Lab 6 - Normality and Sample Properties 1

Nick Sumpter (Edited by Eddie-Williams Owiredu & Guy Twa)

2023-09-06

- Today's Lab
- Loading Packages and Data
- The Normality Assumption
    - Assessing Normality Graphically
    - Assessing Normality Mathematically
- Independent Practice

Today's Lab

Last week, we were introduced to plotting statistical data with `ggplot` and calculating basic statistics. We were also introduced to the Central Limit Theorem which states that regardless of the distribution of a variable in a population, if you repeatedly sampled enough individuals from this population and calculated the sample means, the distribution of the sample means would be normally distributed.

This week, we will learn how to test the assumptions of normality. These assumptions are common to almost all parametric tests and thus you will be using these tests a lot throughout the remainder of this course.

Loading Packages and Data

For today's lab we will introduce the `stat.desc` function from the `pastecs` package. You will need to install these packages prior to loading them. We will also continue to use the `tidyverse` and `sciplot` packages.

```
install.packages("car")
install.packages("pastecs")
```

```
library(car)
library(pastecs)
library(sciplot)
library(tidyverse)

theme_set(theme_bw())
```

We will be using the same `microbiome` dataset from last week, except this week's version includes a Sex variable. Again, make sure to set your working directory to the appropriate location before loading in the data. Also, we will load in some custom functions with the `source` function.

```
setwd("/Users/eddie-williamsowiredu/Desktop/grd770_23/Lab6")

load("microbiome2.RData")

source("functions.R")
```

The Normality Assumption

Today we will go over how you would assess whether a variable is approximately normally distributed. This will be used in almost all labs throughout this course, so it would be good to practice this process on your own. Normality is assessed through a combination of methods, and using these methods, it should be possible to make a general statement about whether a variable is roughly normally distributed.

Assessing Normality Graphically

We can assess normality graphically using the following three plots:

