Scott Kobos

Stochastic Project

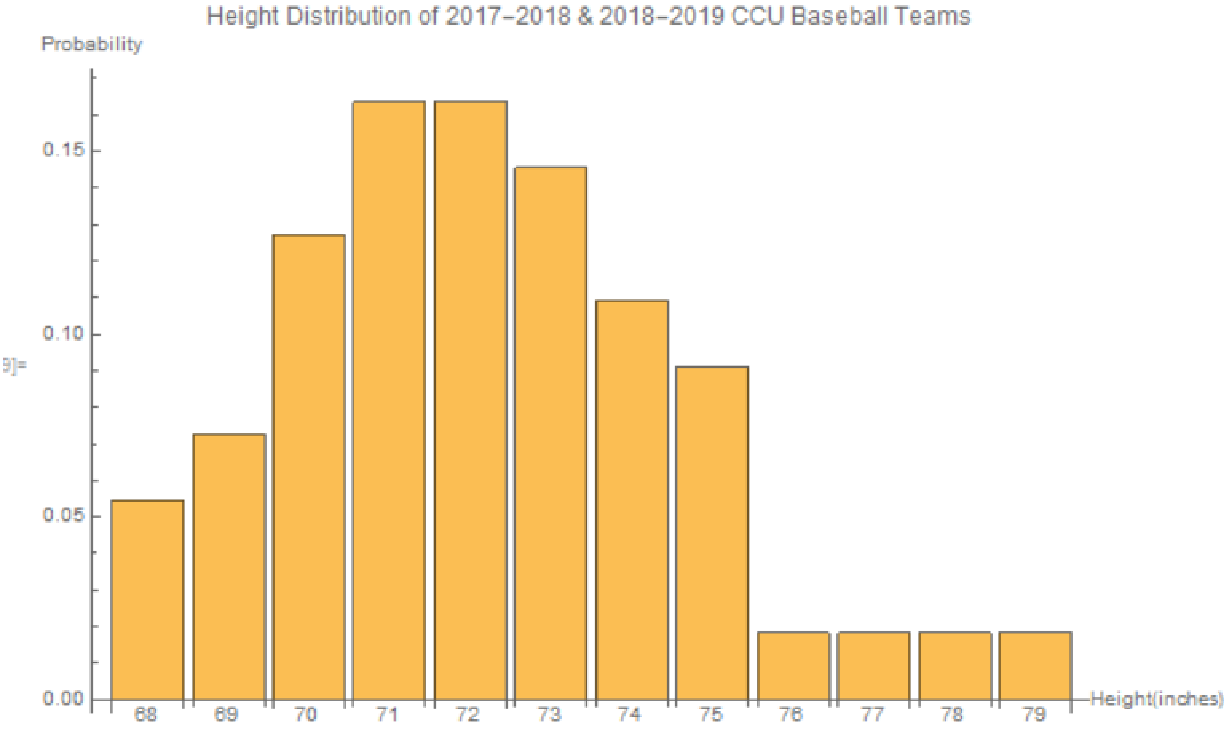
Distribution of Coastal Carolina Baseball's Heights

