############## EVALUATING NORMALITY & VARIANCE ##############

Data are often not normally-distributed, nor do they have homogeneous variance (i.e, normal
curve 1 is wider than normal curve 2). Both these problems violate assumptions of parametric
statistics – the statistics we use this semester. If you analyze data that violate the assumptions,
you may get the wrong answer to your question!

#Download the helicopter data from the course web page:
https://sciences.ucf.edu/biology/d4lab/wp-content/uploads/sites/125/2019/09/helicopter-data.txt
And then start up RStudio, and import the data

Make boxplots to squint at the data ***per Group*** (because there are only 3). Use R code you learned last week.
Is each boxplot centered (i.e., medians are in the middle of boxes, whiskers of same length
above and below)? If so, then each data set *may* be normal.

Are the sizes of boxplots and whisker lengths about the same among IDs?
If so, then variances *may* be homogeneous.

 # Generate a subset for *each* Group.
IMPORTANT: We do this because **we need to evaluate normality For Each Treatment**.
Remember, we wish to compare data sets, where the assumption is that data ***for each data set***
(e.g., Group) are normally distributed and that variances are equal among IDs. Commands here
are shown for only one subset – you need to repeat these for each of the subsets. For example:

```
dataW <- subset(data, data$GROUP == "W")
```

This identifies a subset for Group W in the data file named “data”.

Normality -First we evaluate normality, then homogeneity of variance.

Calculate the mean and SD of **each design**, like this:

```
meanTw <- mean(dataW$Time)
sdTw <- sd(dataW$Time)
```

Calculate and draw the normal curve on the histogram of the data

```
h <- hist(dataW$Time, breaks=10) # make histogram
xfit <- seq(min(dataW$Time), max(dataW$Time), length=100)
yfit <- dnorm(xfit, mean=mean(dataW$Time),   sd=sd(dataW$Time))
yfit <- yfit*diff(h$mids[1:2])*length(dataW$Time)
# the above 3 lines make data for a normal curve to match
lines(xfit, yfit, col="blue", lwd=2) # plots that curve
```

Are the Group data each looking normal? Any that are not so normal-ish?

Now let's make a QQ plot (aka normality plot) of the data)

```
qqnorm(dataW$Time)
qqline(dataW$Time)
```

Lastly, we run a stats test on normailty:

```
shapiro.test(dataW$Time)
```

Homogeneity of Variance – this assumption is even more important for parametric statistics
than normality (you might often hear that statistics are “robust to violations of assumptions”).
That only goes so far, and you should stretch that boundary less for homogeneity of variance
than for normality.

Two tests are common for homogeneity of variance: Bartlett's and Levene's. Bartlett's works
well if data are normal, but not if data are non-normal. Levene's test is more robust to
heterogeneity of variance than Bartlett's. Thus

choose the right test based on normality tests above.

To run Bartlett's test simply enter

```
bartlett.test(Time ~ GROUP)
```

To run Levene's test, first install and then load the package “car” (for Companion to Applied Regression – nothing to do with cars data) then enter

```
fGROUP <- factor(GROUP)
```

This command ensures that R knows groups are factors. Then enter

```
leveneTest(Time ~ fGROUP)
```

So how do our data look? Specifically, can we assume Groups are normally distributed?
And do Groups have homogeneous variance?
What about another treatment? Try something simple like Body Width or Fold.

You have now evaluated important statistical assumptions graphically and by statistical tests – these tools will be important for many analyses hereafter.