Data Science 1
Analysis of Categorical Data
Part 1

Ann Maharaj

Analysis of Categorical Data: Part 1

Chi-Square Test for Proportion: One Population

- When we previously performed tests for the proportion of one population and for the equality of proportions of two populations, we used the normal approximation of the binomial distribution.

- The normal distribution can be used to approximate the probabilities for certain types of binomial distribution problems.

- As the sample size gets large, a binomial distribution approaches the normal distribution in shape regardless of the value of the population proportion π.

- This phenomenon occurs much faster (as n increases) when π is near 0.5.

- If the sample size is small, the normal approximation of the binomial distribution will not work and we have to use an alternative method for testing for the population proportion.

Chi-Square Test for Proportion: One Population

- In obtaining the test statistic for testing for proportion of a single population, we make use of the fact that the mean and variance of a variable X that follows a binomial distribution is $E(X) = n\pi$ and $Var(X) = n\pi(1 - \pi)$, respectively.

- Hence given a sample size n and sample proportion $p = \frac{X}{n}$, and the requirement that $n\pi > 5$ and $n(1 - \pi) > 5$, the sample proportion follows a normal distribution.

$$p \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right).$$

- Hence, the standardised sample proportion follows a standard normal distribution.

$$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0, 1).$$

Chi-Square Test for Proportion: One Population

- We used Z as the test statistic for testing

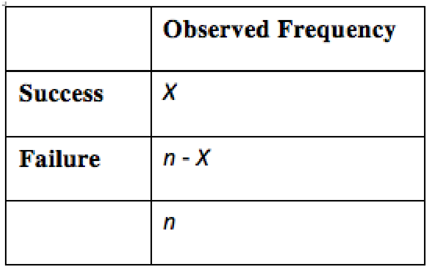
$$\begin{aligned} H_0 &: \pi = \pi_0 & H_1 &: \pi \neq \pi_0 \\ H_0 &: \pi = \pi_0 & H_1 &: \pi > \pi_0 \\ H_0 &: \pi = \pi_0 & H_1 &: \pi < \pi_0 \end{aligned}$$

- Provided $n\pi > 5$ and $n(1-\pi) > 5$, we can also perform these hypotheses using a **chi-square test** as an alternative to the Z test.

- **Assumptions**
    - A random sample is selected for the population of interest.
    - Each observation can be categorised as one of two mutually exclusive types.

Chi-Square Test for Proportion: One Population

- In general, the data is set up as follows:

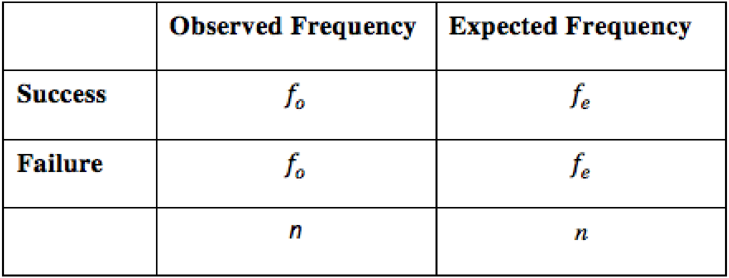


	Observed Frequency
Success	X
Failure	n - X
	n

- X is the number associated with the category of interest, referred to as the number of successes and n is the sample size.

- In order to use a chi-square test, we have to obtain frequencies that are expected under the null hypothesis.

Chi-Square Test for Proportion: One Population



	Observed Frequency	**Expected Frequency**
Success	f_o	f_e
Failure	f_o	f_e
	n	n

- f_o is the observed frequency
- f_e is the expected frequency, where for
- Success: $f_e = n\pi$, Failure: $f_e = n(1-\pi)$,

Chi-Square Test for Proportion: One Population

Test Statistic

- The chi-square test statistic compares the observed and expected frequencies and it is

$$\chi^2 = \sum_{all\ \ cells} \frac{(f_o - f_e)^2}{f_e}$$

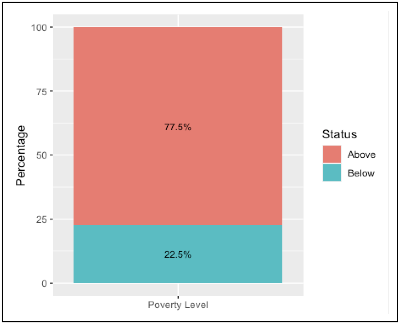
- The test statistic follows a chi-square distribution with 1 degree of freedom.
- Under the null hypothesis, the closer the observed frequencies are to the expected frequencies, the less likely we are to reject the null hypothesis.
- We reject the null hypothesis if the p-value of the test is less than a given level of significance.