Introduction

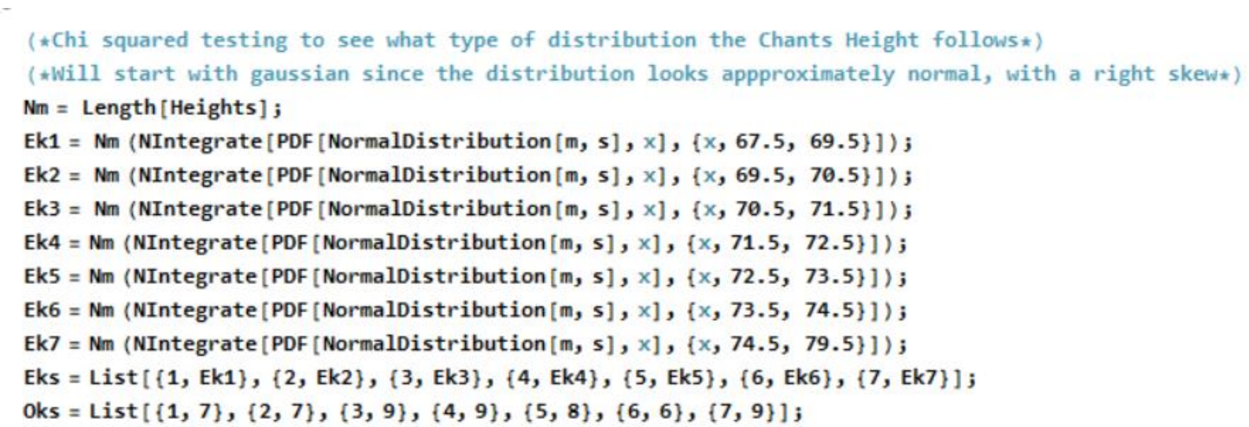
The purpose of this project was to perform distribution fitting calculations on data collected by the student. The heights of the 2017-2018 and 2018-2019 Coastal Carolina Baseball teams and the 2018-2017 team on its' own were taken and Chi squared testing was performed to identify the distribution that the data best fit. The work is relevant due to the data fitting being a major topic in the third and final section of the course material, Stochastic Analysis. It was hypothesized that the distribution of these heights would be normal, which was proven by the testing to be incorrect, but only due to sample size.

Procedure & Analysis Methods

The data collection was a simple process. The heights of the teams are kept on the rosters on the Coastal Carolina Athletics website and are easily accessible. The heights were recorded and stored in a list in Mathematica for use in the data fitting. The data was then used to create a bar chart for visualization.

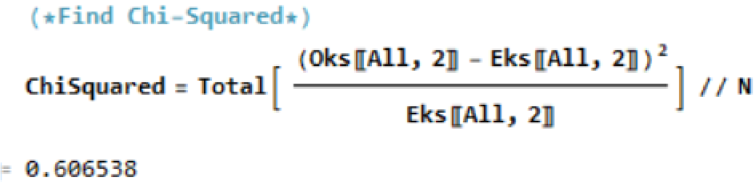


The data appears to be normal with a right skew. After the data was recorded and charted, the testing for the fit of a normal distribution began. Starting with the combined data for both teams, the calculations for Chi Squared are shown in the screenshot of the Mathematica code below.



```
(*Chi squared testing to see what type of distribution the Chants Height follows*)
(*Will start with gaussian since the distribution looks appproximately normal, with a right skew*)
Nm = Length[Heights];
Ek1 = Nm (NIntegrate[PDF[NormalDistribution[m, s], x], {x, 67.5, 69.5}]);
Ek2 = Nm (NIntegrate[PDF[NormalDistribution[m, s], x], {x, 69.5, 70.5}]);
Ek3 = Nm (NIntegrate[PDF[NormalDistribution[m, s], x], {x, 70.5, 71.5}]);
Ek4 = Nm (NIntegrate[PDF[NormalDistribution[m, s], x], {x, 71.5, 72.5}]);
Ek5 = Nm (NIntegrate[PDF[NormalDistribution[m, s], x], {x, 72.5, 73.5}]);
Ek6 = Nm (NIntegrate[PDF[NormalDistribution[m, s], x], {x, 73.5, 74.5}]);
Ek7 = Nm (NIntegrate[PDF[NormalDistribution[m, s], x], {x, 74.5, 79.5}]);
Eks = List[{1, Ek1}, {2, Ek2}, {3, Ek3}, {4, Ek4}, {5, Ek5}, {6, Ek6}, {7, Ek7}];
Oks = List[{1, 7}, {2, 7}, {3, 9}, {4, 9}, {5, 8}, {6, 6}, {7, 9}];
```

In this code the variables are Nm, the number of data points, Ek#, is the expected number of data points in the bin of number, #. The Oks is the observed number of data points in each bin, with the bins being from 67.5-69.5, 69.5-70.5, 70.5-71.5, 72.5-73.5,73.5-74.5, and then 74.5-79.5, for a total of 7 bins. Following this is the actual chi-squared calculation; where we take the Sum of the square of the difference of the observed and expected number of data points in each bin, followed.

