KJPipho_PS_7

Krista Pipho

October 26, 2018

Question 1. How robust are two sample t tests to violations of assumptions?

T tests assume normality and equality of variance. The robustness of the test describes how much violations of those assumptions change the likelyhood of type 1 or type 2 errors. Robustness can also depend on sample size and integrety of the data collection.

```
# Here we input a number of data sets, either randomly generated or explicetly specified. We
 know that all should have a mean of zero.
small_norm<-rnorm(5,0,1)
small_var<-rnorm(5,0,10)
large_norm<-rnorm(500,0,1)
large_var<-rnorm(500,0,10)
Left_skew <- c(0.80,1.3,-0.58, 0.30,0.50)
Right_skew <- c(-0.84,-0.015,-0.74,0.31,-0.26)
```

One assumption is that the two samples will have equal variance

```
# Here we use both the Welch and two sample t tests to compare small data sets. These two hav
e different variances, violating the equality of variance assumption

t.test(small_norm,small_var, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  small_norm and small_var
## t = 1.164, df = 4.1099, p-value = 0.3075
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.431121 20.826501
## sample estimates:
##  mean of x  mean of y
## -0.2274939 -6.4251840
```

```
t.test(small_norm,small_var, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  small_norm and small_var
## t = 1.164, df = 8, p-value = 0.278
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.080785 18.476166
## sample estimates:
##  mean of x  mean of y
## -0.2274939 -6.4251840
```

Using the Welch test we get a p value of .06703, greater than .05. We avoid erronіously rejecting the null hypothesis, but just barely.

Using the regular t-test (var.equal = TRUE) we get a p value of .03993, less than .05. Violation of the equality of variance assumption has led us to falsely reject the null hypothesis

Does size of the data set play a role in sensitivity to this violation?

```
# Here we use both the Welch and two sample t tests to compare larger data sets. These also h
ave different variances, violating the equality of variance assumption
t.test(large_norm,large_var, var.equal = FALSE)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  large_norm and large_var
## t = -0.34363, df = 508.83, p-value = 0.7313
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.0674747  0.7496466
## sample estimates:
##   mean of x   mean of y 
## -0.13614433  0.02276974
```

```
t.test(large_norm,large_var, var.equal = TRUE)
```

```
## 
##  Two Sample t-test
## 
## data:  large_norm and large_var
## t = -0.34363, df = 998, p-value = 0.7312
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.066414  0.748586
## sample estimates:
##   mean of x   mean of y 
## -0.13614433  0.02276974
```

Using these much larger data sets the p-values for the two tests are much higher. The difference in variance is less dangerous with the larger sample. Nearly all of the returned values (including canfidance intervals and sample estimates) are virtually identical between the two tests. The advantage of the Welch test dissapears as the sample grows larger.

Another assumption is that the samples are both normally distributed

```
# Here we use both the Welch and two sample t tests to compare skewed data sets. One is skewe
d left, the other right. This setup violates the assumption of normality

t.test(Left_skew,Right_skew, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  Left_skew and Right_skew
## t = 2.0407, df = 7.1488, p-value = 0.07979
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1189305  1.6649305
## sample estimates:
## mean of x mean of y
##     0.464    -0.309
```

```
t.test(Left_skew,Right_skew, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Left_skew and Right_skew
## t = 2.0407, df = 8, p-value = 0.07559
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1004917  1.6464917
## sample estimates:
## mean of x mean of y
##     0.464    -0.309
```