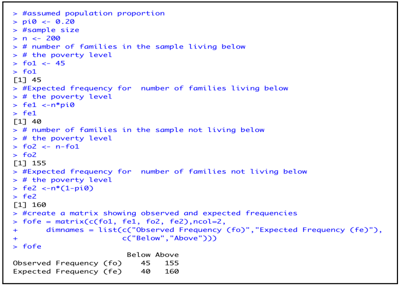
Example: Chi-Square Test for Proportion - One Population

- A census conducted 5 years ago in a particular country, recorded that 20% of families in a community located in a large outer western suburb of a major city, lived below the poverty level.
- To determine if this percentage has changed, a random sample of 200 families in this community is studied, and 45 are found to be living below the poverty level.
- Does this finding indicate if there is sufficient evidence at the 5% level of significance that the current percentage differs from the percentage of families who lived below the poverty level 5 years ago.

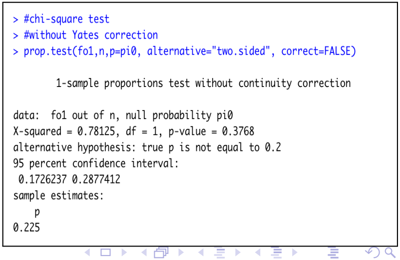
$\pi = 0.20$, $X = 45$, $n = 200$, $\alpha = 0.05$

Example: Chi-Square Test for Proportion - One Population





```
> #assumed population proportion
> pi0 <- 0.20
> #sample size
> n <- 200
> # number of families in the sample living below
> # the poverty level
> fo1 <- 45
> fo1
[1] 45
> #Expected frequency for  number of families living below
> # the poverty level
> fe1 <-n*pi0
> fe1
[1] 40
> # number of families in the sample not living below
> # the poverty level
> fo2 <- n-fo1
> fo2
[1] 155
> #Expected frequency for  number of families not living below
> # the poverty level
> fe2 <-n*(1-pi0)
> fe2
[1] 160
> #create a matrix showing observed and expected frequencies
> fofe = matrix(c(fo1, fe1, fo2, fe2),ncol=2,
+      dimnames = list(c("Observed Frequency (fo)","Expected Frequency (fe)"),
+                      c("Below","Above")))
> fofe
                        Below Above
Observed Frequency (fo)    45   155
Expected Frequency (fe)    40   160
```



```
> #chi-square test
> #without Yates correction
> prop.test(fo1,n,p=pi0, alternative="two.sided", correct=FALSE)

	1-sample proportions test without continuity correction

data:  fo1 out of n, null probability pi0
X-squared = 0.78125, df = 1, p-value = 0.3768
alternative hypothesis: true p is not equal to 0.2
95 percent confidence interval:
 0.1726237 0.2877412
sample estimates:
    p
0.225
```

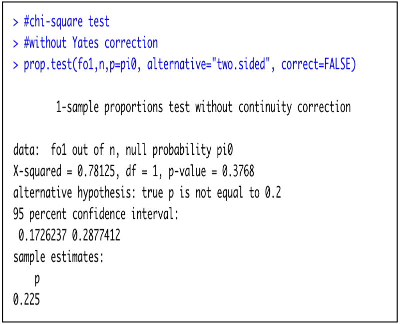
- 22.5% of families in the sample live below the poverty level.
- Both expected frequencies are > 5.
- The chi-square test statistic is obtained as follows:

$$\chi^2 = \frac{(45-40)^2}{40} + \frac{(155-160)^2}{160} = 0.781$$

Example: Chi-Square Test for Proportion - One Population

1. Hypotheses:
   $H_0 : \pi = 0.20$
   $H_1 : \pi \neq 0.20$
2. Level of significance: $\alpha = 0.05$
3. Test statistic: $\chi^2 = 0.781$
4. P-Value: p-value $= 0.377$
5. Decision Rule: Reject H_0 if p-value < 0.05
6. Conclusion: Since p-value > 0.05, do not reject H_0 at the 5% level and indeed any reasonable level of significance.

