R Intermediate Short Course
Session 5 - Hypothesis Testing (Independence Test)

The University of the West Indies, St Augustine

Friday 22rd July 2022
(5:00pm - 7:00pm)
(online)

Chi-Squared Distribution

Random variables may be approximately distributed as **Chi-Square**.

A chi-squared distribution is a result of the sum of squared standard normal distributions. It is skewed to the right.

If $X \sim \chi^2_v$ then,
E(X) = v and Var(X) = 2v
where v = df (degree of freedom)

The two common tests using the chi-squared distribution are:

1. Goodness of fit test.
To test deviation between expected and observed (one way analysis)
2. Test for independence in a contingency table (two way analysis)

Both methods involve hypothesis testing.

The general hypotheses for a chi-squared independence test is:

H_0: independent
vs
H_1: not independent (or dependent)

We reject the null hypothesis, H_0 if:

1. test statistic $>$ critical value
2. p-value $<$ level of significance, α

The test statistic χ^2 and p-value is given in the output of a chi-squared test in R while the critical value $\chi^2_{\alpha,\mathsf{df}}$ can be found by:

```
qchisq(1 - alpha, df)
```

Performing a chi-squared test in R

```
install.packages("MASS")
library(MASS)
chisq.test(data)
```

The type of chi-squared test conducted i.e. one-way analysis or two-way analysis is dependent on the arrangement of the data (whether arranged manually or in a data file before reading into R).

One way analysis

Example 1: In a study to compare the incidence of concussions among athletes, the following data was obtained:

Sport	Baseball	Basketball	Football	Soccer
Frequency	11	26	45	68

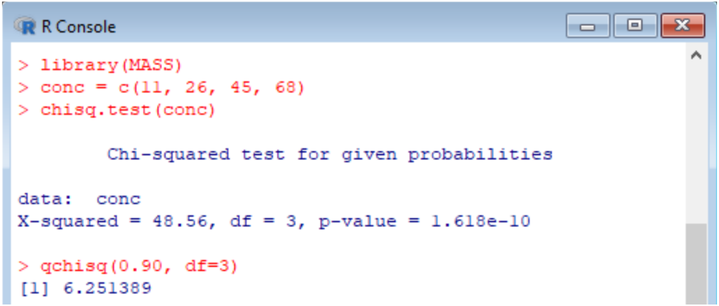
Test at a 10% level of significance whether the number of concussions depends on the sport played by athletes.

Required to test,

H_0: The number of concussions is independent of the sport played
H_1: The number of concussions depends on the sport played

```
> library(MASS)
> conc = c(11, 26, 45, 68)
> chisq.test(conc)
> qchisq(0.90, df=3)
```

R Output:



```
> library(MASS)
> conc = c(11, 26, 45, 68)
> chisq.test(conc)

        Chi-squared test for given probabilities

data:  conc
X-squared = 48.56, df = 3, p-value = 1.618e-10

> qchisq(0.90, df=3)
[1] 6.251389
```

Since $\chi^2 > \chi^2_{0.10,3}$ i.e. $48.56 > 6.251$ and p-value $< \alpha$ i.e. $1.618 \times 10^{-10} < 0.10$ we reject H_0 at a 10% level of significance. The number of concussions depends on the sport played by athletes.

Example 2: Using the *mtcars* built-in data set, determine at a 2% level of significance whether the number of carburetors (carb) depends on the type of vehicle.

Required to test,

H_0: The number of carburetors is independent of the vehicle
H_1: The number of carburetors depends on the vehicle.

We first need to find the frequency of carburetors in the data set. This can be done using the `table` command i.e.

```
> carb = table(mtcars$carb)
```

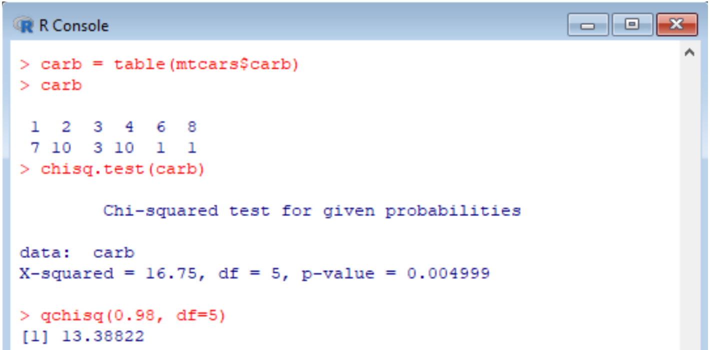
Chi-squared test:

```
> chisq.test(carb)
```

