Lab 4

2024-09-18

Stat 5000Lab #4
Fall 2024 Due Tue Sep 24th Name:

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:

  ```
  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)
```

```
##            BFstat       pval
## results: 3.380434 0.08253377
```

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:

  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)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.0.2
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
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
```

```
## # A tibble: 2 x 8
##   trt   Y_mean Y_med  Y_sd Y_IQR Y_skew Y_kurt Y_excess
##   <fct>  <dbl> <dbl> <dbl> <dbl>  <dbl>  <dbl>    <dbl>
## 1 0       7.98  7.15  2.75  4.95  0.156   1.47   -1.53
## 2 1      13.2  12.2   4.46  6.48  0.513   2.01   -0.987
```

```
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-Wilk normality test
##
## data:  guinea_teeth$growth[guinea_teeth$trt == "1"]
## W = 0.89274, p-value = 0.182
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  guinea_teeth$growth[guinea_teeth$trt == "0"]
## W = 0.89, p-value = 0.1696
```

```
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)
```

```
##
##  Two Sample t-test
##
## data:  growth by trt
## t = -3.1697, df = 18, p-value = 0.005304
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -8.729738 -1.770262
## sample estimates:
## mean in group 0 mean in group 1
##            7.98           13.23
```

- 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)
```

```
##
##  Welch Two Sample t-test
##
## data:  growth by trt
## t = -3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -8.780943 -1.719057
## sample estimates:
## mean in group 0 mean in group 1
##            7.98           13.23
```

- 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)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  growth by trt
## W = 19.5, p-value = 0.02319
## alternative hypothesis: true location shift is not equal to 0
```

```
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")))
                       )
```

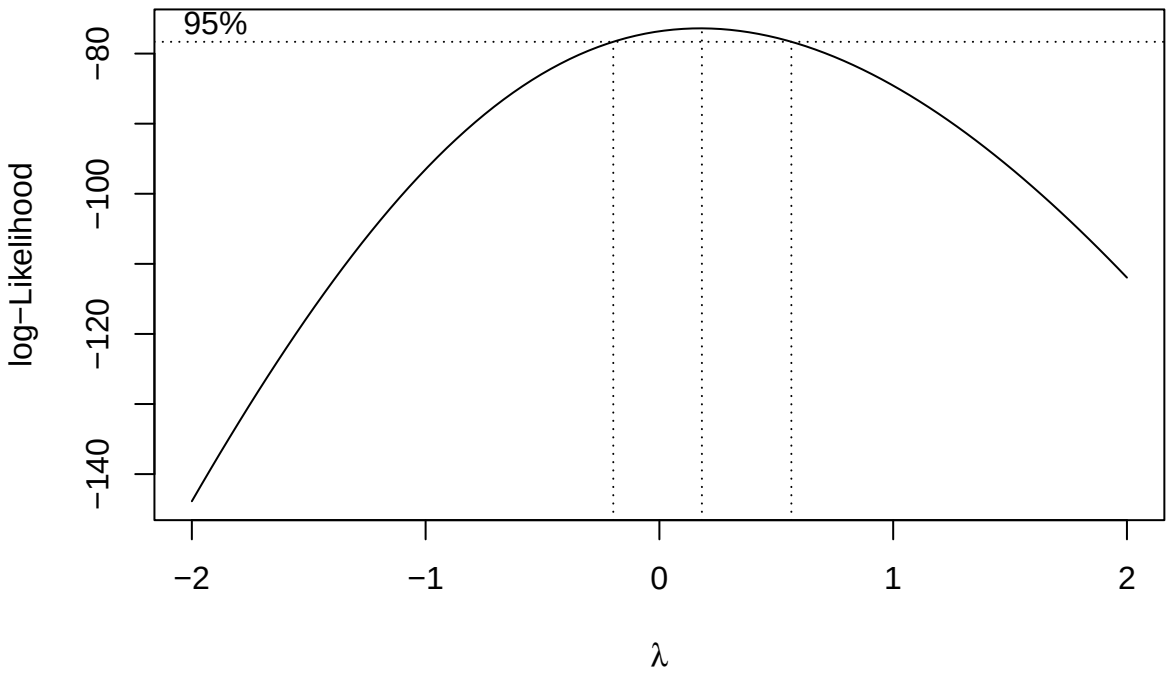
- Now, transform the response variable using the appropriate function in R:
  - Square-root transformation: `X = sqrt(minn_radon$radon)`
  - Log transformation: `X = log(minn_radon$radon)`
  - Arcsin-root transformation: `X = asin(sqrt(minn_radon$radon))`
  - Box-Cox transformation $\left(X = \frac{Y^{\lambda} - 1}{\lambda}\right)$:

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
bct <- boxcox(lm(radon~county, data=minn_radon))
```




```
lambda <- bct$x[which.max(bct$y)]
X=(minn_radon$radon^lambda-1)/lambda
```

