# Midterm2

Keith Hickman

November 13, 2017

## Problem 1

*Boxes of cereal are advertised as having a net weight of 8 ounces. The weights of boxes are assumed to be normally distributed. A new cereal box-filling machine is purchased, and we wish to be sure that on average, it puts at least the correct amount of cereal in the boxes. We randomly select 16 boxes, and find they have weights with sample mean 8.10 ounces and sample standard deviation 0.20 ounces.*

(a) Suppose your data included all the box weights. How would you check the normal distribution assumption? Explain what would indicate a violation of this as- sumption. (Note: You do not have to perform the check on the given data.)

There are several ways to check an assumption of normality, including qqnorm or density kernel plots, boxplots, or summary statistics.

(b) Perform a test of the null hypothesis that the average net weight of the boxes of cereal produced by the new machine is less than or equal to 8 ounces, against the alternative that it is greater than 8 ounces. Test at level $\alpha = 0.05$. (Clearly define the parameter you're estimating and the hypotheses you're testing, and give a full and substantive conclusion).

Because this is a one-sample location problem, the parameter we're testing here is the mean $\mu$. We have a normally distributed variable, and the hypothesis specifies a direction, so we'll perform a one-tailed t-test. The null hypothesis is given as $H_0: \mu_0 \leq 8$ and the alternative $H_1: \mu_1 > 8$.

Parameters:

```
n <- 16
s <- .2
mu.0 <- 8
mu.1 <- 8.1
```

Because the alternative hypothesis is "greater than", we use `1-pnorm(t)` to calculate the p-value. The t-statistic and p-value:

```
t <- (mu.1 - mu.0) / (s/sqrt(n))
t

## [1] 2
```

```r
1-pnorm(t)
```

```
## [1] 0.02275013
```

The p-value of .0228 is well under our $\alpha$ of .05, and we can probably safely reject the null, at least for this sample. However, n = 16 is still relatively small, so further testing is probably a good idea. Additionally, if we want to know whether the machine puts too much cereal in the boxes, I can perform a two-tailed test.

## Problem 2

*To test whether my friend's fish Googly had psychic powers, I wrote R code to display two windows. I entered either Left or Right depending on which way Googly was facing. Then the random number generator in R selected either the left or the right window, with probability 0.5 for each, in which to display a star. Let p be the probability Googly guesses correctly on a given trial (assume this is constant.) In 80 trials, Googly correctly guessed the window with the star 41 times.*

(a) Using mathematical notation, write down null and alternative hypotheses for a one-sided test.

Here, the null hypothesis $H_0: p \leq .5$, where the alternative hypothesis is $H_1: P > .5$ where $P$ is the proportion of observation correct selections or guesses. Since we have a binomial distribution, we're dealing with proportion as our parameter of interest.

(b) If the test statistic is the number of correct guesses (41) in 80 trials, write down R code to find the P-value of a one-sided test.

R code for finding the p-value for 41/80 guesses correct:

```r
1 - pbinom(40, 80, .5)
```

```
## [1] 0.4555361
```

(c) Even without R, we can see that Googly's success rate was close to its expected value under the null, so the one-tailed P-value will be close to 0.5. State your conclusion about the fish's psychic powers.

The p-value of .45 here doesn't tell us anything that we don't already know - Googly likely doesn't have psychic powers; this is a result we could expect by random methods as well.

(d) Continued from part (b). If you only known that the R code dbinom(40, 80,0.5) gives the number 0.0889, how would you find the exact P-value of the test? (Hint: use the property of a symmetric probabilty distribution.)

I would multiply the result by 2 giving `2 * (dbinom(40,80,.5))` to obtain both tails of the probability distribution.

```r
2 * (dbinom(40, 80, .5))
```

```
## [1] 0.1778558
```

## Problem 3

The file snoqualmie14.txt contains the daily precipitation (in inches) in Snoqualmie Falls, WA, for a random sample of 365 days. After saving the file to your computer, you can load it into R by entering the command:

```
rainfall = scan(file.choose())
```

and then selecting the file.

