# Types of Hypothesis Test

**T-test**

This is the test you have previously used but you might not know what it does. A t-test looks at the means of data sets and tests for the likelihood there is any real difference between them.

To perform a t-test the data should be

* Numerical data (either continuous or ordinal)
* Data should be normally distributed
  + Check that the mean and median are close
  + A Histogram of each to show a bell shaped distribution
  + A Shapiro-Wilks tests where the resultant p-value is >0.05
* Ideally the data should be randomly selected

There are also different types of T-Test

**Two Sample T-Tests**

This is done on two independent samples of data. Two-sample t-Tests compare the means of precisely two groups, no more and no less. Typically you perform this test to determine whether two **population** means are different.

For example,

1. Do students who learn using Method A have a different mean score than those who learn using Method B?

2. Testing out to see if medication has an effect on patients resting heart rate if you were comparing patients who took the medications against patients who didn’t take the medication.

The standard form tests the following hypotheses;

Null hypothesis: the two population means are equal.

Alternative hypothesis: the two population means are not equal.

If the 𝜌−𝑣𝑎𝑙𝑢𝑒 is less than the *significance* level (e.g. 0.05), you can reject the null hypothesis. The difference between the two means is statistically significant. The sample provides strong enough evidence to conclude that the two population means are different.

A p-value of 0.05 means there is a 5% chance that there is no difference between the true means.

**Example:** For the following use the data set *Two Sample T Test.csv*.

Upload the data into R Studio.

(a) Look at the data and make sure it is numerical

(b) Check that the data is normally distributed, write what you found below.

|  |
| --- |
| > summary(Column\_1)  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  7.229 8.864 9.794 9.560 10.501 11.406 30  > summary(Column\_2)  Min. 1st Qu. Median Mean 3rd Qu. Max.  6.096 8.596 9.844 9.656 10.660 12.191  Mean and median are close together.  hist(Column\_1)  hist(Column\_2)    > shapiro.test(Column\_1)  Shapiro-Wilk normality test  data: Column\_1  W = 0.9407, p-value = 0.2471  > shapiro.test(Column\_2)  Shapiro-Wilk normality test  data: Column\_2  W = 0.97536, p-value = 0.3767  Since both p values are above 0.05 then we assume the data is normally distributed. |

You will now perform a Hypothesis test.

(c) What is the null hypothesis of the test?

|  |
| --- |
| Since there is no context to the question we are just looking at a relationship between the data.  Null hypothesis: the two population means are equal.  Alternative hypothesis: the two population means are not equal. |

(d) Using the command *t.test(x,y)* perform a hypothesis test. State the p-value and the result below.

|  |
| --- |
| > t.test(Column\_1,Column\_2)  Welch Two Sample t-test  data: Column\_1 and Column\_2  t = -0.26852, df = 37.081, p-value = 0.7898  alternative hypothesis: true difference in means is not equal to 0  95 percent confidence interval:  -0.8197071 0.6278566  sample estimates:  mean of x mean of y  9.560168 9.656093 |

**Example 2:** Group Body Fat Percentages  
The sample data is from a group of men and women who worked out at the gym three times a week for a year. Their percentage of body fat is recorded in the table below.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Men | 13.3 | 6.0 | 20.0 | 8.0 | 14.0 | 19.0 | 18.0 | 25.0 | 16.0 | 24 | 15.0 | 1.0 | 15.0 |
| Women | 22 | 16 | 21.7 | 21 | 30 | 26 | 12 | 21.2 | 28.0 | 23 |  |  |  |

Record the data into an excel spreadsheet and save as a csv file

Upload the data into R studio.

(a) Using R studio, show if the data is normally distributed. Report your findings below.

|  |
| --- |
| Men<-c(13.3,6,20,8,14,19,18,25,16,24,15,1,15)  Women<-c(22,16,21.7,21,30,26,12,21.2,28.0,23)  summary(Men)  Min. 1st Qu. Median  1.00 13.30 15.00  Mean 3rd Qu. Max.  14.95 19.00 25.00  summary(Women)  Min. 1st Qu. Median  12.00 21.05 21.85  Mean 3rd Qu. Max.  22.09 25.25 30.00  > hist(Men)  > hist(Women)  > shapiro.test(Men)  Shapiro-Wilk normality  test  data: Men  W = 0.95845, p-value =  0.7296  > shapiro.test(Women)  Shapiro-Wilk normality  test  data: Women  W = 0.95313, p-value =  0.7056 |

(b) State the null and alternative hypothesis referring to them as Ho and H1.