There is not sufficient evidence to conclude that the percentage of families living below the poverty level has changed.



```
> #chi-square test
> #without Yates correction
> prop.test(fo1,n,p=pi0, alternative="two.sided", correct=FALSE)

        1-sample proportions test without continuity correction

data:  fo1 out of n, null probability pi0
X-squared = 0.78125, df = 1, p-value = 0.3768
alternative hypothesis: true p is not equal to 0.2
95 percent confidence interval:
 0.1726237 0.2877412
sample estimates:
    p
0.225
```

This outcome corresponds to that from the Z-test for proportion in an earlier module.

Binomial Test for Proportion: One Population

- When testing that the proportion of a population is equal to a particular value, we make use of a statistic based on the normal approximation of the binomial distribution.

- This approximation is dependent on the sample size.

- When we have a small sample size, this normal approximation of the binomial distribution will not work, hence we cannot use the Z test statistic or the chi-square test statistic to test for the population proportion.

- We have to use the binomial distribution to obtain exact probabilities in determining the p-value of the test.

Binomial Test for Proportion: One Population

Assumptions: Properties of the Binomial Distribution

- Each observation is classified into one of two mutually and collectively exhaustive categories, usually called success or failure.
- The probability of an item being classified as a success, π, is constant from observation to observation. Thus, the probability of an observation being classified as a failure , 1- π, is also constant for all observations.
- The outcome (i.e., success or failure) of any observation is independent of the outcome of any other observation.
- To ensure independence, the observations can be randomly be selected either from an infinite population without replacement or from a finite population with replacement.

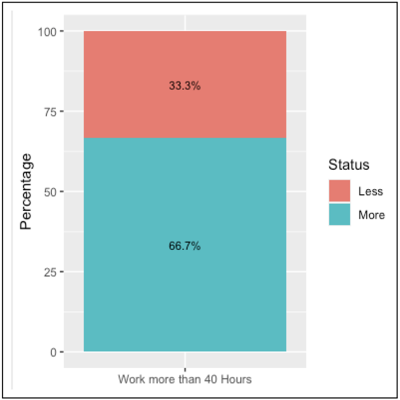
Binomial Test for Proportion: One Population

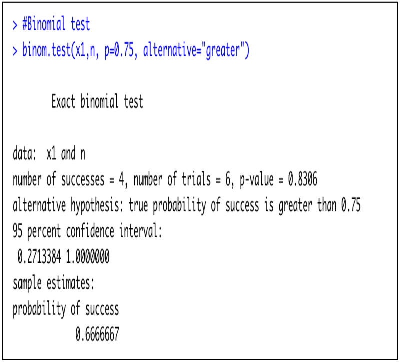
- Let X be the number of successes in a sample. X is used as the test statistic.
- The p-value is computed based on the binomial distribution as the probability of observing X successes for a given probability of success, π_0 in the null hypothesis.

Example: Binomial Test for Proportion - One Population

- Suppose that from a previous survey it was inferred that 75% of people work more than 40 hours a week.
- From a pilot survey of a random sample of 6 people in a small office, 4 indicated that they work more than 40 hours. Based on this sample, can it be concluded that the population proportion is more than 0.75 at the 5% level of significance?
- In this case, the n $= 6$, $\pi = 0.75$ and $n\pi = 4.5 < 5$. Even if $n\pi > 5$, the sample is far too small to assume the binomial distribution in this case approximates the normal distribution. Therefore, since we cannot use the Z statistic or the chi-square test statistic to perform the test for proportion, we will use the binomial test.

Example: Binomial Test for Proportion - One Population





```
> #Binomial test
> binom.test(x1,n, p=0.75, alternative="greater")

        Exact binomial test

data:  x1 and n
number of successes = 4, number of trials = 6, p-value = 0.8306
alternative hypothesis: true probability of success is greater than 0.75
95 percent confidence interval:
 0.2713384 1.0000000
sample estimates:
probability of success
             0.6666667
```

67% of people in the office work more than 40 hours.

