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).

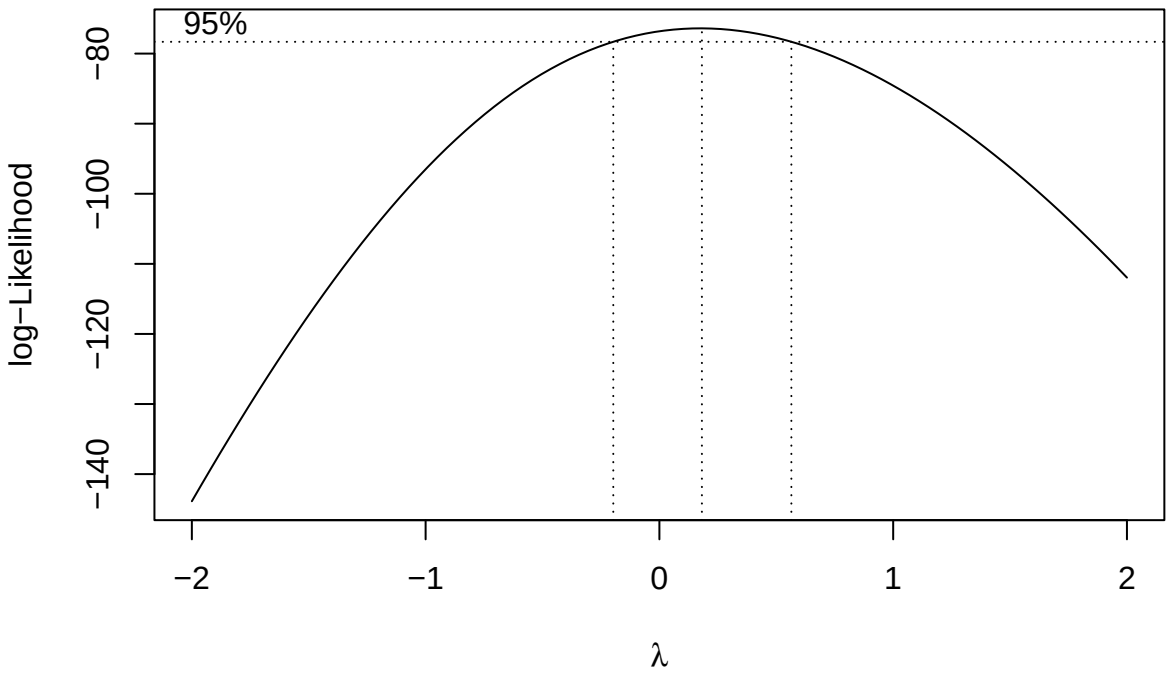
The following examples show how to assess the assumptions necessary to perform the traditional t-test (with equal variances) for difference in mean teeth growth in R. The full R script is provided in the `teeth_Lab4.r` file posted in Canvas.

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

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:

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

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




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

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.

```
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:"))
}
```

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

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

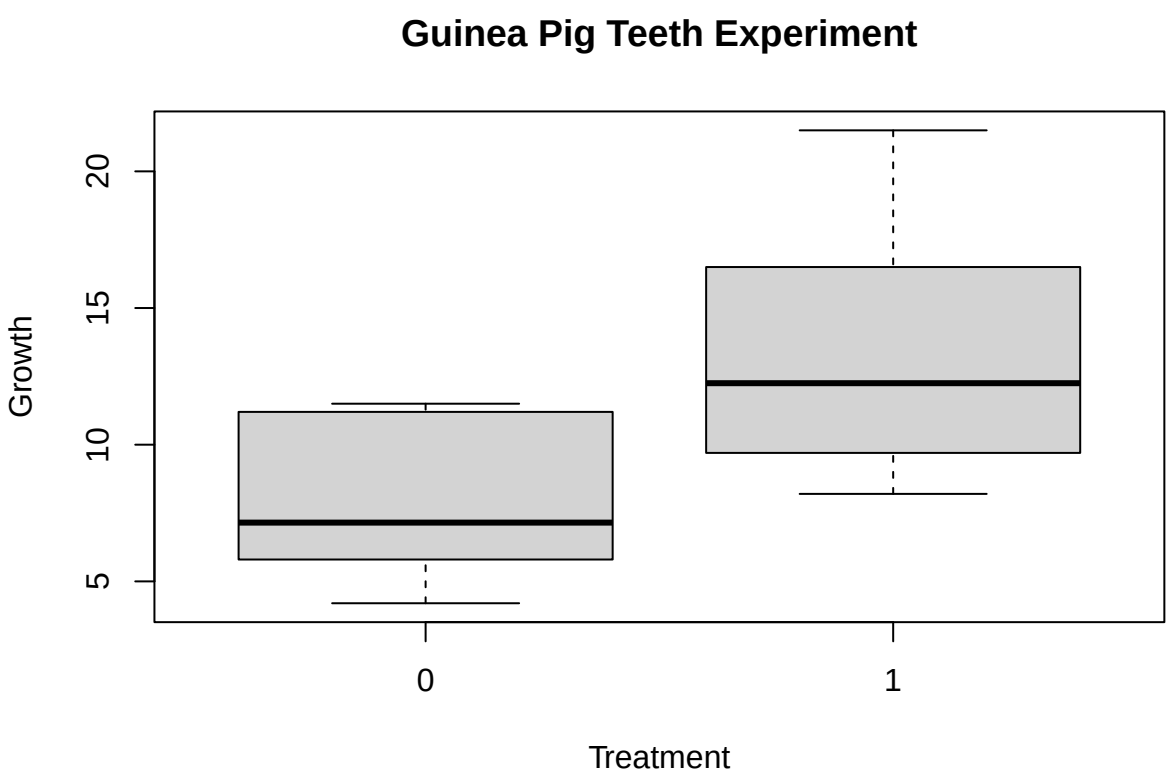
```
var.test(
  x=guinea_teeth$growth[guinea_teeth$trt=="1"],
  y=guinea_teeth$growth[guinea_teeth$trt=="0"],
  alternative="greater")
```

```
##
##  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
```