1. Histograms
2. Q-Q plots
3. Boxplots

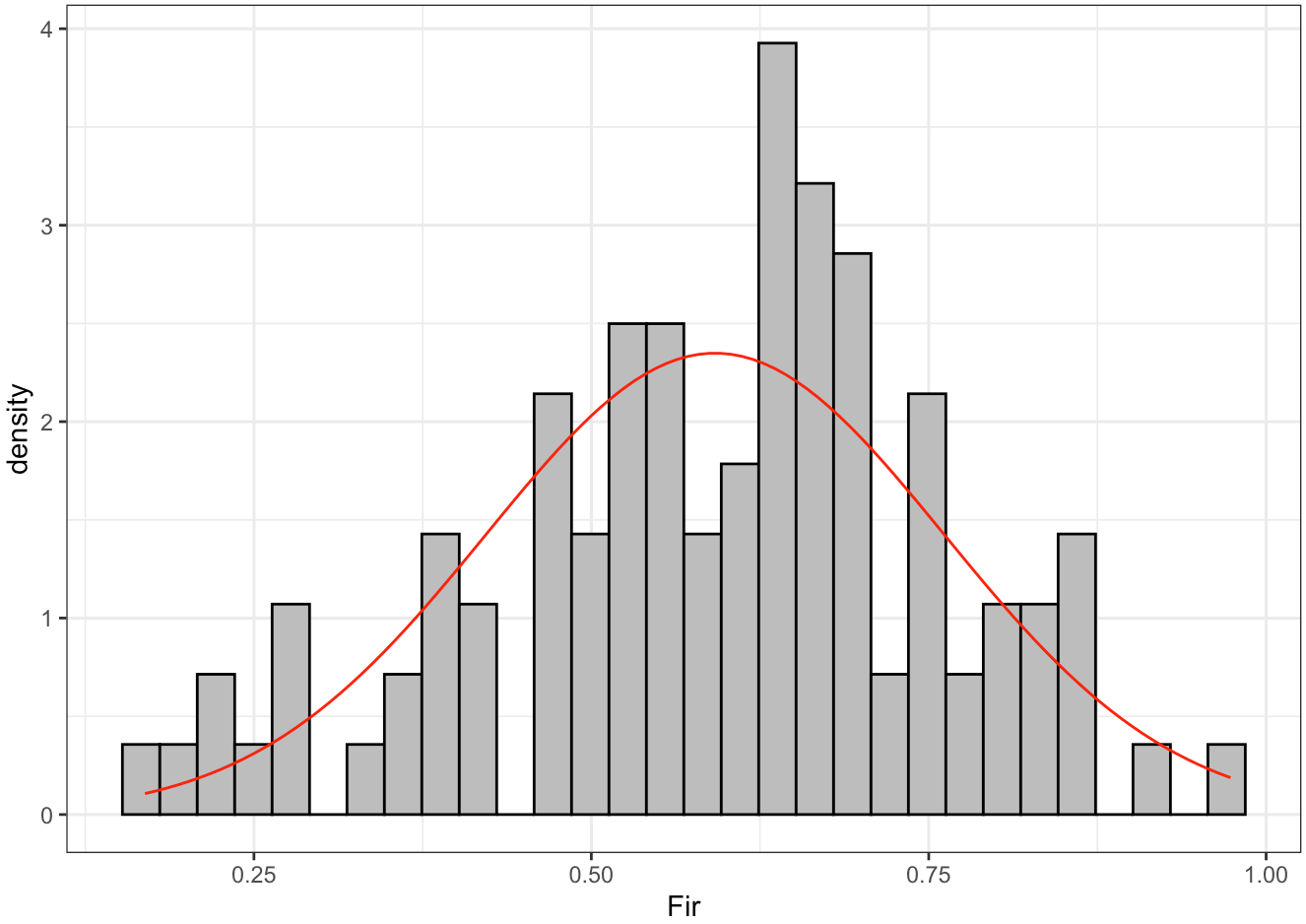
Histograms

Beginning with the `Fir` variable, we will produce a histogram as we did last week, though we will tell it to plot density on the y-axis rather than the counts. This is done by setting the y aesthetic to `..density..`.

To overlay the normal distribution on this histogram, we will use the `stat_function` command, which has the following form: `stat_function(fun, args)`. The `fun` argument specifies the function to be plotted, where `dnorm` is the function that creates a normal distribution (using density on the y-axis). Then, `args` is a `list` of arguments to be passed to `fun`. In this case, `dnorm` needs the mean and sd for our variable `Fir`.