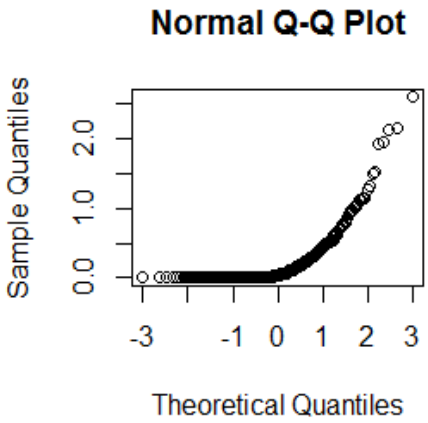
```
rainfall <- read.csv("C:\\Users\\khickman\\Desktop\\Personal\\IUMSDS\\StatsS520\\Module12\\snoqualmie14.txt")
## I used read.csv because everytime I knit to PDF, r markdown asks me to choose the file.

rainfall <- rainfall[1]

summary(rainfall)
```

```
##        X0
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0400
##  Mean   :0.2062
##  3rd Qu.:0.2600
##  Max.   :2.6000
```

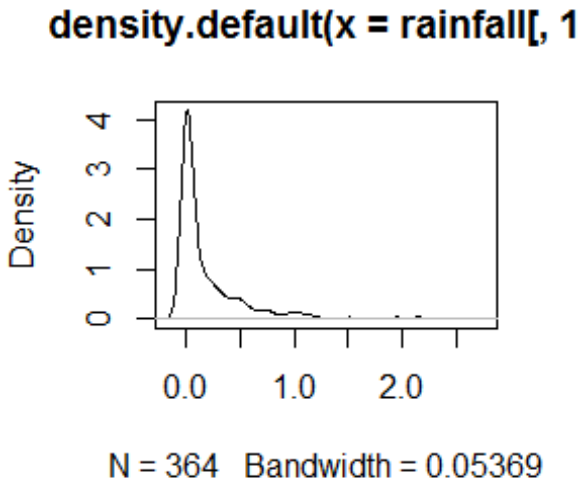
(a) Show that the data does not come from a normally distributed population. Include a graph to support your answer.

We'll use the qqnorm and density kernels to analyze normality for this distribution.

```
qqnorm(rainfall[,1])
```




```
plot(density(rainfall[,1]))
```




This variable is clearly right-skewed and non-normal. We should therefore avoid t-tests and standard normal if possible. We could potentially select a non-parametric test. The `qqnorm` plot indicates that many values are at or near zero, but also that there are a substantial number of higher values. The density kernel also indicates a right-skewed distribution.

(b) The mean annual rainfall in Seattle is 37.7 inches per year. Test (at level $\alpha = 0.05$) the hypothesis that the mean rainfall in Snoqualmie Falls is different from the mean rainfall in Seattle. (Clearly define the parameter you're estimating and the hypotheses you're testing, and give a full and substantive conclusion).

The parameter of interest is the difference in mean rainfall between the two regions. The null hypothesis is that there is no difference in rainfall between the two locales. Let $\mu_1$ be

the rainfall in Seattle, and $\mu_2$ be the rainfall in Snoqualmie. Let $\Delta = \mu_1 - \mu_2$. Thus $H_0: \Delta = 0$. The alternative, which we are trying to show, is stated $H_1: \Delta \neq 0$.

Additionally, we can assume the variables are independant, though we would need additional data to determine this for certain. Additionally, we don't know the variance for the Seattle rainfall, so we can't do a Students t-test, as the populations are neither normal.

Since our hypothesis does not specify a direction, a two-tailed test is appropriate. Even without the test, it's clear that with a mean of 37.7 in Seattle and .26 in Snoqualmie, we're way off from the null.

```
mean.seattle <- 37.7
mean.sno <- mean(rainfall[,1])
n <- length(rainfall[,1])
n

## [1] 364

mean.sno

## [1] 0.2061538

t.stat <- (mean.sno - 0) / sd(rainfall[,1])/sqrt(364)
t.stat

## [1] 0.02932443
```

The p-value is:

```
1 - pt(t.stat, n-1)

## [1] 0.488311
```

The p-value of .4883 is much higher than our significance level, indicates that there is no evidence against the null using significance level $\alpha = .05$. However, as mentioned above, given the non-normal distribution, t-test will probably give very bad results.

