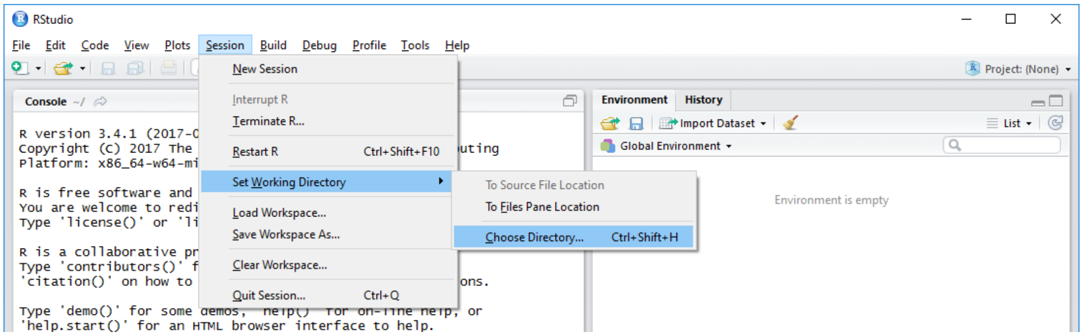
BICF Nano Course: introduction to statistics

Read and check the data

1. Set the working directory



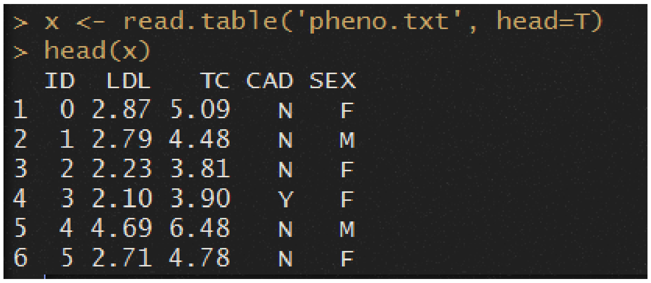
```
R version 3.4.1 (2017-0
Copyright (C) 2017 The                         uting
Platform: x86_64-w64-mi

R is free software and
You are welcome to redi
Type 'license()' or 'li

R is a collaborative pr
Type 'contributors()' f
'citation()' on how to                         ons.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
```

2. Read the data file ‘pheno.txt’

```
x <- read.table('pheno.txt', head=T)
```

3. Check the first few lines of the data

```
head(x)
```



```
> x <- read.table('pheno.txt', head=T)
> head(x)
  ID  LDL   TC CAD SEX
1  0 2.87 5.09   N   F
2  1 2.79 4.48   N   M
3  2 2.23 3.81   N   F
4  3 2.10 3.90   Y   F
5  4 4.69 6.48   N   M
6  5 2.71 4.78   N   F
```

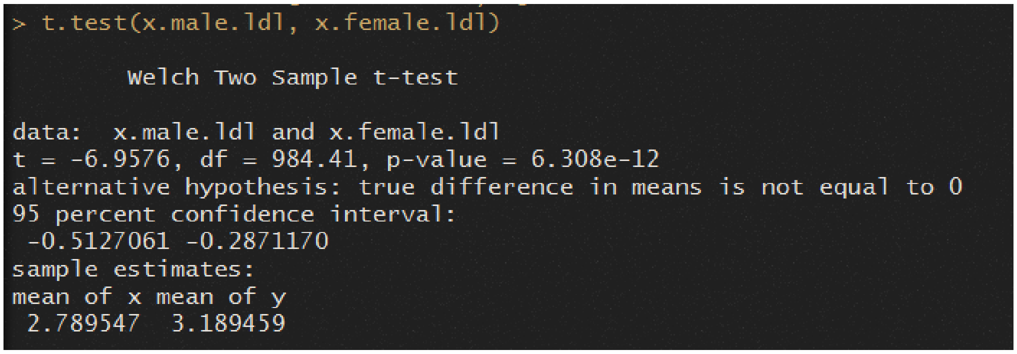
T-test

1. Question: Are there significant difference of LDL levels between the male and female?
2. Slice the data

```
x.male.ldl <- x[x$SEX == 'M', 2]
x.female.ldl <- x[x$SEX == 'F', 2]
```

