Data Science 1
Analysis of Categorical Data
Part 2

Ann Maharaj

Analysis of Categorical Data: Part 2

1. Analysis of Categorical Data: Independent Populations
2. Chi-Square Test for Equal Proportions: > 2 Populations
   - Example
3. Chi-square Test of Independence
   - Example

Introduction

We continue analysing categorical data for independent populations but now extend this analysis to:

- More than two populations where each observation can be categorised as one of two mutually exclusive types.
- Two or more populations where each observation can be categorised as one of more than two mutually exclusive types.

Chi-Square Test for Equal Proportions: More Than Two Populations

- The chi-square test is extended to compare the proportions of more than two independent populations.
- The letter c is used to represent the number of independent populations under consideration.
- The contingency table now has 2 rows and c columns.

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

Chi-Square Test for Equal Proportions: More Than 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: More Than Two Populations

- **Hypotheses**

  $H_0 : \pi_1 = \pi_2 = ... = \pi_c$

  $H_1 : Not \ \ all \ \ \pi_j \ \ are \ \ equal, \ \ j = 1, 2, ..., c$

- If the null hypothesis is true, it implies that the column and row variables are independent of each other.

- **Test Statistic**

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

  follows a chi-square distribution with $(r-1)(c-1) = (c-1)$ degrees of freedom

- **Standardised Residual**

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

Chi-Square Test for Equal Proportions: More Than Two Populations

Determination of Expected Frequencies

- If the null hypothesis is true and the proportions are equal across all c populations, then the sample proportions calculated from each of the c 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 which is an extension of the overall proportion estimate for the two population case is

$$\overline{p} = \frac{X_1 + X_2 + ... + X_c}{n_1 + n_2 + ... + n_c} = \frac{X}{n}$$

- The expected frequency, f_e, of 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, of the cells in the second row of the contingency table, is obtained by multiplying the sample size of each group by 1 - $\overline{p}$.

Chi-Square Test for Equal Proportions: More Than 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**	.	**c**	**Totals**	
Successes	X_1	X_2	.	X_c	X	***Row Total***
Failures	$n_1 - X_1$	$n_2 - X_2$	.	$n_c - X_c$	$n - X$	***Row Total***
Totals	n_1 ***Column Total***	n_2 ***Column Total***		n_c ***Column Total***	n	

Chi-Square Test for Equal Proportions: More Than 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 will need to test the equality of two proportions at a time and then use Fisher's Exact Test if necessary.

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

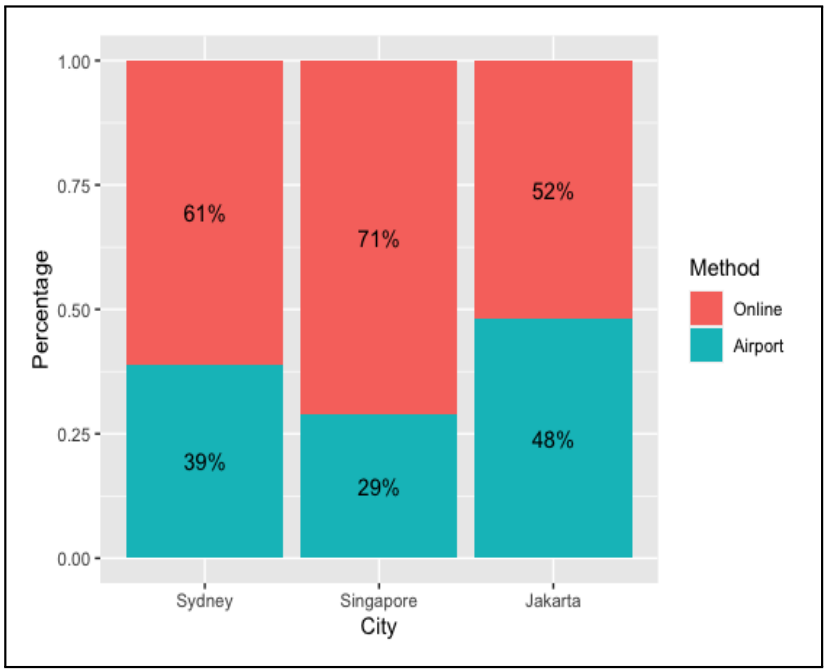
Airlines are constantly trying to find new ways to cut costs and improve efficiency. In recent years, as access to the internet has become more widespread, travelers are being given the option to check-in online for international flights prior to their arrival at the airport. This avoids long airport queues and also reduces the number of staff required. XXX Airlines encourage online check-in by sending reminder emails to its passengers 24 hours prior to their departure time. For operational planning purposes XXX Airlines is examining whether or not there is a difference between the use of online check-in at three of its airports, Sydney, Singapore and Jakarta. The check-in patterns of three samples of different passengers departing from the three cities are given below.