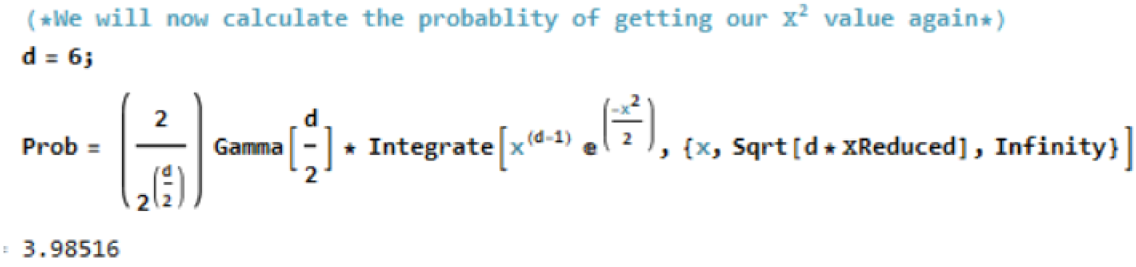


```
(*Find Chi-Squared*)
```

$$\text{ChiSquared} = \text{Total}\left[\frac{(\text{Oks}[[\text{All}, 2]] - \text{Eks}[[\text{All}, 2]])^2}{\text{Eks}[[\text{All}, 2]]}\right] \mathbin{//} \text{N}$$

```
0.606538
```

With a chi-squared value of .607, we can say that the expected distribution agrees with the data. Since chi-squared is less than the number of bins, 7. Next was the calculation of the degrees of freedom. This calculation was simple and does not need a screenshot of Mathematica code. The degrees of freedom are determined as the difference of number of bins vs constraints. With 7 bins and 1 constraint, the degrees of freedom were 6. With 6 degrees of freedom, and a chi-squared value of .607, we can once again state that the data agrees with the distribution since chi-squared is less than the degrees of freedom. Following is another simple calculation of the Reduced Chi-Squared value. The reduced value is simple the original chi-squared divided by the degrees of freedom. The reduced chi-squared value was found to be .101, being less than one we could once again state that the data agrees with the distribution. The final step in determining the fit of the data was to calculate the probability that we get the same chi-squared value again with another set of data. This calculation was as follows:



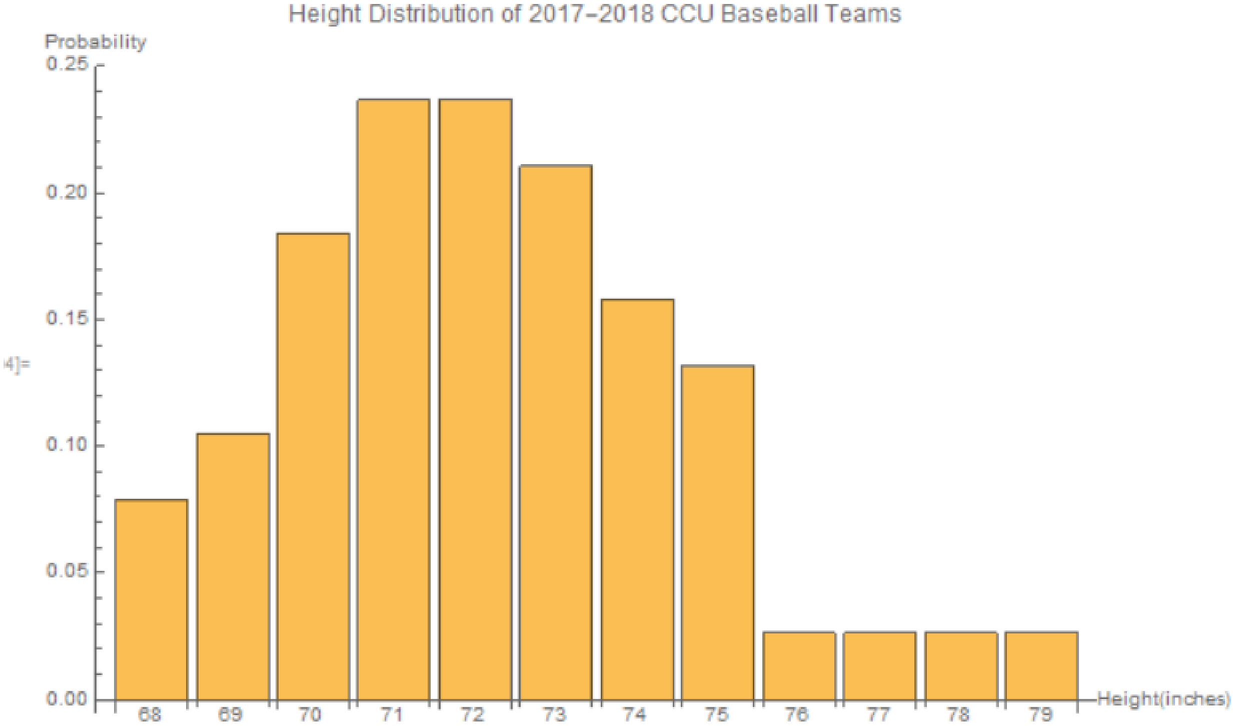
```
(*We will now calculate the probablity of getting our X² value again*)
d = 6;
```

