Directions: Complete the exercises below. When you are finished, turn in any required files online in Canvas, then check-in with the Lab TA for dismissal.

Diagnosing Assumptions in R

Consider an experiment on the effects of vitamins on the growth of guinea pig teeth. The file, `guinea_teeth.csv` (posted in Canvas) contains the length of teeth (`growth`) for guinea pigs whose diets were randomly assigned for vitamin C supplements (`trt`), either by ascorbic acid or orange juice (treatment labels 0 and 1, respectively).

The following examples show how to assess the assumptions necessary to perform the traditional t-test (with equal variances) for difference in mean teeth growth in R. The full R script is provided in the `teeth_Lab4.r` file posted in Canvas.

- First, load in the data using the *Import Dataset* tool in R Studio. Be sure to change the variable type on the `trt` column to "factor" and enter "0,1" as the levels.

```
library(readr)
guinea_teeth <- read_csv("guinea_teeth.csv", col_types = cols(trt =
                                    col_factor(levels = c("0", "1"))))
```

- To check the equal variance assumption, there are three possible (non-graphical) methods, as discussed in lecture:

  1. The ratio of sample standard deviations using the `sd()` function in R:

  ```
  sd(guinea_teeth$growth[guinea_teeth$trt=="1"])/
      sd(guinea_teeth$growth[guinea_teeth$trt=="0"])
  ```

  Remember that you may need to switch the order to place the larger value in the numerator.

  2. The F-test using the `var.test()` function in R:

  ```
  var.test(x=guinea_teeth$growth[guinea_teeth$trt=="1"],
           y=guinea_teeth$growth[guinea_teeth$trt=="0"],
           alternative="greater")
  ```

  Be sure to put the group with the largest sample standard deviation as `x`, the group with the smallest sample standard deviation as `y`, and specify that you want the right-tailed test using the `alternative="greater"` option. Recall the null hypothesis is equal variances.

  3. The Brown-Forsythe test does not have a corresponding built-in function in R, so we must write our own function, `BF.var.test()`:

  ```
  BF.var.test <- function(dat.response, dat.treatment){
    n1 = length(dat.response[dat.treatment==levels(dat.treatment)[1]])
    n2 = length(dat.response[dat.treatment==levels(dat.treatment)[2]])
    M = c(rep(median(dat.response[dat.treatment==levels(dat.treatment)[1]]), n1),
          rep(median(dat.response[dat.treatment==levels(dat.treatment)[2]]),n2))
    Z = abs(c(dat.response[dat.treatment==levels(dat.treatment)[1]],
  ```

```
                dat.response[dat.treatment==levels(dat.treatment)[2]]) - M)
        G = c(dat.treatment[dat.treatment==levels(dat.treatment)[1]],
            dat.treatment[dat.treatment==levels(dat.treatment)[2]])
        df = length(Z)-2
        BFstat = (t.test(Z~G, var.equal=T)$statistic)^2
        pval = pf(BFstat, 1, df, lower.tail=F)
        return(data.frame(BFstat=BFstat, pval=pval, row.names="results:"))
    }
```

To use the function, you need to specify the response variable and then the treatment/group variable:

```
BF.var.test(guinea_teeth$growth, guinea_teeth$trt)
```

The test will output the corresponding F-statistic first and then the p-value. Note that the null hypothesis is equal variances.

- Graphical displays can be used to assess both the equal variance and normal assumptions:

  1. A side-by-side boxplot of the response values within each group can be obtain using the `boxplot()` function:

     ```
     boxplot(guinea_teeth$growth ~ guinea_teeth$trt, xlab="Treatment",
             ylab="Growth", main="Guinea Pig Teeth Experiment")
     ```

     These are useful for assessing both assumptions of equal variance and normality.

  2. Histograms of the response values within each group can be obtain using the `hist()` function:

     ```
     par(mfrow=c(2,1))
     hist(guinea_teeth$growth[guinea_teeth$trt=="1"],
          main="Treatment = orange juice", xlab="Growth")
     hist(guinea_teeth$growth[guinea_teeth$trt=="0"],
          main="Treatment = ascorbic acid", xlab="Growth")
     ```

     These are useful for assessing both assumptions of equal variance and normality.

  3. Normal Q-Q plots of the response values within each group can be obtain using the `qqnorm()` function:

     ```
     par(mfrow=c(1,2))
     qqnorm(scale(guinea_teeth$growth[guinea_teeth$trt=="1"]),
            main="Treatment = orange juice")
     abline(a=0, b=1, col="red")
     qqnorm(scale(guinea_teeth$growth[guinea_teeth$trt=="0"]),
            main="Treatment = ascorbic acid")
     abline(a=0, b=1, col="red")
     ```

     The `scale()` function is used to "standardize" the response values (by subtracting off the mean and dividing by the standard deviation) so they can be compared to the standard normal quantiles. The `abline()` function is used to create the diagonal reference line. These plots are only useful for assessing the normality assumption.