	City			
Check -in method	Sydney	Singapore	Jakarta	Total
Online	258	375	210	843
Airport	162	155	190	507
Total	420	530	400	1350

Test at the 5% level of significance if the proportion of online check-in is the same across the three cities.

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

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

                    Online   Airport   Sum

Sydney      freq       258       162   420
            p.row      61%       39%     .

Singapore   freq       375       155   530
            p.row      71%       29%     .

Jakarta     freq       210       190   400
            p.row      52%       48%     .

Sum         freq       843       507 1'350
            p.row      62%       38%     .
```



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

- 61% of passengers in Sydney check-in online compared to 39% who check-in at the airport.
- 71% of passengers in Singapore check-in online compared to 29% who check-in at the airport.
- 52% of passengers in Jakarta check-in online compared to 48% who check-in at the airport.

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

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

                   Online   Airport   Sum

Sydney     freq       258       162   420
           p.col      31%       32%   31%

Singapore  freq       375       155   530
           p.col      44%       31%   39%

Jakarta    freq       210       190   400
           p.col      25%       37%   30%

Sum        freq       843       507 1'350
           p.col        .         .     .
```

We observe from the the percentage of column totals that:

- Of the passengers who check-in online, 44% are in Singapore, followed by 31% in Sydney and 25% in Jakarta.
- Of the passengers who check-in at the airport , 37% are in Jakarta, followed by 32% in Sydney and 31% in Singapore.

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

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

                  Online   Airport     Sum

Sydney      freq     258       162     420
            exp   262.267   157.733       .
            res    -0.263     0.340       .

Singapore   freq     375       155     530
            exp   330.956   199.044       .
            res     2.421    -3.122       .

Jakarta     freq     210       190     400
            exp   249.778   150.222       .
            res    -2.517     3.245       .

Sum         freq     843       507   1'350
            exp        .         .       .
            res        .         .       .
```

```
> #chi-square test
> prop.test(airtab, correct=FALSE)

        3-sample test for equality of proportions without continuity
        correction

data:  airtab
X-squared = 32.66, df = 2, p-value = 8.09e-08
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3
0.6142857 0.7075472 0.5250000
```

- The expected frequencies are what is expected under the null hypothesis that the proportion of passengers who check-in online is the same in the three cities.
- For example in the cell (Singapore,Online),the expected frequency is $(843 \times 530)/1350 = 330.956$, and the standardised residual is $\frac{375-330.956}{\sqrt{330.956}} = 2.421$.
- In this case the |standardised residuals| of all the cells except (Sydney, Online) and Sydney, Airport) are greater that 2, while the cell (Jakarta, Airport) with |standardised residuals| $= 3.245$ contributes more to the outcome of the test than any other.

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

π_1, π_2 and π_3 : population proportions associated with online check-in, in Sydney, Singapore and Jakarta, respectively.

1. Hypotheses:
   $H_0 : \pi_1 = \pi_2 = \pi_2$
   H_1: Not all π_j are equal, $j = 1, 2, 3$
2. Level of significance: $\alpha = 0.05$
3. Test statistic: $\chi^2 = 32.660$
4. P-value: $p - value = 0.000$
5. Decision Rule: Reject H_0 if p-value < 0.05
6. Conclusion: Since p-value < 0.05, reject H_0 at the 5% level and indeed any reasonable level of significance.
   There is sufficient evidence to conclude that the proportion of passengers who check-in online for their flights across the three cities are different.

