Lecture 9:
Chi-square test

Lecturer: Ksenia Kasianova
xeniakasianova@gmail.com

January 22, 2024

Plan

Plan

1) Contingency tables

2) Chi-square test

3) Partial correlation

Plan

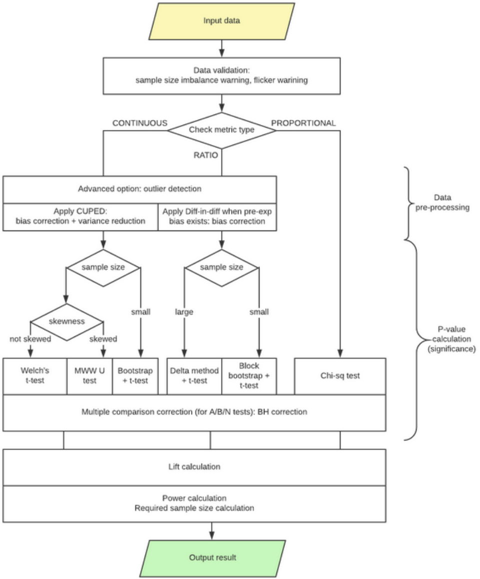


Figure: Uber's statistics engine is used for A/B/N experiments and dictated by fixed horizon

Cross-tabulation

While a frequency distribution describes one variable at a time, a cross-tabulation describes two or more variables simultaneously.

Cross-tabulation

While a frequency distribution describes one variable at a time, a cross-tabulation describes two or more variables simultaneously.

Cross-tabulation results in tables which reflect the **joint distribution** of two or more variables with a limited number of categories or distinct values, for example:

Internet usage	Gender		Row total
	Male	Female	
Light	5	10	15
Heavy	10	5	15
Column total	15	15	

Cross-tabulation

Since two variables have been cross-classified, percentages could be computed either *column-wise*, based on column totals, or *row-wise*, based on row totals. The general rule is to compute the percentages in the direction of the **independent variable**, across the dependent variable.

Cross-tabulation

Since two variables have been cross-classified, percentages could be computed either *column-wise*, based on column totals, or *row-wise*, based on row totals. The general rule is to compute the percentages in the direction of the **independent variable**, across the dependent variable.

(i) Column-wise percentages: internet usage is dependent variable (meaningful, gender is consistent)

	Gender	
Internet usage	Male	Female
Light (1)	33.3%	66.7%
Heavy (2)	66.7%	33.3%
Column totals	100%	100%

Cross-tabulation

Since two variables have been cross-classified, percentages could be computed either *column-wise*, based on column totals, or *row-wise*, based on row totals. The general rule is to compute the percentages in the direction of the **independent variable**, across the dependent variable.

(i) Column-wise percentages: internet usage is dependent variable (meaningful, gender is consistent)

	Gender	
Internet usage	Male	Female
Light (1)	33.3%	66.7%
Heavy (2)	66.7%	33.3%
Column totals	100%	100%

(ii) Row-wise percentages: gender is dependent variable (not meaningful)

	Gender total		
Internet usage	Male	Female	Row total
Light (1)	33.3%	66.7%	100%
Heavy (2)	66.7%	33.3%	100%

Perks and limitations of cross-tabulations

Cross-tabulations are popular for the following reasons.

Perks and limitations of cross-tabulations

Cross-tabulations are popular for the following reasons.

- **Ease of comprehension** -i.e. cross-tabulation analysis and results can be easily interpreted and understood by managers who have little statistical knowledge.

Perks and limitations of cross-tabulations

Cross-tabulations are popular for the following reasons.

- **Ease of comprehension** -i.e. cross-tabulation analysis and results can be easily interpreted and understood by managers who have little statistical knowledge.

- **Versatility** - i.e. a series of cross-tabulations may provide greater insights into a complex phenomenon than a single multivariate analysis.

Perks and limitations of cross-tabulations

Cross-tabulations are popular for the following reasons.

- **Ease of comprehension** -i.e. cross-tabulation analysis and results can be easily interpreted and understood by managers who have little statistical knowledge.

- **Versatility** - i.e. a series of cross-tabulations may provide greater insights into a complex phenomenon than a single multivariate analysis.

- **Clarity** - i.e. the clarity of interpretations provides a stronger link between the research results and managerial action.

Perks and limitations of cross-tabulations

Cross-tabulations are popular for the following reasons.

- **Ease of comprehension** -i.e. cross-tabulation analysis and results can be easily interpreted and understood by managers who have little statistical knowledge.

- **Versatility** - i.e. a series of cross-tabulations may provide greater insights into a complex phenomenon than a single multivariate analysis.