## Problem 4

In one year in the United States, 4.247 million babies were born. Of these, 2.173 million were male and 2.074 million were female. With very few exceptions (e.g. identical twins), the sexes of the babies are independent, so we can use the binomial distribution to model the number of babies that are female. Let p be the probability that a random (future) newborn is female.

(a) (2 points) What percentage of the babies were female? (To get credit for this question, you must give your answer as a percentage and you must round appropriately.)

```
p <- (2.074 / 4.247) * 100
p
```

```
## [1] 48.83447
```

Expressed as a percentage, $p$ = 48.83% of babies were female in that year.

(b) (3 points) Suppose we wish to test the null hypothesis that the probability a baby is female is 50%. Write down null and alternative hypotheses in mathematical notation for this test.

The null hypothesis that $p$ = .50 can be stated as follows: $H_0: p = .50$. The alternative is $H_1: p \neq .50$

(c) (10 points) Find the P-value, using both Binomial probability and Normal approximation. (Carefully show your work and R codes.)

Here, we are concerned with finding the true value of $p$. To start, I would like to find the p-value if 2.07 out of 4.24 million babies born were female assuming that the true probability is 50%. The p-value is calculated under the null as follows:

```
pbinom(2074000, 4247000, .5)
```

```
## [1] 0
```

The p-value is essentially 0, indicating that if the true female birth proportion were 50%, this would be an extremely unlikely value. Thus, we can safely reject the null.

(d) (2 points) Using the Central Limit Theorem, find a 95% confidence interval for the probability that a birth is female.

Because we have such a large sample size, the Central Limit Theorem allows us to assume normality, the 95% confidence interval for a binomial distribution can be stated established by adding and subtracting a standard error term to observed $p$:

```
p <- .48883
lower_bound <- p - qnorm(.975) * sqrt(p * (1 - p)) / sqrt(4247000)
upper_bound <- p + qnorm(.975) * sqrt(p * (1 - p)) / sqrt(4247000)
upper_bound
```

```
## [1] 0.4893054
```

```
lower_bound
```

```
## [1] 0.4883546
```

Thus, a 95% confidence interval for $p$ is very narrow in terms of percentage,, from 48.83546% to 48.93054%.

(e) Explain what this confidence interval means without using the word "confident."

The confidence interval indicates that in 95% of the samples (years?), the proportion of babies born who are female will fall between the upper and lower bounds of 48.84% to 48.93%.

(f) The P-value for your test in part (c) is basically zero. From this and your confidence interval, write in a sentence your conclusion about the probability that a random newborn is female.

A very small p-value means that we can safely reject the null, and proceed with an understanding that the actual proportion of babies born who are female is closer to 48.83%

## Problem 5

*The basketball player Steph Curry sometimes shoots free throws with his mouthguard in his mouth, and sometimes shoots free throws with his mouthguard outside of his mouth. His free throw statistics for one season were: - Free throws with mouthguard in: 110 completed, 13 missed (89.4%) - Free throws with mouthguard out: 198 completed, 16 missed (92.5%).*

His observed free throw completion rate was slightly higher when his mouthguard was outside his mouth. However, we should check whether the difference could be plausibly explained as luck.

(a) Using the Central Limit Theorem, find an approximate 95% confidence interval for the probability that Curry completes a free throw with his mouthguard in. Give a numerical answer.

Assuming normality, the 95% confidence interval for a binomial distribution can be stated established by adding and subtracting a standard error term to observed $p$:

```
p5.in <- .894
lower_bound <- p5.in - qnorm(.975) * sqrt(p5.in * (1 - p5.in)) / sqrt(4247000
)
upper_bound <- p5.in + qnorm(.975) * sqrt(p5.in * (1 - p5.in)) / sqrt(4247000
)
upper_bound

## [1] 0.8942928

lower_bound

## [1] 0.8937072
```

The lower and upper bounds of our confidence interval are .8937 and .8943 respectively, meaning that 95% of the time, Steph Curry will make between 89.37% and 89.43% of his freethrows with his mouthguard in.