Example: Binomial Test for Proportion - One Population

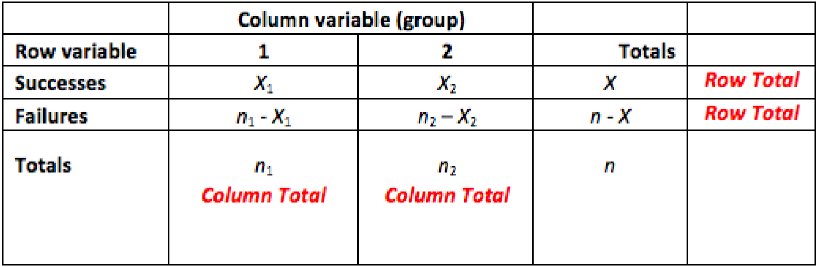
1. Hypotheses:
   $H_0 : \pi = 0.75$
   $H_1 : \pi > 0.75$
2. Level of significance: $\alpha = 0.05$
3. Test statistic: $X = 4$.
4. P-Value: p-value $= 0.831$
5. Decision Rule: Reject H_0 if p-value < 0.05
6. Conclusion: Since p-value > 0.05, do not reject H_0 at the 5% level and indeed any reasonable level of significance.

There is not sufficient evidence to conclude that that greater than 75% of people work more than 40 hours a week.

Analysis of Categorical Data: Independent Populations

- If we are interested in comparing the counts of categorical responses between 2 independent groups or populations, we develop a two-way cross-classification table to display the frequency of occurrence of successes and failures for each group.

- This table is called a contingency table. We can describe the 2 x 2 contingency table in this general form.



	Column variable (group)			
Row variable	**1**	**2**	**Totals**	
Successes	X_1	X_2	X	***Row Total***
Failures	$n_1 - X_1$	$n_2 - X_2$	$n - X$	***Row Total***
Totals	n_1 ***Column Total***	n_2 ***Column Total***	n	

Chi-Square Test for Equal Proportions: Two Populations

Assumptions

- Random samples are selected from independent populations of interest.
- Each observation can be cross-classified as one of two mutually exclusive types.

Chi-Square Test for Equal Proportions: Two Populations

Hypotheses

$$H_0 : \pi_1 = \pi_2 \quad H_1 : \pi_1 \neq \pi_2$$

- π_1: proportion associated with the first population, π_2: proportion associated with the second population.
- If the null hypothesis is true, it implies that the column and row variables are independent of each other.

Chi-Square Test for Equal Proportions: Two Populations

Test Statistic

- The chi-square test statistic compares the observed and expected frequencies:

$$\chi^2 = \sum_{all\ \ cells} \frac{(f_o - f_e)^2}{f_e}$$

- f_o is the observed frequency in a particular cell of a contingency table.
- f_e is the expected frequency, in a particular cell of a contingency table if the null hypothesis is true.
- The test statistic follows a chi-square distribution with (r-1)(c-1) degree of freedom. In the case of the 2 x 2 contingency table, $r = 2$, $c = 2$; hence, this test statistic has one degree of freedom.
- The closer the observed frequencies are to the expected frequencies under the null hypothesis, the less likely are we to reject the null hypothesis.

Chi-Square Test for Equal Proportions: Two Populations

Determination of Expected Frequencies

- If the null hypothesis is true, the proportion of successes in the two populations will be equal.

- Then the sample proportions calculated from each of the two groups would differ from each other only by chance, and each would provide an estimate of the common population parameter, π.

- The pooled estimate of π is
$$\overline{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

- The expected frequency, f_e, for each cell pertaining to success, i.e., the cells in the first row of the contingency table, is obtained by multiply the sample size of each group by $\overline{p}$.

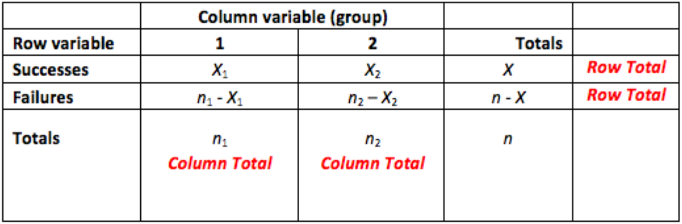
- The expected frequency, f_e, for each cell pertaining to failure, i.e., the cells in the second row of the contingency table, is obtained by multiply the sample size of each group by 1 - $\overline{p}$.