- **Clarity** - i.e. the clarity of interpretations provides a stronger link between the research results and managerial action.

- **Simplicity** - i.e. cross-tabulation analysis is simple to conduct and appealing to the less sophisticated researcher.

Cross-tabulation is *seldom used* in computations involving more than three variables, since the interpretation becomes quite complex.

Also, since the number of cells increases multiplicatively, maintaining an adequate number of participants in each cell becomes problemat·ic. Consequently, the statistics computed could be *unreliable*.

Three variables

Often the introduction of a *third variable* clarifies the initial association (or lack of it) observed between two variables.

Three variables

Often the introduction of a *third variable* clarifies the initial association (or lack of it) observed between two variables.

1. It can refine the association observed between the two original variables.

Three variables

Often the introduction of a *third variable* clarifies the initial association (or lack of it) observed between two variables.

1. It can refine the association observed between the two original variables.

2. It can indicate no association between the two variables, although an association was initially observed. In other words, the third variable indicates that the initial association between the two variables was spurious.

Three variables

Often the introduction of a *third variable* clarifies the initial association (or lack of it) observed between two variables.

1. It can refine the association observed between the two original variables.

2. It can indicate no association between the two variables, although an association was initially observed. In other words, the third variable indicates that the initial association between the two variables was spurious.

3. It can reveal some association between the two original variables, although no association was initially observed. In this case, the third variable reveals a suppressed association between the first two variables: a suppressor effect.

Three variables

Often the introduction of a *third variable* clarifies the initial association (or lack of it) observed between two variables.

1. It can refine the association observed between the two original variables.

2. It can indicate no association between the two variables, although an association was initially observed. In other words, the third variable indicates that the initial association between the two variables was spurious.

3. It can reveal some association between the two original variables, although no association was initially observed. In this case, the third variable reveals a suppressed association between the first two variables: a suppressor effect.

4. It can indicate no change in the initial association.

Refined initial relationship

	Marital status	
Purchase of luxury branded clothing	Married	Unmarried
High	31%	52%
Low	69%	48%
Column	100%	100%
Number of participants	700	300

Refined initial relationship

	Marital status	
Purchase of luxury branded clothing	Married	Unmarried
High	31%	52%
Low	69%	48%
Column	100%	100%
Number of participants	700	300

52% of unmarried participants fell in the high-purchase category, as opposed to 31% of the married participants. Before concluding that unmarried participants purchase more luxury branded clothing than those who are married, a third variable, the buyer's gender, was introduced into the analysis.

	Gender			
Purchase of luxury branded clothing	Male marital status	Female marital status		
	Married	Unmarried	Married	Unmarried
High	35%	40%	25%	60%
Low	65%	60%	75%	40%
Column	100%	100%	100%	100%
Number of participants	400	120	300	180

Refined initial relationship

Purchase of luxury branded clothing	Marital status: Married	Marital status: Unmarried
High	31%	52%
Low	69%	48%
Column	100%	100%
Number of participants	700	300

52% of unmarried participants fell in the high-purchase category, as opposed to 31% of the married participants. Before concluding that unmarried participants purchase more luxury branded clothing than those who are married, a third variable, the buyer's gender, was introduced into the analysis.

Purchase of luxury branded clothing	Gender			
	Male marital status	Female marital status		
	Married	Unmarried	Married	Unmarried
High	35%	40%	25%	60%
Low	65%	60%	75%	40%
Column	100%	100%	100%	100%
Number of participants	400	120	300	180

In the case of females, 60% of the unmarried participants fall in the high-purchase category compared with 25% of those who are married. On the other hand, the percentages are much closer for males.

Refined initial relationship

	Marital status	
Purchase of luxury branded clothing	Married	Unmarried
High	31%	52%
Low	69%	48%
Column	100%	100%
Number of participants	700	300

52% of unmarried participants fell in the high-purchase category, as opposed to 31% of the married participants. Before concluding that unmarried participants purchase more luxury branded clothing than those who are married, a third variable, the buyer's gender, was introduced into the analysis.

	Gender			
Purchase of luxury branded clothing	Male marital status	Female marital status		
	Married	Unmarried	Married	Unmarried
High	35%	40%	25%	60%
Low	65%	60%	75%	40%
Column	100%	100%	100%	100%
Number of participants	400	120	300	180

In the case of females, 60% of the unmarried participants fall in the high-purchase category compared with 25% of those who are married. On the other hand, the percentages are much closer for males.

Hence, the introduction of gender (third variable) has **refined the relationship** between marital status and purchase of luxury branded clothing (original variables).