Critical value:

```
> qchisq(0.98, df=5)
```

R Output:



```
> carb = table(mtcars$carb)
> carb

 1  2  3  4  6  8
 7 10  3 10  1  1
> chisq.test(carb)

        Chi-squared test for given probabilities

data:  carb
X-squared = 16.75, df = 5, p-value = 0.004999

> qchisq(0.98, df=5)
[1] 13.38822
```

Since $\chi^2 > \chi^2_{0.02,5}$ i.e. $16.750 > 13.388$ and p-value $< \alpha$ i.e. $0.005 < 0.02$ we reject H_0 at a 2% level of significance.

The number of carburetors depends on the type of vehicle.

Two way analysis

Example 1: A study conducted to determine the preference of exercise times for 65 adults produced the following contingency table:

	Morning	Evening	Night
Male	6	21	9
Female	13	10	6

Using a test for independence with $\alpha = 0.05$, test the hypothesis that there is no relationship between exercise time preference and gender.

Required to test,

H_0: There is no relationship between exercise preference time and gender
H_1: There is a relationship between exercise preference time and gender

or H_0: Exercise preference time is independent of gender
H_1: Exercise preference time depends on gender

We must arrange the data in the format of a contingency table. The data can be entered manually in R or entered in a data file which can then be read/imported into R.

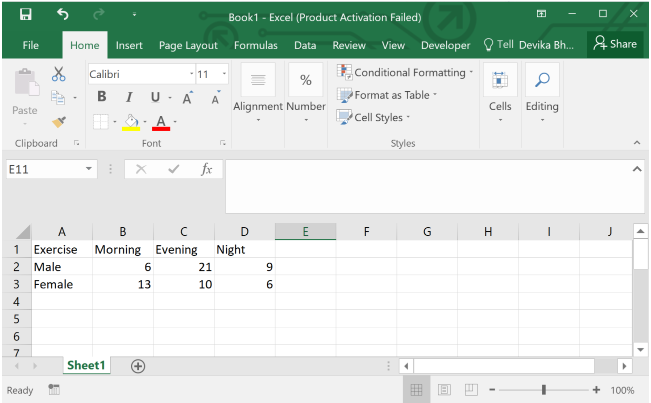
Option 1: Manually

```
> Male = c(6, 21, 9)
> Female = c(13, 10, 6)
> table = rbind(Male,Female)
> colnames(table) = c("Morning","Evening","Night")
> table
       Morning  Evening  Night
Male        6       21      9
Female     13       10      6
```

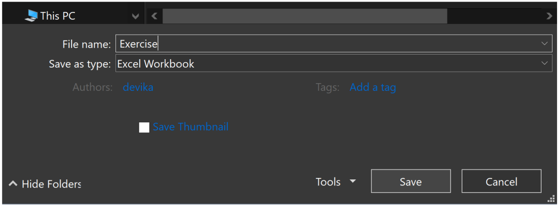
Note that the matrix command can also be used:

```
> table = matrix(c(6,21,9,13,10,6), nrow=2, ncol=3,
          byrow=T)
```

Option 2: Creating a data file - example excel workbook



	A	B	C	D
1	Exercise	Morning	Evening	Night
2	Male	6	21	9
3	Female	13	10	6

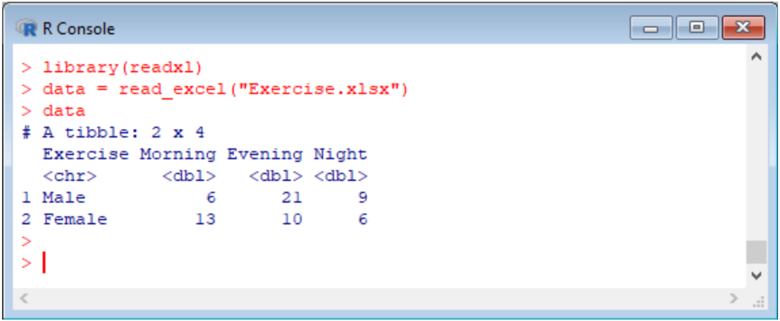


File name: Exercise

Save as type: Excel Workbook

Authors: devika

Tags: Add a tag



```
> library(readxl)
> data = read_excel("Exercise.xlsx")
> data
# A tibble: 2 x 4
  Exercise Morning Evening Night
  <chr>      <dbl>   <dbl> <dbl>
1 Male           6      21     9
2 Female        13      10     6
>
> 
```

Note that a chi-squared test on this data results in an error in R. This is because excel reads the first column as data points.