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

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

```
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 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, which is to say both groups of guinea pigs vary equally in the growth of teeth.

Justification: The above box plots show obvious differences in skewness, which would lead us to consider our assumption of equal variances to be violated. Furthermore, use of the F-test provides some evidence to consider rejecting the null hypothesis of no difference in variance to support the alternative hypothesis that there is some difference in the variances of the two groups. This is further corroborated through the additional variance test detailed in the user-defined function.

Though the p-values for the respective statistical tests are around the 0.08 range, I would nonetheless support the interpretation of the above figures to rejecting the hypothesis of equal variance in favor of an alternative hypothesis of the two guinea pig groups having a difference in their population variances.

Worth noting as well that the ratio of sample standard deviations is 1.62, which according to slides should have “little impact”, though this is illustrative of there being differences between the two groups.

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.

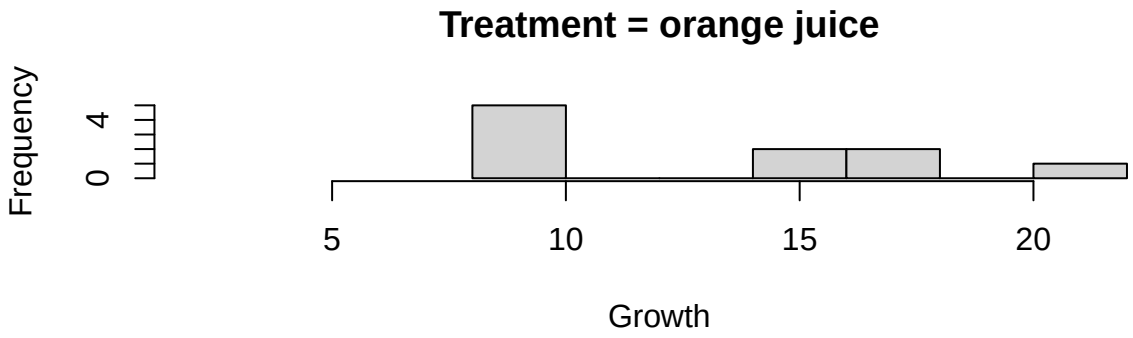
```
library(tidyverse)
```

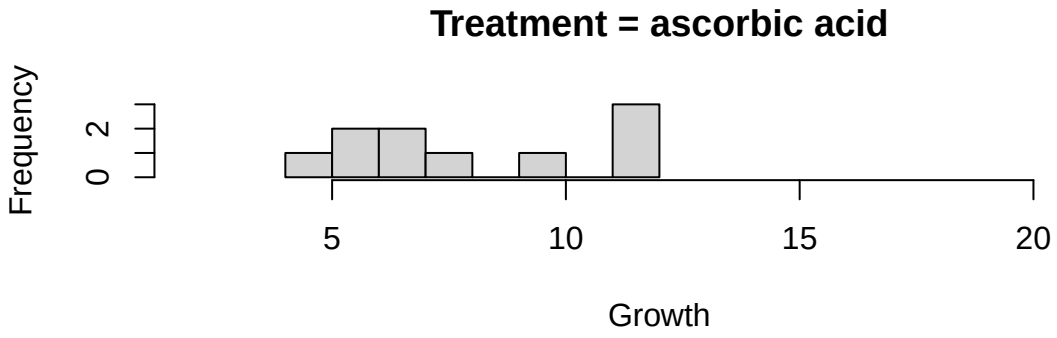
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.0.2
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
## 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
```

```
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
  )
```