|  |
| --- |
| Ho – There is no difference in the average body fat of men and women  H1 – there is is difference in the average body fat of men and women. |

(c) Perform a Hypothesis test. State the p-value and confidence interval and interpret what that means.

|  |
| --- |
| > t.test(Men,Women)  Welch Two Sample t-test  data: Men and Women  t = -2.817, df = 20.989,  p-value = 0.01033  alternative hypothesis: true difference in means is not equal to 0  95 percent confidence interval:  -12.417816 -1.869876  sample estimates:  mean of x mean of y  14.94615 22.09000  Since p<0.05 this is a statistically significant result. It means that we can reject the null hypothesis and conclude that in all likelihood there is a difference in the average amount of body fat in men and women. |

**Paired Sample T-Test**

A paired t-Test is also called a correlated pairs t-Test or dependent samples t-Test. Dependent samples are essentially connected, they are tests on the same person or thing say, one year apart. For example, tests on the same group of people before and after a fitness regime; or a comparison of two different measurements/treatments are applied to the same subjects.

To apply the paired t-test to test for differences between paired measurements, the following assumptions need to hold;

* Subjects must be independent.
* Each pair must be obtained from the same source (i.e. a person’s weight before and after.
* The data must be normally distributed.

**Example** 1.

Calculate a paired t-Test for the following:

Pupils in a Maths class sit a test and their scores are noted. After 6 weeks of supported study, there is another test and, again, their scores are noted.

|  |  |  |
| --- | --- | --- |
| Student | Before | After |
| 1 | 3 | 20 |
| 2 | 3 | 13 |
| 3 | 3 | 13 |
| 4 | 12 | 20 |
| 5 | 15 | 29 |
| 6 | 16 | 32 |
| 7 | 17 | 23 |
| 8 | 19 | 29 |
| 9 | 23 | 25 |
| 10 | 24 | 15 |
| 11 | 32 | 30 |

(a) State the null and alternative hypothesis

|  |
| --- |
| Ho – There will be no difference in the average test scores from before and after the supported study.  H1 – There will be a difference in the average test score before and after the supported study. |

We will use the paired t-Test to see if the supported study sessions affected the test scores.

Input the data into R studio.

Using the command *t.test(x,y, paired = TRUE)* perform a hypothesis test.

State what the p-value and interpret what this means in the context of the data.

|  |
| --- |
| > t.test(Before,After, paired=TRUE)  Paired t-test  data: Before and After  t = -2.5567, df = 10,  p-value = 0.02853  alternative hypothesis: true difference in means is not equal to 0  95 percent confidence interval:  -10.8886115 -0.7477522  sample estimates:  mean of the differences  -5.818182  Since p <0.05 this is a statistically significant result. It means we can reject the null hypothesis and conclude that in all likelihood the supported study has had an impact on the test scores of the pupils. |

What does this mean for our test? It means our sample data shows a higher mean test score after help than before help (supported study). The significant result suggests that we can conclude that the same is true at the population level. We can infer that the population mean of the test score after help is greater than the population mean of the test score before help.

**Example 2.** For the following use the *Crime* preinstalled data set from R studio.

Look at the variables *CrimeRate* and *CrimeRate10*.

These show the amount of offences for every million people.

(a) Test these for normal distribution.

|  |
| --- |
| > summary(CrimeRate)  Min. 1st Qu. Median Mean 3rd Qu. Max.  45.5 82.7 103.0 102.8 120.7 161.8  > summary(CrimeRate10)  Min. 1st Qu. Median Mean 3rd Qu. Max.  26.50 76.35 103.50 102.07 130.25 178.20  > hist(Crime)  Error in hist(Crime) : object 'Crime' not found  > hist(CrimeRate)  > hist(CrimeRate10)  > shapiro.test(CrimeRate)  Shapiro-Wilk normality test  data: CrimeRate  W = 0.98632, p-value = 0.8508  > shapiro.test(CrimeRate10)  Shapiro-Wilk normality test  data: CrimeRate10  W = 0.98194, p-value = 0.6736 |

(b) State the null and alternative hypothesis for performing a statistical test.

|  |
| --- |
| Ho – There is no difference in the amount of crime 10 years apart.  H1 – There is a difference in the amount of crime 10 years apart. |

(c) Perform a t-test. State the p-value and confidence interval and interpret what this means.

|  |
| --- |
| > t.test(CrimeRate,CrimeRate10, paired=TRUE)  Paired t-test  data: CrimeRate and CrimeRate10  t = 0.47082, df = 46, p-value = 0.64  alternative hypothesis: true difference in means is not equal to 0  95 percent confidence interval:  -2.418114 3.894710  sample estimates:  mean of the differences  0.7382979 |

**Mann-Whitney/Wilcox Test**

An alternative test for paired samples that can be considered is called the Wilcoxon (signed-rank) test. This test can be useful as it **does not** require the data to be normally distributed or for the standard deviations (variances) to be similar.

