Lab-3-Part-1

Overview of the content

This guided lab will cover Hypothesis Testing. There are 3 examples:

1. The bags of flour example, where a sample mean is tested vs a population mean
2. An example where a proportion of patients cured with a treatment is compared to the overall proportion of cured patients with no therapy.
3. Starting from data (using the Iris data) we test:

(i) The mean of one column of the sample data against a population mean
(ii) Is the column of data normally distributed
(iii) Looking at two types of flowers - testing whether the variance is the same in both species for petal length.

1. Hypothesis testing the mean from a sample vs population mean

This is the example covered in the lecture

- Bags of flour are supposed to contain 2kg on average.
- A random sample of 20 bags found to have a mean weight of 1.97 Kg, with a standard deviation of 0.1kg
- Is the flour bagging machine working correctly?

The hypotheses to test are $H_0 : \mu = 2$ and $H_1 : \mu \neq 2$

This is how to perform this hypothesis test in R

Computing the test statistic and in this case we should use t as the sample is small

$t_{obs} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$

```
test.statistic<-(1.97-2)/(0.1/sqrt(20))
```

Finding the p-value for the two sided test

```
pt(test.statistic, df=19)*2
```

```
## [1] 0.1955287
```

In this case the p-value is above 0.05, therefore there is not enough evidence to reject H_0. In other words if the population mean is 2kg, then the chances of sample like the one we have or more extereme is 20%. So it is not so rare that we start to doubt the Null Hypothesis.

There is another way of using this information to test the hypothesis. Instead of finding the p-value for the test statistic we find the critical value that corresponds to the confidence level we want.

We are testing at 95% confidence level on both sides, so we want to find out what is the value on the x axis that corresponds to have 95% of the probability between it. This corresponds to the value of having 5% of the probability at the extremes. We can therefore use the qt function to obtain that for one side at the 0.975 (or we could also have used the 0.025 level)

```r
qt(0.025, df=19)
```

```
## [1] -2.093024
```

Lets remind ourselves of the value of the test statistic:

```r
test.statistic
```

```
## [1] -1.341641
```

The value of the test statistic is less extreme that this value, so this (as expected) supports the conclusion not to reject H_0.

What would happen if the sample size wass now $n = 60$?

We could use the Z test and we would need to re-compute the test statistic to include the new value for n.

perform the same hypothesis test using the Z distribution: use *pnorm* and *qnorm*

```r
test.statistic2<-(1.97-2)/(0.1/sqrt(60))
test.statistic2
```

```
## [1] -2.32379
```

Now lets find the probability of a value such as the test statistic obtained (-2.323) or more extreme

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

```
## [1] 0.01006838
```

Now this is significant.

---

2. Hypothesis test sample proportion vs population

A new drug therapy is tested. Of 50 patients in the study, 43 had no recurrence in their illness after 18 months. With no drug therapy, the expected percentage of no recurrence would have been 75%.

Test at the 5% significance level the hypothesis that the **proportion** of patients with no recurrence has increased with the new therapy

The hypothesis to be tested is: $H_0 : \pi = 0.75$ $H_1 : \pi \geq 0.75$

in the study (our sample data) the value of p is:

```r
43/50
```

```
## [1] 0.86
```

So 86% of patients had no recurrence.

We can use the test statistic:

$Z = \frac{p-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$

The observed test statistic is:

```r
den<-sqrt((0.75*0.25)/50)
p.test.stat<-(0.86-0.75)/den
p.test.stat
```

```
## [1] 1.796292
```

Now we can find the probability assuming a normal distribution (as this is a large sample n*p >5)

```
pnorm(p.test.stat)
```

```
## [1] 0.963776
```

This is a one way test so we need to find the probability of a value greater than our test statistic. This means that we should look at

```
1-pnorm(p.test.stat)
```

```
## [1] 0.03622401
```

This tells us that such a value is very unlikely given H0, so we have evidence to reject H_0 and to conclude that the therapy is beneficial.

If we are testing with α =0.05 one sided then the critical value is

```
qnorm(0.95)
```

```
## [1] 1.644854
```

Which is smaller (less extreme) than the test statistic from the sample therefore there is evidence to reject H_0

But what is the p-value for this test statistic

```
1-pnorm(p.test.stat)
```

```
## [1] 0.03622401
```

---

3. Hypothesis Testing starting from raw data

We may have a sample of data available to use for hypothesis testing. The Iris data is available in R

In order to access it:

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

This data has 4 measurments for each flower, and for each flower we also know what species it is.

If we wanted to see how many species there are in the data and what they are:

```
table(iris$Species)
```

```
##
##     setosa versicolor  virginica
##         50         50         50
```

Now we are going to look at the mean petal length for all flowers.

In our sample in order to compute the mean we can use:

```r
mean(iris$Petal.Length)
```

```
## [1] 3.758
```

(i) The hypothesis we are asked to test is: mean petal length for all flowers = 4.0

In this case $H_0 : \mu = 4$ and $H_1 : \mu \neq 4$

This is a two sided hypothesis test so we will be looking to compare the value of the test statistic from the sample to the corresponding value at $p = 0.025$ if the test is two sided with an overall significance level of 0.95.