Chi-Square Test for Equal Proportions: Two Populations

Determination of Expected Frequencies

- The procedure to obtaining the expected frequencies is identical to:

$$Expected\ \ frequency = \frac{RowTotal\ \ x\ \ ColumnTotal}{n}$$



	Column variable (group)			
Row variable	**1**	**2**	**Totals**	
Successes	X_1	X_2	X	**Row Total**
Failures	n_1 - X_1	$n_2 - X_2$	n - X	**Row Total**
Totals	n_1 **Column Total**	n_2 **Column Total**	n	

- 1st row/1st column:

$$f_e = \frac{n_1 \times X}{n}$$

- 2nd row/1st column:

$$f_e = \frac{n_1 \times (n - X)}{n}$$

Chi-Square Test for Equal Proportions: Two Populations

Residuals

- **Unstandardised Residual**: The difference between an observed value and the expected value.
- A positive residual indicates that there are more cases in the cell than there would be if the row and column variables were independent.
- A negative residual indicates that there are less cases in the cell than there would be if the row and column variables were independent.
- **Standardised Residual**: Residual divided by its standard deviation

$$z = \frac{f_o - f_e}{\sqrt{f_e}}$$

- Standardised residuals have mean 0 and standard deviation 1.
- If a cell has a large standardized residual: > 2 or < -2, $\implies$ that cell contributes more to the outcome of the test.

Chi-Square Test for Equal Proportions: Two Populations

- The chi-square test used in this context is a special case of the chi-square test of independence.
- The requirement of the chi-square test is that the expected frequency of each cell must be at least 5.
- If this requirement is not met, we use an alternative test, viz., Fisher‘s Exact Test.

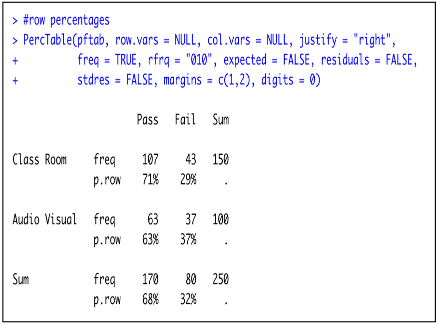
Example: Chi-Square Test for Equal Proportions - Two Populations

A lecturer who prefers innovative teaching methods wishes to compare the effectiveness of teaching Statistics by the traditional classroom lecture method and by the extensive use of audio-visual aids. To do so, 100 students were selected at random from a class of 250 and were assigned to audio-visual instruction. The remaining 150 students are taught Statistics in classroom lectures. At the end of the semester all 250 students are given a test. The number of students from each group who passed the test is given in the following table.

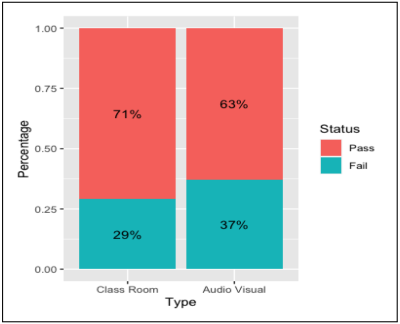
	Classroom Lecture	Audio-Visual Instruction
Pass	107	63
Fail	43	37
Total	150	100

Do the data support the hypothesis that a better pass rate is achieved using the classroom lecture method, than is achieved using the audio-visual method?

Example: Chi-Square Test for Equal Proportions: Two Populations



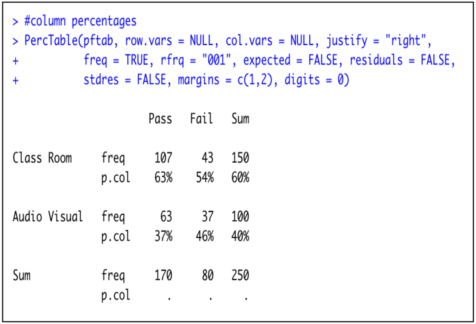
```
> #row percentages
> PercTable(pftab, row.vars = NULL, col.vars = NULL, justify = "right",
+           freq = TRUE, rfrq = "010", expected = FALSE, residuals = FALSE,
+           stdres = FALSE, margins = c(1,2), digits = 0)

                      Pass  Fail  Sum

Class Room    freq     107    43  150
              p.row    71%   29%    .

Audio Visual  freq      63    37  100
              p.row    63%   37%    .

Sum           freq     170    80  250
              p.row    68%   32%    .
```



We observe from the bar chart and the percentage of row totals that:

- Of those students who were instructed in the traditional classroom method, 71% passed the test compared to 29% who failed.
- Of those students who were instructed in the audio visual method, 63% passed the test compared to 37% who failed.