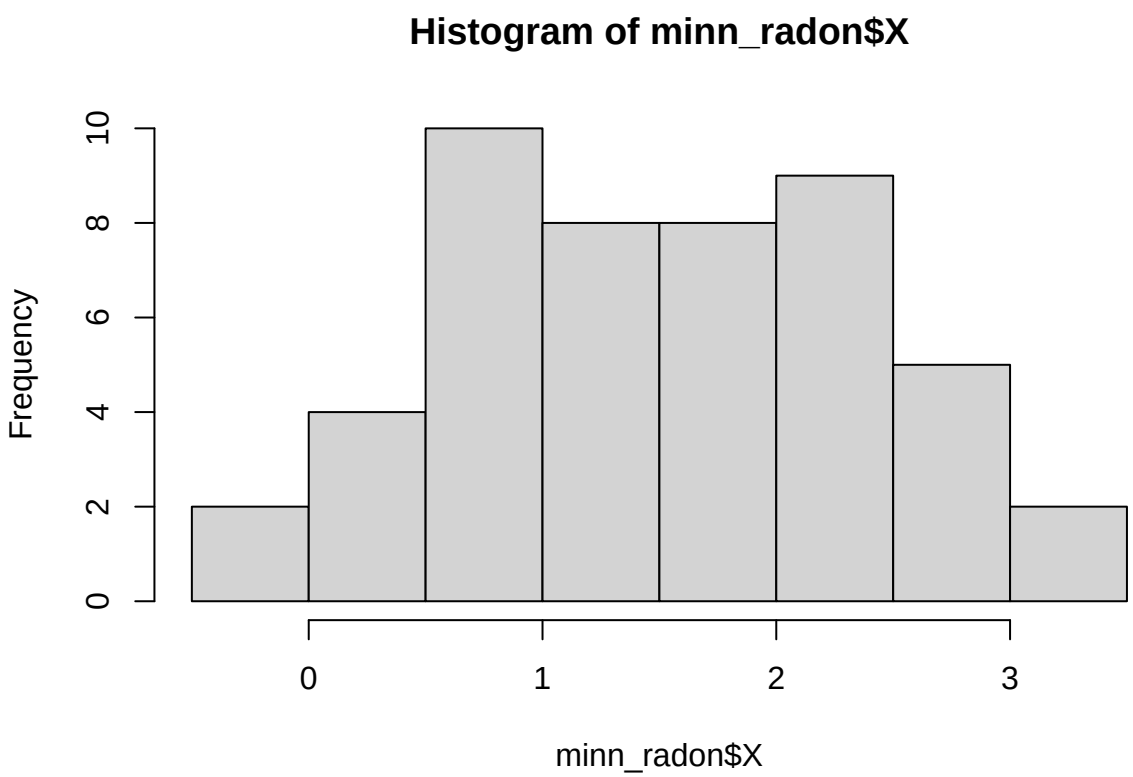
```
    If you have not already installed the ‘MASS‘ package, you can do
    so using the code:

                install.packages("MASS")
```

and append it to the original dataset:

```
minn_radon = cbind(minn_radon, X)
# minn_radon
```

```
hist(minn_radon$X)
```


Histogram of minn_radon$X
Frequency
0
2
4
6
8
10
0
1
2
3
minn_radon$X

Assignment

1.