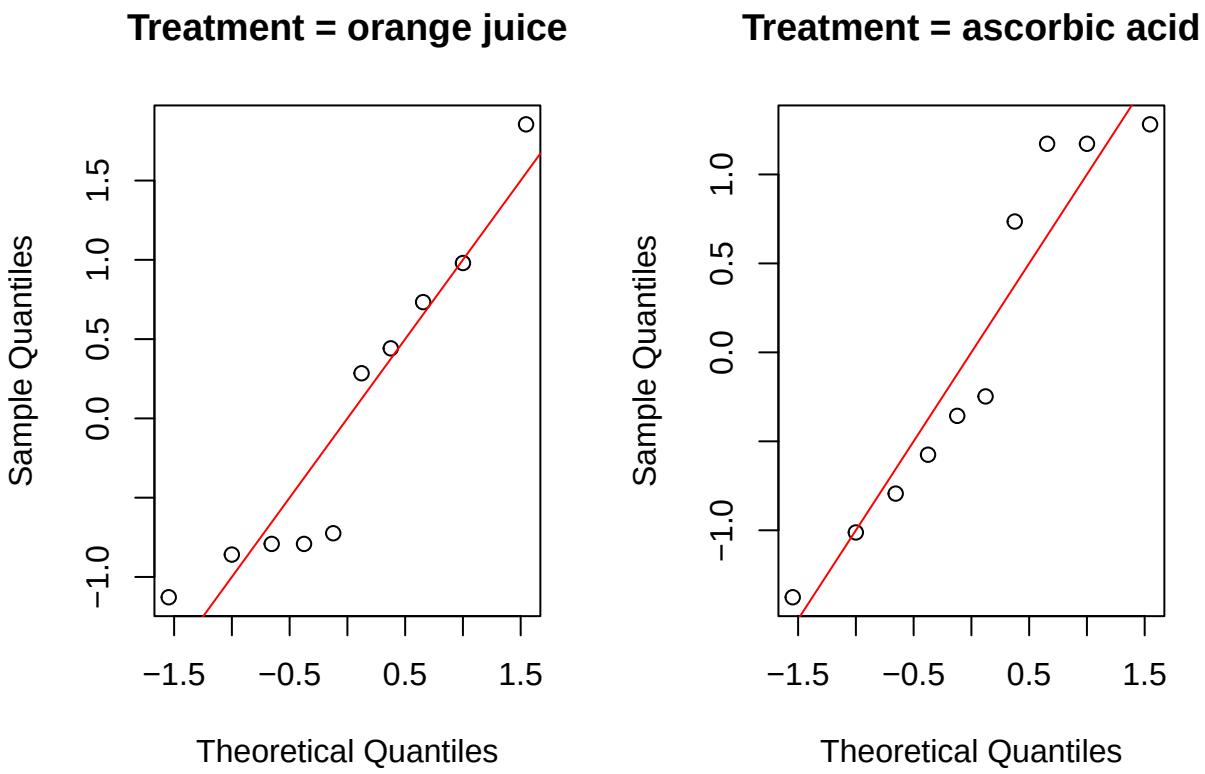

```
shapiro.test(guinea_teeth$growth[guinea_teeth$trt=="1"])
```