Example: Chi-Square Test for Equal Proportions - Two Populations

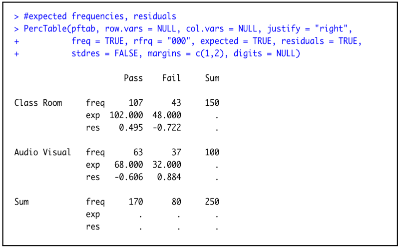


```
> #column percentages
> PercTable(pftab, row.vars = NULL, col.vars = NULL, justify = "right",
+           freq = TRUE, rfrq = "001", expected = FALSE, residuals = FALSE,
+           stdres = FALSE, margins = c(1,2), digits = 0)

                          Pass  Fail  Sum

Class Room       freq      107    43  150
                 p.col     63%   54%  60%

Audio Visual     freq       63    37  100
                 p.col     37%   46%  40%

Sum              freq      170    80  250
                 p.col       .     .    .
```

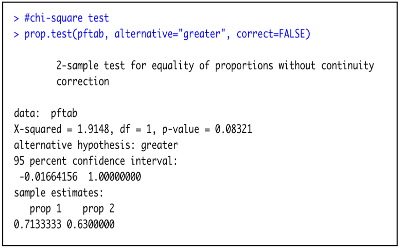
We observe from the percentage of column totals that:

- Of those students who passed the test, 63% were instructed in the traditional classroom method compared to 37% who were instructed in the audio visual method.
- Of those students who failed the test, 54% were instructed in the traditional classroom method compared to 46% who were instructed in the audio visual method.

Example: Chi-Square Test for Equal Proportions - Two Populations



```
> #expected frequencies, residuals
> PercTable(pftab, row.vars = NULL, col.vars = NULL, justify = "right",
+           freq = TRUE, rfrq = "000", expected = TRUE, residuals = TRUE,
+           stdres = FALSE, margins = c(1,2), digits = NULL)

                       Pass     Fail      Sum

Class Room     freq     107       43      150
               exp  102.000   48.000        .
               res    0.495   -0.722        .

Audio Visual   freq      63       37      100
               exp   68.000   32.000        .
               res   -0.606    0.884        .

Sum            freq     170       80      250
               exp        .        .        .
               res        .        .        .
```



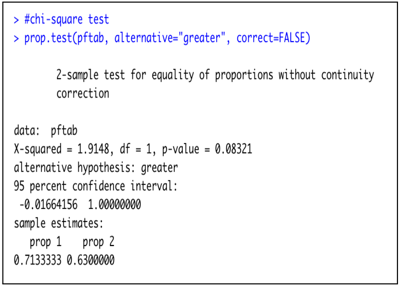
```
> #chi-square test
> prop.test(pftab, alternative="greater", correct=FALSE)

        2-sample test for equality of proportions without continuity
        correction

data:  pftab
X-squared = 1.9148, df = 1, p-value = 0.08321
alternative hypothesis: greater
95 percent confidence interval:
 -0.01664156  1.00000000
sample estimates:
   prop 1    prop 2
0.7133333 0.6300000
```

- The expected frequencies are what are expected under the null hypothesis that the population proportions associated with the classroom lecture and audio-visual instruction methods are equal.
- For example, in the cell (Pass, Class Room), the expected frequency is (150 x 170)/250 = 102.
- Note that residuals generated are also referred to a *Pearsons residuals* which we refer to as standardised residuals, for example, in the cell (Pass, Class Room), the standardised residual $z = \frac{f_o - f_e}{\sqrt{f_e}} = \frac{107 - 102}{\sqrt{102}} = 0.495$.
- In this case none of the $|$standardised residuals$| > 2$; hence no one cell contributes more to the the outcome of the test than any other.
- The chi-square statistic $\chi^2 = 0.495^2 + (-0.606)^2 + (-0.722)^2 + (.884)^2 = 1.915$

Example: Chi-Square Test for Equal Proportions - Two Populations

π_1 and π_2: population proportions associated with the classroom lecture and audio-visual instruction methods, respectively.