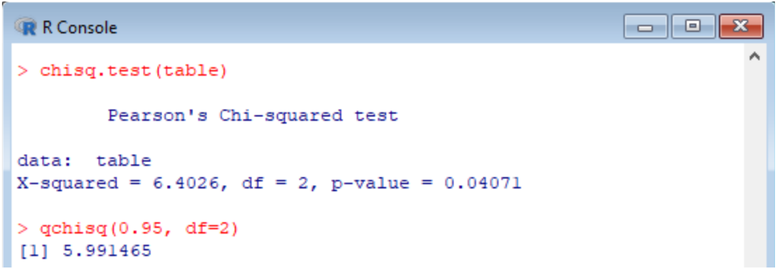
We therefore need to exclude the first column. This can be done by:

```
> table = data[,2:4]
```

Chi-squared test: `> chisq.test(table)`

Critical value: `> qchisq(0.95, df=2)`

R Output:



```
> chisq.test(table)

        Pearson's Chi-squared test

data:  table
X-squared = 6.4026, df = 2, p-value = 0.04071

> qchisq(0.95, df=2)
[1] 5.991465
```

Since $\chi^2 > \chi^2_{0.05,2}$ i.e. $6.403 > 5.991$ and p-value $(= 0.041) < 0.05$ we reject H_0 at a 5% level of significance. There is a significant relationship between exercise preference time and gender.

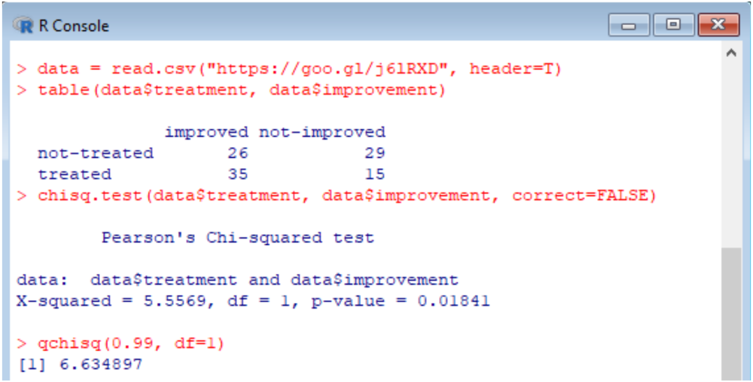
Example 2: Using the data set provided in *https://goo.gl/j6lRXD* (dim(dataset)=150,3) conduct an appropriate test of independence at $\alpha = 0.01$.

Required to test,

H_0: Improvement is independent of treatment
H_1: Improvement is dependent on treatment

```
> data = read.csv("https://goo.gl/j6lRXD",
    header=T)
> table(data$treatment, data$improvement)
> chisq.test(data$treatment, data$improvement,
    correct=FALSE)
> qchisq(0.99, df=1)
```

Notice that here we have to include `correct = T/F` for Yates continuity correction. Excluding this will assume `correct = TRUE`.



```
> data = read.csv("https://goo.gl/j61RXD", header=T)
> table(data$treatment, data$improvement)

               improved not-improved
  not-treated        26           29
  treated            35           15
> chisq.test(data$treatment, data$improvement, correct=FALSE)

        Pearson's Chi-squared test

data:  data$treatment and data$improvement
X-squared = 5.5569, df = 1, p-value = 0.01841

> qchisq(0.99, df=1)
[1] 6.634897
```

Since $\chi^2 < \chi^2_{0.01,1}$ i.e. $5.557 < 6.635$ and p-value $(= 0.018) > 0.01$ we do not reject H_0 at a 1% level of significance. Improvement does not depend treatment.

Correlation Tests

We can also test for a significant relationship between two variables using:

1. Pearson's correlation test (*normal distributions*) or
2. Spearman's correlation test (*non-parametric distributions*)

Pearson's Correlation Test

Pearson's correlation coefficient gives the strength of the relationship between two numeric variables. This test can only be performed if both variables are normally distributed. As such it is known as a parametric correlation test i.e. it depends on the distribution of the data.

