Lab 3 Part 2 - a solution

Isabel Sassoon

```
#loading any libraries to be used in this notebook
library(ggplot2)
```

Hypothesis testing

Part 1

A company produces synthetic diamonds that have an average weight of 0.5 carat. An experiment has been conducted to evaluate a new process for producing synthetic diamonds. Six diamonds have been generated by the new process, with recorded weights of 0.46, 0.61, 0.52, 0.48, 0.57 and 0.54 carat. It is essential that the new process produces diamonds with a weight in excess of 0.50 carat. Do the six diamond weight measurements present sufficient evidence to indicate that the average weight of the diamonds produced by the process is in excess of 0.5 carat?

Test at the 1% significance level.

X is the diamond weights

```
x<-c(0.46, 0.61, 0.52, 0.48, 0.57, 0.54)
```

The hypothesis H0 μ =0.5 vs H1 $\mu > 0.5$

The hypothesis test can be performed by either - finding the critical value from the t distribution that is compared to the test statistic obtained from the sample - finding the p-value for the test statistic

Below are both approaches:

The critical value (using t due to sample size)

```
qt(0.99, df=5)
```

```
## [1] 3.36493
```

The test statistic computation:

```
test.st<-(mean(x)-0.5)/(sd(x)/sqrt(6))
```

We can also compute the p-value that corresponds to the test statistic, but recall we want the probability of a value equivalent to the sample mean or greater. So using the cumulative distribution function (of t) we need to look at 1 - pt(test.st, df)

```
1-pt(test.st, 5)
```

```
## [1] 0.1227025
```

Conclusion - fail to reject H : 0 as the test statistic is smaller than the critical value AND the p-value computed from the test statistic is 0.122 which is much larger than 0.01 (our desired significance level)

Hypothesis Testing Part 2

Breast Cancer data

Part 2

(a) Read in the data set breastCancer_Wissonsin.csv

```
breast.cancer <- read.csv("data/BreastCancer_Wisconsin.csv")
```