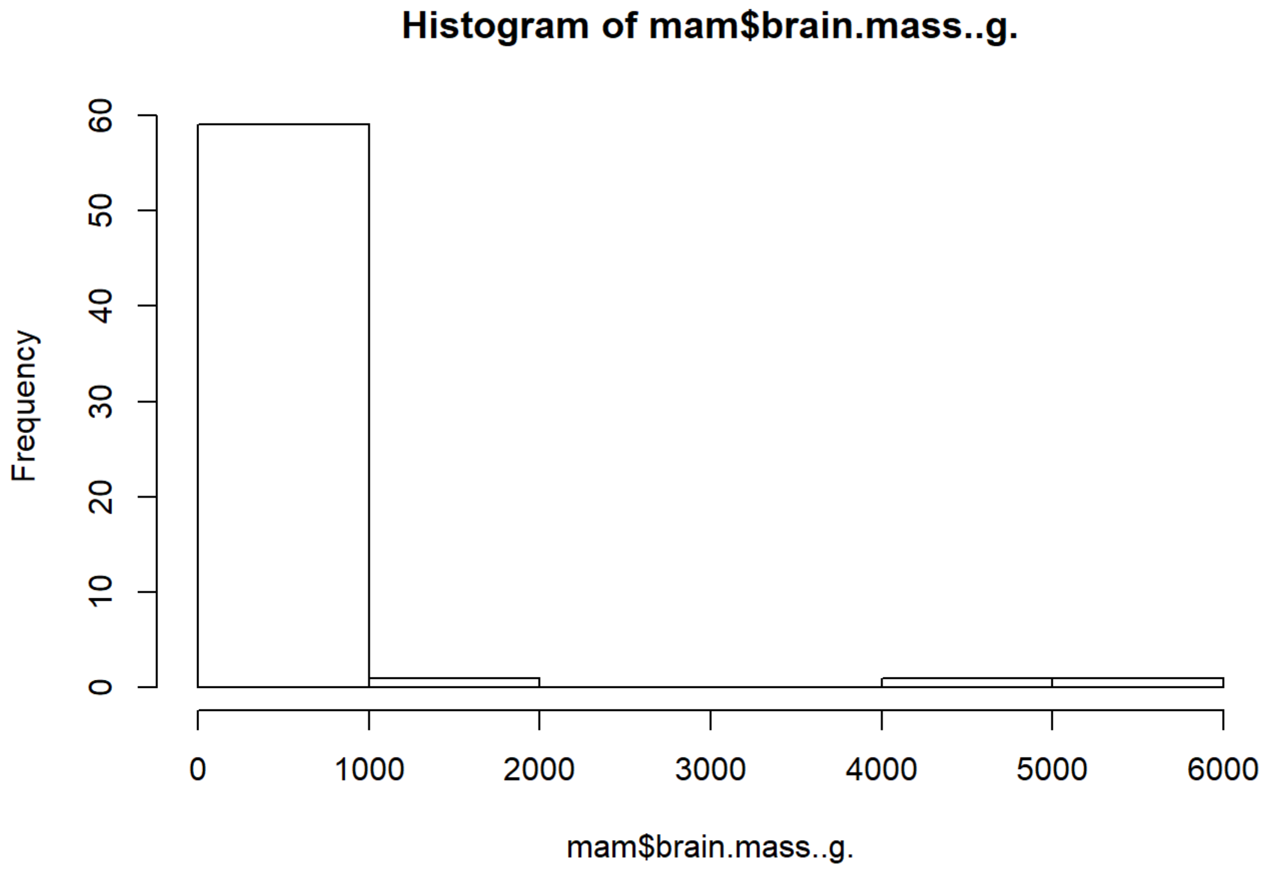
Again, both p values are worryingly close to .05. Neither actually causes an eronious rejection of the null. In this case, the p value for the Welch test is higher than that for the two sample test. The Welch test may have a slight advantage in robustness to violation of the normality assumption.

Question 2.

Are the brain sizes of mammals normally distributed?

a.

