Donavan & Britani
MA523-70
Project Part 2

1. Let X be the random variable WHIP (Walks Plus Hits Per Inning Pitched) for all starting pitchers who threw more than 20 innings in the 2017 season.
   Then $X \sim N(1.343, 0.0476)$.

2. Goodness of Fit Test at $\alpha = 0.05$:
   $H_0: X \sim N(1.343, 0.0476)$
   $H_1: X \nsim N(1.343, 0.0476)$

   The test statistic is $\mathcal{X}^2 = 11.545$ ($\Rightarrow p - value = 0.11655$)
   The critical value is $\mathcal{X}^2_{0.95;\ df=7} = 14.067$
   Since the test statistic 11.545 < the critical value 14.067 (or since the p-value 0.11655 > 0.05), we fail to reject the null hypothesis. Therefore, we conclude that $X \sim N(1.343, 0.0476)$ at the 0.05 level of significance.

3. Pdf: $f(x) = \frac{e^{-\frac{(x-1.343)^2}{2(0.0476)}}}{\sqrt{2\pi(0.0476)}}, 0 \leq x \leq \infty$

   Cdf: $F(x) = \text{◊}\left(\frac{x-1.343}{0.0476}\right)$

   Empirical Cdf: $$G(x) = \begin{cases} 0, & x < 0.688 \\ 1/109, & 0.688 \leq x < 0.906 \\ 19/218, & 0.906 \leq x < 1.125 \\ 101/218, & 1.125 \leq x < 1.343 \\ 92/109, & 1.343 \leq x < 1.561 \\ 106/109, & 1.561 \leq x < 1.779 \\ 217/218, & 1.779 \leq x < 1.997 \\ 1, & x \geq 1.997 \end{cases}$$

   MGF: $M(t) = e^{1.343t + \frac{0.076t^2}{2}}$

   $E(X) = 1.343$

   $E(X) = 1.343$

4. Our sample consisted of all starting pitchers in the 2017 season who threw more than 20 innings, and we hypothesized that the random variable WHIP followed an approximately normal distribution with $\mu = 1.343$ and $\sigma^2 = 0.0476$. We performed a Goodness of Fit

Test at the 0.05 level of significance on the data to see if the data followed the normal distribution with these parameters. Since the p-value > 0.05, we failed to reject the null hypothesis and therefore concluded the null hypothesis that the random variable WHIP followed an approximately normal distribution with $\mu = 1.343$ and $\sigma^2 = 0.0476$.

The following table shows the values of the pdf, cdf, and empirical cdf at 8 intervals:

WHIP Range	Pdf	Cdf	Empirical Cdf
Less than 0.688	0.001349898	0.001349898	0
0.688 to 0.906	0.021400234	0.022750132	0.009174312
0.906 to 1.125	0.135905122	0.158655254	0.087155963
1.125 to 1.343	0.341344746	0.5	0.463302752
1.343 to 1.561	0.341344746	0.841344746	0.844036697
1.561 to 1.779	0.135905122	0.977249868	0.972477064
1.779 to 1.997	0.021400234	0.998650102	0.995412844
Greater than 1.997	0.001349898	1	1

The close values of the cdf and empirical cdf further support our claim that WHIP followed an approximately normal distribution with $\mu = 1.343$ and $\sigma^2 = 0.0476$.

Finally, we found the MGF to be $M(t) = e^{1.343t+\frac{0.076t^2}{2}}$, $E(X) = 1.343$, and $E(X) = 1.343$.

5. We limited our sample to starting pitchers that through more than 20 innings in the 2017 season. Keeping all starting pitchers in the sample caused the distribution to be skewed right, as many pitchers that throw very few innings have a higher WHIP. Not only does the low number of innings pitched often give extreme values, but it may also be more likely for pitchers with extremely high WHIPs to throw less innings; poor performance in the beginning may lead to fewer opportunities to redeem themselves in the future. However, limiting the required number of innings pitched to a number much greater than 20 also caused problems, as it made the sample size much smaller and therefore more difficult to be approximated by the normal distribution. Because of this, requiring that the pitchers have more than 20 innings pitched does not necessarily have any meaning; limiting the sample size to pitchers who met this requirement just happened to appear to follow the normal distribution with the stated parameters.

6. The most difficult part of this project was finding data that would follow a distribution. We found many data sets that were close in expected and observed values in almost all intervals of a random variable (or values of a random variable), but were relatively far apart on one or two intervals. We had more difficulties finding data than we expected in the beginning, but finding our own data also allowed us to look at data on subjects in which we were interested and understood. Looking into data that we had a good understanding of made the other parts of the project easier and allowed us to give better interpretations of the data.