```
> #chi-square test
> prop.test(airtab, correct=FALSE)

        3-sample test for equality of proportions without continuity
        correction

data:  airtab
X-squared = 32.66, df = 2, p-value = 8.09e-08
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3
0.6142857 0.7075472 0.5250000
```

This is equivalent to concluding that there is a relationship between city and method of flight check-in.

Chi-square Test of Independence

- In the previous section we tested for equality of proportions between populations using the chi-square test.
- The objective was to make comparisons and evaluate differences between the proportions of success at various levels.
- In effect, we have actually been performing a test of independence between the column and row variables with the restriction that the row variable has only two levels (success and failure).
- In general, for the chi-square test of independence, there are two factors of interest (variables), where each observation can be categorised as one of more than two mutually exclusive types.

Chi-square Test of Independence

- For a contingency table that has r rows and c columns, we can generalise the chi-square test as a test of independence for two categorical variables A and B.

		Variable A				
		1	2	.	c	**Row Total**
Variable B	1	f_{11}	f_{12}	.	f_{1c}	$n_{1.}$
	2	f_{21}	f_{22}	.	f_{2c}	$n_{2.}$
	.					
	r	f_{r1}	f_{r2}	.	f_{rc}	$n_{r.}$
	Column Total	$n_{.1}$	$n_{.2}$	.	$n_{.c}$	n

- In general we have r rows and c columns: a $r \times c$ contingency table.
- $f_{ij}, i = 1, 2, ..., r; j = 1, 2, ..., c$ is the observed frequency of each cell.

Chi-square Test of Independence

Assumptions

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

Chi-square Test of Independence

- **Hypotheses**
  H_0 : The two categorical variables are independent (i.e., there is no relationship between them)
  H_1: The two categorical variables are dependent (i.e., there is a relationship between them)

- If the null hypothesis is true, it implies that the column and row variables are independent of each other.

- **Test Statistic**

$$\chi^2 = \sum_{all\ \ cells} \frac{(f_0 - f_e)^2}{f_e}$$

follows a chi-square distribution with $(r-1)(c-1)$ degrees of freedom

Chi-square Test of Independence

		Variable A				
		1	2	.	c	**Row Total**
Variable B	1	f_{11}	f_{12}	.	f_{1c}	$n_{1.}$
	2	f_{21}	f_{22}	.	f_{2c}	$n_{2.}$
	.					
	r	f_{r1}	f_{r2}	.	f_{rc}	$n_{r.}$
	Column Total	$n_{.1}$	$n_{.2}$	.	$n_{.c}$	n

- **Expected Frequencies**

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

- **Standardised Residual**

$$z = \frac{f_0 - f_e}{\sqrt{f_e}}$$

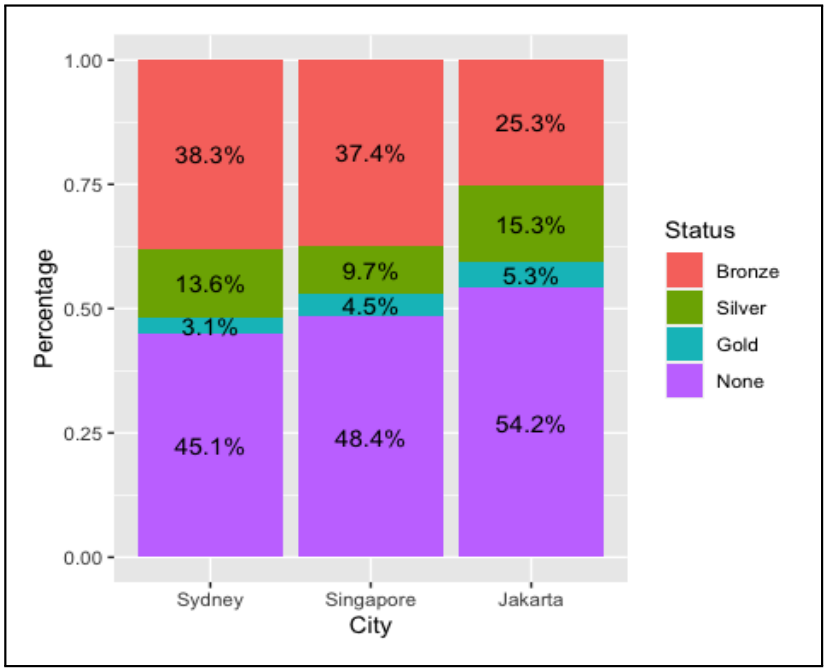
Example: Chi-square Test of Independence

Consider again the previous example which was on the analysis of airport check-in methods for an airline that operates out of Sydney, Singapore and Jakarta. Now for those who checked in at the airport, a second factor was examined to determine if passenger departure experience contributed to intercity difference. The measure of experience used was the passengers' frequent flyer status with the airline's loyalty program. We are going to examine the relationship between the passengers who checked in at the airport in the three cities, and their frequent flyer status

	City			
Frequency Flyer Status	Sydney	Singapore	Jakarta	Total
Bronze	62	58	48	168
Silver	22	15	29	66
Gold	5	7	10	22
None	73	75	103	251
Total	162	155	190	507

Example: Chi-square Test of Independence

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

                  Bronze   Silver   Gold   None   Sum

Sydney     freq       62       22      5     73   162
           p.row     38%      14%     3%    45%     .

Singapore  freq       58       15      7     75   155
           p.row     37%      10%     5%    48%     .

Jakarta    freq       48       29     10    103   190
           p.row     25%      15%     5%    54%     .

Sum        freq      168       66     22    251   507
           p.row     33%      13%     4%    50%     .
```