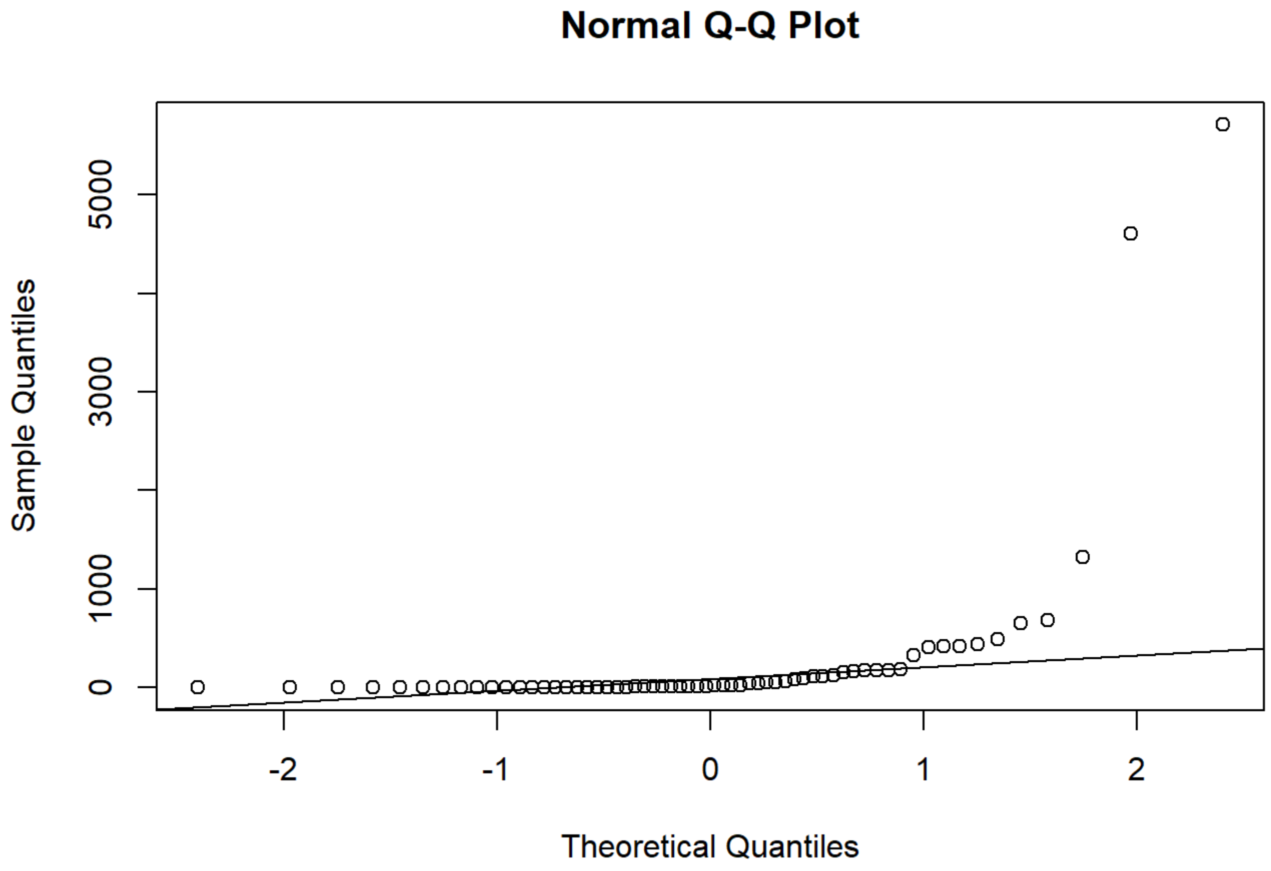
```
# Reads in the mammal data set and saves it as mam
mam <- read.csv("C:\\Users\\xenon\\Desktop\\R Studio 2018\\mammals.csv")

# Draws a histogram of the brain mass values
hist(mam$brain.mass..g.)
```



At a glance this does not appear normally distributed. Lets look at the qq plot to see how well it matches a normal distribution, and do a formal test for normality with shapiro.

b.

```
# Draws the qq plot and a line representing the ideal perfect match to a normal distribution
qqnorm(mam$brain.mass..g.)
qqline(mam$brain.mass..g.)
```