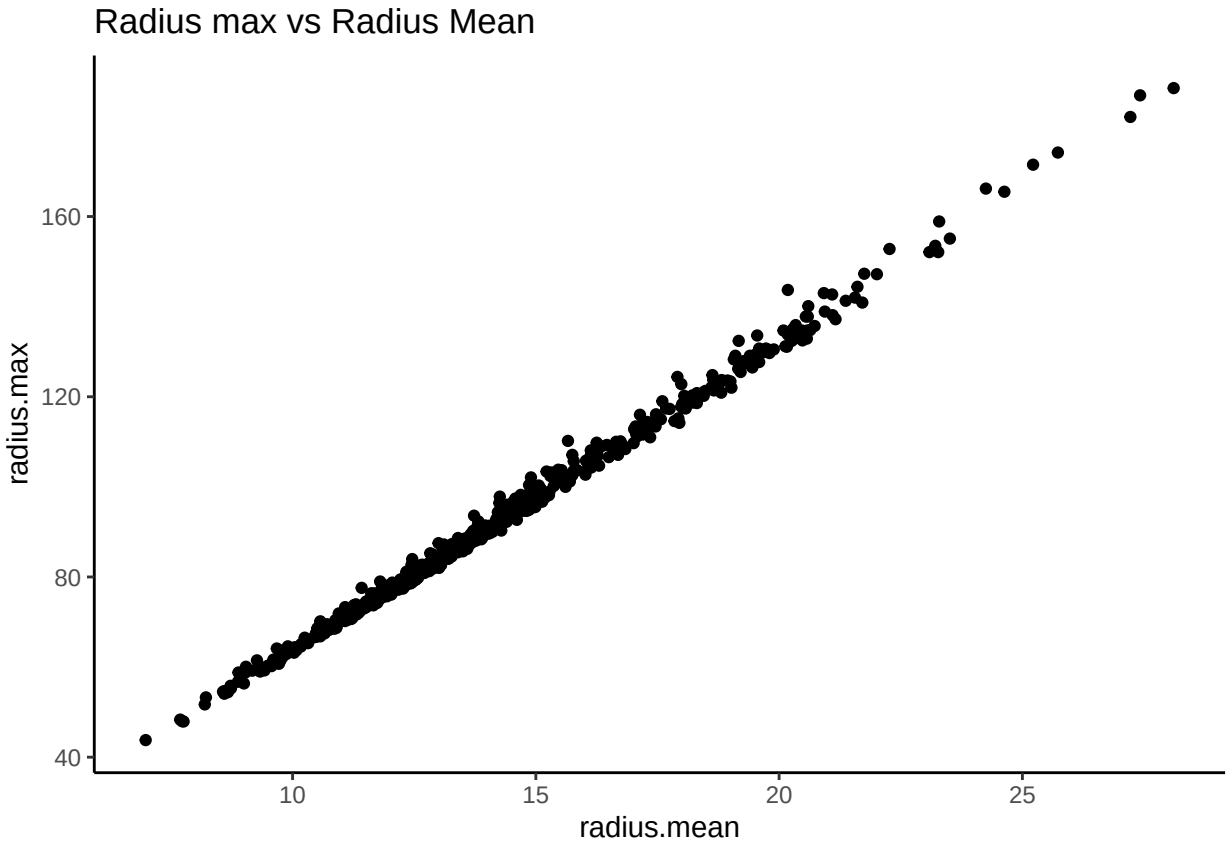
(b) Explore the data and the variables

```
summary(breast.cancer)
```

```
##   diagnosis          radius.mean       radius.sd       radius.max
##  Length:569         Min.   : 6.981   Min.   : 9.71   Min.   : 43.79
##  Class :character   1st Qu.:11.700   1st Qu.:16.17   1st Qu.: 75.17
##  Mode  :character   Median :13.370   Median :18.84   Median : 86.24