$$\text{Prob} = \left(\frac{2}{2^{\left(\frac{d}{2}\right)}}\right) \text{Gamma}\left[\frac{d}{2}\right] * \text{Integrate}\left[x^{(d-1)}\, e^{\left(\frac{-x^2}{2}\right)}, \{x, \text{Sqrt}[d * \text{XReduced}], \text{Infinity}\}\right]$$

```
3.98516
```

The variables we are already familiar with, d, the degrees of freedom, XReduced, being the reduced chi-squared value. With a probability of just 3.99%, there was significant reason to reject the distribution. So, although each previous step accepted the distribution, this step does not, so the data is not normal.

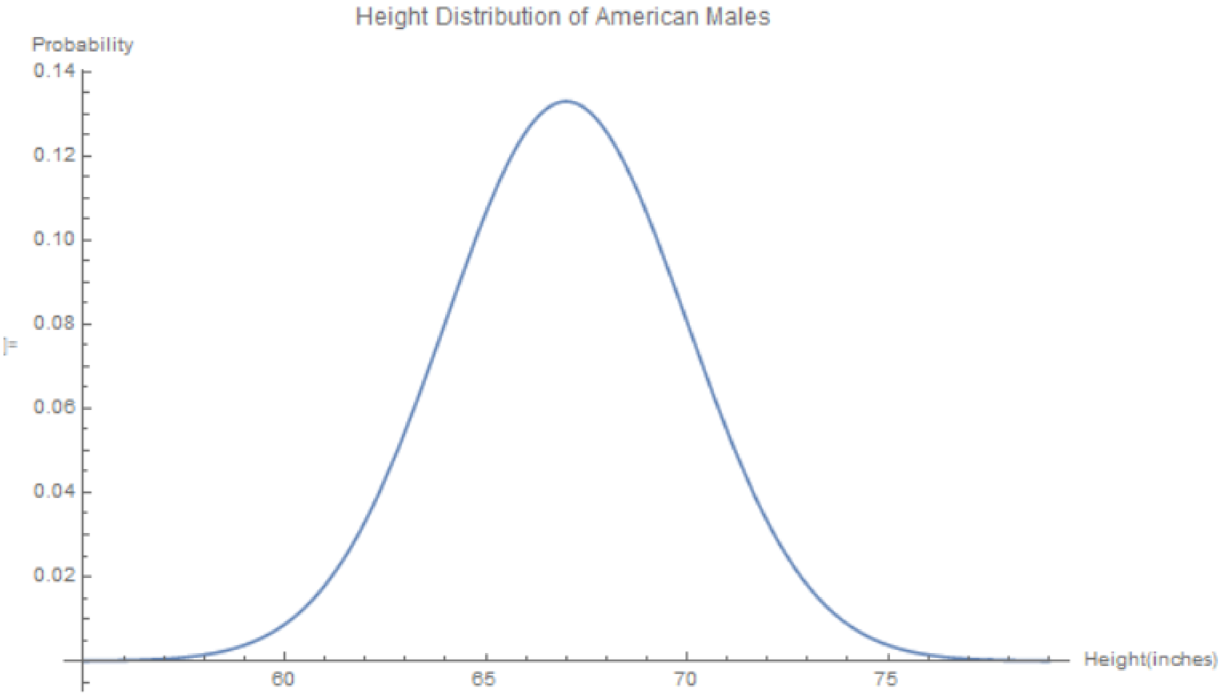
The same procedure was followed for the calculation of the fit for the 2017-2018 team and can be found in the appendix with the entire Mathematica code, and the bar chart from this data is as shown:



The results were similar and the bar chart for the 17-18 team looked approximately normal as well. The distribution seemed to be a good fit until the probability was calculated. The probability for the 2017-2018 team data was a mere 1.28%.

Analyzing Results

For this data fitting, our data did not fit with the normal distribution due to the low probability of getting the chi-squared value again. However, this does not mean that the data does not follow a normal distribution. The data for the combined 2017-2018 and 2018-2019 teams looked very close to being normal, while the data for the 2017-2018 team also appeared to be normal. One of two key differences are that the probability calculated for the 2017-2018 and 2018-2019 teams combined had a probability of 3.99% which the probability for just the 2017-2018 team had a probability of 1.28%. The other key difference is the sample size for each set of data. For the 2017-2018 team had a sample size of 38, while the combined 2017-2018 and 2018-2019 teams had a sample size of 55. What this tells us is that while both sets failed to match perfectly with the distribution, it shows that as the sample size increased, the data fit better with the distribution. This tells us that the fault is not that the data does not follow the normal distribution, but rather that the sample size was not large enough to fully normalize the distribution. This claim can be validated by the distribution of American Male Heights that can be seen below and follows the normal distribution.



What this show us is that as the sample size increases to an extremely large number of data points, in this case the American male population, the data normalizes and begins to follow the distribution, proving that our data would follow the distribution if more samples had been taken.

Conclusion

The goal of this project was to determine which distribution the data for the heights of the 2017-2018 and 2018-2019 Coastal Carolina Baseball teams followed and we did just that. Although our data was proven to not follow the distribution, it was shown that the data would have, if the sample size had been greater, and error in the sampling, not the fit of the distribution. Another interesting calculation that can be performed using this data, to put it all into context, is to see just how much taller the Coastal Carolina Baseball teams are compared to the average American male. As can be found at the end of the code in the Appendix, the average height of a Coastal Baseball player was found to be 1.67 standard deviations away from the average American male height, which is a pretty significant height advantage.

Appendix

Begins on next page

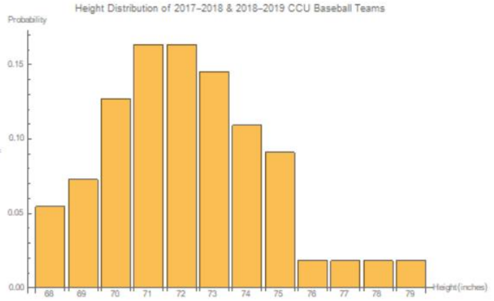
Data Fitting for the 2017-2018 and 2018-2019 Coastal Carolina Baseball Teams

```
In[1063]:=
Heights = List[68, 68, 68, 69, 69, 69, 69, 70, 70, 70, 70, 70, 70, 70, 71, 71, 71, 71, 71, 71, 71, 71, 71, 72, 72, 72, 72, 72, 72, 72, 72, 72, 73, 73, 73, 73, 73, 73, 73, 73, 74, 74, 74, 74, 74, 74, 75, 75, 75, 75, 75, 76, 77, 78, 79];
Sort[Tally[Heights]]
n = Tally[Heights] ⟦All, 2⟧;
```