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

```
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
```

```
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: We have some evidence to reject the assumption that our data is normally distributed. From both the box plots and summary statistics we note there are some deviations in each groups distributions from the normal. Of particular note: Medians and Means for each of the groups are not equal, which is further supported by the finding of skewness in both groups. There is also excess (negative) kurtosis, meaning the tails of the distributions is not normal.

However, it is worth noting that the samples in question are fairly small, so the deviations from the normal are understandable to a certain extent (there are only 20 total observations, 10 for each group of guinea pigs). Despite this, the statistical test for normality via Shapiro-Wilk is below the 0.20 p-value range. We will see this further when testing for differences in distribution, but it is worth noting here that the visuals above generally do not look normally distributed. Particularly for the quantile plots, we see at least half of the observations in each group do not correspond to the expected distribution given from the empirical (theoretical) qualtiles of a normal distribution.

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?

The W statistic for the Wilcoxon rank-sum test is the same as the corresponding sum of the ranks in the first group, which is $W = 19.5$.

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

The p-value for this test is f-value:= 0.02319. However this is the one-tailed so to get the two-tailed we multiply this value by 2, giving us p-value:= $0.02319 * 2 = 0.04638$

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

In the context of p-value is the likelihood we observed the results we did given the null hypothesis. Where: Null Hypothesis: Distribution of response variable (teeth growth) is the same for both groups (of guinea pigs), and Alternative Hypothesis: Distribution of response variable (teeth growth) is different between the two groups (of guinea pigs).

So we observe a fairly low p-value, below the 0.05 threshold, meaning we calculate the probability of observing the results we did under the null hypothesis being true (no difference in distributions between groups), to be under 0.05. Said differently, we have evidence to support the alternative hypothesis that there is a difference in the distributions of teeth growth of guinea pigs with an ascorbic acid diet compared to the distribution of values for guinea pigs treated with an orange juice diet. Given the data in support of the hypothesis that there are differences between the two groups we may further extend the details to say this is support in favor of believing there are differences in median teeth growth between the two groups (in addition to other summary statistics).

3.

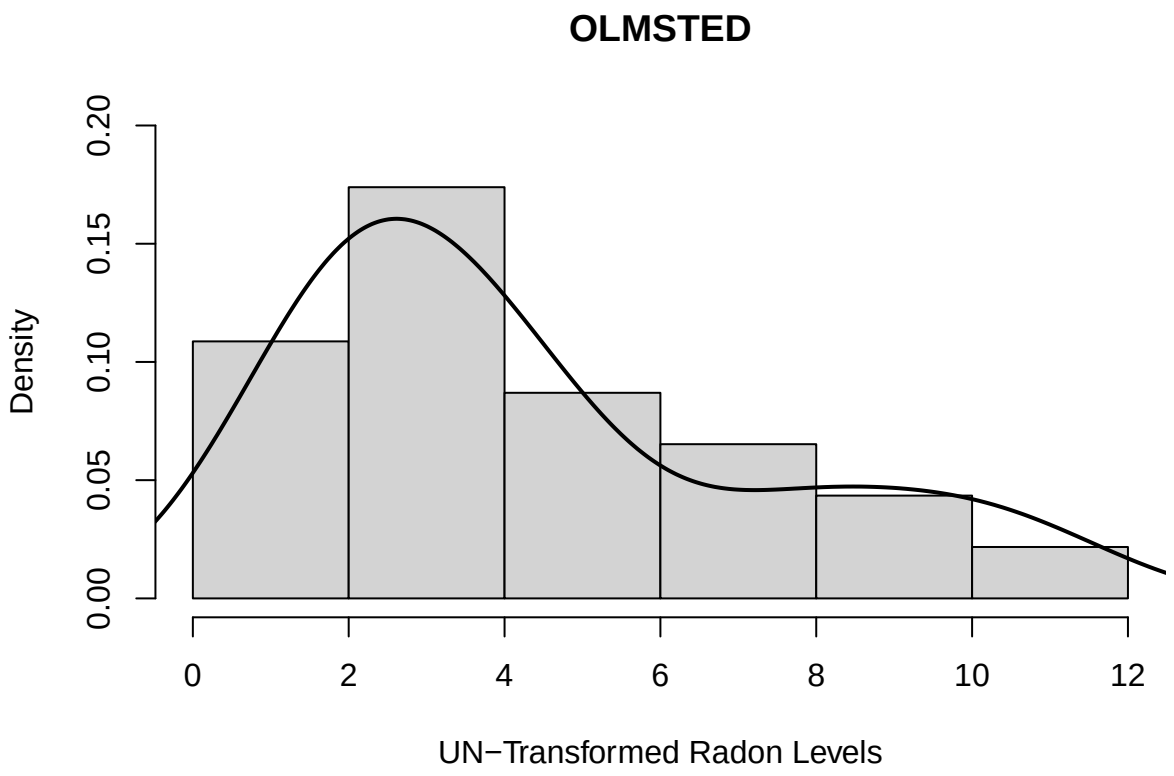
Use R to complete the following exercises for the radon study:

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

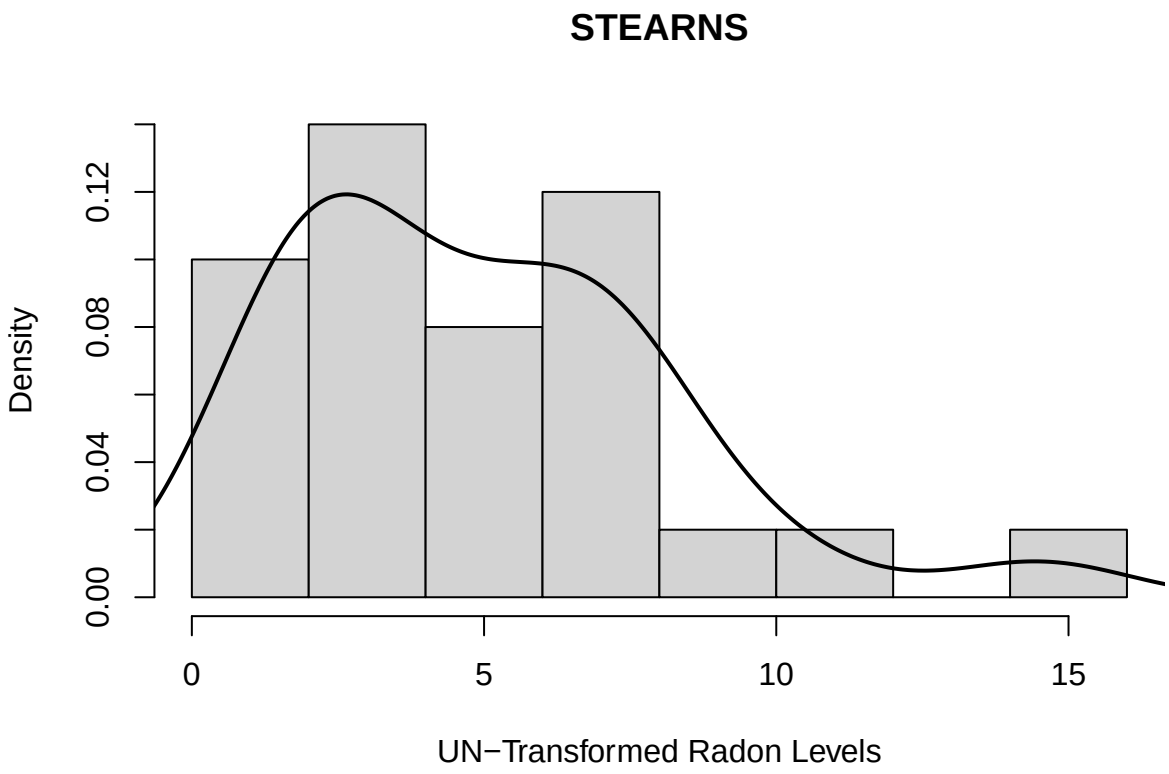
We use the Box-Cox transformation to transform the data into something (more) normally distributed for radon concentration levels.

Justification for using this specific transformation was due to having skewed data, as evidenced in the below histogram(s).

