CM 2062 - Statistical Computing with R
Lab Sheet 11

Binomial Test in R

A binomial test compares a sample proportion to a hypothesized proportion. **Recommended when sample size is small.**

The test has the following null and alternative hypotheses:

H_0: P = p (the population proportion P is equal to some value p)

H_1: P $\neq$ p (the population proportion P is not equal to some value p)

The test can also be performed with a one-tailed alternative that the true population proportion is greater than or less than some value p.

Syntax:

```
binom.test(x, n, p=0.5,
           alternative = c("two.sided", "less", "greater"),
           conf.level = 0.95)
```

Where:
x: number of successes
n: number of trials
p: probability of success on a given trial
alternative: indicates the alternative hypothesis and must be one of "two.sided", "greater" or "less". You can specify just the initial letter.
conf.level: confidence level for the returned confidence interval.

Example 1

You want to determine whether or not a die lands on the number "3" during 1/6 of the rolls so you roll the die 24 times and it lands on "3" a total of 9 times. Determine if the die actually lands on "3" during 1/6 of rolls.

```
> binom.test(9, 24, 1/6)
```

```
        Exact binomial test

data:  9 and 24
number of successes = 9, number of trials = 24,
p-value = 0.01176
alternative hypothesis: true probability of success is not equal to 0.1666667
95 percent confidence interval:
 0.1879929 0.5940636
sample estimates:
probability of success
                 0.375
```

The p-value of the test is 0.01176. Since this is less than 0.05, we can reject the null hypothesis and conclude that there is evidence to say the die does not land on the number "3" during 1/6 of the rolls.

Example 2

You want to determine whether or not a coin is less likely to land on heads compared to tails so you flip the coin 30 times and find that it lands on heads just 11 times. Determine if the coin is actually less likely to land on heads compared to tails.

```
> binom.test(11, 30, 0.5, alternative="less")

        Exact binomial test

data:  11 and 30
number of successes = 11, number of trials =
30, p-value = 0.1002
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.5330863
sample estimates:
probability of success
             0.3666667
```

The p-value of the test is 0.1002. Since this is not less than 0.05, we fail to reject the null hypothesis. We do not have sufficient evidence to say that the coin is less likely to land on heads compared to tails.

One Sample Proportion Test (One proportion z-test)

The One proportion Z-test is used to compare an observed proportion to a theoretical one, when there are only two categories. **It can be used when sample size is large (n > 30). It uses a normal approximation to binomial.**

Syntax:

```
prop.test(x, n, p = NULL,
          alternative = c("two.sided", "less", "greater"),
          conf.level = 0.95, correct = TRUE)
```

Example 1

A shop makes widgets with 80% effectiveness. They implement a new system that they hope will improve the rate of effectiveness. They randomly select 50 widgets from a recent production run and find that 46 of them are effective. Determine if the new system leads to higher effectiveness.

```
> # One Sample Proportion Test
> prop.test(46, 50, 0.8, alternative="greater", correct = FALSE)

        1-sample proportions test without continuity
        correction

data:  46 out of 50, null probability 0.8
X-squared = 4.5, df = 1, p-value = 0.01695
alternative hypothesis: true p is greater than 0.8
95 percent confidence interval:
 0.8333021 1.0000000
sample estimates:
   p
0.92
```

Note:
R does not provide the z score, but it can be derived by noting that the z score is the square root of the chi square.

The p-value of the test is 0.01695. Since this is less than 0.05, we reject the null hypothesis. We have sufficient evidence to say that the new system produces effective widgets at a higher rate than 80%.

Example 2

Suppose we want to know whether or not the proportion of residents in a certain county who support a certain law is equal to 60%. To test this, we collect the following data on a random sample: residents who support law is 64 and the sample size is 100.

```
> prop.test(x=64, n=100, p=0.60, correct = FALSE)

        1-sample proportions test without continuity
        correction

data:  64 out of 100, null probability 0.6
X-squared = 0.66667, df = 1, p-value = 0.4142
alternative hypothesis: true p is not equal to 0.6
95 percent confidence interval:
 0.5423540 0.7272878
sample estimates:
   p
0.64
```

Since the p-value (0.475) is greater than 0.05, we fail to reject the null hypothesis. We do not have sufficient evidence to say that the proportion of residents who support the law is different from 0.60.

Exercise 1

Suppose in a coin tossing, the chance to get a head or tail is 50%. In a real case, we have 100 coin tossings, and get 48 heads, is our original hypothesis true?

Exercise 2

Imagine a pigeon was given a 2AFC task (two-alternative forced-choice task) to discriminate between pictures of circular shapes and angular shapes. On each trial the pigeon is shown two pictures, one circular and one angular. They are rewarded when they correctly peck on the circular shape. Pigeon A received 100 trials, and pecked the correct circular shape on 65% of the trials. Test whether the pigeon is rewarded.

Exercise 3

We have a population of mice containing half male and half females. Some of these mice (n = 160) have developed spontaneous cancer, including 95 males and 65 females. We want to know, whether cancer affects more males than females?

Two Samples Proportion Test (Two-Proportions Z-Test)

The Two Proportion Z-test is used to conduct a hypothesis test about the difference between the proportions of two populations.