If the data is normally distributed and the standard deviations close together then it is recommended that a t-test is used.

**Example:** A comparison of two methods A and B for measuring physical fitness, a random sample of eight people was assessed by both method. Their scores were recoded as follows.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Subject | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Method A | 11.2 | 8.6 | 6.5 | 17.3 | 14.3 | 10.7 | 67.8 | 13.3 |
| Method B | 10.4 | 12.1 | 9.1 | 15.6 | 16.7 | 10.7 | 12.8 | 34.5 |

(a) Is the data suitable for a T test? State your reason below.

|  |
| --- |
| > summary(MethodA)  Min. 1st Qu. Median Mean 3rd Qu. Max.  6.50 10.18 12.25 18.71 15.05 67.80  > summary(MethodB)  Min. 1st Qu. Median Mean 3rd Qu. Max.  9.10 10.62 12.45 15.24 15.88 34.50  > hist(MethodA)  > hist(MethodB)  > shapiro.test(MethodA)  Shapiro-Wilk normality test  data: MethodA  W = 0.58171, p-value = 9.702e-05  > shapiro.test(MethodB)  Shapiro-Wilk normality test  data: MethodB  W = 0.70819, p-value = 0.002787 |

(b) State the null and alternative hypothesis below

|  |
| --- |
| Ho – There is no difference between the average score of either Method  H1 – There is a difference in the average score for both the methods |

(c) Perform a hypothesis test to show if there is any significant difference between the two methods of measuring fitness.

|  |
| --- |
| > wilcox.test(MethodA,MethodB, paired=TRUE)  Wilcoxon signed rank test with continuity  correction  data: MethodA and MethodB  V = 10, p-value = 0.5541  alternative hypothesis: true location shift is not equal to 0 |

**Exercise: T-testing**

1. A tennis coach wants to determine whether a new racquet improves the speed of his pupil’s serves (faster serves are considered better). He tests a group of 9 children to discover their service speed with their current racquet and with the new racquet. The results are shown in the table below.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Child | A | B | C | D | E | F | G | H | I |
| Speed with current racquet | 58 | 68 | 49 | 71 | 80 | 57 | 46 | 57 | 66 |
| Speed with new racquet | 72 | 81 | 52 | 59 | 75 | 72 | 48 | 62 | 70 |

(a) State the null and alternative hypothesis.

Ho – There is no difference in the average speed for the different racquets

H1 – There is a difference in the average speed for the different racquets

(b) Perform a statistical analysis to see if buying a new racquet is guaranteed to increase your service speed. Make sure to confirm which type of test you will use and state the reasons for this.

> summary(Racquet1)

Min. 1st Qu. Median

46.00 57.00 58.00

Mean 3rd Qu. Max.

61.33 68.00 80.00

> summary(Racquet2)

Min. 1st Qu. Median

48.00 59.00 70.00

Mean 3rd Qu. Max.

65.67 72.00 81.00

> hist(Racquet1)

> hist(Racquet2)

> shapiro.test(Racquet1)

Shapiro-Wilk normality

test

data: Racquet1

W = 0.96516, p-value =

0.8505

> shapiro.test(Racquet2)

Shapiro-Wilk normality

test

data: Racquet2

W = 0.94568, p-value =

0.6427

Since the data is normally distributed and it is paired then a paired t test is going to be used

> t.test(Racquet1,Racquet2, paired=TRUE)

Paired t-test

data: Racquet1 and Racquet2

t = -1.4489, df = 8,

p-value = 0.1854

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-11.229961 2.563294

sample estimates:

mean of the differences

-4.333333

Since p>0.05 then we cannot reject the null hypothesis meaning we don’t know if changing to a new racquet will actually improve someone’s ball speed.

2. Petra noticed that one of her apple trees grew in the shade and other did not. She wanted to find out if apples from the tree in the shade weighed less than those in the sun. She picked 9 random apples from each tree and weighted them.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Tree in Shade | 75 | 82 | 93 | 77 | 85 | 78 | 91 | 83 | 92 |
| Tree Not in Shade | 74 | 81 | 95 | 79 | 95 | 82 | 93 | 88 | 90 |

(a) Is the data paired?