3. Run t-test

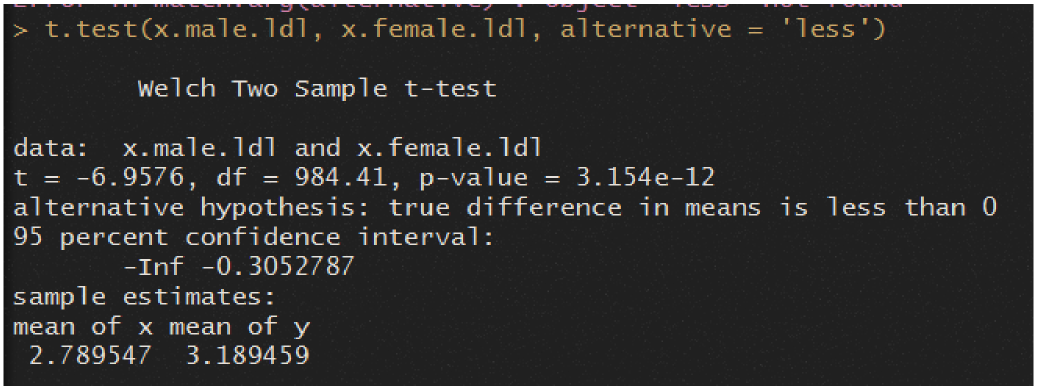
```
t.test(x.male.ldl, x.female.ldl)
```



```
> t.test(x.male.ldl, x.female.ldl)

        Welch Two Sample t-test

data:  x.male.ldl and x.female.ldl
t = -6.9576, df = 984.41, p-value = 6.308e-12
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5127061 -0.2871170
sample estimates:
mean of x mean of y 
 2.789547  3.189459 
```

4. Question: Is the LDL levels in male lower than the LDL levels in female?
5. One-sided t-test

```
t.test(x.male.ldl, x.female.ldl, alternative = 'less')
```

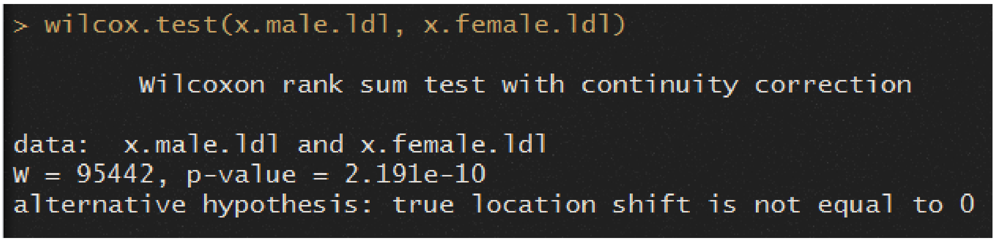


```
> t.test(x.male.ldl, x.female.ldl, alternative = 'less')

        Welch Two Sample t-test

data:  x.male.ldl and x.female.ldl
t = -6.9576, df = 984.41, p-value = 3.154e-12
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.3052787
sample estimates:
mean of x mean of y
 2.789547  3.189459
```

Wilcoxon rank-sum test

1. Question: Are there significant difference of LDL levels between the male and female?
2. Run wilcox.test

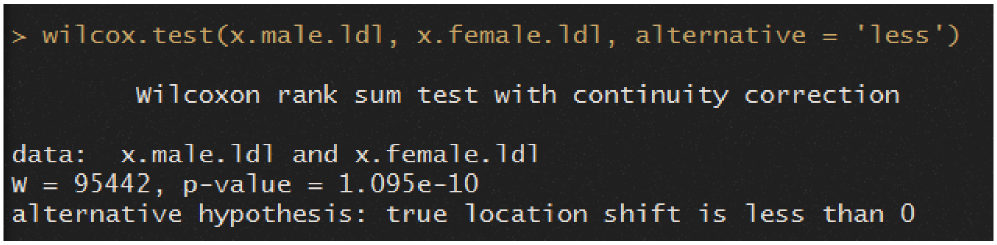
```
wilcox.test(x.male.ldl, x.female.ldl)
```



```
> wilcox.test(x.male.ldl, x.female.ldl)

        Wilcoxon rank sum test with continuity correction

data:  x.male.ldl and x.female.ldl
W = 95442, p-value = 2.191e-10
alternative hypothesis: true location shift is not equal to 0
```

3. Question: Is the LDL levels in male lower than the LDL levels in female?
4. One-sided wilcox.test

```
wilcox.test(x.male.ldl, x.female.ldl, alternative = 'less')
```

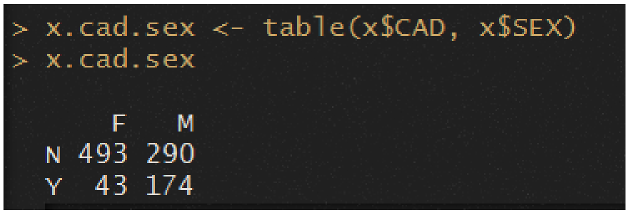


```
> wilcox.test(x.male.ldl, x.female.ldl, alternative = 'less')

        Wilcoxon rank sum test with continuity correction

data:  x.male.ldl and x.female.ldl
W = 95442, p-value = 1.095e-10
alternative hypothesis: true location shift is less than 0
```