$$F = \frac{n}{55};$$

```
Out[1064]= {{68, 3}, {69, 4}, {70, 7}, {71, 9}, {72, 9}, {73, 8}, {74, 6}, {75, 5}, {76, 1}, {77, 1}, {78, 1}, {79, 1}}

In[1067]:=
m = Mean[Heights] // N
s = StandardDeviation[Heights] // N

Out[1067]= 72.1818

Out[1068]= 2.44261

In[1069]:=
chants = BarChart[F, AxesLabel → {"Height (inches)", "Probability"}, PlotLabel → "Height Distribution of 2017-2018 & 2018-2019 CCU Baseball Teams", ChartLabels → {"68", "69", "70", "71", "72", "73", "74", "75", "76", "77", "78", "79"}]
```

Out[1069]=



```
In[1070]:=
(*Chi squared testing to see what type of distribution the Chants Height follows*)
(*Will start with gaussian since the distribution looks appproximately normal, with a right skew*)
Nm = Length[Heights];
Ek1 = Nm (NIntegrate[PDF[NormalDistribution[m, s], x], {x, 67.5, 69.5}]);
Ek2 = Nm (NIntegrate[PDF[NormalDistribution[m, s], x], {x, 69.5, 70.5}]);
Ek3 = Nm (NIntegrate[PDF[NormalDistribution[m, s], x], {x, 70.5, 71.5}]);
```

```
(*Chi squared testing to see what type of distribution the Chants Height follows*)
(*Will start with gaussian since the distribution looks appproximately normal, with a right skew*)
Nm = Length[Heights];
Ek1 = Nm (NIntegrate[PDF[NormalDistribution[m, s], x], {x, 67.5, 69.5}]);
Ek2 = Nm (NIntegrate[PDF[NormalDistribution[m, s], x], {x, 69.5, 70.5}]);
Ek3 = Nm (NIntegrate[PDF[NormalDistribution[m, s], x], {x, 70.5, 71.5}]);
Ek4 = Nm (NIntegrate[PDF[NormalDistribution[m, s], x], {x, 71.5, 72.5}]);
Ek5 = Nm (NIntegrate[PDF[NormalDistribution[m, s], x], {x, 72.5, 73.5}]);
Ek6 = Nm (NIntegrate[PDF[NormalDistribution[m, s], x], {x, 73.5, 74.5}]);
Ek7 = Nm (NIntegrate[PDF[NormalDistribution[m, s], x], {x, 74.5, 79.5}]);
Eks = List[{1, Ek1}, {2, Ek2}, {3, Ek3}, {4, Ek4}, {5, Ek5}, {6, Ek6}, {7, Ek7}];
Oks = List[{1, 7}, {2, 7}, {3, 9}, {4, 9}, {5, 8}, {6, 6}, {7, 9}];

In[1080]:=
(*Find Chi-Squared*)
```

$$\text{ChiSquared} = \text{Total}\left[\frac{(\text{Oks}⟦\text{All}, 2⟧ - \text{Eks}⟦\text{All}, 2⟧)^2}{\text{Eks}⟦\text{All}, 2⟧}\right] \mathbin{//} \text{N}$$

```
Out[1080]= 0.606538
```

With a χ^2 value of .607, we can say that the expected distribution agrees with the data. Since χ^2 is less than the number of bins, 7.

```
In[1081]:=
(*We will now find the degrees of freedom. nn is used for number of bins since n was used previously*)
nn = 7;
c = 1;
d = nn - c

Out[1083]= 6
```

With 6 degrees of freedom, and a X^2 value of .607, once again we can say that the suggested distribution agrees with our data.

```
In[1084]:=
(*We will no find the reduced X² value*)
```

$$\text{XReduced} = \frac{\text{ChiSquared}}{\text{d}}$$

```
Out[1084]= 0.10109
```

