Name: *Caleb McWhorter — Solutions*
MAT 222
Spring 2019
Excel Lab 2: Ch. 7.1 & 7.2

"For you are life, rarer than a quark and unpredictable beyond the dreams of Heisenberg; the clay in which the forces that shape all things leave their fingerprints most clearly."
–Dr. Manhattan, Watchmen

The U.S. Equal Employment Opportunity Commission posts data about sex-based harassment claims filed with EEOC on their website. The data in `Sexual Harassment Claims.xlsx` was created based on the data reported on this website. Use Excel to answer the following questions based on the created data.

Problem 1: Complete the following:

Number of Months Collected: 12

Average Monthly Filings: 1087.91

Standard Deviation Monthly Filings: 207.48

Problem 2: Construct a 93% confidence interval for the average number of complaints filed. Fill in your data below.

Degrees of Freedom: 11

t^*: 2.007

Lower Interval Value: 967.73

Upper Interval Value: 1208.10

Problem 3: To determine whether the average monthly complaints has decreased since 2017, perform an appropriate hypothesis test using a significance level of 0.05. Be sure to state your null and alternative hypothesis, test statistic, p-value, and conclusion.

$$\begin{cases} H_0 : \mu = 1036 \\ H_a : \mu < 1036 \end{cases}$$

We have test statistics $t = 0.8668$ so that our p-value is 0.7977. Since $p \not< \alpha$, we fail to reject the null hypothesis. Hence, we cannot conclude that the average number of sexual harassment claims has dropped.

Problem 4: Discuss what statistical assumptions are required for the calculations above in Problems 2 & 3. What are the ways one could improve the validity of the computations above? Be sure to use statistical concepts from the course to support your answer.

Since $n < 15$, we need the monthly data to be normally distributed, with no skewness or outliers. Notice the box plot shows the data does not satisfy this. One could improve these calculations by using a large sample size.

The website https://www.baseball-reference.com/ compiles baseball related statistics. The site contains data on over 19,000 players, and contains data spanning over 100 years of baseball. The file `MLB Batting Averages.xlsx` contains the batting averages (BA) from several of the top players from the 1990s and the 2000s. Based on this data, use Excel to answer the following questions.

Problem 5: Complete the following:

Number 1990s Top Players:	35	Number 2000s Top Players:	39
1990s Top BA Average:	0.291	2000s Top BA Average:	0.328
1990s Top BA StDev:	0.0142	2000s Top BA StDev:	0.0161

Problem 6: Construct a 97% confidence interval for the difference between the 1990s top BA and the 2000s top BA. Fill in your data below.

Degrees of Freedom:	34
t^*:	2.562
Lower Interval Value:	0.029
Upper Interval Value:	0.047

Problem 7: To determine whether the BA of the top players has increased from the 1990s to the 2000s, perform an appropriate hypothesis test using $\alpha = 0.01$. Be sure to state your null and alternative hypothesis, test statistic, p-value, and conclusion.

Let μ_1 denote the batting average for the 2000s top players and μ_2 denote the batting average for the 1990s top players.

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_a : \mu_1 - \mu_2 > 0 \end{cases}$$

We have test statistic $t = 10.667$ with degrees of freedom 34 which results in a p-value of ≈ 0. Since $p < \alpha$, we reject the null hypothesis. Thus, we conclude the batting average of the top players most likely increased from the 1990s to the 2000s.

Problem 8: Discuss what statistical assumptions are required for the calculations above in Problems 6 & 7. What are the ways one could improve the validity of the computations above? Be sure to use statistical concepts from the course to support your answer.

Using d.o.f. $= \min(39, 35) = 35$, which is between 15 and 40, we need no skewness or outliers in the dataset. [This is the case, looking at the box plots, so there might be some concern about the accuracy of our conclusions.] Using the more exact degrees of freedom 72, we can use the methods even in the presence of skewness or outliers. Of course, one could choose more equal sample sizes or choose larger sample sizes to improve the strength of the conclusions.