1. Hypotheses:
   $H_0 : \pi_1 - \pi_2 = 0 \quad H_1 : \pi_1 - \pi_2 > 0$
2. Level of significance: $\alpha = 0.05$
3. Test statistic: $\chi^2 = 1.9158$
4. P-value: $p - value = 0.083$
5. Decision Rule: Reject H_0 if p-value < 0.05
6. Conclusion: Since p-value > 0.05, do not reject H_0 at the 5% level of significance. There is insufficient evidence to conclude that the classroom lecture method is superior to the audio-visual instruction method.



```
> #chi-square test
> prop.test(pftab, alternative="greater", correct=FALSE)

	2-sample test for equality of proportions without continuity
	correction

data:  pftab
X-squared = 1.9148, df = 1, p-value = 0.08321
alternative hypothesis: greater
95 percent confidence interval:
 -0.01664156  1.00000000
sample estimates:
   prop 1    prop 2 
0.7133333 0.6300000 
```

However, if we test the hypothesis at the 10% level of significance, we will reject H_0 and conclude that the classroom lecture method could be superior to the audio-visual instruction method.

This outcome corresponds to that from the Z-test for equality of two population proportions in an earlier module.

Fisher‘s Exact Test for Equal Proportions: Two Populations

- The Fisher‘s exact test is used when we want to test for equal population proportions but one or more of the cells has an expected frequency of less than five.

- The chi-square test assumes that each cell has an expected frequency of five or more, but the Fisher’s exact test has no such assumption and can be used regardless of how small the expected frequency is.

- Instead of comparing the observed and expected cell frequencies, Fisher‘s exact test is based on calculations of the probability of the observed cell frequencies using the hypergeometric distribution.

  Refer to the reference section for more details about this distribution.

Fisher‘s Exact Test for Equal Proportions: Two Populations

Assumptions

- Random samples are selected from independent populations of interest.
- Each observation can be categorised as one of two mutually exclusive types.

Fisher‘s Exact Test for Equal Proportions: Two Populations

- What is being tested is that the population odds ratio is equal to 1. This is equivalent to testing equal population proportions.
- Using an exact method, a sample odds ratio using the cell probabilities and a corresponding p-value are determined.

	Column Variable	
Row variable		
Successes	a	b
Failures	c	d

- Given the above set up, the odds ratio is $\frac{ad}{bc}$ and is estimated by using an exact method such as the maximum likelihood method.

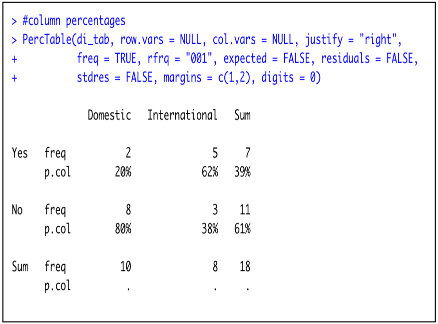
Example: Fisher‘s Exact Test for Equal Proportions -Two Populations

- Suppose a random sample of 10 domestic students and a random sample of 8 international students from a Masters course were asked if they would support continuous online learning. The data is provided below.

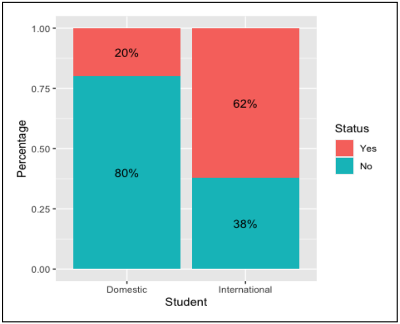
	Domestic	International
Yes	2	5
No	8	3

- We want to test if there is a difference in the proportions of domestic and international students who support continuous online learning.

Example: Fisher‘s Exact Test for Equal Proportions - Two Populations



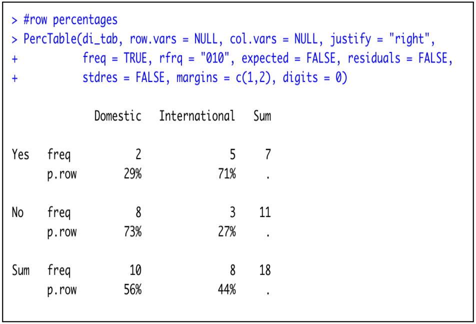
```
> #column percentages
> PercTable(di_tab, row.vars = NULL, col.vars = NULL, justify = "right",
+          freq = TRUE, rfrq = "001", expected = FALSE, residuals = FALSE,
+          stdres = FALSE, margins = c(1,2), digits = 0)

              Domestic  International  Sum

Yes   freq           2              5    7
      p.col        20%            62%  39%

No    freq           8              3   11
      p.col        80%            38%  61%

Sum   freq          10              8   18
      p.col          .              .    .
```



We observe from the bar chart and the percentage of column totals that:

- Of the domestic students, 80% do not support continuous online learning compared to 20% who do.
- Of the international students, 38% do not support continuous online learning compared to 62% who do.