```
hist(minn_radon$radon[minn_radon$county=="OLMSTED"], xlab = "UN-Transformed Radon Levels", main = "OLMS
lines(density(minn_radon$radon[minn_radon$county=="OLMSTED"]), col = "black", lwd = 2)
```

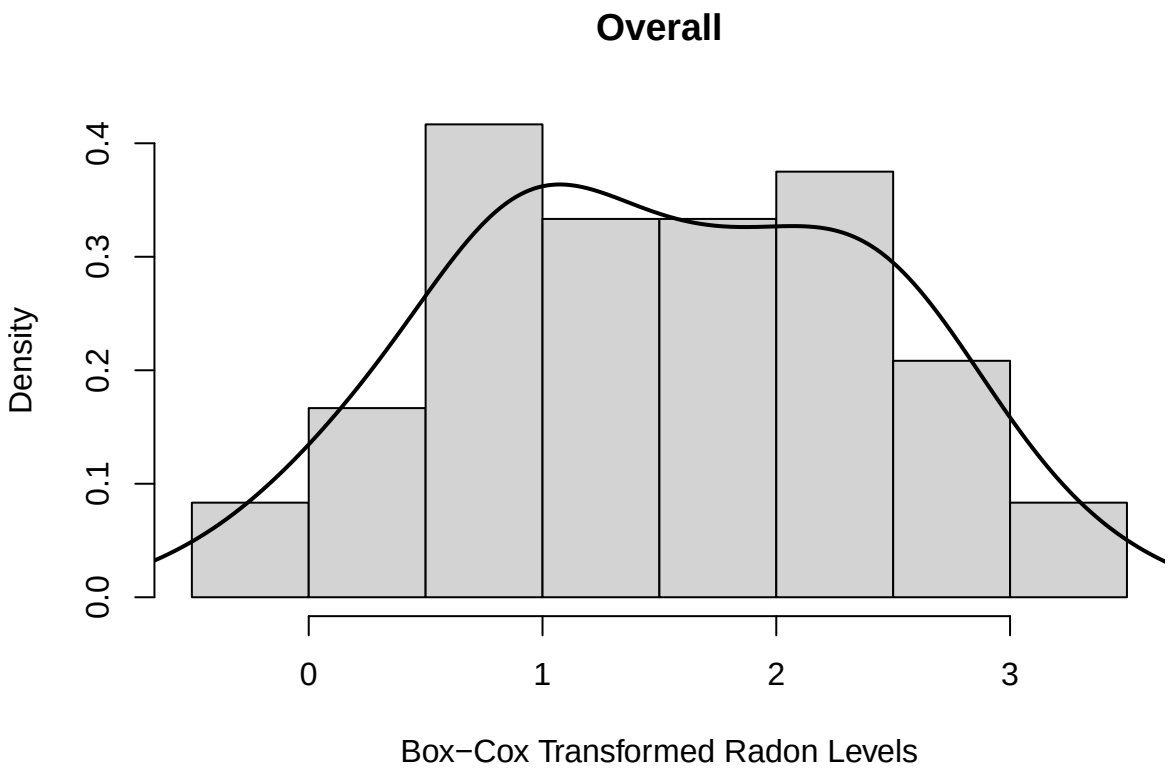