No

(b) Show if the data is normally distributed and if it is suitable for a t test.

> summary(Shade)

Min. 1st Qu. Median Mean 3rd Qu. Max.

75 78 83 84 91 93

> summary(NoShade)

Min. 1st Qu. Median Mean 3rd Qu. Max.

74.00 81.00 88.00 86.33 93.00 95.00

> hist(Shade)

> hist(NoShade)

Histograms are clearly not normally distributed.

(c) State the null and alternative hypothesis

Ho -There is no difference in the weight of apples from trees in the shade against those that were not in the shade.

H1 – There is a difference in the weight of apples from trees in the shade against those that are not in the shade.

(d) Perform a statistical test to show if there is any difference in the weight from trees that grow in shade and those that don’t.

> wilcox.test(Shade,NoShade)

Wilcoxon rank sum test with continuity

correction

data: Shade and NoShade

W = 33, p-value = 0.5359

alternative hypothesis: true location shift is not equal to 0

Since p >0.05 this is a statistically significant result. This means

3. In an attempt to determine if two competing brands of cold medicine contain, on the average, the same amount of acetaminophen, twelve different tablets from each of the two competing brands were randomly selected and tested for the amount of acetaminophen each contains. The results (in milligrams) follow. Use a significance level of 0.01.

Brand A Brand B

517, 495, 503, 491 493, 508, 513, 521

503, 493, 505, 495 541, 533, 500, 515

498, 481, 499, 494 536, 498, 515, 515

State and perform an appropriate hypothesis test.

> BrandA<-c(517,495,503,491,503,493,505,495,498,481,499,494)

> BrandB<-c(493,508,513,521,541,533,500,515,536,498,515,515)

> summary(BrandA)

Min. 1st Qu. Median Mean 3rd Qu. Max.

481.0 493.8 496.5 497.8 503.0 517.0

> summary(BrandB)

Min. 1st Qu. Median Mean 3rd Qu. Max.

493.0 506.0 515.0 515.7 524.0 541.0

> hist(BrandA)

> hist(BrandB)

> shapiro.test(BrandA)

Shapiro-Wilk normality test

data: BrandA

W = 0.95368, p-value = 0.6912

> shapiro.test(BrandB)

Shapiro-Wilk normality test

data: BrandB

W = 0.94807, p-value = 0.609

> t.test(BrandA,BrandB)

Welch Two Sample t-test

data: BrandA and BrandB

t = -3.524, df = 17.705, p-value =

0.002474

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-28.477760 -7.188906

sample estimates:

mean of x mean of y

497.8333 515.6667

Since p<0.05 we cannot reject the null hypothesis and therefore cannot say if there is a difference in the amount of acetaminophen in the two brands of tablet.

4. In an investigation to compare the accuracy of Crackshot and Fastfire12-bore shotguns in clay pigeon shooting, ten competitors each fired 100 shots with each make of gun. Their scores are shown in the table below.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Competitor | A | B | C | D | E | F | G | H | I | J |
| Crackshot | 93 | 99 | 90 | 86 | 85 | 94 | 87 | 91 | 96 | 79 |
| Fastfire | 87 | 91 | 86 | 87 | 78 | 95 | 89 | 84 | 88 | 74 |

Assume the data is normally distributed

(a) State the null and alternative hypothesis

Ho - There is no difference in the accuracy of the shotguns

H1 – There is a difference in the accuracy of the shotguns

(b) Examine the claim that the Crackshot shotgun is more accurate for clay pigeon shooting.

> t.test(Crackshot,Fastfire, paired=TRUE)

Paired t-test

data: Crackshot and Fastfire

t = 3.2768, df = 9, p-value = 0.009578

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

1.26954 6.93046

sample estimates:

mean of the differences

4.1

Since p <0.05 we conclude that the claim is inaccurate as there is so statistical difference between the average accuracy of the two shotguns.

5. Use a Mann-Whitney U test to test if heart rate differs between men and women at the 95% level.

|  |  |
| --- | --- |
| Heart Rate Women | Heart Rate Men |
| 84 | 80 |
| 81 | 74 |
| 80 | 73 |
| 70 | 72 |
| 72 | 78 |
| 69 | 75 |
| 65 | 70 |
| 74 | 74 |
| 80 | 69 |

(a) State the null and alternative hypothesis

H0 – There will be no difference in the heart rate of men and women

H1 – There will be a difference in the heart rate of mean and women

(b) Compare using a Wilcox text to determine if there is a significant difference.