Syntax:

```
prop.test(x, n, p = NULL,
          alternative = c("two.sided", "less", "greater"),
          conf.level = 0.95, correct = TRUE)
```

Example 1

We have two groups of individuals call Group A and Group B where Group A consists with 500 lung cancer patients and Group B consists with 500 healthy individuals. From the Group A lung cancer patients 490 are smokers and Group B healthy individuals 400 are smokers. Test whether the proportions of smokers are the same in the two groups of individuals.

```
 > # Two Sample Proportion Test
> res <- prop.test(x = c(490, 400), n = c(500, 500), correct = FALSE )
> res

        2-sample test for equality of proportions
        without continuity correction

data:  c(490, 400) out of c(500, 500)
X-squared = 82.737, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1428536 0.2171464
sample estimates:
prop 1 prop 2
  0.98   0.80
```

The p-value of the test is less than the significance level 0.05. We can conclude that the proportion of smokers is significantly different in the two groups.

Example 2

In the built-in data set named "quine" package" in "MASS" package, children from an Australian town is classified by ethnic background, gender, age, learning status and the number of days absent from school. Assuming that the data in quine follows the normal distribution, find the 95% confidence interval estimate of the difference between the female proportion of Aboriginal students and the female proportion of Non-Aboriginal students.

```
> library(MASS)
> head(quine)

  Eth Sex Age Lrn Days
1   A   M  F0  SL    2
2   A   M  F0  SL   11
3   A   M  F0  SL   14
4   A   M  F0  AL    5
5   A   M  F0  AL    5
6   A   M  F0  AL   13
```

Eth indicates whether the student is Aboriginal or Not ("A" or "N"), and the column Sex indicates Male or Female ("M" or "F").

```
> table(quine$Eth, quine$Sex)

    F  M
A  38  31
N  42  35
```

In R, we can tally the student ethnicity against the gender with the **table** function. As the result shows, within the Aboriginal student population, 38 students are female. Whereas within the Non-Aboriginal student population, 42 are female.

```
> prop.test(table(quine$Eth, quine$Sex), correct=FALSE)

        2-sample test for equality of proportions
        without continuity correction

data:  table(quine$Eth, quine$Sex)
X-squared = 0.0040803, df = 1, p-value = 0.9491
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1564218  0.1669620
sample estimates:
   prop 1    prop 2
0.5507246 0.5454545
```

The 95% confidence interval estimate of the difference between the female proportion of Aboriginal students and the female proportion of Non-Aboriginal students is between -15.6% and 16.7%.

Exercise 1

Let's say we have two groups of student A and B. Group A with an early morning class of 400 students with 342 female students. Group B with a late class of 400 students with 290 female students. Use a 5% alpha level. Test whether the observed proportion of Females in group A is greater than the observed proportion of Females in group B.

Chi-Square Test of Independence (Chi-Square Test for Association)

A Chi-Square Test of Independence is used to determine whether or not there is a significant association between two categorical variables.

H_0: The two variables are independent.
H_1: The two variables are not independent.

or

H_0: There is no association between two variables.
H_1: There is an association between two variables.

Example 1
Suppose we want to know whether or not gender is associated with political party preference. We take a simple random sample of 500 voters and survey them on their political party preference. The following table shows the results of the survey.

	Republican	Democrat	Independent	Total
Male	120	90	40	250
Female	110	95	45	250
Total	230	185	85	500

```
> data <- matrix(c(120, 90, 40, 110, 95, 45), ncol=3, byrow=TRUE)
> colnames(data) <- c("Rep","Dem","Ind")
> rownames(data) <- c("Male","Female")
> data

       Rep Dem Ind
Male   120  90  40
Female 110  95  45
```

```
> chisq.test(data)


        Pearson's Chi-squared test

data:  data
X-squared = 0.86404, df = 2, p-value = 0.6492
```

Since the p-value (0.6492) of the test is not less than 0.05, we fail to reject the null hypothesis. This means we do not have sufficient evidence to say that there is an association between gender and political party preference.

Note:
If you have raw data, you can create a contingency table from the raw data using the "table" function and then use chisq.test function.

Example 2
Let's read the "titanic" csv file in to R and check whether there any association between Gender and the Survival.

```
> setwd("F:/KDU_work/CM 2062/My notes/Week 11")
> Titanic <- read.csv("titanic.csv")
> head(Titanic)

> ls(Titanic)
[1] "Age"
[2] "Fare"
[3] "Name"
[4] "Parents.Children.Aboard"
[5] "Pclass"
[6] "Sex"
[7] "Siblings.Spouses.Aboard"
[8] "Survived"

> data <- table(Titanic$Survived, Titanic$Sex)
> data

    female male
  0     81  464
  1    233  109

> chisq.test(data)
```

```
        Pearson's Chi-squared test with Yates'
        continuity correction

data:  data
X-squared = 258.39, df = 1, p-value < 2.2e-16
```

Since p value is less than 0.05, there is a significant evidence to suggest that there is and association between Gender and the Survival.

Exercise 1

Consider the "iris" data set in R.

1. Add a variable call "Size" to the iris data frame using below condition. If the length of the petal is less than the median length of the petal then Size is Small, otherwise Size is Big.

2. Test whether there any association between Size and the Species.