##                     Mean   :14.127   Mean   :19.29   Mean   : 91.97
##                     3rd Qu.:15.780   3rd Qu.:21.80   3rd Qu.:104.10
##                     Max.   :28.110   Max.   :39.28   Max.   :188.50
```

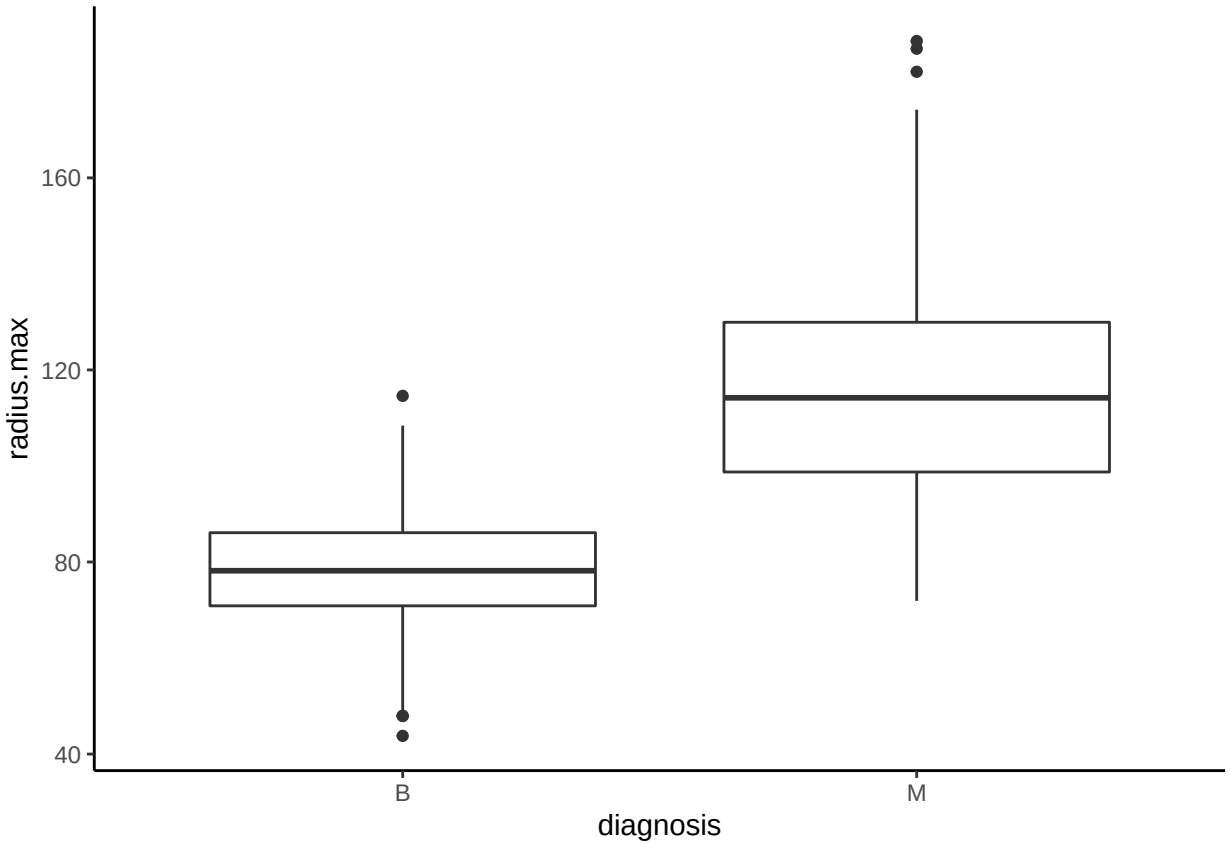
```
#example using ggplot

ggplot(data = breast.cancer, aes(x=radius.mean, y=radius.max)) + geom_point() + theme_classic() + ggtitl
```