(b) Using the Central Limit Theorem, find an approximate 95% confidence interval for the probability that Curry completes a free throw with his mouthguard out.

```
p5.out <- .925
lower_bound <- p5.out - qnorm(.975) * sqrt(p5.out * (1 - p5.out)) / sqrt(4247
000)
upper_bound <- p5.out + qnorm(.975) * sqrt(p5.out * (1 - p5.out)) / sqrt(4247
```

```
000)
upper_bound

## [1] 0.9252505

lower_bound

## [1] 0.9247495
```

The lower and upper bounds of our confidence interval are .9247 and .9252 respectively, meaning that 95% of the time, Steph Curry will make between 92.47% and 92.52% of his freethrows with his mouthguard out.

(c) (3 points) Suppose we wish to test the null hypothesis that Curry's probability of completing a free throw is the same with his mouthguard in as it is with his mouthguard out. The P-value for such a test is 0.33. What does this P-value tell you? Explain.

The null hypothesis of $H_0: \Delta = 0$ where $\mu_1$ is mouthguard in and $\mu_2$ is mouthguard out and $\Delta = \mu_1 - \mu_2$...we can't reject the null with this p-value. A p-value of .33 would indicate that if the alternative were true, we would be seeing some very unusual data.

## Problem 6

*Rosene (1950) studied how quickly hairs on radish roots absorbed water when they were immersed. For each of eleven radishes, she measured the rate of influx of water for a young root hair and an old root hair on that radish. The data is given below.*

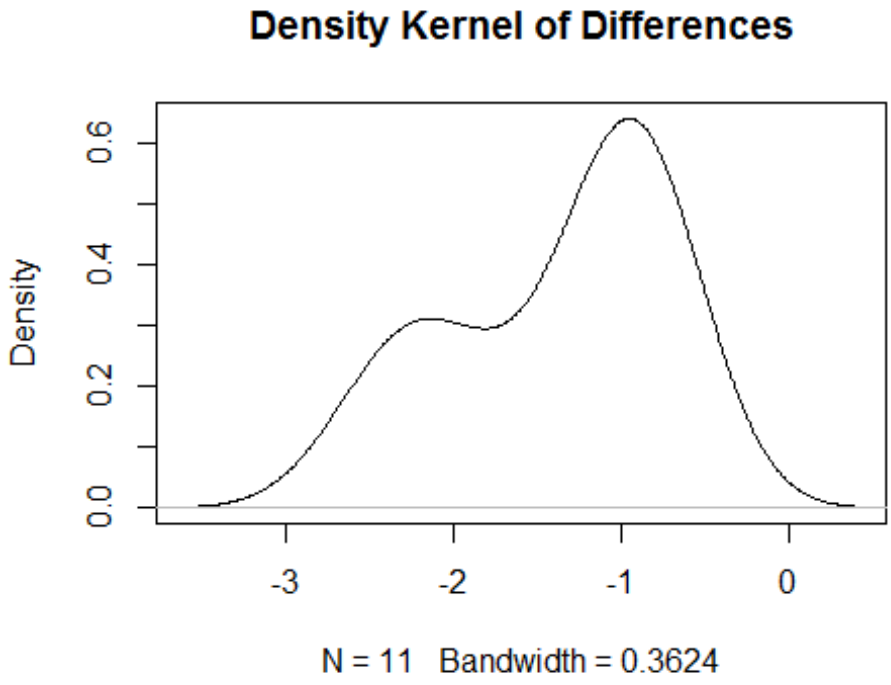
(a) Explain what test we should consider to use based on the data and context, and why. (Hint: 1-sample or 2-sample test, z-test of t-test, etc)

Our dataset is a paired sample; our unit of measurement is a pair of radish hairs. Thus we have a two-sample test. The unit of measure is water absorption rate, and the measurements were taken once each. We are interested in $\Delta$ where $\Delta$ is the difference in absorption rates between young and old radishes. We do not know the variance of the entire population, and we do not have a large $n$, therefore we should perform a t-test.