Normality Test

We can test whether variables are normally distributed using the Shapiro-Wilk test:

H_0: The data is normally distributed
H_1: The data is not normally distributed

where H_0 is rejected at a given level of significance if p-value $< \alpha$

Shapiro-Wilk test in R:

```
shapiro.test(variable)
```

Once the variables are normally distributed i.e. we fail to reject H_0 at α, the **pearson's correlation test** can be performed:

H_0: $\rho = 0$ vs H_1: $\rho \neq 0$

where ρ is the population correlation coefficient estimated by person's correlation coefficient r.

We reject H_0 at a given level of significance when,

1. confidence interval does not contain 0
2. p-value $< \alpha$

Pearson's correlation test in R:

```
cor.test(variable1, variable2, conf.level=CL,
    method = "pearson")
```

Spearman's Correlation Test

If the variables of interest are not normally distributed, spearman's correlation test can be performed (non-parametric). The spearman correlation coefficient is computed using the rank of both variables.

The hypotheses tested are,

H_0: $\rho = 0$ vs H_1: $\rho \neq 0$

where ρ is the population correlation coefficient and H_0 is rejected when p-value $< \alpha$ for a given level of significance.

Spearman's correlation test in R:

```
cor.test(variable1, variable2, method = "spearman",
    exact=F)
```

The argument `exact = FALSE` is included to remove a warning message which states that the exact p-value cannot be computed with ties; a tie refers to more than one data point having the same value.

The `conf.level` argument is removed since this test does not generate a confidence interval.

Note that the R codes for performing both pearson's and spearman's correlation test assumes a two tailed test i.e. `two.sided` is the default.

Example 1: The *malaria.csv* data set provided shows the antibody level by age of 100 children exposed and unexposed to malaria. The "malaria" variable is a scale vector where 0 indicates no exposure to malaria and 1 indicates exposure.

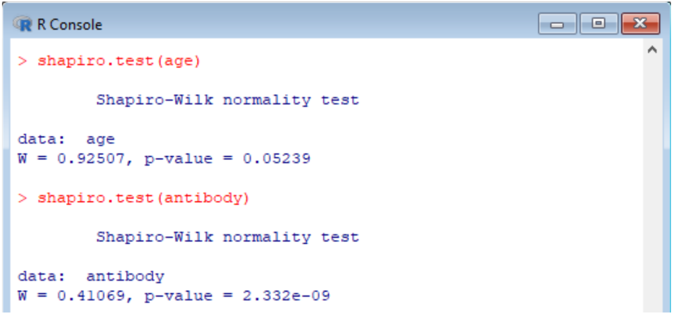
By first sorting and creating an appropriate data set, test at a 5% level of significance whether there is a significant relationship between age and antibody level for children exposed to malaria.

STEP 1: Read the data file into R and create a separate variable for malaria $= 1$ (exposure).

```
> malaria = read.csv("malaria.csv", header=T)
> library(data.table)
> data = as.data.table(malaria)
> exposed = data[which(data$malaria == 1)]
> attach(exposed)
```

STEP 2: Test for normality of variables "age" and "antibody".

```
> shapiro.test(age)
> shapiro.test(antibody)
```



```
> shapiro.test(age)

        Shapiro-Wilk normality test

data:  age
W = 0.92507, p-value = 0.05239

> shapiro.test(antibody)

        Shapiro-Wilk normality test

data:  antibody
W = 0.41069, p-value = 2.332e-09
```

STEP 3: Deduce the appropriate correlation test and perform at $\alpha = 0.05$.

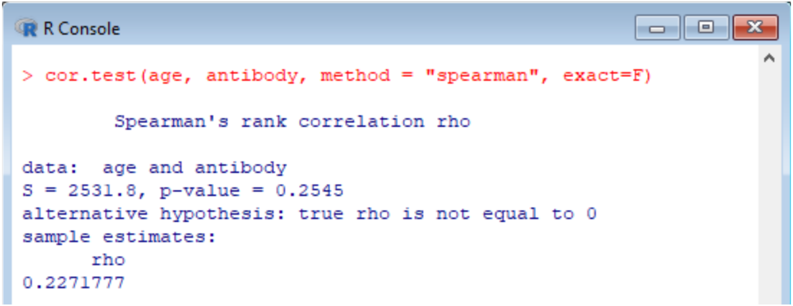
The normality test for age produces a p-value $= 0.052 > 0.05$. We therefore fail to reject H_0 concluding that age is normally distributed.