Now let's create a histogram with a normal distribution overlaid on top for reference. What we are looking for is that the general shape of the histogram does not deviate significantly from the plotted normal distribution.

```
ggplot(data = microbiome2, mapping = aes(x = Fir)) +
  geom_histogram(mapping = aes(y = ..density..), bins = 30, fill = 'gray', color = 'blac
k') +
  stat_function(fun = dnorm,
                args = list(mean = mean(microbiome2$Fir), sd = sd(microbiome2$Fir)),
                color = 'red')
```



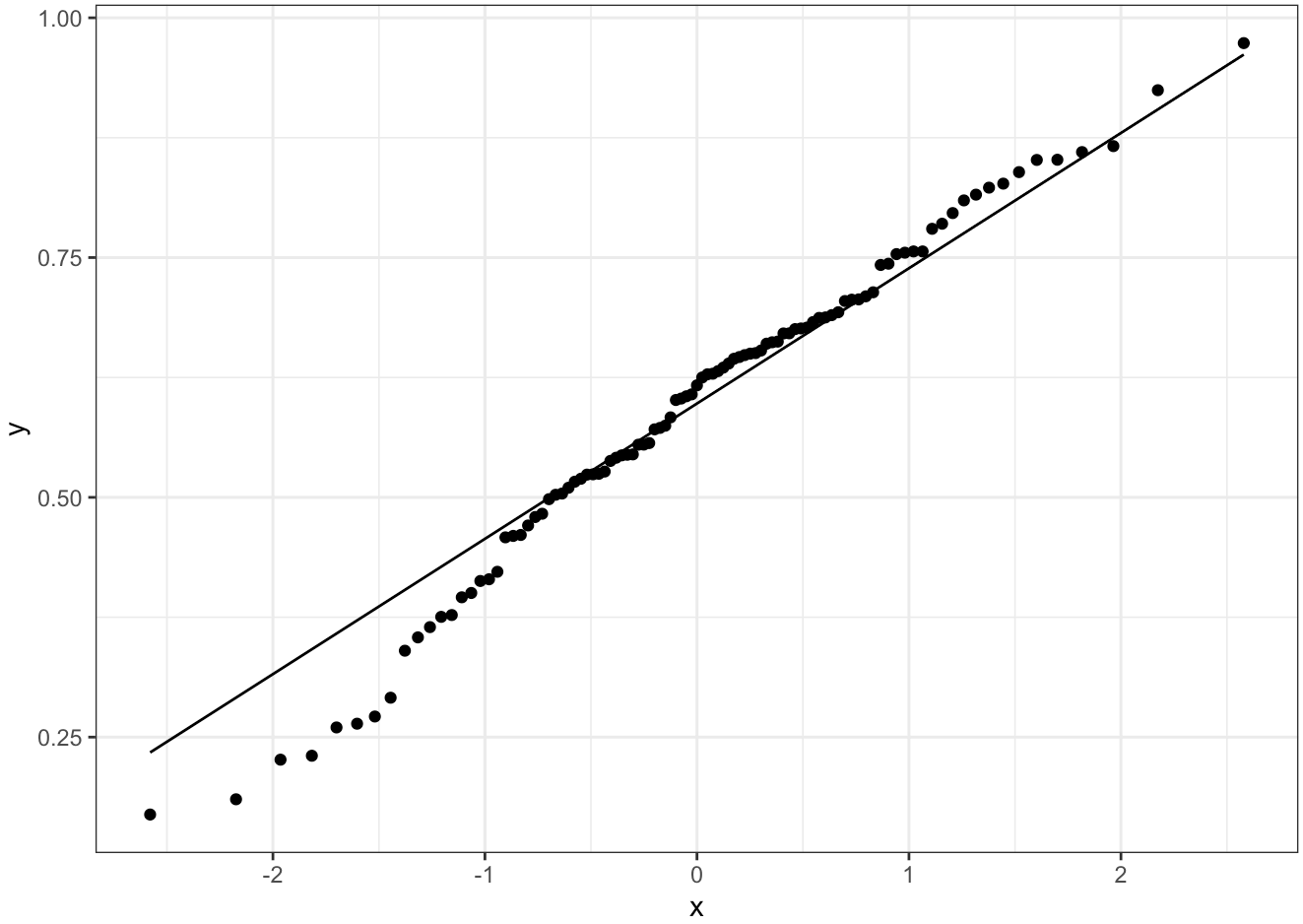
Based on this histogram, does the distribution of the `Fir` data seem normal? If so, why?