```
hist(minn_radon$radon[minn_radon$county=="STEARNS"], xlab = "UN-Transformed Radon Levels", main = "STEA
lines(density(minn_radon$radon[minn_radon$county=="STEARNS"]), col = "black", lwd = 2)
```

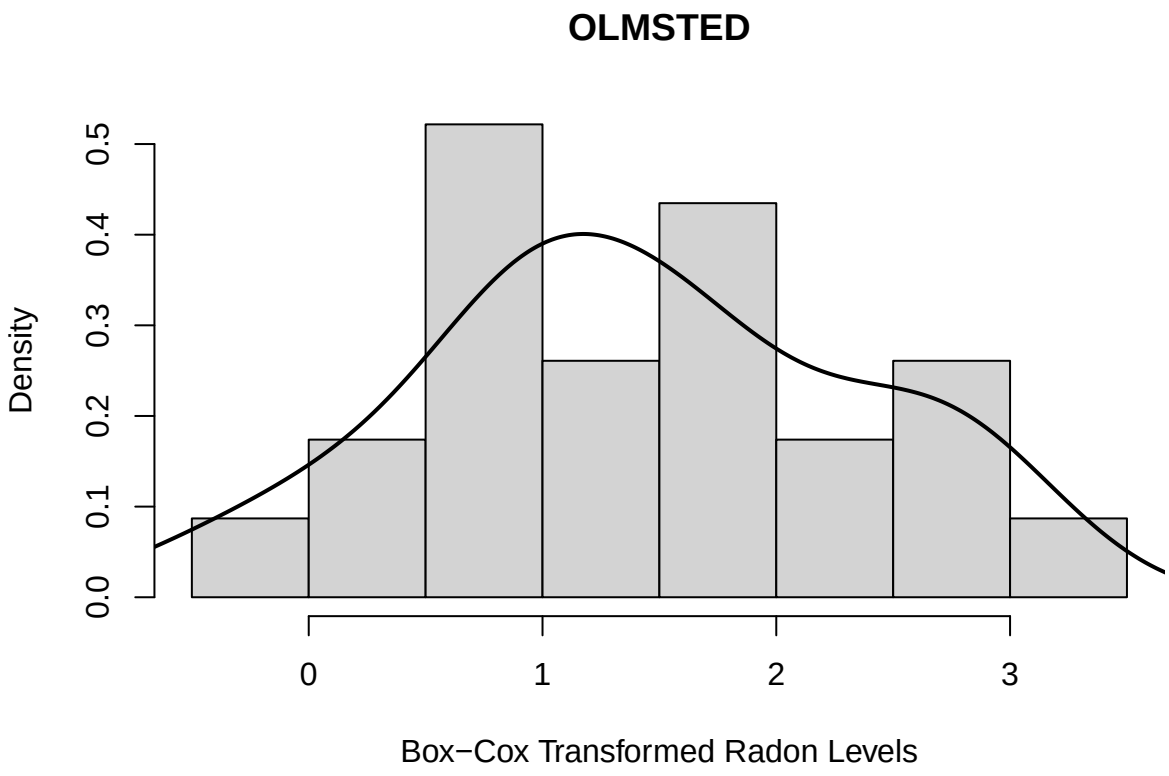


We verify this transformation normalized the data by checking the following:

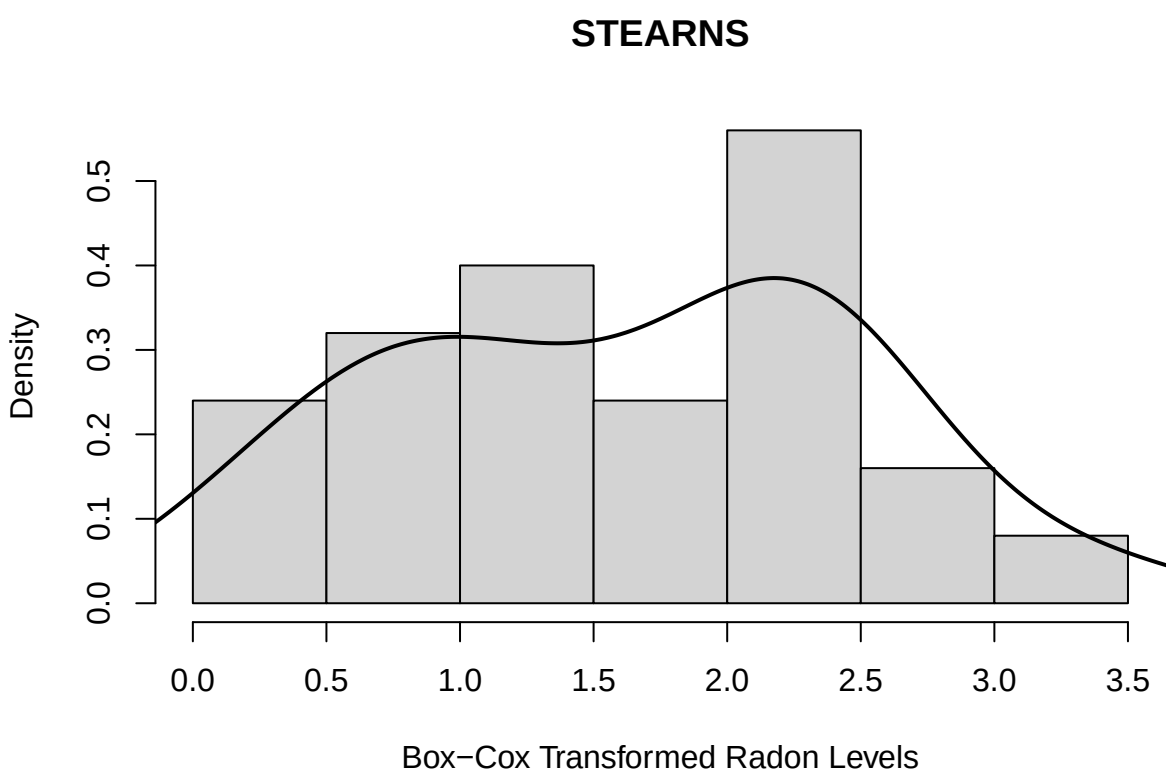
```
hist(minn_radon$X, xlab = "Box-Cox Transformed Radon Levels", main = "Overall", freq = FALSE)
lines(density(minn_radon$X), col = "black", lwd = 2)
```




