Problem 1

You batter bell-lieve it

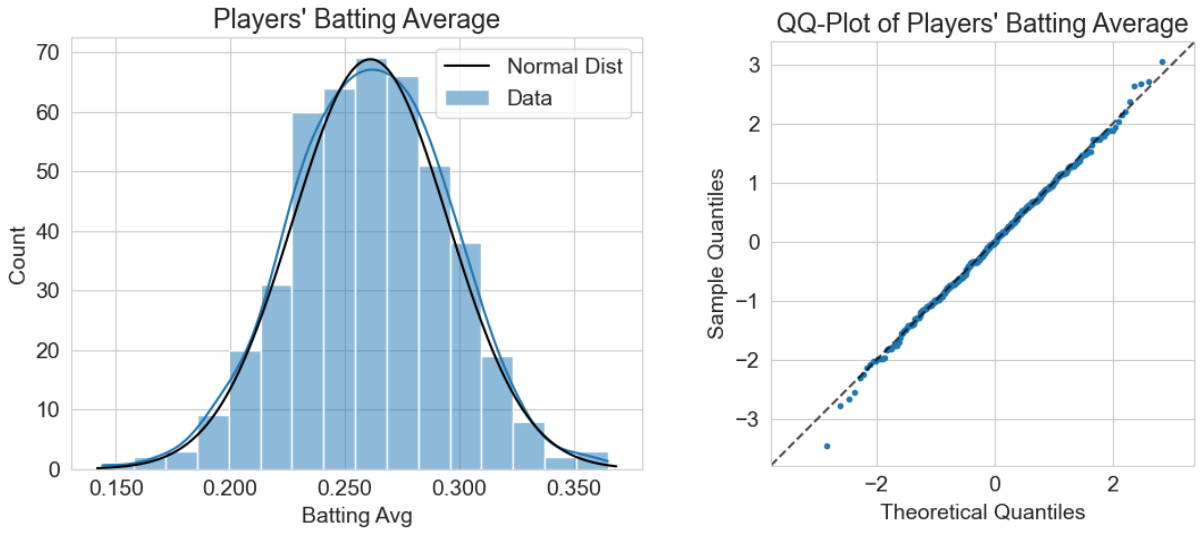
Use the file `BattingAverages.csv`, containing batting averages for all players with at least 100 at bats for the 2009 season, for the following questions. Assume this is a random sample rather than a census.

(a) Are the batting averages data (`BattingAvg`) approximately normally distributed? Use both graphical and analytical methods to make your argument.

Yes, they are approximately normally distributed. I first inspected the distribution of the players' batting averages using a histogram, as show in Figure 1a. The kernel density estimate (KDE) is also plotted with the histogram and both resemble a normal distribution. The black line on the plot is a normal distribute with with the mean (0.261) and standard deviation (0.033) calculated from the data. It is scaled to be a similar hieght to the distribution data. The data matches very closely to the normal distribution.

Figure 1b shows the QQ-plot, which also supports the normality of the distribution. The sample quantiles follow the expected theoretical quantiles except for at the tails of the distribution.

As a final check, I also ran an Omnibus test (because there were 446 data points) and the resulting p-value was 0.788. The null hypothesis for this test is that it is normally distributed. With such a high p-value, we fail to reject the null hypothesis. This also supports the visual indication of normality.



(a) (b)

Figure 1

(b) Is the mean value of batting averages greater than .265? Perform a test to find out. Use $\alpha = 0.05$ for your test.

Since we are attempting to make a determination about the mean of a single sample, I will use a one-sample t test. Let's check assumptions.

- Quantitative data: This is obviously true.
- Data is iid: Each player's batting average is independent of all others. We assume they are from the same normal distribution.
- Population distribution is normal: We already discussed in part (a) that the distribution is approximately normal.
- Large sample size: $n = 446$

For this test the null hypothesis is the mean value of the batting averages is equal to .265 and the alternate hypothesis is that the mean value is greater than .265.

$$H_0 : \mu = .265 \qquad H_a : \mu > .265$$

Since the alternate hypothesis is only that the mean is greater, it is a one sided test. The result of the one-sample t test gave a p-value of 0.991. Because $p > \alpha$ we fail to reject the null hypothesis. We can not conclude that the mean value of batting averages is greater than .265.