(b) Use a normal probability plot of the differences (old minus young) and the sample size to explain why we should be hesitant do such a test (the one you proposed in the previous question) on these differences (old minus young).

```
radishes <- read.csv("C:\\Users\\khickman\\Desktop\\Personal\\IUMSDS\\StatsS5
20\\Module12\\radishes.csv")
radishes

##    Radish  Old Young
## 1       A 0.89  2.13
## 2       B 0.49  1.16
## 3       C 0.91  2.60
## 4       D 0.80  1.58
## 5       E 0.56  1.53
```

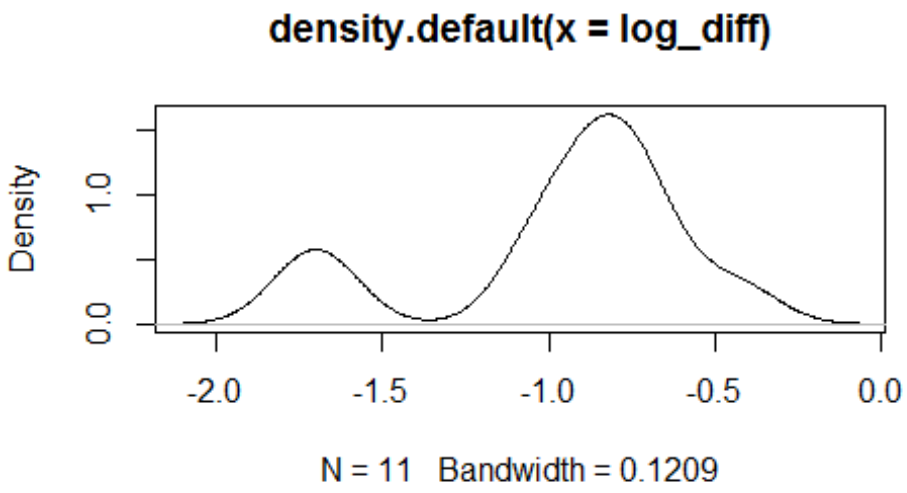
```
## 6       F 0.79  1.70
## 7       G 0.47  2.67
## 8       H 0.50  2.64
## 9       I 1.08  2.19
## 10      J 1.65  2.54
## 11      K 1.94  4.46

radishes$difference <- radishes$Old - radishes$Young
plot(density(radishes$difference),main="Density Kernel of Differences")
```




As shown in the density kernel above, the distribution of differences is non-normal, left-skewed, and probably not independent. The t-test may be misleading, as it assumes a normal underlying distrubtion and independence. Additionally, if we had a large $n$, we could use either the t-test or normal distribution (via the Central Limit Theorem) to overcome a non-normal underlying population.

(c) Explain why we should not take the logs of these differences (old minus young.)
Examining the logged `difference` variable:

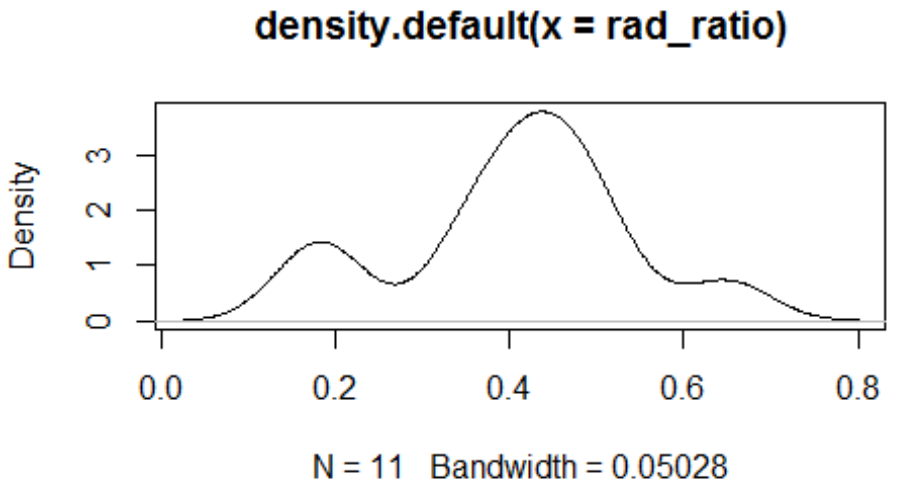
```
log_diff <- log(radishes$Old) - log(radishes$Young)
log_diff

##  [1] -0.8726558 -0.8617699 -1.0498221 -0.6805684 -1.0050862 -0.7663506
##  [7] -1.7371011 -1.6639261 -0.7069405 -0.4313888 -0.8324608

plot(density(log_diff))
```