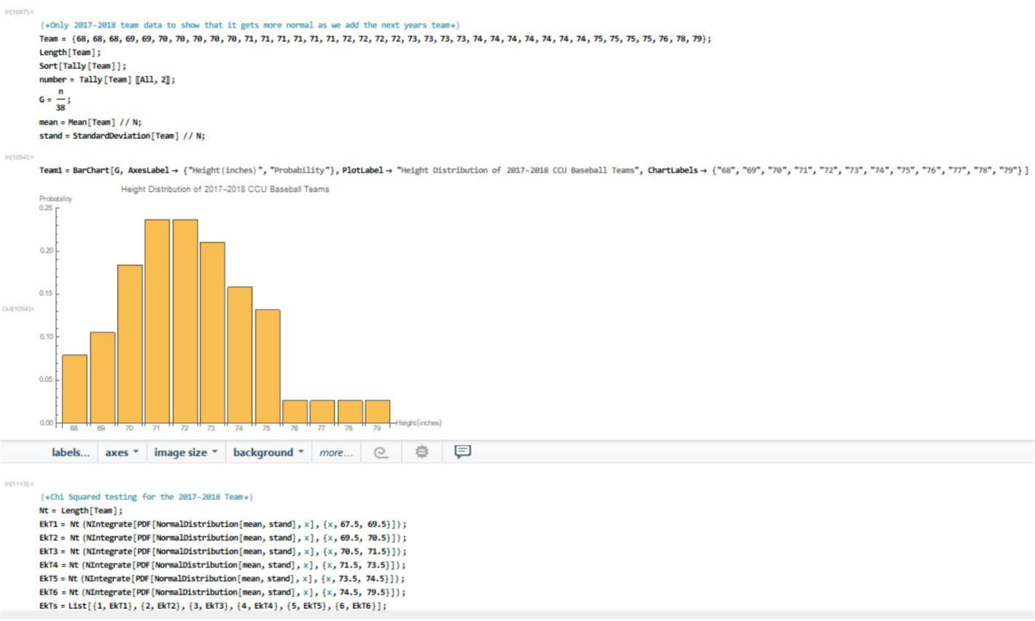
With a XReduced value of .101, which is less than 1, we can once again say that the suggested distribution agrees with our data.

```
In[1085]:=
(*We will now calculate the probablity of getting our X² value again*)
d = 6;
```

$$\text{Prob} = \left(\frac{2}{2^{\left(\frac{d}{2}\right)}}\right) \text{Gamma}\left[\frac{d}{2}\right] * \text{Integrate}\left[x^{(d-1)}\, e^{\left(\frac{-x^2}{2}\right)}, \{x, \text{Sqrt}[d * \text{XReduced}], \text{Infinity}\}\right]$$

```
Out[1086]= 3.98516
```

Data Fitting for the 2017-2018 Coastal Carolina Baseball Teams



In[1087]:=

```
(*Only 2017-2018 team data to show that it gets more normal as we add the next years team*)
Team = {68, 68, 68, 69, 69, 70, 70, 70, 70, 70, 71, 71, 71, 71, 71, 71, 72, 72, 72, 72, 73, 73, 73, 73, 74, 74, 74, 74, 74, 74, 74, 75, 75, 75, 75, 76, 78, 79};
Length[Team];
Sort[Tally[Team]];
number = Tally[Team] [[All, 2]];
G = n/38;
mean = Mean[Team] // N;
stand = StandardDeviation[Team] // N;
```

In[1094]:=

```
Team1 = BarChart[G, AxesLabel → {"Height(inches)", "Probability"}, PlotLabel → "Height Distribution of 2017-2018 CCU Baseball Teams", ChartLabels → {"68", "69", "70", "71", "72", "73", "74", "75", "76", "77", "78", "79"}]
```