> HRWomen<-c(84,81,80,70,72,69,65,74,80)

> HRMen<-c(80,74,73,72,78,75,70,74,69)

> wilcox.test(HRWomen,HRMen)

Wilcoxon rank sum test with continuity

correction

data: HRWomen and HRMen

W = 44.5, p-value = 0.7559

alternative hypothesis: true location shift is not equal to 0

Since p>0.05 this means we cannot reject the null hypothesis and conclude there is no statistical difference between the heart rate of men and women.

**Chi-Square Test for Independence**

This test is used to determine if there is a relationship between two categorical variables. The chi-square test evaluates is there is a significant association between the categories of the two variables e.g. between gender and whether they smoke or not.

For the Chi-square test of independence we need two variables.

Our idea is that the variables are not related (null hypothesis);

Null hypothesis: the variables are not related

*Alternative hypothesis: the variables are related*

For a valid test, we need:

* Data values that are a simple random sample from the population of interest.
* Two categorical or nominal variables. (Nominal identifies the category, e.g. 1 for male, 2 for female).

Look at the following.

**Example:**

In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among community college students, four year college students, and nonstudents. In the table below a sample of volunteers and the number of hours they work per week is shown below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Type of Volunteer | 1-3 Hours | 4-6 Hours | 7-9 Hours | Total |
| Community College Students | 111 | 96 | 48 | 225 |
| Four Year College Students | 96 | 133 | 61 | 290 |
| Non Students | 91 | 150 | 53 | 294 |
| Total | 298 | 379 | 162 | 839 |

(a) For a Chi Square test there needs to be two categorical variables, state the categorical variables below

|  |
| --- |
| Type of Volunteer (community college, 4 year college, non students)  Hours they worked in intervals (1-3 Hours, 4-6 hours, 7-9 hours) |

(b) Input the data into R Studio

*Extension: Do so directly and add Row and Column titles.*

*R version 4.1.1 (2021-08-10) -- "Kick Things"*

*Copyright (C) 2021 The R Foundation for Statistical Computing*

*Platform: x86\_64-w64-mingw32/x64 (64-bit)*

*R is free software and comes with ABSOLUTELY NO WARRANTY.*

*You are welcome to redistribute it under certain conditions.*

*Type 'license()' or 'licence()' for distribution details.*

*R is a collaborative project with many contributors.*

*Type 'contributors()' for more information and*

*'citation()' on how to cite R or R packages in publications.*

*Type 'demo()' for some demos, 'help()' for on-line help, or*

*'help.start()' for an HTML browser interface to help.*

*Type 'q()' to quit R.*

*> volunteers<-matrix(c(111,96,48,96,133,61,91,150,53))*

*> volunteers*

*[,1]*

*[1,] 111*

*[2,] 96*

*[3,] 48*

*[4,] 96*

*[5,] 133*

*[6,] 61*

*[7,] 91*

*[8,] 150*

*[9,] 53*

*> volunteers<-matrix(c(111,96,48,96,133,61,91,150,53), nrow=3)*

*> volunteers*

*[,1] [,2] [,3]*

*[1,] 111 96 91*

*[2,] 96 133 150*

*[3,] 48 61 53*

*> volunteers<-matrix(c(111,96,48,96,133,61,91,150,53), nrow=3, byrow=TRUE)*

*> volunteers*

*[,1] [,2] [,3]*

*[1,] 111 96 48*

*[2,] 96 133 61*

*[3,] 91 150 53*

*> rownames(volunteers)<-c("Community College Students","Four Year College","Non Students")*

*> volunteers*

*[,1] [,2] [,3]*

*Community College Students 111 96 48*

*Four Year College 96 133 61*

*Non Students 91 150 53*

*> colnames(volunteers)<-c("1-3 Hours", "4-6 Hours", "7-9 Hours")*

*> volunteers*

*1-3 Hours 4-6 Hours 7-9 Hours*

*Community College Students 111 96 48*

*Four Year College 96 133 61*

*Non Students 91 150 53*

(c) State the Null and Alternative Hypothesis in the box below.

|  |
| --- |
| Ho – there is no relationship between the number of hours a volunteer works and what kind of volunteer they are.  H1 – there is a relationship between the number of hours a volunteer works and what kind of volunteer they are. |

(d) Perform a Chi-Square test for independence to see if there is a relationship between the variables. State the p-value in the box below and interpret the result.

|  |
| --- |
| *> chisq.test(volunteers)*  *Pearson's Chi-squared test*  *data: volunteers*  *X-squared = 12.991, df = 4, p-value = 0.01132* |

**Example 2:** A cinema looks into if there is a relationship between if customers buy snacks and what kind of movie they attend.