This looks much further from a normal distribution. We could take the square root of the differences, but I believe that would produce a less than desirable result as well.

(d) Instead of using the differences, we can look at the ratio: old divided by young. This ratio looks to come from a much closer to normal distribution. Write R code to find a 90% confidence interval for the average value of this ratio.

```
rad_ratio <- radishes$Old / radishes$Young
rad_ratio

##  [1] 0.4178404 0.4224138 0.3500000 0.5063291 0.3660131 0.4647059 0.1760300
##  [8] 0.1893939 0.4931507 0.6496063 0.4349776

plot(density(rad_ratio))
```




R Code for a 90% confidence interval:

```r
mean(rad_ratio)
```

```
## [1] 0.4064055
```

```r
lower_rad <- mean(rad_ratio) - qnorm(.95) * sd(rad_ratio) / sqrt(11)
upper_rad <- mean(rad_ratio) + qnorm(.95) * sd(rad_ratio) / sqrt(11)
lower_rad
```

```
## [1] 0.3387247
```

```r
upper_rad
```

```
## [1] 0.4740863
```

There's a slightly narrower interval than we would see under 95% confidence here, as we're saying that 90% of the time, the mean value of the ratio of the `Old / Young` variables would fall between .338 and .474.

## Problem 7

*The basketball player Stephen Curry was the NBA's Most Valuable Player for the 2015 / 16 season. He is known for being very good at shooting (throwing the basketball into the hoop) from long distance. The file* `currydist.txt` *in the Data folder on Canvas contains data on the 1,598 shots he attempted during the 2015/16 season. Of these shots, 805 of them were successful. For the purpose of this analysis, we treat the data as random independent samples (this isn't quite true but is close enough.)*

Loading the data into R:

```r
currydist <- read.csv("currydist.csv")
summary(currydist)
```

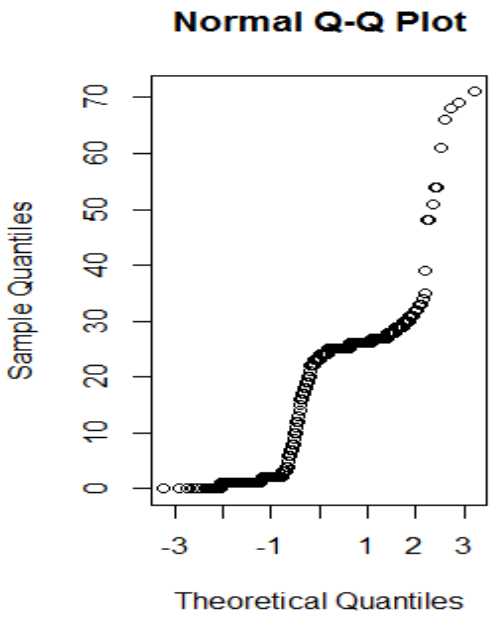
```
##     distance      venue          made        
##  Min.   : 0.00   Away:823   Min.   :0.0000  
##  1st Qu.: 4.00   Home:775   1st Qu.:0.0000  
##  Median :23.00              Median :1.0000  
##  Mean   :17.37              Mean   :0.5038  
##  3rd Qu.:25.00              3rd Qu.:1.0000  
##  Max.   :71.00              Max.   :1.0000
```

```r
str(currydist)
```