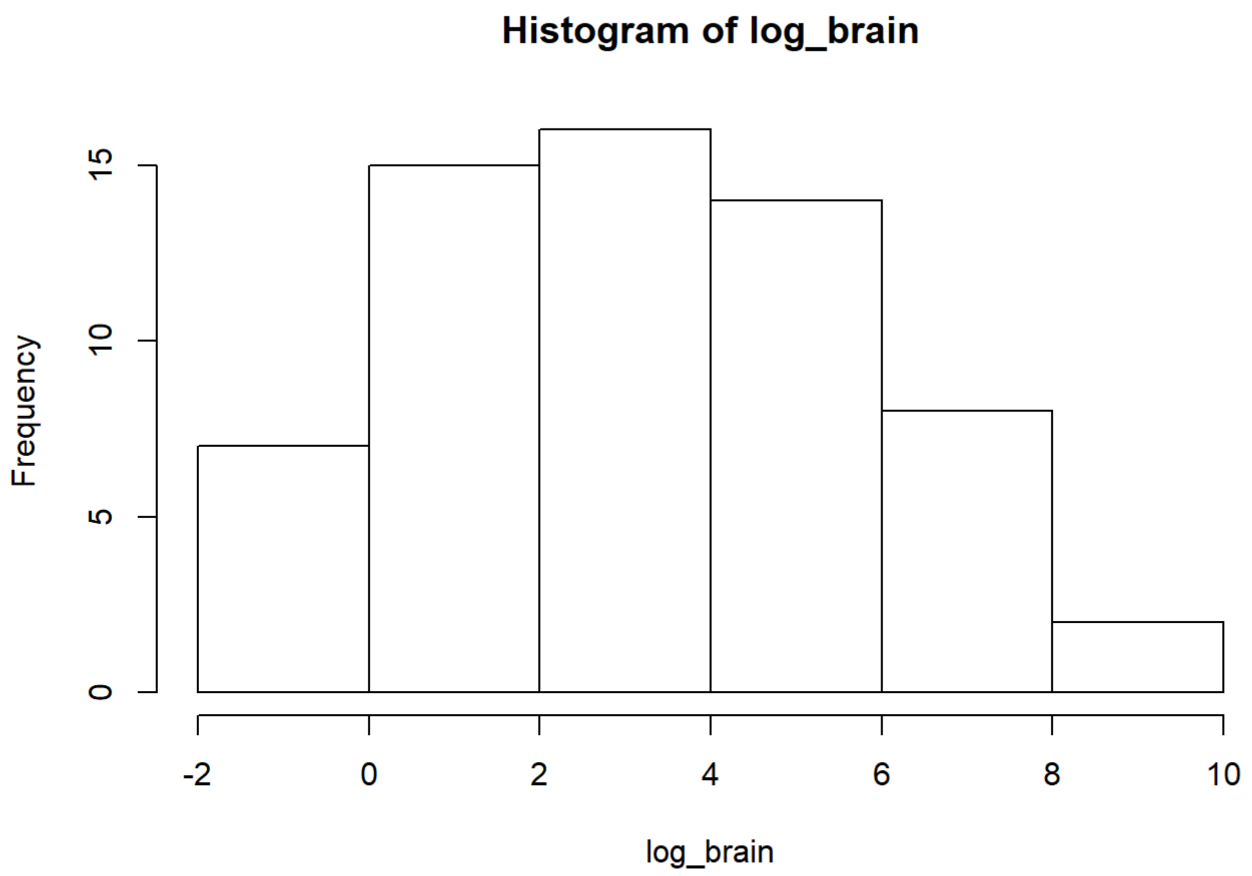
```
# Performs a shapiro formal test of normality.
shapiro.test(mam$brain.mass..g.)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mam$brain.mass..g.
## W = 0.31228, p-value = 1.677e-15
```