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

- Highest percentage of passengers are those with no frequency flyer status.
- Lowest percentage passengers are those with gold frequent flyer status.

Example: Chi-square Test of Independence

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

                    Bronze   Silver   Gold   None   Sum

Sydney      freq        62       22      5     73   162
            p.col      37%      33%    23%    29%   32%

Singapore   freq        58       15      7     75   155
            p.col      35%      23%    32%    30%   31%

Jakarta     freq        48       29     10    103   190
            p.col      29%      44%    45%    41%   37%

Sum         freq       168       66     22    251   507
            p.col        .        .      .      .     .
```

We observe from the percentage of column totals, for passengers who have:

- No frequency flyer status, the highest percentage come from Jakarta, followed by Singapore and then by Sydney.
- Bronze frequency flyer status, the highest percentage come from Sydney, followed by Singapore and then by Jakarta.
- Silver frequency flyer status, the highest percentage come from Jakarta, followed by Sydney and then by Singapore.
- Gold frequency flyer status, the highest percentage come from Jakarta, followed by Singapore and then by Sydney.

Example: Chi-square Test of Independence

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

                    Bronze   Silver   Gold   None    Sum

Sydney      freq        62       22      5     73    162
            exp     53.680   21.089  7.030 80.201      .
            res      1.136    0.198 -0.765 -0.804      .

Singapore   freq        58       15      7     75    155
            exp     51.361   20.178  6.726 76.736      .
            res      0.926   -1.153  0.106 -0.198      .

Jakarta     freq        48       29     10    103    190
            exp     62.959   24.734  8.245 94.063      .
            res     -1.885    0.858  0.611  0.921      .

Sum         freq       168       66     22    251    507
            exp          .        .      .      .      .
            res          .        .      .      .      .
```

```
> #chi-square test
>
> chisq.test(fftab, correct=FALSE)

        Pearson's Chi-squared test

data:  fftab
X-squared = 10.311, df = 6, p-value = 0.1121
```

- The expected frequencies are what is expected under the null hypothesis that the proportion of passengers with a specific frequency flyer status is the same across three cities.
- For example in the cell (Sydney, Bronze), the expected frequency is $(168 \times 162)/507 = 53.680$
  and the standardised residual
  $\frac{62-53.680}{\sqrt{53.680}} = 1.136.$
- In this case the $|\text{standardised residuals}| < 2$ of all the cells; hence no one cell contributes more to the outcome of the test than any other.

Example: Chi-square Test of Independence

1. Hypotheses:
   H_0: There is no relationship between the city and frequent flyer status of those passengers who check in at the airport.

   H_1: There is a relationship between the city and frequent flyer status of those passengers who check in at the airport.
2. Level of significance: $\alpha = 0.05$
3. Test statistic: $\chi^2 = 10.311$
4. P-value: $p - value = 0.1121$
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.

```
> #chi-square test
>
> chisq.test(fftab, correct=FALSE)

        Pearson's Chi-squared test

data:  fftab
X-squared = 10.311, df = 6, p-value = 0.1121
```

There is not sufficient evidence to conclude that there is a relationship between city and frequent flyer status of those passengers who check in at the airport.