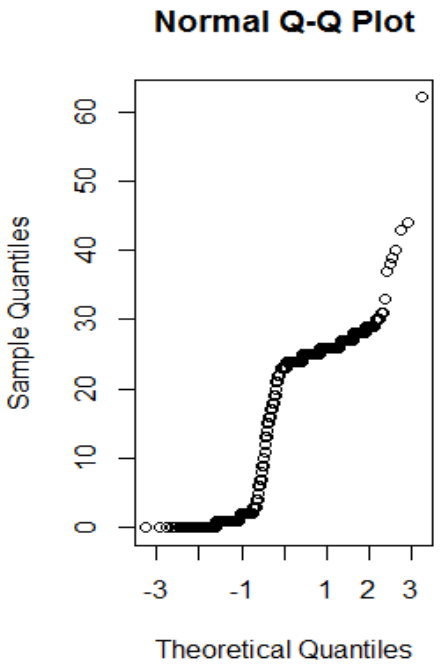
```
## 'data.frame':    1598 obs. of  3 variables:
##  $ distance: int  3 26 2 27 0 25 22 28 25 27 ...
##  $ venue   : Factor w/ 2 levels "Away","Home": 2 2 2 2 2 2 2 2 2 2 ...
##  $ made    : int  1 0 0 1 1 1 1 1 1 0 ...
```

(a) Create two vectors in R: one containing the distances for Curry's shots at home, and one containing the distances for Curry's shots away. Remember that we use square brackets for subsetting in R:

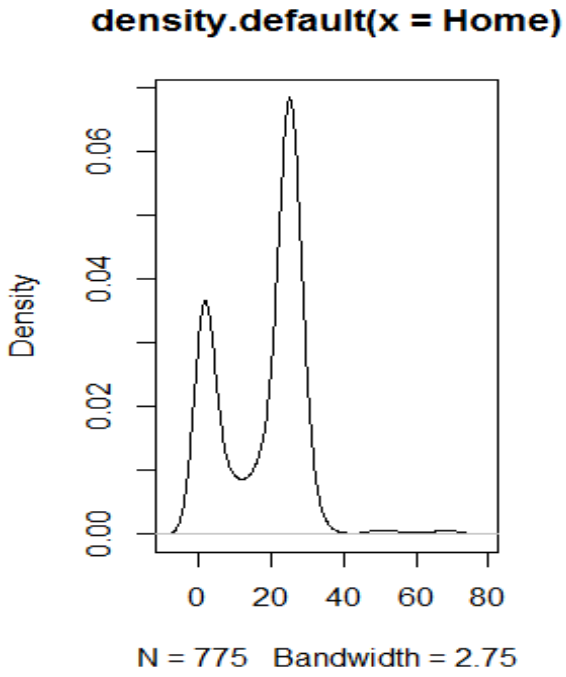
```
Home <- currydist$distance[currydist$venue=="Home"]
Away <- currydist$distance[currydist$venue=="Away"]
summary(Home)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    4.00   23.00   17.83   26.00   71.00

summary(Away)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    3.00   23.00   16.94   25.00   62.00

##Plots to show distribution characteristics:
qqnorm(Home)
```



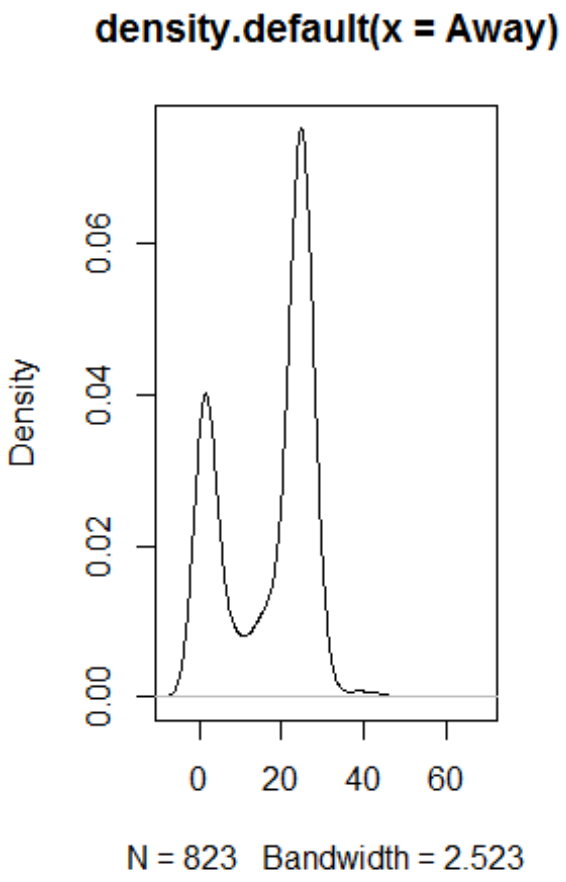

```
qqnorm(Away)
```




```
#Probability Density plots:
plot(density(Home))
```




```
plot(density(Away))
```