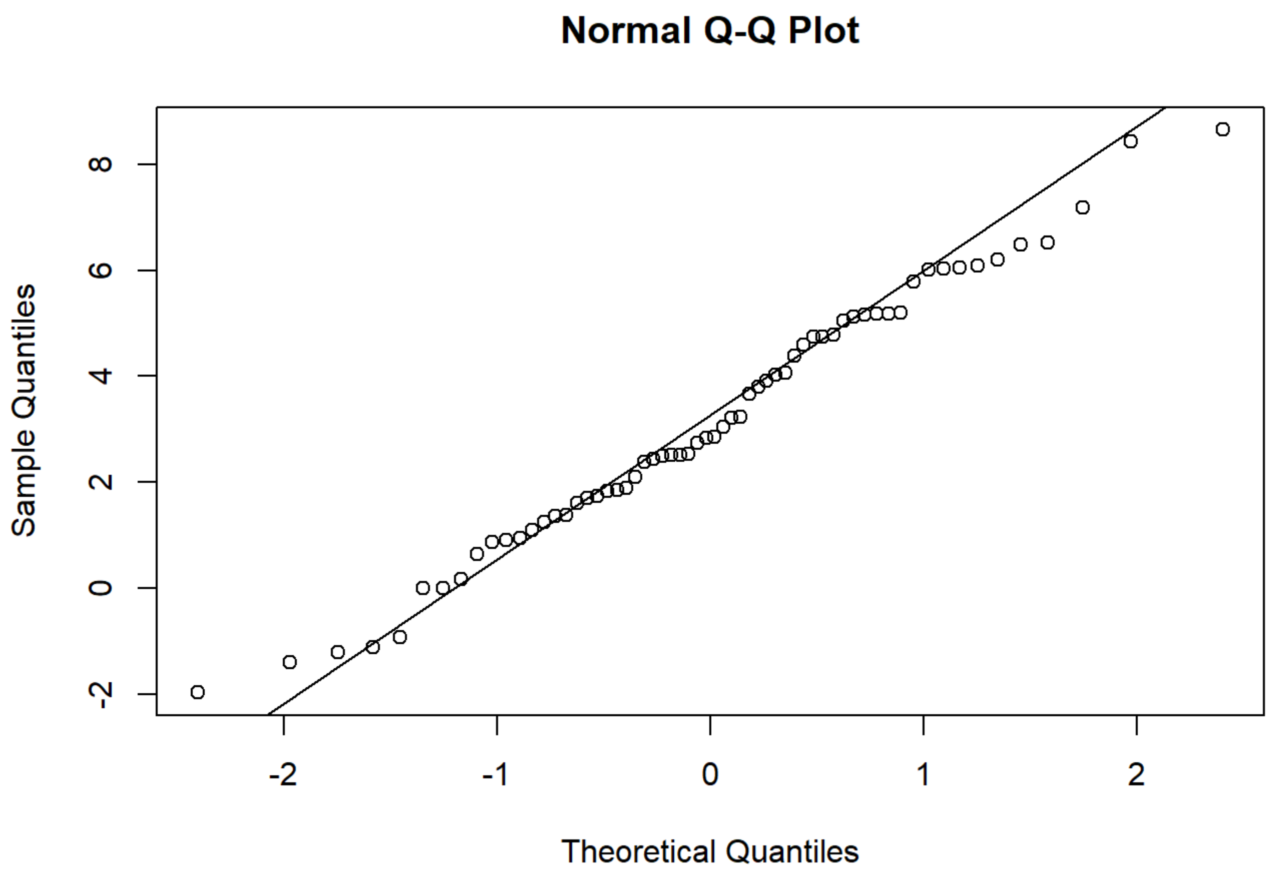
The distribution deviates wildly from normal at the right of the graph, getting very large. This does not look like a normal distribution because it is not linear.The p value of the wilk test is miniscule, and we can confidantly reject the null hypothesis of a normal distribution.Both examinations tell us the same thing. How can we transform the data to make it mor normal? The deviation appears exponential, so we can try a log transform.

c.

```
# Log transforms the brain mass values
log_brain <- log(mam$brain.mass..g.)
# Makes a histogram of the transformed values
hist(log_brain)
```



```
# Creates a qq plot to visualize how it compares to the line of an ideal normal distribution
qqnorm(log_brain)
qqline(log_brain)
```



```
# Performes a formal test of normality
shapiro.test(log_brain)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log_brain
## W = 0.98589, p-value = 0.697
```

The new distribution appears normal both in a histogram and a normal qq plot. The formal test of normality yeilds a p value of .697, which is greater than .05. this means we cannot reject the null hypothesis that it is normal.

Question 3.

In PS4 I investigated Humorous length. Is the variance equal between the male humorus length set and the female humorus length set?

```
# Reads in the bumpus data and saves it as bumpus
bumpus <- read.csv("C:\\Users\\xenon\\Desktop\\R Studio 2018\\bumpus.csv")

# Subsets the bumpus data into a male set and a female set
Male <- subset(bumpus, bumpus$sex == "m")
Female <- subset(bumpus, bumpus$sex == "f")

# Uses a two sample t-test to see if the varriances are equal between the humerus lengths in
 these groups
t.test(Male$length.humerus.in.,Female$length.humerus.in.,var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Male$length.humerus.in. and Female$length.humerus.in.
## t = 1.8836, df = 134, p-value = 0.06178
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0003846059  0.0157690769
## sample estimates:
## mean of x mean of y
## 0.7347126 0.7270204
```

p-value is .06178 which is greater than .05. We cannot reject the null hypothesis that viariance is equal between male and female humorus length.

Question 4. Is the distribution of snake tooth assymmetry scores normally distributed?

```
# Reads in the snake data set and saves it as SES
SES <- read.csv("C:\\Users\\xenon\\Desktop\\R Studio 2018\\snail-eating snakes.csv")

# Performes a formal test of normality on the assymmetry score data
shapiro.test(SES$Asymmetry.score)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  SES$Asymmetry.score
## W = 0.89948, p-value = 0.1562
```

We cannot reject the null hypothesis that this data is normally distributed.