The normality test for antibody however produces a p-value much smaller than 0.05. Therefore we reject H_0 concluding that antibody is not normally distributed.

Since age is normally distributed but antibody is not, we must use spearman's correlation test.

```
> cor.test(age, antibody, method = "spearman",
    exact=F)
```

R Output : (see next slide)



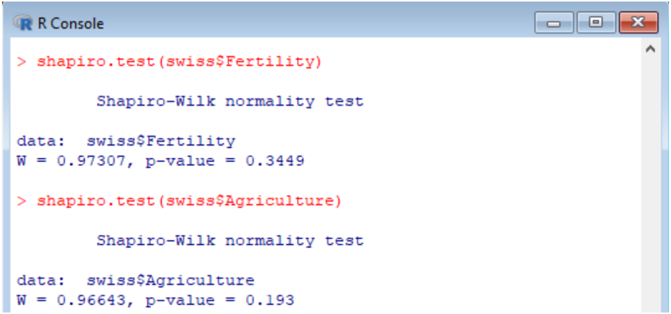
```
> cor.test(age, antibody, method = "spearman", exact=F)

        Spearman's rank correlation rho

data:  age and antibody
S = 2531.8, p-value = 0.2545
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2271777
```

The p-value $= 0.255 > 0.05$. We cannot reject H_0 at a 5% level of significance ($\rho = 0$). There is no significant relationship between the age and antibody level of children exposed to malaria.

Example 2: Using the swiss built-in data in R, conduct a correlation test at $\alpha = 0.10$ for the relationship between the variables "Fertility" and "Agriculture".

STEP 1: Observe the data and test for the normality of the variables

```
> shapiro.test(swiss$Fertility)
> shapiro.test(swiss$Agriculture)
```



```
> shapiro.test(swiss$Fertility)

        Shapiro-Wilk normality test

data:  swiss$Fertility
W = 0.97307, p-value = 0.3449

> shapiro.test(swiss$Agriculture)

        Shapiro-Wilk normality test

data:  swiss$Agriculture
W = 0.96643, p-value = 0.193
```

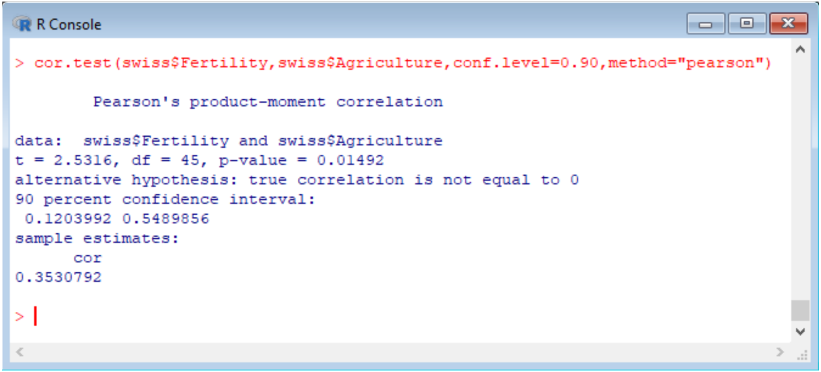
STEP 2: Deduce and perform the appropriate correlation test.

The p-value form both normality tests above are greater than $\alpha = 0.10$. This means that we do not reject H_0 at a 1% level of significance. The variables "Fertility" and "Agriculture" are normally distributed.

Since both variables are distributed normally, the pearson's correlation test must be used.

```
> cor.test(swiss$Fertility, swiss$Agriculture,
    conf.level=0.90, method="pearson")
```

R Output : (see next slide)



```
> cor.test(swiss$Fertility,swiss$Agriculture,conf.level=0.90,method="pearson")

        Pearson's product-moment correlation

data:  swiss$Fertility and swiss$Agriculture
t = 2.5316, df = 45, p-value = 0.01492
alternative hypothesis: true correlation is not equal to 0
90 percent confidence interval:
 0.1203992 0.5489856
sample estimates:
      cor
0.3530792

> 
```

The 90% confidence interval (0.120, 0.549) does not contain 0 and the p-value $(= 0.015) < 0.10$. We can reject H_0 at $\alpha = 0.10$ to conclude that there is a significant relationship between Fertility and Agriculture $(\rho \neq 0)$.