(b) Is there enough evidence to show that the (population) average distance of Curry's shots at home is different from the (population) average distance of Curry's shots away? Perform an appropriate significance test and find an appropriate confidence interval. Clearly define the parameter you're estimating and the hypotheses you're testing, and give a full and substantive conclusion (i.e. not just Reject or Don't reject.)

First, let's analyze the distribution of our newly-minted variables `Home` and `Away`. From the summary numbers, they look close to normal. We probably won't have many gross outliers, as there's only so much distance one man can shoot from.

Let the mean Home distance (`Home`) be $\mu_1$ and mean Away distance be $\mu_2$. Let $\Delta$ represent $\mu_1 - \mu_2$. The parameter of interest is $\Delta$.

We have a two-sample location problem, and we do not know the population standard deviation. (Someone does, but we don't have that data here). We have a relatively large $n$ at 1598, therefore we can use a Welch's Two Sample T-test, even though the underlying populations are not normal. Additionally, we can rely on the Central Limit Theorem to allow us to assume normality.

Under the null hypothesis, there would be no difference between the distances at Home and Away. Stated mathematically: $H_0: \Delta = 0$, and the Alternative, $H_1: \Delta \neq 0$. Let $\alpha = .1$

```
p7delta.hat <- mean(Home) - mean(Away)
p7delta.hat

## [1] 0.8917963
```

The difference between the means $\Delta$ is .89. We can perform a t-test by plugging in the two variables Home and Away:

```
s1 <- sd(Home)
s2 <- sd(Away)
n1 <- length(Home)
n2 <- length(Away)
std.error <- sqrt(s1^2/n1 + s2^2/n2)

Tstat <- p7delta.hat/std.error
Tstat

## [1] 1.595331

p7df <- (s1^2/n1 + s2^2/n2) ^2 / ((s1^2/n1)^2/(n1-1) + (s2^2/n2)^2/(n2-1))
p7df

## [1] 1567.814
```

Again, we're testing the null that $\Delta = 0$ with a t-statistic of 1.593, so we need to perform a two-tailed t-test.

```
2 * (1 - pt(abs(Tstat), p7df))

## [1] 0.1108397
```

And, to verify that I got the calculations correct:

```
t.test(Home,Away)

##
##  Welch Two Sample t-test
##
## data:  Home and Away
## t = 1.5953, df = 1567.8, p-value = 0.1108
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2046775  1.9882702
## sample estimates:
## mean of x mean of y
##  17.83226  16.94046
```

Everything looks the same. Here, my p-value of .111 is higher than my significance level of .1, so we can say with a good degree of confidence that there is not enough evidence to reject the null with this sample.

Now I can proceed with a confidence interval by taking plus or minus a quantile value from mean difference. I may set a lower confidence interval (say 80%), unless I'm a bookie in Vegas; Steph Curry's shot distances is typically not life-or-death stuff. Let's go with an 80% confidence interval.

```
p7upper <- p7delta.hat + qt(0.90, df=p7df) * std.error
p7lower <- p7delta.hat - qt(0.90, df=p7df) * std.error
p7upper

## [1] 1.608491

p7lower

## [1] 0.1751021
```

(c) The significance test you did requires assumptions. Show that the assumptions are met, or, if they are not met, explain why they can be overlooked.

The t-test assumes a normal underlying distribution. In our dataset, examining the sample of all shots distances, only Home and only Away, none of the distributions are normal. We can overlook this, however, because we have a rather large $n$ at $\sim$1500 observations, thus the Central Limit Theorem kicks in.