(c) Is there a difference between batting averages in the National League and American League (column `League`)? Use $\alpha = 0.05$ for your test.

The batting averages of the National League (NL) and American League (AL) can be treated as two separate samples. From the sample data the mean of the batting averages for each league (to three decimal places) are both 0.261. For this I will use a two-sample t test with the null hypothesis that there is no difference between the mean batting averages for the NL and AL. The alternate hypothesis is that there is a difference between the mean batting average for each league.

$$H_0 : \mu_{AL} - \mu_{NL} = 0 \qquad H_a : \mu_{AL} - \mu_{NL} \neq 0$$

The assumptions for this test are essentially the same as the single sample test. I checked the sample sizes for each league and they are 240 and 206 for the NL and AL respectively. Both of these are still large enough for the central limit theorem to apply. We could also make an assumption about pooled variance, but I did not do this for the test I ran.

Unlike the first hypothesis test, this is a two-sided test because we are looking for a difference that could be one greater or less than the other. The p-value for this test was 0.922. Once again $p > \alpha$ so we fail to reject the null hypothesis that there is no difference between the mean batting averages of each league.

Problem 2

Cowbell usage must have increased...

A researcher studying true body temperature in adult humans collected the data in `BodyTemp.csv` in degrees Fahrenheit.

(a) Is body temperature approximately normally distributed? Use graphical and analytical methods to make your argument.

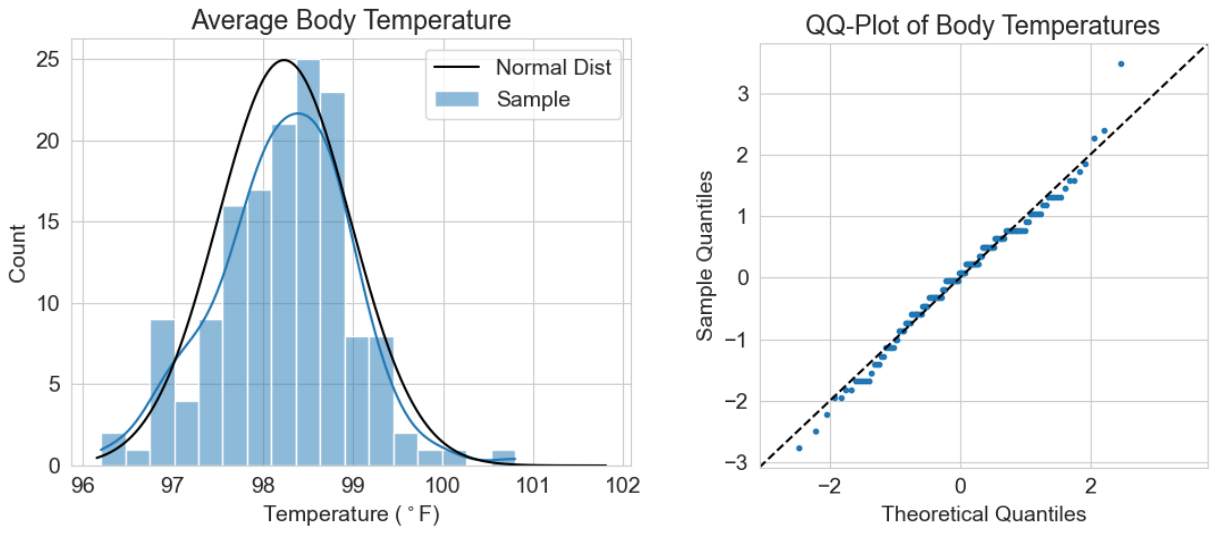
Figure 2a shows the distribution of body temperatures with a histogram and the KDE (in blue). There is some resemblance of a normal distribution but it appears like it is skewed to the left. The black line on the plot is a normal distribution with a mean (98.2°F) and standard deviation (0.737°F) calculated from the sample data. It is scaled to be close to the max histogram count. You can tell from this that the mean is to the left of the mode of the data, indicating it is left skewed.