We should compute the test statistic using:

$z = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$

```r
mean.petal.length<-mean(iris$Petal.Length)
sd.petal.length<-sd(iris$Petal.Length)
sqrt.n<-sqrt(150)
iris.test.statistic<-(mean.petal.length-4)/sd.petal.length/sqrt.n
```

We can now compare the likelihood of this test statistic based on the normal distribution

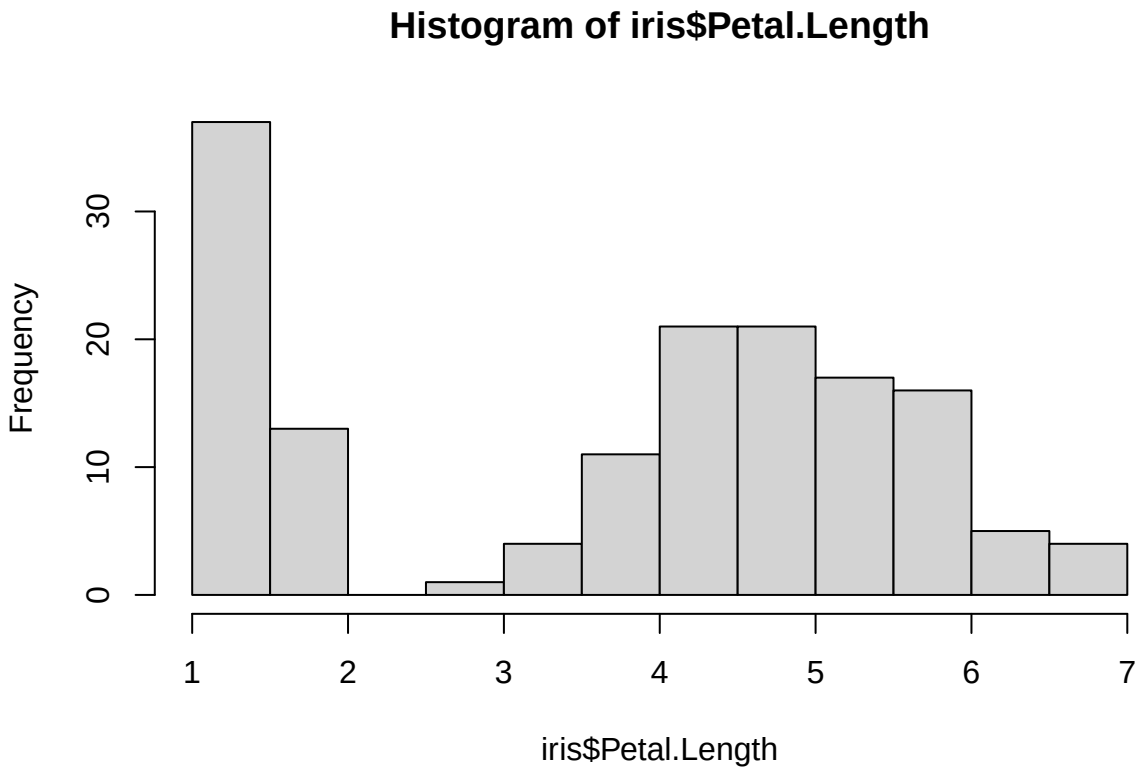
```r
pnorm(iris.test.statistic)
```

```
## [1] 0.4955347
```

This means that there is almost a 50% chance of getting a sample mean like ours or more extreme. So we do not reject the Null Hypothesis in this case.

(ii) Now lets look at testing for normality in one of the columns of data

```r
hist(iris$Petal.Length)
```



This simple histogram does not look very normally distributed

```
shapiro.test(iris$Petal.Length)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  iris$Petal.Length
## W = 0.87627, p-value = 7.412e-10
```

This is a very small p-value, we reject the null hypothesis that petal length for all flower species in this sample is normally distributed.

Now we want to see how hypothesis testing works when we have two samples

(iii) Next we may want to check differences in variance between flower species:

Firstly we should split the data by flower type.

```
setosa.flowers<-subset(iris, iris$Species=="setosa")
```

Notice that this code above is case sensitive - if you use ==“Setosa” it wont return any flowers!

The same can be done for other flowers

```
versicolor.flowers<-subset(iris, iris$Species=="versicolor")
virginica.flowers<-subset(iris, iris$Species=="virginica")
```

We can now compare two variances one from each of two different flower species

Lets look at petal length and find the variance of petal length for both setosa and virginical flowers

```
var(setosa.flowers$Petal.Length)
```

```
## [1] 0.03015918
```

```
var(virginica.flowers$Petal.Length)
```

```
## [1] 0.3045878
```

The test statistic is obtained by dividing the variances and then using the F - distribution

```
var.setosa<-var(setosa.flowers$Petal.Length)
var.virginica<-var(virginica.flowers$Petal.Length)
f.st<-var.setosa/var.virginica
f.st
```

```
## [1] 0.0990164
```

This ratio looks very far from 1 so using the a built in function in R to perform the variance test

```
var.test(setosa.flowers$Petal.Length,virginica.flowers$Petal.Length )
```

```
##
##  F test to compare two variances
##
## data:  setosa.flowers$Petal.Length and virginica.flowers$Petal.Length
## F = 0.099016, num df = 49, denom df = 49, p-value = 1.875e-13
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.05618945 0.17448557
## sample estimates:
## ratio of variances
##          0.0990164
```

We confirm that the p-value is very small so such a ratio is very unlikely (if we assume H_0 of equal variances). We can conclude that the variances are different.

If we want to compare the means using the apropriate hypothesis test

```
t.test(setosa.flowers$Petal.Length,virginica.flowers$Petal.Length)
```

```
##
##  Welch Two Sample t-test
##
## data:  setosa.flowers$Petal.Length and virginica.flowers$Petal.Length
## t = -49.986, df = 58.609, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.253749 -3.926251
## sample estimates:
## mean of x mean of y
##     1.462     5.552
```

R has adjusted for the unequal variances, and the t-test conculded with a p-value that is very small. We can reject the Null Hypothesis H_0. The means are different.