(a) State the Null and Alternative Hypothesis

|  |
| --- |
| Ho – there is no relationship between what kind of movie a customer goes to see and the number of snacks they buy.  H1 – There is a relationship between what kind of movie a customer goes to see and the number of snacks they buy. |

The table is written below

|  |  |  |
| --- | --- | --- |
| Type of Movie | Snacks | No Snacks |
| Action | 50 | 75 |
| Comedy | 125 | 175 |
| Family | 90 | 30 |
| Horror | 45 | 10 |

(b) Input the data into R Studio

> MovieSnacks<-matrix(c(50,75,125,175,90,30,45,10))

> MovieSnacks

[,1]

[1,] 50

[2,] 75

[3,] 125

[4,] 175

[5,] 90

[6,] 30

[7,] 45

[8,] 10

> MovieSnacks<-matrix(c(50,75,125,175,90,30,45,10),byrow =T, nrow = 4)

> MovieSnacks

[,1] [,2]

[1,] 50 75

[2,] 125 175

[3,] 90 30

[4,] 45 10

> rownames(MovieSnacks)<-c("Action","Comedy","Family","Horror")

> MovieSnacks

[,1] [,2]

Action 50 75

Comedy 125 175

Family 90 30

Horror 45 10

> colnames(MovieSnacks)<-c("Snacks bought","No Snacks bought")

> MovieSnacks

Snacks bought No Snacks bought

Action 50 75

Comedy 125 175

Family 90 30

Horror 45 10

(c) State the two categorical variables

|  |
| --- |
| Genre of film  If they bought snacks or not |

(d) Perform a Chi Square Test and interpret the result below, remember to state the p-value and if it is a significant result.

|  |
| --- |
| > chisq.test(MovieSnacks)  Pearson's Chi-squared test  data: MovieSnacks  X-squared = 65.012, df = 3, p-value = 4.987e-14  This is a statistically significant result since p < 0.05. This means we can reject the null hypothesis and conclude in all likelihood that people who go to certain types of movies do buy more snacks. |

**Chi Square Test for Goodness of Fit**

Another use of the chi-squared test is for goodness-of-fit. This test is used to find out how the observed value of a given categorical outcome differs from the expected value. The term ‘goodness-of-fit’ is used to compare the observed sample distribution with the expected probability distribution.

For the goodness-of-fit test, we need one variable. We also need an idea (hypothesis) about how that variable is distributed. Here are a couple of examples.

**Example:** We have bags of sweets with five flavours in each bag. The bag should contain an equal number of pieces of each flavour. The idea we’d like to test is that the proportions of five flavours in each bag are the same.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Red | Yellow | Green | Blue | Orange |
| Observed | 7 | 11 | 16 | 7 | 9 |
| Expected | 10 | 10 | 10 | 10 | 10 |

A Chi Square test will look at these numbers and assuming that they should all appear the same what is the probability that this could happen if everything is fair i.e. there is an even chance of getting each sweet.

Null Hypothesis: There is no difference between the amounts of each sweet

Alternate Hypothesis: There is a difference between the amounts of each sweet.

We will look at a similar example later on.

**Example 1:** 200 rolls of a die result in the following distribution

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 |
| 28 | 30 | 22 | 31 | 38 | 51 |

(a) How many of each should you have expected?

|  |
| --- |
| 200 / 6 = 33 each |

(b) State the null and alternative hypothesis

|  |
| --- |
| Ho – there is no statistical difference in the results of what should occur against what has occurred, i.e. the dice is fair.  H1 – there is a statistical difference between what should occur and what has occurred, i.e the dice is unfair (weighted to one side) |

(c) Perform a Chi Square Test to see if we can conclude if the die is fair.

|  |
| --- |
| > dice<-c(28,30,22,31,38,51)  > chisq.test(dice)  Chi-squared test for given probabilities  data: dice  X-squared = 15.22, df = 5, p-value = 0.009463  Since p<0.05 we can reject the null hypothesis that the dice is fair and conclude that in all likelihood the dice is unfair. |

**Example 2:** A college reports that 50% of the students in its statistics classes are freshmen, 30% are sophomores, 10% are juniors, and 10% are seniors. A simple random sample of 65 such students has the following breakdown.