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

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

Description: Our assumption is that guinea pigs (our units for the study) have independence both within and between the two groups, and that all units for a particular group belong to the same population (identically distributed). In the context of this study, we have no reason to assume this is violated, as our participants are all guinea pigs taken from the same pool of guinea pigs, and their treatments are randomly assigned, e.g. we have no reason to believe the total pool of guinea pigs was composed of two distinct types of guinea pigs or some underlying condition existing within the population that would warrant the believe of more than one population being contained within each sample.

2. 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.

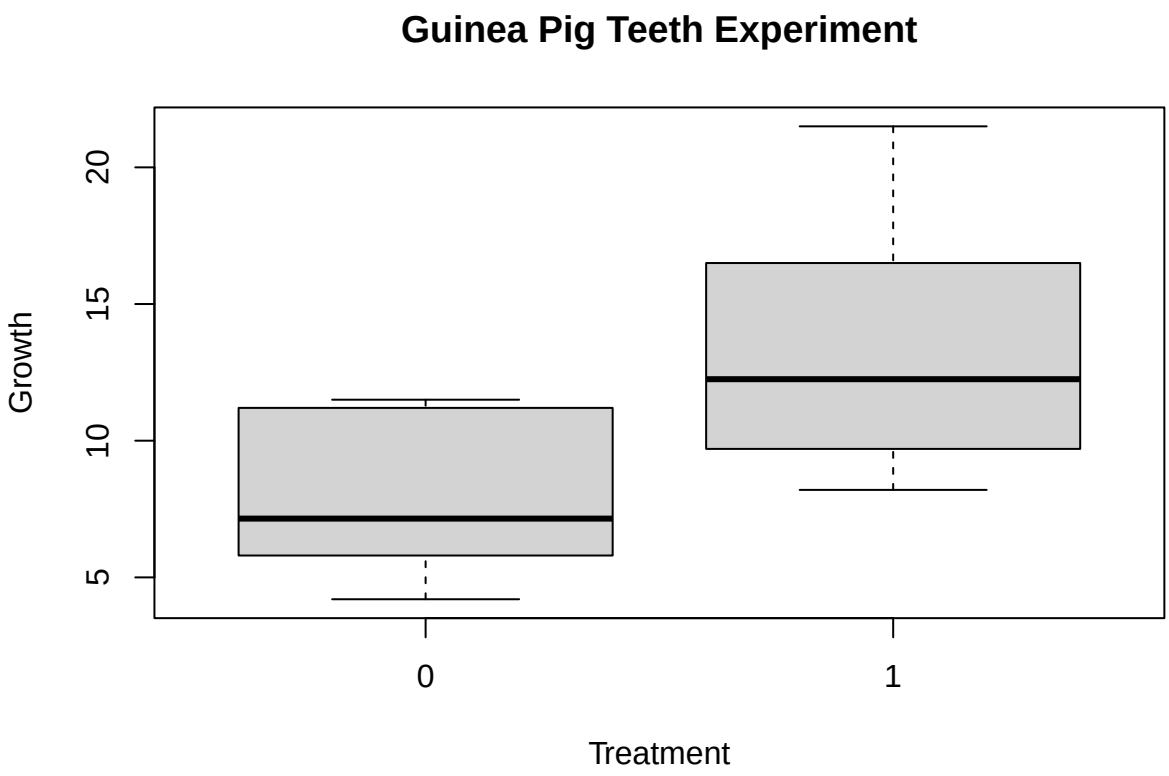
```
sd(guinea_teeth$growth[guinea_teeth$trt=="1"])/
sd(guinea_teeth$growth[guinea_teeth$trt=="0"])
```

```
## [1] 1.623699
```

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

```
##
##  F test to compare two variances
##
## data:  guinea_teeth$growth[guinea_teeth$trt == "1"] and guinea_teeth$growth[guinea_teeth$trt == "0"]
## F = 2.6364, num df = 9, denom df = 9, p-value = 0.08245
## alternative hypothesis: true ratio of variances is greater than 1
## 95 percent confidence interval:
##  0.8293452       Inf
## sample estimates:
## ratio of variances
##             2.6364
```