```
hist(minn_radon$X[minn_radon$county=="OLMSTED"], xlab = "Box-Cox Transformed Radon Levels", main = "OLMS
lines(density(minn_radon$X[minn_radon$county=="OLMSTED"]), col = "black", lwd = 2)
```




```
hist(minn_radon$X[minn_radon$county=="STEARNS"], xlab = "Box-Cox Transformed Radon Levels", main = "STE
lines(density(minn_radon$X[minn_radon$county=="STEARNS"]), col = "black", lwd = 2)
```

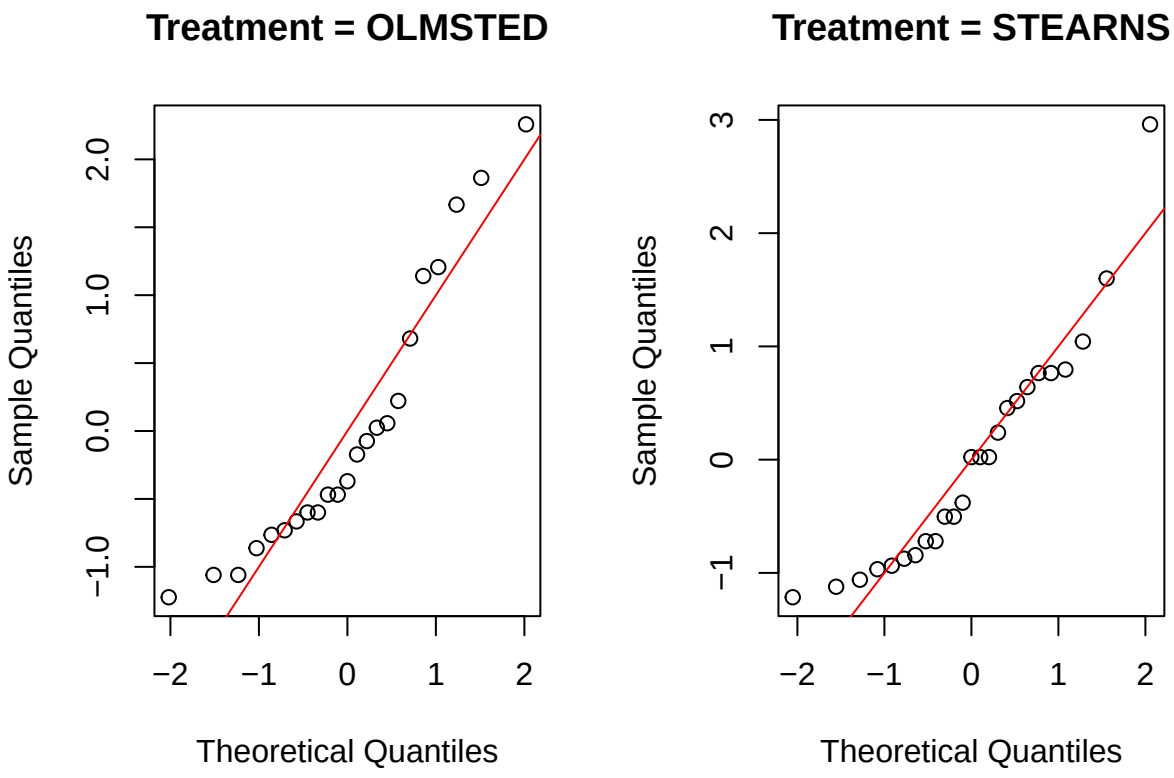


Which appears, approximately, normal.

However, if we look at the radon levels before and after using Quantile plots we have:

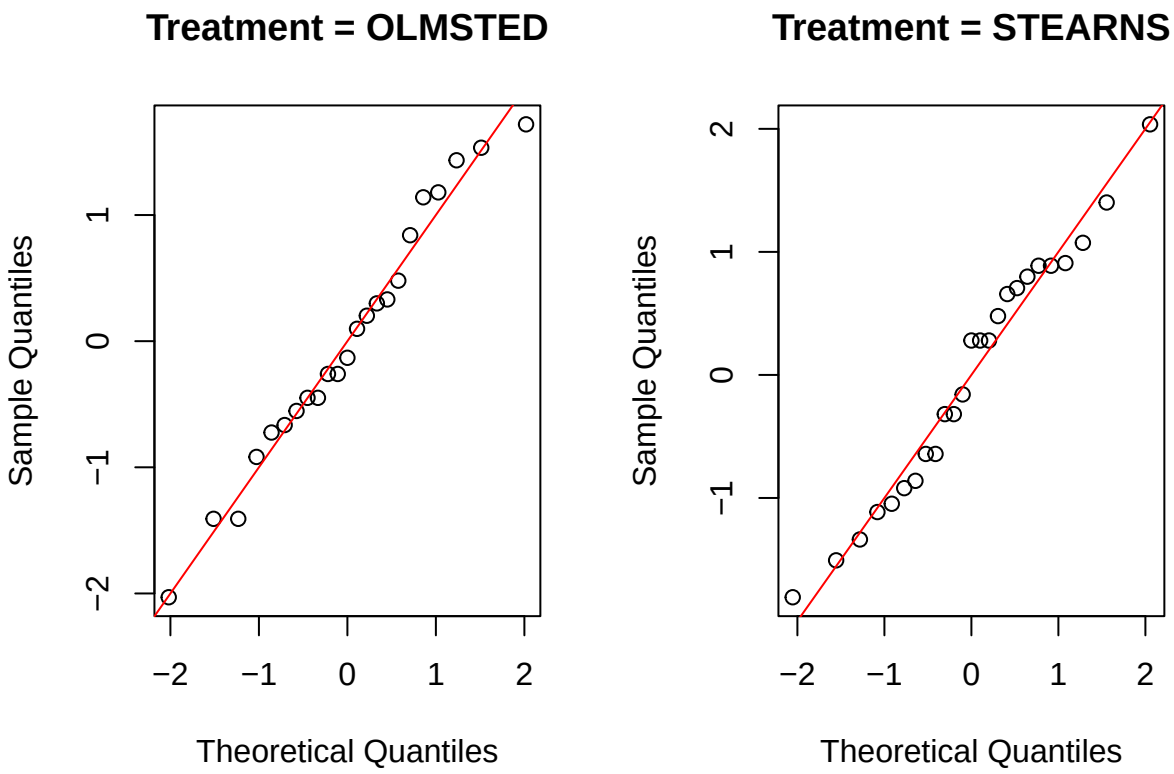
Before Transformation