Also look at plots between continuous and discrete, for example:

```
ggplot(data=breast.cancer, aes(x=diagnosis, y=radius.max)) + geom_boxplot() + theme_classic()
```



There appears to be a difference between the maximum radius between Beningn and Malignant Tumors. Benign ones have a smaller median value, and a smaller variance.

(c) Compute the proportion of tumors classified as benign (b)

```
table(breast.cancer$diagnosis)
```

```
##
##   B   M
## 357 212
```

```
357/length(breast.cancer$diagnosis)
```

```
## [1] 0.6274165
```

In the sample the proportion of benign tumours is 0.627.

Clinical research has found that the proportion of benign tumours in the population from which this sample data is drawn from is 0.6. This clinical research is used as our H_0.

Is the proportion in this sample data supportive of this proportion in the population? Write out the hypothesis to test and run them in R. Is H_0 rejected?

Answer: $H_0 : p = 0.6$ $H_1 : p > 0.6$

$Z = \frac{p-\pi}{\sqrt{\pi(1-\pi)/n}}$

```
denominator<-sqrt(0.6*0.4/569)
test.statistic<-(0.6274165-0.6)/denominator
test.statistic
```

```
## [1] 1.334942
```

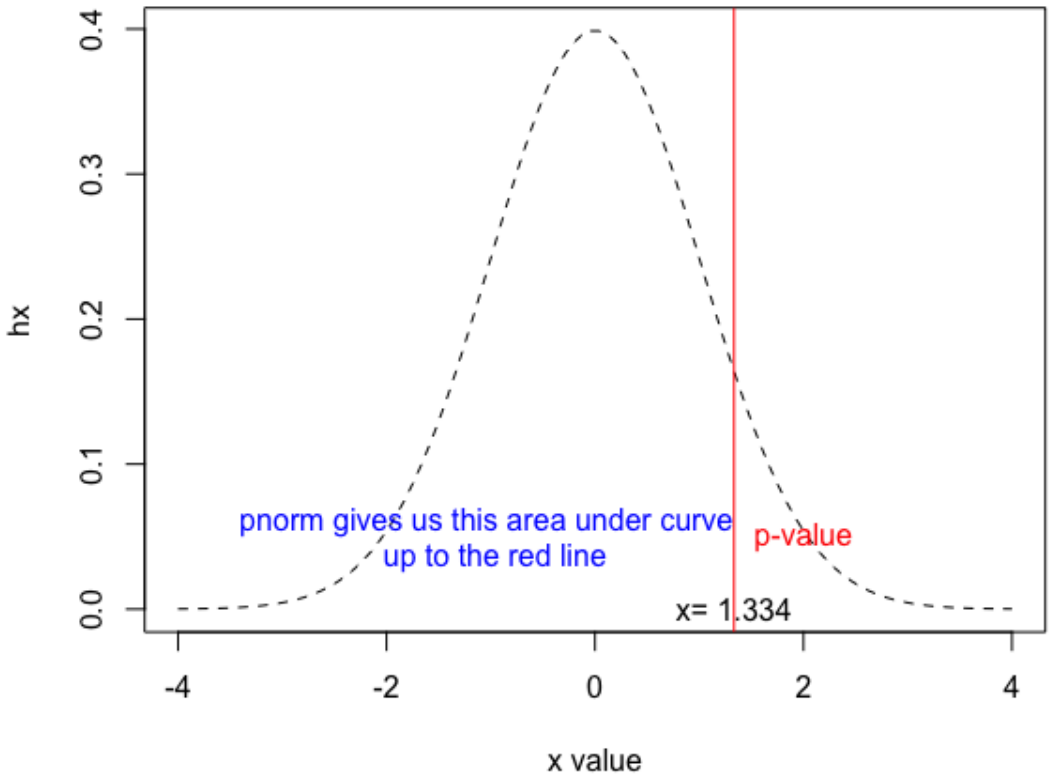


Figure 1: The test statistic on the Normal distribution

As we are comparing an alternative hypothesis of greater than - we look at the top part (x values greater than our test statistic) of the normal curve and want to know what the probability to obtain a test statistic value equal or more extreme to the one calculated here.

Figure 1 (above) shows the normal distribution and where our computed test statistic falls on the x axis. The p-value is the area under the curve of this density plot, from the red vertical line and above (from x=red line and larger values of x). This corresponds to the probability (in this distribution) that we would obtain a test statistic with a value of 1.334 or greater. But using pnorm(1.334) will give us the area under the curve up to the red line. Therefore to find the area under the curve from x=1.334 and above we need 1-pnorm(x).

Remember that the pnorm(x) function gives us the cumulative distribution function for the normal distribution. This is the probability of getting a value that is equal to or less than x.

```
pnorm(test.statistic)
```

```
## [1] 0.9090524
```

In our case we are wanting to find the probability of a value equal to or greater than our test statistic, so we need the opposite probability.

```
1-pnorm(test.statistic)
```

```
## [1] 0.0909476
```

This is our p-value, and it is larger then 0.05 and therefore we cannot reject the NULL Hypothesis H_0.

THIS IS EXTRA TO THE WORKSHEET

What would happen if clinical research stated that the percentage of beningn tunours is 0.7, (not 0.6). Then we may want to test this set of hypotheses:

$H_0 : \pi = 0.7$ and $H_1 : \pi < 0.7$

Our sample proportion is smaller than the one in the population.

$Z = \frac{p-\pi}{\sqrt{\pi(1-\pi)/n}}$

```
denominator2<-sqrt(0.7*0.3/569)
test.statistic2<-(0.6274165-0.7)/denominator2
test.statistic2
```

```
## [1] -3.778195
```

In this case pnorm(test statistic2) will give us the p-value. See in Figure 2 - we want the area under the curve up to a value of x=-3.78.

```
pnorm(test.statistic2)
```

```
## [1] 7.898452e-05
```

This value is very small (<0.0001) so in this case we would reject the null hypothesis H_0.

THIS IS ALSO EXTRA TO THE WORKSHEET

What would happen if we wanted to test the following hypothesis:

$H_0 : \pi = 0.7$ and $H_1 : \pi \neq 0.7$

The calculation for the test statistic are the same as we did for the previous test, where we computed test.statistic2. The computation of the test statistic does not depend on the "direction" of the hypothesis being tested.