|  |  |  |  |
| --- | --- | --- | --- |
| Freshman | Sophomore | Junior | Senior |
| 28 | 24 | 9 | 4 |

Using a statistical test show whether there is any reason to doubt the percentages recorded by the college.

|  |
| --- |
| > students<-c(28,24,9,4)  > props<-c(0.5,0.3,0.1,0.1)  > chisq.test(students,props)  Chi-squared test for given probabilities  data: students  X-squared = 3.5846, df = 3, p-value = 0.31 |

**Exercise: Chi Square Testing**

1. The members of a sports team are interested in whether the weather has an effect on results. They play 50 matches, with the following results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **Weather** | | Total |
| Good | Bad |
| **Result** | Win | 12 | 4 | 16 |
| Draw | 5 | 8 | 13 |
| Lose | 7 | 14 | 21 |
| Total | | 24 | 26 | 50 |

Formulate suitable null and alternative hypotheses. Then perform a Hypothesis test to show if the weather has any effect on the result of the games.

Ho – There is no statistical difference between the results in good or bad weather.

H1 – There is a statistical difference between the results in good or bad weather.

> Table2<-matrix(c(12,5,7,4,8,14), nrow=3)

> Table2

[,1] [,2]

[1,] 12 4

[2,] 5 8

[3,] 7 14

> chisq.test(Table2)

Pearson's Chi-squared test

data: Table2

X-squared = 6.9568, df = 2, p-value = 0.03086

Since p <0.05 we reject the null hypothesis and conclude that there is a statistical difference between good and bad weather.

2. A poker-dealing machine is supposed to deal cards at random, as if from an infinite deck.

In a test, you counted 1600 cards, and observed the following:

Spades - 404

Hearts- 420

Diamonds -400

Clubs - 376

Could it be that the suits are equally likely? Or are these discrepancies too much to be random?

> cards<-c(404,420,400,376)

> chisq.test(cards)

Chi-squared test for given probabilities

data: cards

X-squared = 2.48, df = 3, p-value = 0.4789

Since p>0.05 we cannot reject the null hypothesis

3. A staff member of an emergency medical service wishes to determine whether the number of accidents is equally distributed during the week. A week is selected at random, and the following data were obtained.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Day | Mon | Tues | Wed | Thurs | Fri | Sat | Sun |
| No. of Accidents | 28 | 32 | 15 | 14 | 38 | 43 | 19 |

Is there evidence to reject the hypothesis that the number of accidents is equally distributed throughout the week? State any null or alternative hypothesis you may have used and interpret the p-value.

Ho – There is no statistical difference between the days of the week and how many accidents occur (a day of the week does ot have more accidents than any other)

H1 -There is a statistical difference between a day of the week and how many accidents occur. (Some days are more likely to have accidents than others)

> NoOfAccidents<-c(28,32,15,14,38,43,19)

> chisq.test(NoOfAccidents)

Chi-squared test for given probabilities

data: NoOfAccidents

X-squared = 28.889, df = 6, p-value = 6.385e-05

Since p<0.05 this means we can reject the null hypothesis and coclude that there is a statistical difference between days of the week and how many accidents there are. This means that there are certain days of the week when accidents are more likely to occue.

4. The chair of the history department of a college hypothesizes that the final grades are distributed as 40% A’s, 30% B’s, 20% C’s, 5% D’s, 5% F’s. At the end of the semester the following grades were earning.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Grade | A | B | C | D | F |
| Number | 88 | 70 | 44 | 12 | 10 |

Perform a Hypothesis test to show is if the History is perform as expected or do some changes need to be made?

> chisq.test(Grades,props4)

Pearson's Chi-squared test

data: Grades and props4

X-squared = 15, df = 12, p-value = 0.2414

Warning message:

In chisq.test(Grades, props4) : Chi-squared approximation may be incorrect

Since p >0.05 we cannot reject the null hypothesis and conclude that History might not have performed as has been hypothesised.

5. University students may be interested in whether or not their degree have any effect on starting salaries after graduation. Suppose that 300 recent graduates were surveyed as to their majors in college and their starting salaries after graduation. Below are the data. Conduct a hypothesis test.

| **Degree** | **< $50,000** | **$50,000 - $68,999** | **$69,000 +** |
| --- | --- | --- | --- |
| English | 5 | 20 | 5 |
| Engineering | 10 | 30 | 60 |
| Nursing | 10 | 15 | 15 |
| Business | 10 | 20 | 30 |
| Psychology | 20 | 30 | 20 |

> Salaries

Less than 50,000 Between 50,000 and 68,999

English 5 20

Engineering 10 30

Nursing 10 15

Business 10 20

Psycology 20 30

Greater than 69000

English 5

Engineering 60

Nursing 15

Business 30

Psycology 20

> chisq.test(Salaries)

Pearson's Chi-squared test

data: Salaries

X-squared = 33.546, df = 8, p-value = 4.91e-05

Since p<0.05 we reject the null hypothesis and conclude that there is a relationship between what degree people get and how high their starting salary is.