```
par(mfrow=c(1,2))
qqnorm(scale(minn_radon$radon[minn_radon$county=="OLMSTED"]),
  main="Treatment = OLMSTED")
abline(a=0, b=1, col="red")
qqnorm(scale(minn_radon$radon[minn_radon$county=="STEARNS"]),
  main="Treatment = STEARNS")
abline(a=0, b=1, col="red")
```



After Box-Cox Transformation

```
par(mfrow=c(1,2))
qqnorm(scale(minn_radon$X[minn_radon$county=="OLMSTED"]),
  main="Treatment = OLMSTED")
abline(a=0, b=1, col="red")
qqnorm(scale(minn_radon$X[minn_radon$county=="STEARNS"]),
  main="Treatment = STEARNS")
abline(a=0, b=1, col="red")
```



We do see that this transformation has made progress in normalizing the data (many more values are close to the plotted quantile line as well as more values being nearer to the 0-th theoretical quantile).

One Other Thing

```
shapiro.test(minn_radon$radon[minn_radon$county=="STEARNS"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  minn_radon$radon[minn_radon$county == "STEARNS"]
## W = 0.90521, p-value = 0.02386
```

```
shapiro.test(minn_radon$radon[minn_radon$county=="OLMSTED"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  minn_radon$radon[minn_radon$county == "OLMSTED"]
## W = 0.88894, p-value = 0.01506
```

```
shapiro.test(minn_radon$X[minn_radon$county=="STEARNS"])
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  minn_radon$X[minn_radon$county == "STEARNS"]
## W = 0.97044, p-value = 0.6563
```

```
shapiro.test(minn_radon$X[minn_radon$county=="OLMSTED"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  minn_radon$X[minn_radon$county == "OLMSTED"]
## W = 0.97612, p-value = 0.8308
```

Comparing the Shapiro-Wilk normality test between the groups of transformed and untransformed response variables similarly supports the conclusion that the Cox transformation normalized the data.

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
```

With a p-value of 0.4541, we do not have substantive evidence to reject the null hypothesis, which is that the difference in average levels of radon (more on this shortly, because this is not just your typical "radon levels"!) between Olmsted and Stearns county is zero.

However, it is worth noting that this test was done on the Box-Cox transformed response variable (radon levels), so we should instead say that we observe a p-value of 0.4541 and a 95% Confidence Interval of (-0.733, 0.333) provide evidence to not reject the null hypothesis and support the observation that the difference in **average Box-Cox transformed levels of radon** between Olmsted and Stearns county is zero.

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