Out[1094]=

In[1113]:=

```
(*Chi Squared testing for the 2017-2018 Team*)
Nt = Length[Team];
EkT1 = Nt (NIntegrate[PDF[NormalDistribution[mean, stand], x], {x, 67.5, 69.5}]);
EkT2 = Nt (NIntegrate[PDF[NormalDistribution[mean, stand], x], {x, 69.5, 70.5}]);
EkT3 = Nt (NIntegrate[PDF[NormalDistribution[mean, stand], x], {x, 70.5, 71.5}]);
EkT4 = Nt (NIntegrate[PDF[NormalDistribution[mean, stand], x], {x, 71.5, 73.5}]);
EkT5 = Nt (NIntegrate[PDF[NormalDistribution[mean, stand], x], {x, 73.5, 74.5}]);
EkT6 = Nt (NIntegrate[PDF[NormalDistribution[mean, stand], x], {x, 74.5, 79.5}]);
EkTs = List[{1, EkT1}, {2, EkT2}, {3, EkT3}, {4, EkT4}, {5, EkT5}, {6, EkT6}];
```

In[1122]:= (*Calculate Chi Squared for the 17-18 Team*)

$$\text{TeamChiSquared} = \text{Total}\left[\frac{(\text{OkTs}[[\text{All}, 2]] - \text{EkTs}[[\text{All}, 2]])^2}{\text{EkTs}[[\text{All}, 2]]}\right] \text{ // N}$$

Out[1122]= 2.83585

With a Team Chi Squared value of 2.84, we can say that the data agrees with the distribution since the value is less than the number of bins (6).

In[1123]:=

```
(*Now to calculate the degrees of freedom*)
bins = 6;
constraints = 1;
Df = bins - constraints
```

Out[1125]= 5

With 5 degrees of freedom and a Team Chi Squared value of 2.84 we can once again say that our data agrees with the distribution.

In[1126]:=

(*Now to calculate the ChiReduced value*)

$$\text{TeamChiReduced} = \frac{\text{TeamChiSquared}}{\text{Df}}$$

Out[1126]= 0.56717

With a Team Chi Reduced value less than 1, at .567, we can once again say that the data agrees with the distribution.

In[1127]:=

(*Now to calculate the probability of getting these values again*)

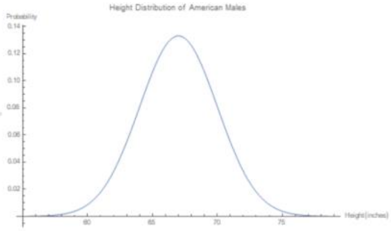
$$\text{Probabil} = \left(\frac{2}{2^{\left(\frac{\text{Df}}{2}\right)}}\right)\text{Gamma}\left[\frac{\text{Df}}{2}\right] * \text{Integrate}\left[x^{(\text{Df}-1)}\, e^{\left(\frac{-x^2}{2}\right)}, \{x, \text{Sqrt}[\text{Df} * \text{TeamChiReduced}], \text{Infinity}\}\right]$$

Out[1127]= 1.28167

With a probability of just over one percent we must state that there is a significant reason to reject this distribution.

Distribution of Average American Male Heights

```
gaussian = Plot[PDF[NormalDistribution[67, 3], x], {x, 55, 79}, PlotLabel → "Height Distribution of American Males", AxesLabel → {"Height(inches)", "Probability"}]
```



As can be seen by this graph, the overall distribution of American Male heights is normal with a mean of 67 inches (5 feet 7 inches) and a standard deviation of 3 inches. Compared to the distribution of the Coastal baseball teams, the baseball teams have a much higher mean of 72.18 inches and a standard deviation of 2.44 inches. The mean height of 72 inches for the Coastal baseball team is 5 inches greater than the mean of the American Male and would fall 1.667 standard deviations away from the mean in the distribution of American Male height.

```
μAm = 67;
σAm = 3;
N[1 - CDF[NormalDistribution[μAm, σAm], μAm + 5/3 σAm]] * 100
```

```
4.77904
```

This calculation shows the probability of an individual being 1.667 standard deviations away from the mean American Male height. The mean height of Coastal Baseball players is in the top 4.78 percent of American Male height.