The QQ-plot in Figure 2b also indicates slight left skewness. It is most noticeable in the lower left corner where the sample quantiles are more negative than the theoretical quantiles the farther from zero they are. It appears to hava a similar pattern on the positive side initially, but the last several data points do not follow that trend. Despite all of this the data is not far from a normal distribution, so I think it is approximately normal enough.

The sample size for this data set is $n = 148$. I used an Omnibus test and the result was a p value of 0.316. This also supports the claim that the distribution is not far from normal.

(b) Is the mean body temperature equal to 98.6°F?

To test this I will use a one-sample t test. The assumptions are valid because the data is quantitative, the data is reasonably iid, the sample distribution is approximately normal, and it is a large sample size. Since 98.6°F is regarded as the normal body temperature, I would like to be fairly confident in a result indicating it is not the mean. I will use a significance of $\alpha = 0.05$.



(a) (b)

Figure 2

The null hypothesis for this test is that the mean of the body temperatures is equal to 98.6°F. The alternate hypothesis is that the mean of the body temperatures is not equal to 98.6°F. This will be a two-sided test.

$$H_0 : \mu = 98.6 \qquad H_a : \mu \neq 98.6$$

The result of the t test is a p value of 1.24×10^{-8}, which is essentially zero. In this case $p < \alpha$ so we reject the null hypothesis in favor of the alternate hypothesis. The mean of the body temperatures is not 98.6°F.

(c) For the α you selected, what is the power to detect a difference of 0.2°F? Assume the population variance is equal to the sample variance.
Remember that effect size is difference to detect divided by standard deviation. See the progress check for examples.

To solve for the power we need the significance (α), the detectable difference (δ) and the number of observations (n). These values are:

$$\alpha = 0.05$$
$$\delta = 0.2$$
$$n = 148$$

To solve using the stats model in python the detectable difference is converted into an effect size by dividing the detectable difference by the standard deviation,

$$\text{effect size} = \frac{\delta}{\sigma}$$

The standard deviation of the sample is 0.737, which gives an effect size of 0.271. The result is a power of 0.906.

(d) Create a plot showing how α (x-axis) affects the power to detect a difference of 0.2°F (y-axis).

I created the plot by creating an array of α values from 0.01 to 0.99. For each value I calculated the corresponding power. The result is shown in Figure 3. The power initially increases rapidly towards 1

as α increases. As the power increases there will be less false negative (Type II Errors). As α increases, the confidence decreases and there will be more false positives (Type I Errors). Holding everything else constant we can trade off confidence for power, as show in Figure 3, to achieve a balance in acceptable errors of each type.

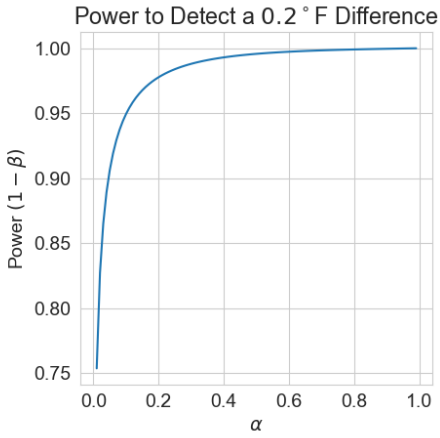


Figure 3

(e) Is there a difference between body temperature in males and females?

I split the data into two samples, one for male and female. The mean body temperatures from the sample data were 98.1°F and 98.4° for males and females respectively. The same assumptions above still apply and the sample size for each is $n = 74$, which is large enough for the central limit theorem to apply.

I then ran a two-sample, two-tailed t test to determine if there was a difference. I used the same significance value ($\alpha = 0.05$) as before. The null hypothesis is that there is no difference between the mean body temperatures in males and females. The alternate hypothesis is that there is a difference between the mean body temperatures.

$$H_0 : \mu_M - \mu_F = 0 \qquad H_a : \mu_M - \mu_F \neq 0$$

The result of the test was a p value of 0.006. Since $p < \alpha$, we reject the null hypothesis in favor of the alternate hypothesis. There is a difference between the means of the body temperatures for males and females.