Fisher's exact test

1. Question: Is the gender associated with cardiovascular disease (CAD)?
2. Generate the contingency table

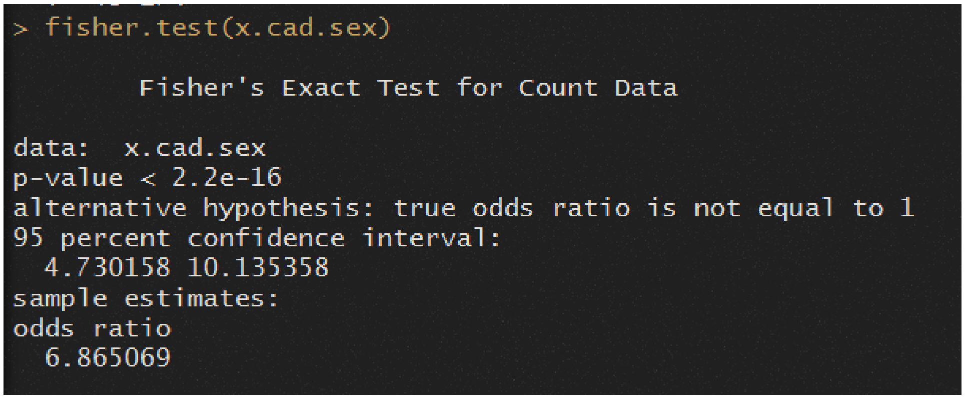
```
x.cad.sex <- table(x$CAD, x$SEX)
x.cad.sex
```



```
> x.cad.sex <- table(x$CAD, x$SEX)
> x.cad.sex

      F   M
  N 493 290
  Y  43 174
```

3. Run fisher.test

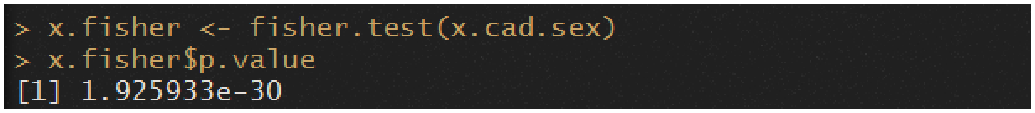
```
fisher.test(x.cad.sex)
```



```
> fisher.test(x.cad.sex)

        Fisher's Exact Test for Count Data

data:  x.cad.sex
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  4.730158 10.135358
sample estimates:
odds ratio
  6.865069
```

4. Get the exact p-value

```
x.fisher <- fisher.test(x.cad.sex)
x.fisher$p.value
```

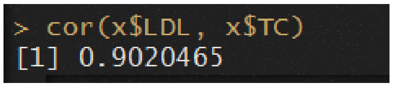


```
> x.fisher <- fisher.test(x.cad.sex)
> x.fisher$p.value
[1] 1.925933e-30
```

Correlation

1. Question: Is the LDL correlated with TC?
2. Estimate Pearson correlation coefficience

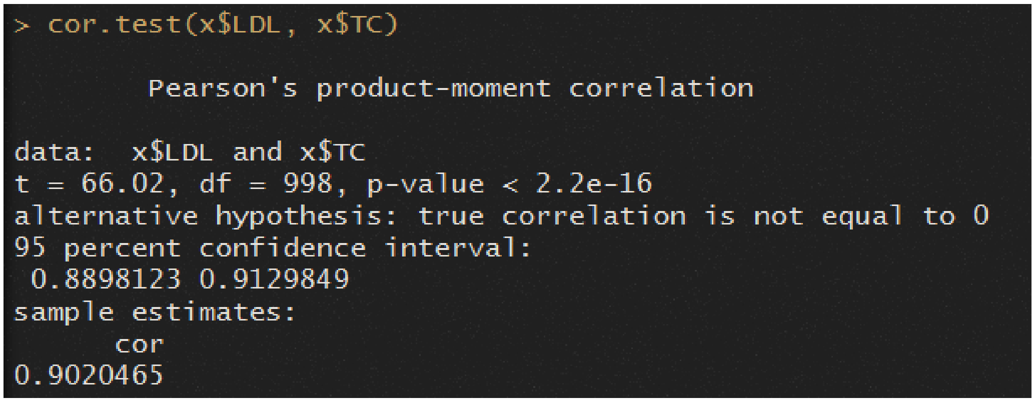
```
cor(x$LDL, x$TC)
```



```
> cor(x$LDL, x$TC)
[1] 0.9020465
```

3. Test the significance of the correlation between LDL and TC

```
cor.test(x$LDL, x$TC)
```



```
> cor.test(x$LDL, x$TC)

        Pearson's product-moment correlation

data:  x$LDL and x$TC
t = 66.02, df = 998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8898123 0.9129849
sample estimates:
      cor
0.9020465
```