**Two Proportion Z-Test**

A **two proportion z-test** is used to test for a difference between two population proportions.

Suppose we want to know if there is a difference in the proportion of residents who support a certain law in county A compared to the proportion who support the law in county B.

Since there are thousands of residents in each county, it would take too long and be too costly to go around and survey every individual resident in each county.

Instead, we might take a simple random sample of residents from each county and use the proportion in favour of the law in each sample to estimate the true difference in proportions between the two counties:

**Example 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. We want to know, whether the proportions of females are the same in the two groups of the student? Here let’s use **prop.test()**.

|  |
| --- |
| ***Syntax:*** *prop.test(x, n, p = NULL, alternative = “two.sided”, correct = TRUE)*  ***Parameters:******x*** *= number of successes and failures in data set.* ***n*** *= size of data set.* ***p*** *= probabilities of success. It must be in the range of 0 to 1.* ***alternative*** *= a character string specifying the alternative hypothesis.* ***correct*** *= a logical indicating whether Yates’ continuity correction should be applied where possible.*  *> prop.test(x = c(342,290), n = c(400,400))*  *2-sample test for equality of proportions with*  *continuity correction*  *data: c(342, 290) out of c(400, 400)*  *X-squared = 19.598, df = 1, p-value = 9.559e-06*  *alternative hypothesis: two.sided*  *95 percent confidence interval:*  *0.07177443 0.18822557*  *sample estimates:*  *prop 1 prop 2*  *0.855 0.725*  *Since p <0.05 we reject the null hypothesis and conclude that there is a difference between the two classes. The difference being that 85.5% of people in the first class were women and 72.5% of people in the second class were women.* |

**Example 2:** It was conjectured that more women than men live to one hundred years old.

1. State a null and alternative hypothesis

|  |
| --- |
| Ho – There is no statistical difference in the amount of men or women reaching the ages of 100.  H1 – There is a statistical difference in the amount of men and women reaching the ages of 100. |

(b) An audit found 23 out of 9,500 women and 11 out of 10,000 men were centenarians.

In the *sample*, did more women live to 100?

|  |
| --- |
| > prop.test(x = c(23,11), n = c(9500, 10000))  2-sample test for equality of  proportions with continuity  correction  data: c(23, 11) out of c(9500, 10000)  X-squared = 4.1553, df = 1, p-value =  0.04151  alternative hypothesis: two.sided  95 percent confidence interval:  3.574941e-05 2.606356e-03  sample estimates:  prop 1 prop 2  0.002421053 0.001100000  Since p<0.05 we reject the null hypothesis and conclude there is a difference between the amount of men and women that live past 100. Looking at the results 0.24% of women compared to 0.11% of men. |

**Sampling error**

Statistics is the art of using a sample to try and understand a population. It should therefore be understood that any conclusions about the population drawn from the sample are necessarily approximate, and could be different if a different sample was chosen.

There are two types of error:

**Sampling error** occurs if the sample is not typical of the population, and accidentally introduces some bias (a problem with the design of the sample)

**Non-sampling error** occurs if data is incorrectly recorded, or there are gaps in the data as it was not possible to sample fully (a problem with the execution of the sample)

A common cause of sampling error is if the sample is performed all at one time.

Sampling error is minimised by taking a large sample, although this is time consuming so instead there is an attempt to try and describe the uncertainty, for example with confidence intervals.

**Question:** A survey at a local Post Office last Tuesday found that 85% of respondents agreed that

*“Young people have much less respect for their elders than they used to”.*

This was then used to claim that in the whole country people don’t respect their elders.

(a) State two possible causes of sampling error

The sample was not taken from all the population.

(b) State two possible causes of non-sampling error

Data being input incorrectly

*(c) The sample is very limited. Because the sample was made at just one location it does not reflect the whole population well. The sample was also just made at just one time.*

*(d) Respondents may not have filled in the survey honestly. The figure of 85% may have been reached by a counting error. Some surveys may have got lost.*