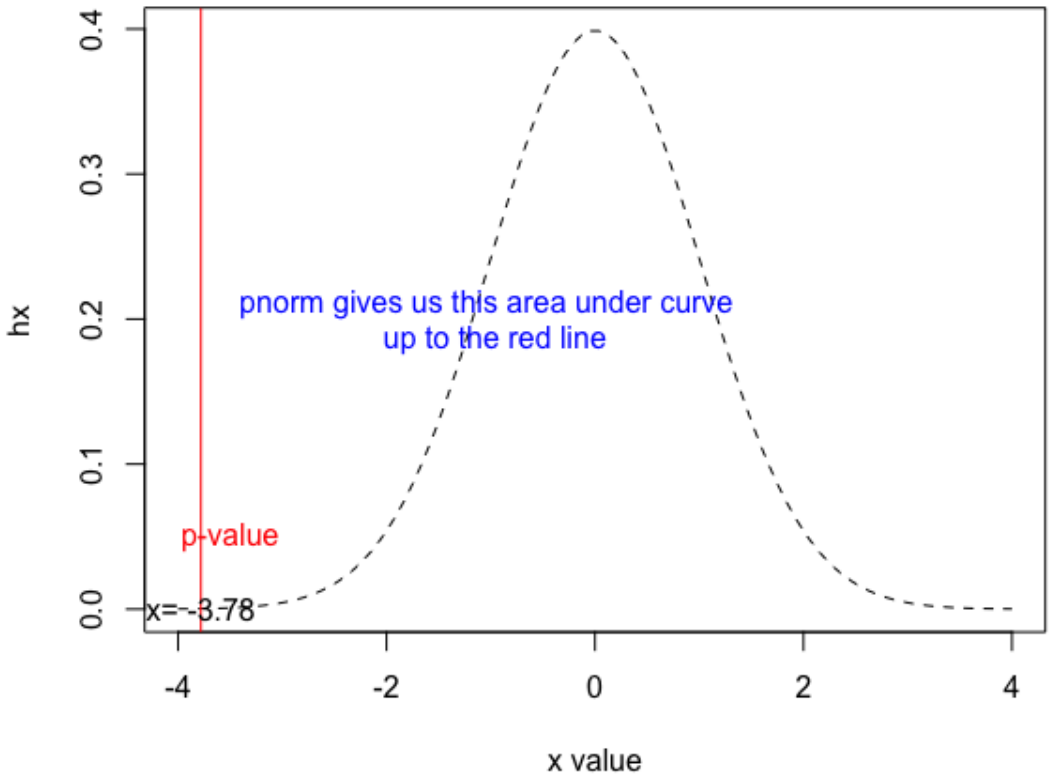


Figure 2: The test statistic on the Normal distribution other

BUT we need to find the probability on the normal distribution that we obtain a value more extreme in both directions. In other words we need both ends. See Figure 3.