Example: Fisher‘s Exact Test for Equal Proportions - Two Populations

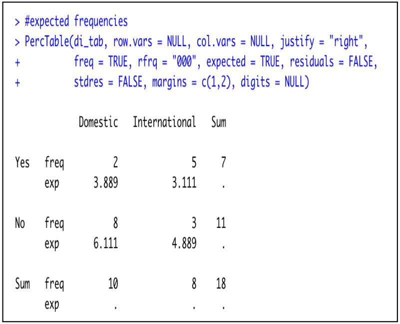


```
> #row percentages
> PercTable(di_tab, row.vars = NULL, col.vars = NULL, justify = "right",
+           freq = TRUE, rfrq = "010", expected = FALSE, residuals = FALSE,
+           stdres = FALSE, margins = c(1,2), digits = 0)

              Domestic   International   Sum

Yes   freq           2               5     7
      p.row        29%             71%     .

No    freq           8               3    11
      p.row        73%             27%     .

Sum   freq          10               8    18
      p.row        56%             44%     .
```

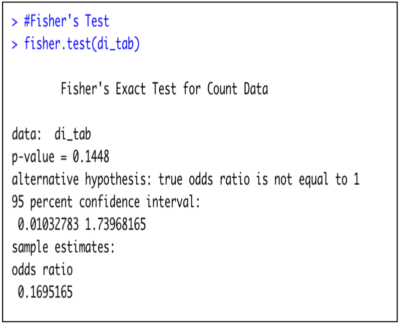
We observe from the the percentage of row totals that:

- Of the students who support continuous online learning, 29% are domestic compared to 71% who are international.
- Of the students who do not support continuous online learning, 73% are domestic compared to 27% who are international.

Example 4



```
> #expected frequencies
> PercTable(di_tab, row.vars = NULL, col.vars = NULL, justify = "right",
+          freq = TRUE, rfrq = "000", expected = TRUE, residuals = FALSE,
+          stdres = FALSE, margins = c(1,2), digits = NULL)

             Domestic  International  Sum

Yes  freq          2              5    7
     exp       3.889          3.111    .

No   freq          8              3   11
     exp       6.111          4.889    .

Sum  freq         10              8   18
     exp           .              .    .
```



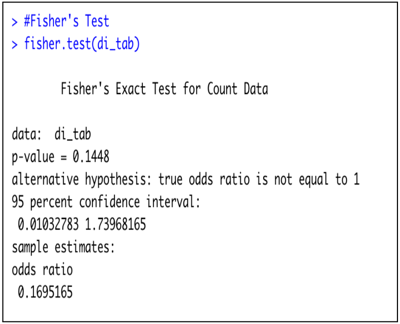
```
> #Fisher's Test
> fisher.test(di_tab)

        Fisher's Exact Test for Count Data

data:  di_tab
p-value = 0.1448
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.01032783 1.73968165
sample estimates:
odds ratio
 0.1695165
```

- For example in the cell (Yes, Domestic), the expected frequency is (10 x 7)/18= 3.889.
- Two cells in the contingency table have expected frequencies less than 5; hence the chi-square test for equal population proportions is not applicable. We use Fisher's test which is an exact test.

Example: Fisher's Exact Test for Equal Proportions - Two Populations

π_1 and π_2: population proportions associated with domestic and International students, respectively.

1. Hypotheses:
   $H_0 : \pi_1 - \pi_2 = 0 \quad H_1 : \pi_1 - \pi_2 \neq 0$
2. Level of significance: $\alpha = 0.05$
3. Test statistic: No test statistic since this is an exact test.
4. P-value: $p - value = 0.145$
5. Decision Rule: Reject H_0 if p-value < 0.05
6. Conclusion: Since p-value > 0.05, do not reject H_0 at the 5% level or indeed any reasonable level of significance.



```
> #Fisher's Test
> fisher.test(di_tab)

        Fisher's Exact Test for Count Data

data:  di_tab
p-value = 0.1448
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.01032783 1.73968165
sample estimates:
odds ratio
 0.1695165
```

There is not sufficient evidence to conclude there is any difference in the proportion of domestic and international students in their support for continuous online learning.