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

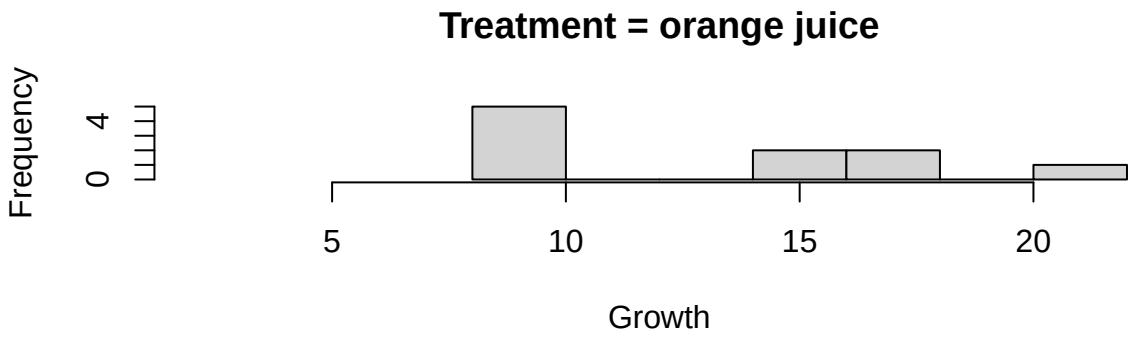


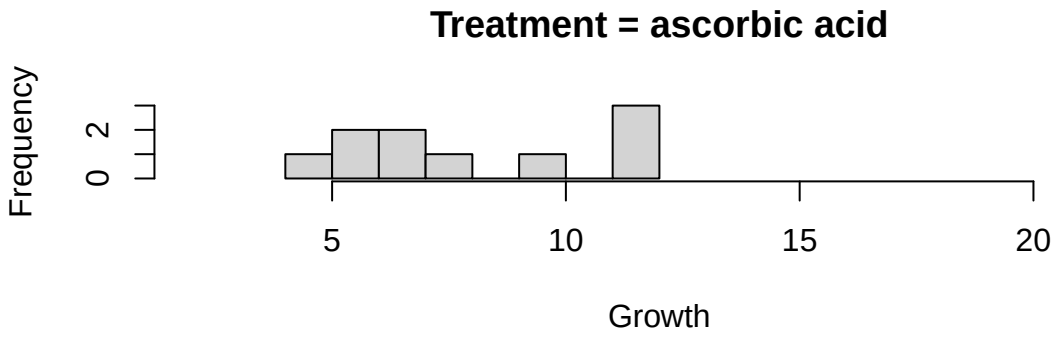
Description: In the context of the study, the equal variance assumption is interpreted as meaning we presume the variability in the growth of teeth between the guinea pigs whose diets were of ascorbic acid or orange juice (treatment labels 0 and 1, respectively) are equal, that is to say both groups of guinea pigs vary equally in the growth of teeth.

Justification:

3. 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.

```
par(mfrow=c(2,1))
hist(guinea_teeth$growth[guinea_teeth$trt=="1"],
  main="Treatment = orange juice",
  xlab="Growth",
  xlim = range(c(2, 22)),
  breaks = 8
  )
hist(guinea_teeth$growth[guinea_teeth$trt=="0"],
  main="Treatment = ascorbic acid",
  xlab="Growth",
  xlim = range(c(2, 22)),
  breaks = 8
  )
```