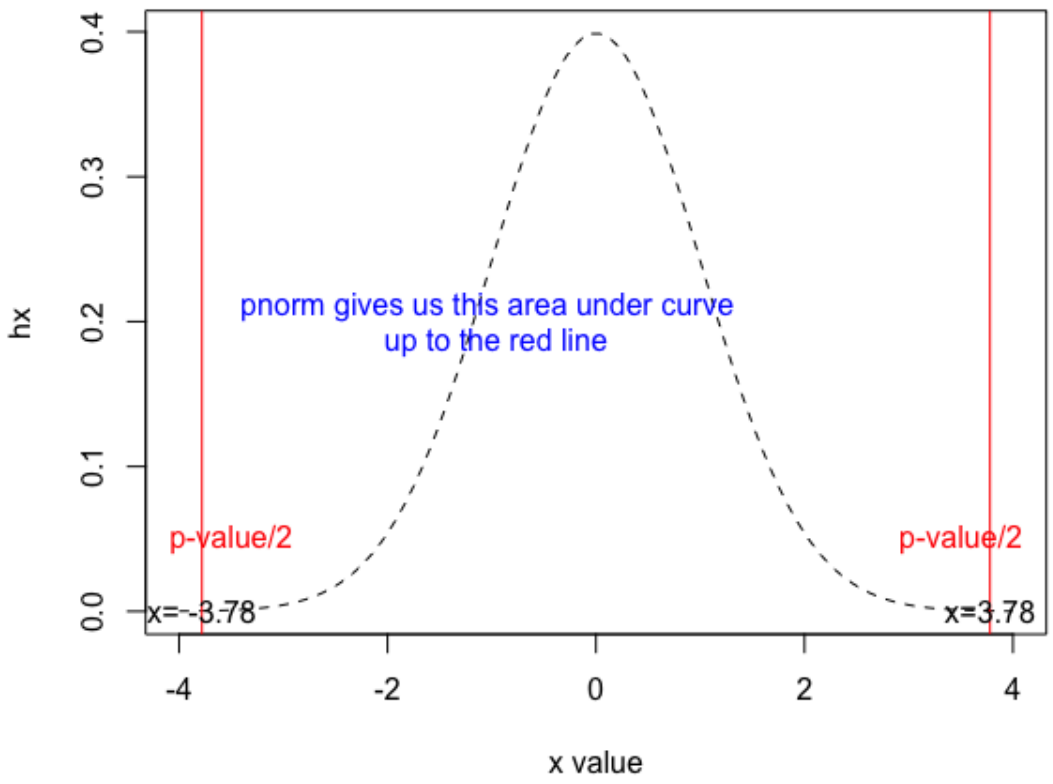


Figure 3: The test statistic on the Normal distribution two sided

So we need the probability (area under curve) for x=-3.78 and below AND x=3.78 and above

```
test.statistic2
```

```
## [1] -3.778195
```

```
#area under curve up to the value of the test statistic (that is negative on the LHS of the mean of the
pnorm(test.statistic2)
```

```
## [1] 7.898452e-05
```

```
#area under the curve from the value of the test statistic (that is positive and on the RHS of the mean
1- pnorm(-test.statistic2)
```

```
## [1] 7.898452e-05
```

As you can see the tails of the distribution are symetric round 0, so the area under the curve for each tail is the same. To obtain the p-value we add them up

```
pnorm(test.statistic2) + 1- pnorm(-test.statistic2)
```

