM	Completed HW 7	0.85	2	1.7	\$50	85
4/11/2015	3:00	CULC	Polly, Rebecca	Final Project	4/20/2015	4:30 PM	Moving slowly on Module 4	0.7	1.5	1.05	\$50	52.5
4/12/2015	7:12 AM	Home	Polly, Rebecca	Final Project	4/20/2015	10:00 AM	Completed HW 8	0.85	3	2.55	\$50	127.5
4/13/2015	2:30 PM	Home	Polly, Rebecca	Final Project	4/20/2015	4:00 PM	Moving slowly on Module 5	0.7	1.5	1.05	\$50	52.5
4/14/2015	11.3	CULC	Polly, Rebecca	Final Project	4/20/2015	1:30 PM	Moving slowly on Module 5	0.8	2	1.6	\$50	80
4/15/2015	11.3	Library	Polly, Rebecca	Final Project	4/20/2015	1:15 AM	Completed Module 5	0.85	1.5	1.275	\$50	63.75
4/16/2015	11.3	CULC	Polly, Rebecca	Final Project	4/20/2015	1:15AM	Started Compiling Project	0.7	1.5	1.05	\$50	52.5
4/17/2015	7:00 PM	Home	Polly, Rebecca	Final Project	4/20/2015	11:00 PM	Project moving slowly but sur	0.85	4	3.4	\$50	170
4/18/2015	3:00 PM	Home	Polly, Rebecca	Final Project	4/20/2015	6:00 PM	Project moving slowly but sur	0.85	3	2.55	\$50	127.5
4/19/2015	4:00 PM	CULC	Polly, Rebecca	Final Project	4/20/2015	5:30 PM	Project moving slowly but sur	0.7	1.5	1.05	\$50	52.5
4/20/2015	8:00 AM	CULC	Polly, Rebecca	Final Project	4/20/2015	2:00 PM	Completed Project	0.7	1.5	1.05	\$50	52.5
Total Billable Ho										40.025	Final Bill:	2001.25

Appendix C: R Code

```

baddata = (threeCone == 9999.00)
> good3Cone = threeCone[ !baddata]
> good3Cone
[1] 7.59 7.16 6.87 7.40 6.91 6.71 6.92 6.79
[9] 7.99 7.18 7.71 7.06 8.01 7.49 7.70 6.86
[17] 7.68 7.38 7.33 7.04 6.89 7.36 7.48 6.79
[25] 7.08 7.62 7.57 6.99 7.10 7.16 7.07 6.95
[33] 6.98 7.88 7.08 6.93 7.25 7.21 6.83 6.79
[41] 7.28 7.91 7.07 7.87

>summary(good3Cone)
  Min. 1st Qu.  Median   Mean 3rd Qu.   Max.
 6.710  6.945  7.160  7.248  7.510  8.010

>sd(good3Cone)
[1] 0.3672479

>boxplot(good3Cone)
>points(1,mean(good3Cone),pch=19,col="red",cex=2)

```

QQ Plot:

```

good3Cone = sort(good3Cone)
n = length(good3Cone)
probs = (1:n)/(n+1)
norm.quant = qnorm(probs,mean(good3Cone),sd(good3Cone))
plot(good3Cone ,sort(norm.quant), ylab="Theoretical Quantiles from a Normal Distribution",
xlab=c("Empirical Quantiles of", "3 Cone drill time (s)"),main="Q-Q Plot of 3-Cone drill times
vs. the Normal distritbution")

```

```
abline(0,1)
```

Histogram with Normal Curve Overlaid:

```
x <- good3Cone
> h<-hist(x, breaks=12, col="red", main="Histogram with Normal Curve of 3 average three drill
times")
> xfit<-seq(min(x),max(x),length=40)
> yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
> yfit <- yfit*diff(h$mids[1:2])*length(x)
> lines(xfit, yfit, col="blue", lwd=1)
```

One sample, one variable t-test:

```
t = -4.5484, df = 43, p-value = 4.376e-05
alternative hypothesis: true mean is not equal to 7.5
90 percent confidence interval:
7.155110 7.341254
sample estimates:
mean of x
7.248182
```

Questions:

Is the true mean time of a 3-cone drill test equal to 7.5 seconds?

```
t.test(x,y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 7.5, conf.level = 0.90)
```

Shapiro Wilk Test:

NFL Draft Prospects: Key Performance Indicators

```
x <- good3Cone
shapiro.test(x)
```

Variable: “weight”

Numerical summary:

```
> summary(weight)
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
186.0 215.5 246.0 255.7 305.8 339.0
```

Histogram against the uniform distribution:

```
hist(weight, breaks = 12)
```

```
lines(c(0,400), c(3,3), col = "red", lwd = 2)
```

Barplot for “race”:

```
z = states$Ethnicity
```

```
t = table(z)
```

```
table(p)
```

```
b = barplot(t, ylab = "Frequency", main = "Distribution of Race in the NFL draft", space = 0.35,
names.arg = c("Black", "White"))
```

Welch two sample t-test for hAll:

```
> t.test(h2015, hAll, alternative= c("greater"),mu=0, paired=FALSE, var.equal=FALSE,
conf.level= 0.95)
```

```
data: h2015 and hAll
```

```
t = 0.6187, df = 84.833, p-value = 0.2689
```

```
alternative hypothesis: true difference in means is greater than 0
```

```
95 percent confidence interval:
```

```
-0.04594236 Inf
```

```
sample estimates:
```

```
mean of x mean of y
```

```
6.240200 6.212979
```

Welch two sample t-test for fortyA:

```
> t.test(fastT, fortyA, alternative= c("greater"),mu=0, paired=FALSE, var.equal=FALSE,  
conf.level= 0.95)
```

Welch Two Sample t-test

data: fastT and fortyA

t = -0.232, df = 44.811, p-value = 0.5912

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

-0.1073416 Inf

sample estimates:

mean of x mean of y

4.557647 4.570676

Appendix D: Acknowledgements

We would like to extend our sincerest gratitude to everyone who has reached out a hand and offered us help. Dr. Waller has been of great assistance and guidance throughout this project. We would also like to thank Douglas Montgomery and George Runger, the authors of our textbook, for imparting their knowledge for our betterment. NFL.com has been incredibly helpful in sorting together data in ways that weren't done before and conducting statistical analysis on them. We would also like to thank our amazing peers and colleagues who listen patiently, offer advice and are always there to lend a hand when things get rough.

Reflection

What have you learned differently or more deeply about doing statistical analysis by doing this project?

Statistical analysis can be applied to answer very vague questions about the world that can have answers with mathematical reasoning. Although this seems to be a very broad and basic understanding, it is in fact the application of the previous sentence that I learned by doing this project. It is different to be able to answer a question like: “hey what do you think the average weight of those football players are” with a statistical approach and answering with numerical values.

What have you learned about yourself as a learner by doing this project?

For this project, I was pressed for time because hell week has pulled me in all directions. However, I learned that I work and learn well under pressure. I love beating the time and that is my biggest motivation for learning. I also learned that sometimes, wanting to beat the clock may not be good enough and the work pressure is too high so it is extremely important to manage time.

What have you learned about yourself as a project partner by doing this project?

As a project partner, I am incredibly flexible and try to communicate well with my partner. Even when we had to change our entire project theme from football to basketball to American football, I was willing to adjust and learn. However, as the time for writing the final report approached all this miscommunication had meant we had very little to start with. I was a little disappointed in the efforts put by my partner because she was very difficult to communicate with and was not ready with her half by the finalization period. But in the end we pulled through and kept the show going!