Yes, the distribution does look fairly normal. The histogram bins peak near the middle of the plot and tail off at each end. The distribution is also unimodal (there appears to be only one peak).

Q-Q Plots

The next plot we will use to assess normality is the Q-Q plot. Q-Q plots are dot plots that compare your data to a standard normal distribution. If your data is roughly normally distributed, most points should fall on the diagonal line. Q-Q plots are made using the `geom_qq` command and the diagonal is plotted using the `geom_qq_line` command. For basic Q-Q plots, the only necessary aesthetic is `sample`.

```
ggplot(data = microbiome2, mapping = aes(sample = Fir)) +
  geom_qq() +
  geom_qq_line()
```



Based on this Q-Q plot, does the distribution of the `Fir` data seem normal? If so, why?

Yes, this plot has most of the dots fall on or near the plotted diagonal, suggesting a normal distribution of the Fir variable. While the dots trail off from the line near the left end, this deviation from the line isn't enough to say the distribution is non-normal. Non-normal data will typically deviate substantially more from the diagonal line, especially at the tails.

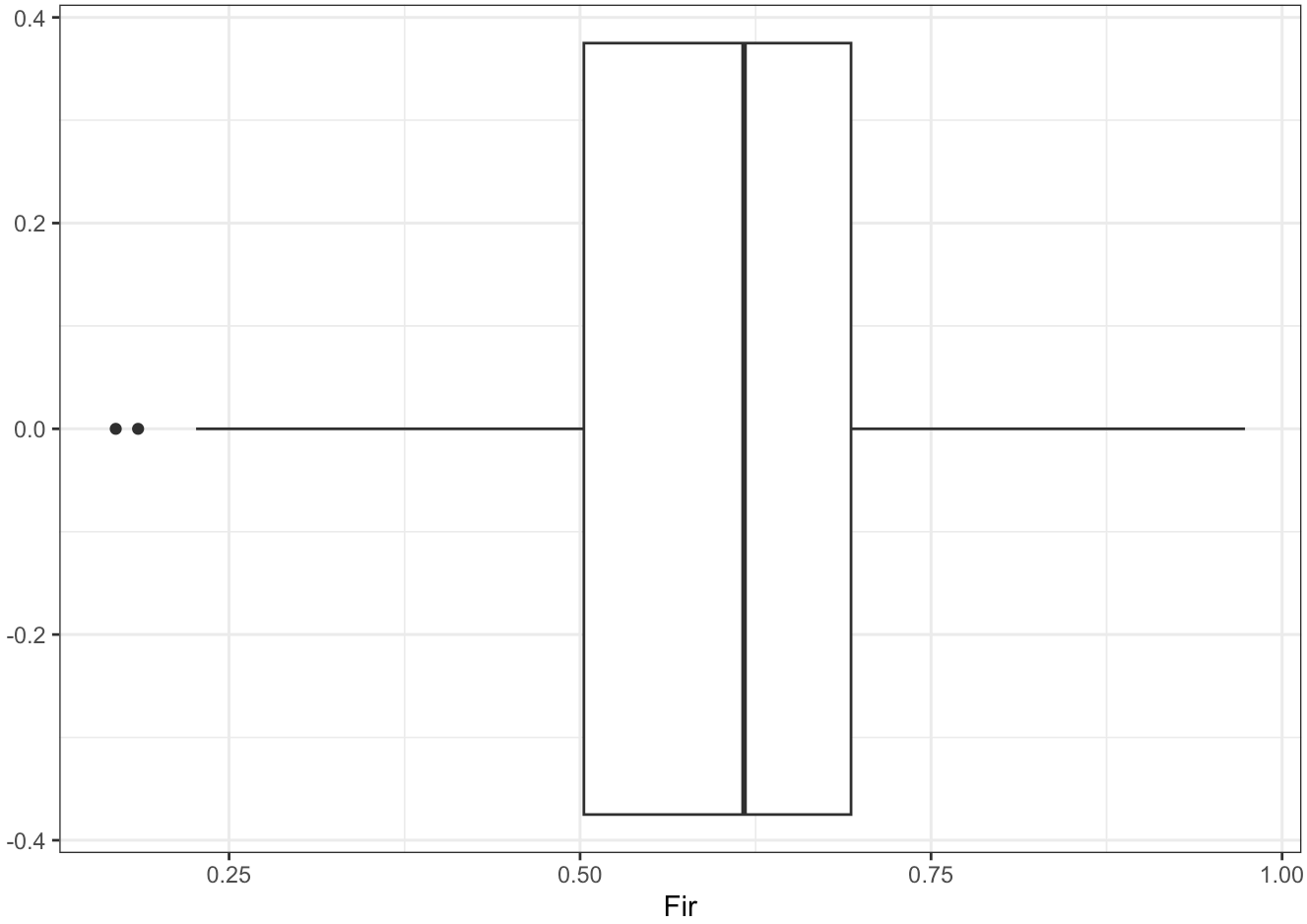
Boxplots

Boxplots are the last type of plot we will use to assess normality. Boxplots show the median, interquartile range, and any outliers for the data. If the data are normally distributed, you would expect the following three criteria to be roughly true:

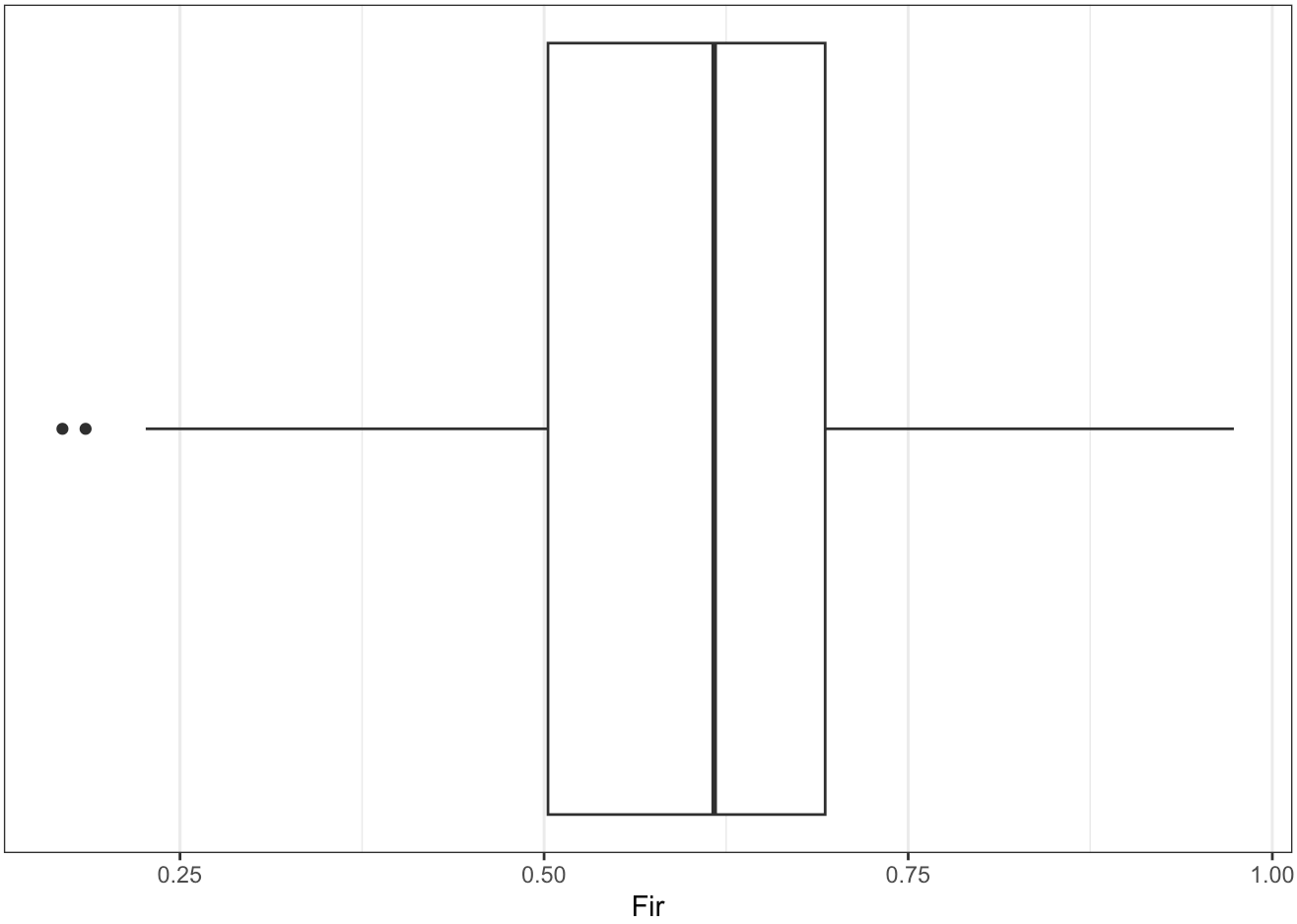
1. The median line should be roughly in the middle of the box
2. The two tails should be roughly the same length
3. There should be few outliers

To create a boxplot, you will use the `geom_boxplot()` function. Note that you can use either the y or the x aesthetic for this, though I prefer using the x aesthetic as it can be more easily compared to a histogram. Ignore the values on the y-axis as this is just a weird ggplot thing. We can tell it to suppress the axis labels and tick marks, along with the background lines for the y-axis with `theme`.

```
ggplot(data = microbiome2, mapping = aes(x = Fir)) +
  geom_boxplot()
```




```
# Remove y-axis stuff
ggplot(data = microbiome2, mapping = aes(x = Fir)) +
  geom_boxplot() +
  theme(axis.ticks.y = element_blank(),
        axis.text.y = element_blank(),
        panel.grid.major.y = element_blank(),
        panel.grid.minor.y = element_blank())
```



Based on this boxplot, does the distribution of the `Fir` data seem normal? If so, why?

Yes, the distribution seems to be fairly normal. The left and right whisker lengths are pretty even and the median line is close to the center of the box. There are two outliers at the left end of the distribution, however since we have a dataset of 101 values, these constitute only 2% of all values. This satisfies the “few outliers” condition for a normal distribution.

Assessing Normality Mathematically

While useful in most cases, sometimes the graphical assessments of normality are not easily interpreted. This is often the case with low sample sizes, where histograms are limited by the number of bins, dot placement on the Q-Q plot can vary wildly without a general trend in the dots being obvious, and boxplot shapes can be less than helpful. In these cases, mathematical calculations of normality can be extremely useful. **To be clear, you should run both graphical and mathematical assessments of normality in all cases where the assumption of normality applies. This will be especially important on your exams.**

Calculating Skew, Kurtosis, and the Shapiro-Wilk Test

Reporting skew and kurtosis for your data can be informative for assessing normality. In order to calculate these measures, we will use the `stat.desc.clean` function, which is a modification of the `stat.desc` function from the `pastecs` package.

`stat.desc.clean` has the following inputs:

- `dataset` : the dataset of interest
- `variable` : the variable of interest
- `...` : any grouping variable of interest for subsetting the data

Let's go ahead and calculate skew and kurtosis values for our `Fir` data:

```
stat.desc.clean(dataset = microbiome2, variable = Fir)
```

```
## # A tibble: 1 × 6
##   skewness skew.2SE kurtosis kurt.2SE normtest.W normtest.p
##      <dbl>    <dbl>    <dbl>    <dbl>      <dbl>      <dbl>
## 1   -0.371   -0.773   -0.194   -0.204      0.982      0.190
```

So we have values but how do we interpret them? The outputs we are concerned with for skewness and kurtosis are: **skewness, skew.2SE, kurtosis, and kurt.2SE**.