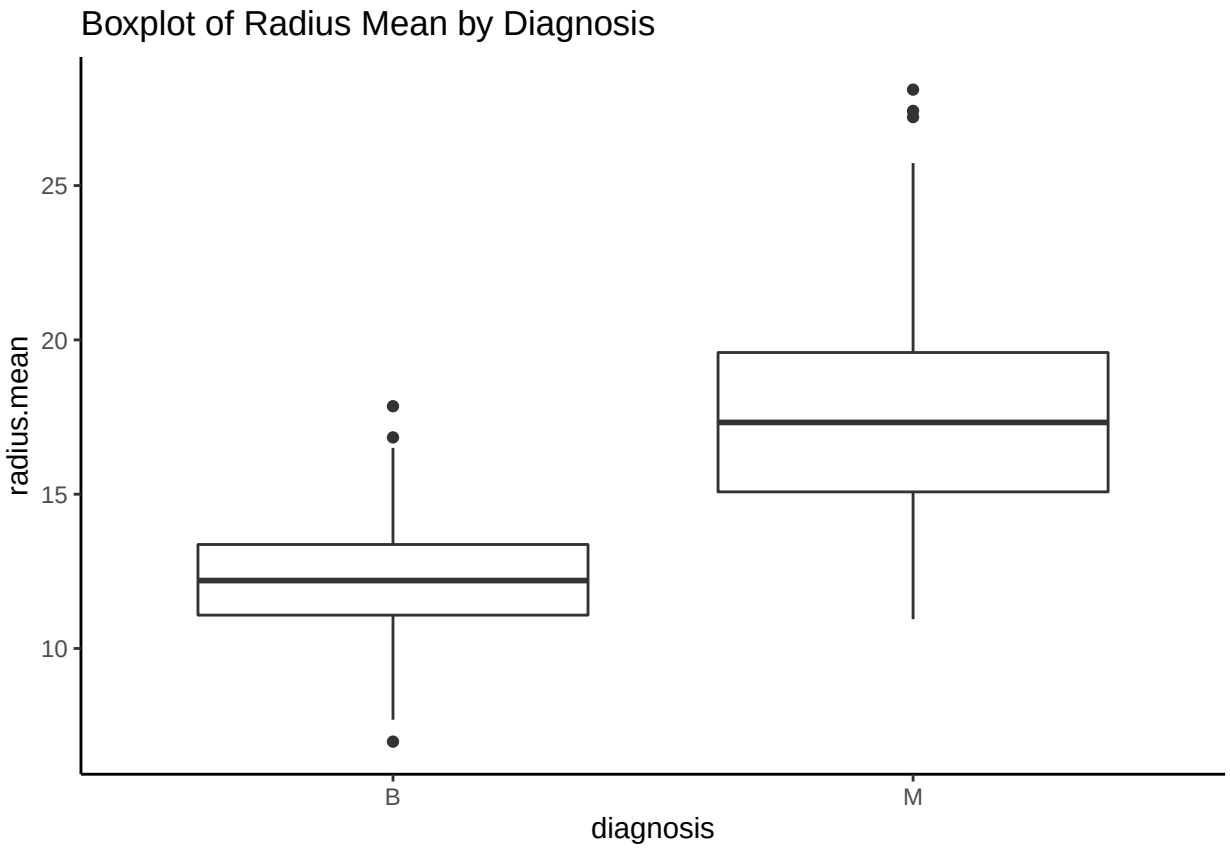
```
## [1] 0.000157969
```

This is a very small p-value and we reject this H_0 two sided.

(e) Is there a difference between the radius mean for benign vs malignant? (test the variances too)

Explore the radius.mean attribute (column) in the data. Do so numerically and graphically. Firstly we can visualize this:

```
ggplot(breast.cancer, aes(x=diagnosis, y=radius.mean)) + geom_boxplot() +
        ggtitle("Boxplot of Radius Mean by Diagnosis") + theme_classic()
```



(f) In order to perform the hypothesis test that compares the radius mean for beningn tunours to the radius mean for malignant tumours the data should be split:

```
#starting by splitting the data by diagnosis
bc.b<-subset(breast.cancer, breast.cancer$diagnosis=="B")
bc.m<-subset(breast.cancer, breast.cancer$diagnosis=="M")
```

Then the variance of each of the samples can be compared

```
var.test(breast.cancer$radius.mean~breast.cancer$diagnosis, alternative="two.sided")
```

```
##
##  F test to compare two variances
##
## data:  breast.cancer$radius.mean by breast.cancer$diagnosis
## F = 0.30883, num df = 356, denom df = 211, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2416864 0.3916480
```

```
## sample estimates:
## ratio of variances
##           0.308825
```

In this case the variances are not equal but in R this will be corrected in the Welch t-test

```
t.test(breast.cancer$radius.mean~breast.cancer$diagnosis, alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  breast.cancer$radius.mean by breast.cancer$diagnosis
## t = -22.209, df = 289.71, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.787448 -4.845165
## sample estimates:
## mean in group B mean in group M
##        12.14652        17.46283
```

Clearly there is a difference in the groups. The p value in this case is very small.

(g) 95% confidence interval for a proportion

The sample proportion of beningn tumours is 0.627

Recall how to compute a confidence interval for a proportion:

$p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$

```
p<-357/length(breast.cancer$diagnosis)
n<-length(breast.cancer$diagnosis)
#the sqrt part of the confidence interval eq
vr<-sqrt(p*(1-p)/n)

ucl<-p+qnorm(0.025)*vr
lcl<-p-qnorm(0.025)*vr

ucl
```

```
## [1] 0.5876899
```

```
lcl
```

```
## [1] 0.6671432
```

The 95% confidence interval is: (0.59, 0.67). Note that it does include the H_0 hypothesis from (d).