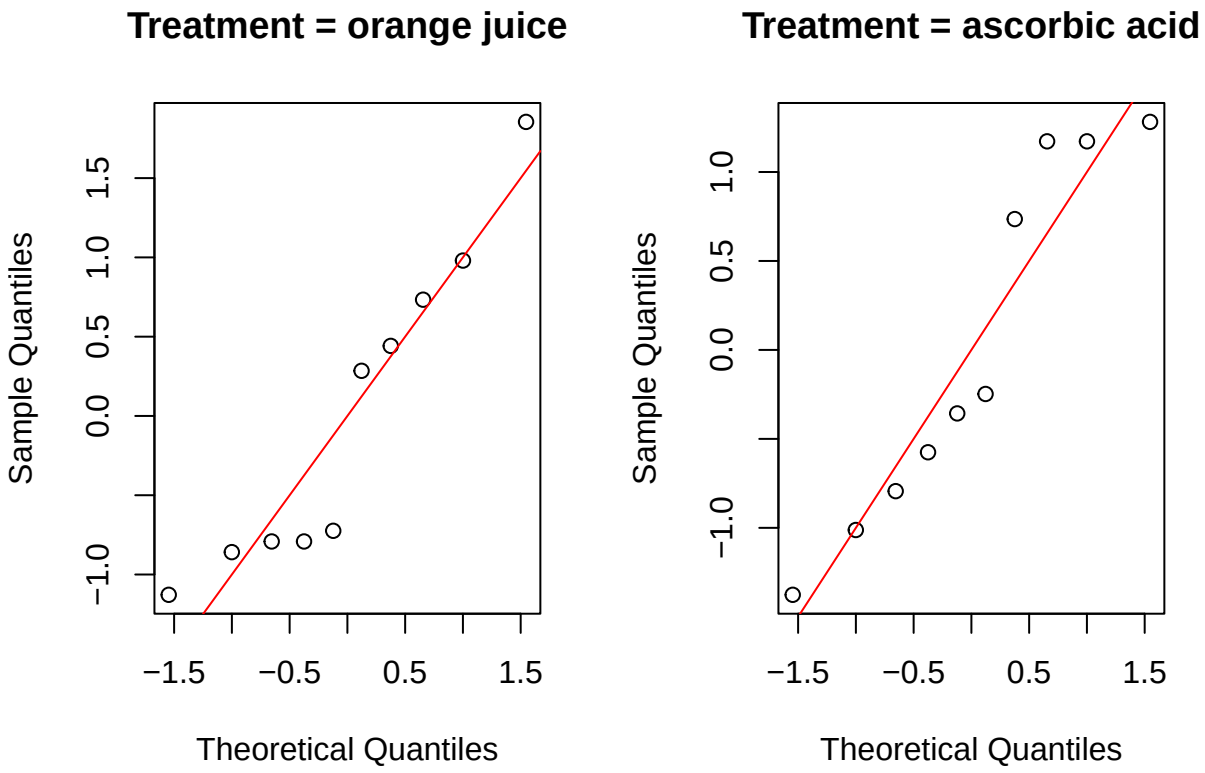

```
shapiro.test(guinea_teeth$growth[guinea_teeth$trt=="1"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  guinea_teeth$growth[guinea_teeth$trt == "1"]
## W = 0.89274, p-value = 0.182
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  guinea_teeth$growth[guinea_teeth$trt == "0"]
## W = 0.89, p-value = 0.1696
```

```r
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")
```



Description: In the context of the study, the normal distribution assumption may be interpreted as meaning that the distribution of the teeth growth of guinea pigs is normally distributed, and that this assumption holds both for the guinea pigs whose diets were of ascorbic acid or orange juice (treatment labels 0 and 1, respectively).

Justification:

2.

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

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

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  growth by trt
## W = 19.5, p-value = 0.02319
## alternative hypothesis: true location shift is not equal to 0
```

1. What is the value of the test statistic W? What is the corresponding sum of the ranks in the first group?

W = 19.5

2. What is the value of the p-value?

p-value = 0.02319

3. Interpret the result of the test in the context of the study.

p-value is the likelihood we observed the results we did given the null hypothesis. Where:

Null Hypothesis: Distribution of response variable is the same for both groups.

Alternative Hypothesis: Distribution of response variable is different between the two groups.

So we have evidence in support of the alternative hypothesis, namely that there is a difference between the two groups of guinea pigs.

3.

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

1. Which transformation of the data will result in normal distributions of radon concentration levels within each county?

Box-Cox transformation

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

```
t.test(X~county, data=minn_radon, var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  X by county
## t = -0.755, df = 46, p-value = 0.4541
## alternative hypothesis: true difference in means between group OLMSTED and group STEARNS is not equal
## 95 percent confidence interval:
##  -0.7333225  0.3332656
## sample estimates:
## mean in group OLMSTED mean in group STEARNS
##              1.418978              1.619007
```

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