- **Skewness** can be either positive or negative. Closer to 0 means less skewed. A positive skewness value means there is a pile-up of data points on the left side of the distribution, while a negative skewness value means they are piled on the right side. In other words, the direction of the longer tail indicates whether it is positive or negative skew.
- **Kurtosis** can be positive or negative. A positive value means the tails drop rapidly (i.e. a pointy distribution), negative values indicate a much more gradual fall (a flatter distribution). Positive = leptokurtic, negative = platykurtic.

For either of these statistics, values further from 0 indicate that the data is further from a normal distribution. Typically, skewness values between -0.5 and 0.5 are approximately symmetric, between 0.5 and 1 (or -0.5 and -1) are slightly skewed, and > 1 (or < -1) are highly skewed. These cutoffs can be applied to the kurtosis values as well.

The **.2SE** values show skewness and kurtosis values divided by 2 SEs. This is essentially converting these statistics to z-scores, which makes it possible to easily assess whether there is significant skewness or kurtosis. If you just divided by 1 SE for each, this would be a true z-score, with anything outside of -1.96 and 1.96 being significant. As these values are further divided by 2, you can simply use a cutoff of -1 to 1, with anything outside of these values indicating either significant skewness or kurtosis. **USE CAUTION WHEN APPLYING THESE**. For large samples (> 200 samples), it is important to look at the shape itself rather than relying on these rules of thumb.

Thus, based on the skewness and kurtosis statistics, does the distribution of the `Fir` data seem normal? If so, why?

Yes, skewness is between -0.5 and 0.5 meaning there is relatively little skew to the data. The skew.2SE value being between -1 and 1 agrees with this. Likewise, kurtosis has a value of -0.19. So while the distribution is slightly platykurtic, it's close enough to 0 to be considered normal. The kurt.2SE value agrees with this assessment, since the value of -0.204 is comfortably between -1 and 1.

`stat.desc.clean` also outputs the results from the Shapiro-Wilk test (`normtest.W` and `normtest.P`). The Shapiro-Wilk test (SW test) is a simple statistical assessment of whether a distribution resembles a normal distribution. The `normtest.W` value corresponds to the test statistic (W) while `normtest.p` is the p value. If p > 0.05, the distribution is not significantly different from a normal distribution. When reporting results of the SW test, do not say the SW test proves the distribution is normal or it is significantly normal if the p-value is > 0.05. All it

shows is that the distribution is not significantly different from a normal distribution, and there is a slight, but meaningful, difference between the two. If $p < 0.05$, say the Shapiro-Wilk test suggests the distribution is significantly different from a normal distribution.

Based on the SW test, does the distribution of the `Fir` data seem normal? If so, why?

Yes, the p-value for the SW test is 0.19, therefore the distribution is not significantly different from a normal distribution.

Taking all of the graphs and tests into account, is `Fir` normally distributed?

Yes, the Fir variable can be considered normal. The histogram peaks near the middle of the distribution, gradually falls off at either end, and appears to be unimodal (only has one peak). The Q-Q plot also shows little evidence of non-normality. Most points are almost directly on the line, those that aren't are fairly close and do not drastically tail off, and the shape of the points is fairly linear. The boxplot shows even-length whiskers and a median line near the center of the box. There are two outliers, but this is a small percentage of the total dataset. The graphical components together suggest that the variable is roughly normally distributed. Likewise, the numerical tests of skewness indicate the data are not significantly skewed. Skewness is between -0.5 and 0.5 while the skew.2SE value is between -1 and 1. The distribution is slightly platykurtic with a kurtosis value of -0.194, but this is much less than -0.5, and the 2SE value is between -1 and 1, so the tails of the distribution fall off fairly normally. Overall, both graphical and numerical tests suggest that this variable is roughly normally distributed, and thus this data would meet the assumption of normality.

Independent Practice

1. Test the 2 assumptions for the `Bacter` data.
    a. Is the data approximately normal?
    b. Is the variance approximately equal between sexes?
2. Test the 2 assumptions for the `Actin` data.
    a. Is the data approximately normal?
    b. Is the variance approximately equal between sexes?
3. Test the 2 assumptions for the `Verru` data.
    a. Are the data approximately normal?
    b. Is the variance approximately equal between sexes?