- Numerical summaries, including the mean, median, standard deviation, interquartile range, skew, kurtosis, and excess kurtosis, can be computed for each group in R using the following

code:

```
library(tidyverse)
library(moments)
numerical_stats <- guinea_teeth |> group_by(trt) |> summarize(
    Y_mean = mean(growth),
    Y_med = quantile(growth, 0.5),
    Y_sd = sd(growth),
    Y_IQR = quantile(growth, 0.75) - quantile(growth, 0.25),
    Y_skew = skewness(growth),
    Y_kurt = kurtosis(growth),
    Y_excess = Y_kurt - 3)
numerical_stats
```

If you have not already installed the `moments` package, you can do so using the code:

```
install.packages("moments")
```

- Statistical tests, such as the Shapiro-Wilk test, can be used to further assess the normality assumption in R using the `shapiro.test()` function:

```
shapiro.test(guinea_teeth$growth[guinea_teeth$trt=="1"])
shapiro.test(guinea_teeth$growth[guinea_teeth$trt=="0"])
```

A summary of results displays the test statistic value and associated p-value.

Remedies in R

Refer to the guinea pig teeth study described above.

- As we discussed during the last lab, the traditional t-test can be conduct in R using the `t.test()` function with the `var.equal` option to true:

```
t.test(growth~trt, data=guinea_teeth, var.equal=T)
```

- If the equal variance assumption does not hold, you can conduct the Welch test with the Satterthwaite approximation by setting the `var.equal` option to false:

```
t.test(growth~trt, data=guinea_teeth, var.equal=F)
```

- If there is reason to believe that neither the equal variance or the normality assumptions of the traditional t-test will not hold, the example R code below will show you how to conduct the Wilcoxon rank-sum test using the `wilcox.test()` function:

```
wilcox.test(growth~trt, data=guinea_teeth, exact=F)
```

You can set the `exact` option to true to compute the exact p-value for the test, or to false to approximate the p-value when the sample size is too large. The function will output a summary including the statistic W, which is equal to the sum of the ranks in the first group minus the quantity $n_1(n_1 + 1)/2$, and the corresponding p-value.

Now, consider another dataset containing the radon concentration levels (`radon`) for a selection of homes in two different counties (`county`) in Minnesota, Olmsted and Stearns. The data are found in the `minn_radon.csv` file posted in Canvas. While exploring this dataset, you should see that the radon concentration levels are non-normal within the counties. One possible remedy for this if the researchers desire a model-based inferential procedure is to find a transformation of the data that will result in normality. The following example will show you how to conduct the transformation in R:

- First, load in the data using the *Import Dataset* tool in R Studio. Be sure to change the variable type on the `county` column to "factor" and enter "OLMSTED,STEARNS" as the levels.

  ```
  library(readr)
  minn_radon <- read_csv("minn_radon.csv",col_types = cols(county =
                                  col_factor(levels = c("OLMSTED", "STEARNS"))))
  ```

- Now, transform the response variable using the appropriate function in R:

  - Square-root transformation: `X = sqrt(minn_radon$radon)`

  - Log transformation: `X = log(minn_radon$radon)`

  - Arcsin-root transformation: `X = asin(sqrt(minn_radon$radon))`

  - Box-Cox transformation $\left(X = \dfrac{Y^{\lambda} - 1}{\lambda}\right)$:

    ```
    library(MASS)
    bct <- boxcox(lm(radon~county, data=minn_radon))
    lambda <- bct$x[which.max(bct$y)]
    X=(minn_radon$radon^lambda-1)/lambda
    ```

    If you have not already installed the `MASS` package, you can do so using the code:

    ```
    install.packages("MASS")
    ```

  and append it to the original dataset: `minn_radon = cbind(minn_radon, X)`

Assignment

1. Use R to assess the assumptions of the traditional t-based inference procedure for the guinea pig teeth study:

   (a) Describe the independent treatment groups assumption in the context of the study. Explain why this assumption is valid.

   (b) Describe the equal variance assumption in the context of the study. Check whether this assumption is appropriate. Justify your response by including all relevant graphs, summary statistics, test results, etc.

   (c) Describe the normal distribution assumption in the context of the study. Check whether this assumption is appropriate. Justify your response by including all relevant graphs, summary statistics, test results, etc.

2. Perform a Wilcoxon rank-sum test to determine if the distributions of teeth lengths are the same for the two treatment groups.

   (a) What is the value of the test statistic W? What is the corresponding sum of the ranks in the first group?

   (b) What is the value of the p-value?

   (c) Interpret the result of the test in the context of the study.

3. Use R to complete the following exercises for the radon study:

   (a) Which transformation of the data will result in normal distributions of radon concentration levels within each county?

   (b) Conduct the traditional t-test (with equal variances) for the transformed data. Interpret the results in the context of the study.

Total: 25 points   # **correct:** _____   %: _____