Refined initial relationship

	Marital status	
Purchase of luxury branded clothing	Married	Unmarried
High	31%	52%
Low	69%	48%
Column	100%	100%
Number of participants	700	300

52% of unmarried participants fell in the high-purchase category, as opposed to 31% of the married participants. Before concluding that unmarried participants purchase more luxury branded clothing than those who are married, a third variable, the buyer's gender, was introduced into the analysis.

	Gender			
Purchase of luxury branded clothing	Male marital status	Female marital status		
	Married	Unmarried	Married	Unmarried
High	35%	40%	25%	60%
Low	65%	60%	75%	40%
Column	100%	100%	100%	100%
Number of participants	400	120	300	180

In the case of females, 60% of the unmarried participants fall in the high-purchase category compared with 25% of those who are married. On the other hand, the percentages are much closer for males.

Hence, the introduction of gender (third variable) has **refined the relationship** between marital status and purchase of luxury branded clothing (original variables).

Unmarried participants are more likely to fall into the high-purchase category than married ones, and this effect is much more pronounced for females than for males.

Initial relationship was spurious

	Education	
Own expensive car	Degree	No degree
Yes	32%	21%
No	68%	79%
Column	100%	100%
Number of participants	250	750

Initial relationship was spurious

	Education	
Own expensive car	Degree	No degree
Yes	32%	21%
No	68%	79%
Column	100%	100%
Number of participants	250	750

The table shows that 32% of those with university degrees own an expensive (more than €80,000) car, compared with 21% of those without university degrees.

Initial relationship was spurious

	Education	
Own expensive car	Degree	No degree
Yes	32%	21%
No	68%	79%
Column	100%	100%
Number of participants	250	750

The table shows that 32% of those with university degrees own an expensive (more than €80,000) car, compared with 21% of those without university degrees.

Conclusion: education influenced ownership of expensive cars.

Initial relationship was spurious

	Education	
Own expensive car	Degree	No degree
Yes	32%	21%
No	68%	79%
Column	100%	100%
Number of participants	250	750

The table shows that 32% of those with university degrees own an expensive (more than €80,000) car, compared with 21% of those without university degrees.

Conclusion: education influenced ownership of expensive cars.

However, income may also be an important factor for determining car ownership.

	Income			
Own expensive car	Low-income education		High-income education	
	Degree	No degree	Degree	No degree
Yes	20%	20%	40%	40%
No	80%	80%	60%	60%
Column totals	100%	100%	100%	100%
Number of participants	100	700	150	50

Initial relationship was spurious

Own expensive car	Education: Degree	Education: No degree
Yes	32%	21%
No	68%	79%
Column	100%	100%
Number of participants	250	750

The table shows that 32% of those with university degrees own an expensive (more than €80,000) car, compared with 21% of those without university degrees.

Conclusion: education influenced ownership of expensive cars.

However, income may also be an important factor for determining car ownership.

	Income			
Own expensive car	Low-income education		High-income education	
	Degree	No degree	Degree	No degree
Yes	20%	20%	40%	40%
No	80%	80%	60%	60%
Column totals	100%	100%	100%	100%
Number of participants	100	700	150	50

The percentages of those with and without university degrees who own expensive cars are the same for each income group.

Initial relationship was spurious

Own expensive car	Education: Degree	Education: No degree
Yes	32%	21%
No	68%	79%
Column	100%	100%
Number of participants	250	750

The table shows that 32% of those with university degrees own an expensive (more than €80,000) car, compared with 21% of those without university degrees.

Conclusion: education influenced ownership of expensive cars.

However, income may also be an important factor for determining car ownership.

Own expensive car	Income: Low-income education		Income: High-income education	
	Degree	No degree	Degree	No degree
Yes	20%	20%	40%	40%
No	80%	80%	60%	60%
Column totals	100%	100%	100%	100%
Number of participants	100	700	150	50

The percentages of those with and without university degrees who own expensive cars are the same for each income group.

When the data for the high-income and low-income groups are examined separately, the association between education and ownership of expensive cars disappears, indicating that the **initial relationship** observed between these two variables was **spurious**.

Reveal suppressed association

	Age	
Desire to travel abroad	Under **45**	**45** or older
Yes	50%	50%
No	50%	50%
Column totals	100%	100%
Number of participants	500	500

Cross-tabulation indicate no association.

Reveal suppressed association

	Age	
Desire to travel abroad	Under **45**	**45** or older
Yes	50%	50%
No	50%	50%
Column totals	100%	100%
Number of participants	500	500

Cross-tabulation indicate no association.

Let's introduced gender as the third variable.

	Gender			
	Male age		Female age	
Desire to travel abroad	Under **45**	**45** or older	Under **45**	**45** or older
Yes	60%	40%	35%	65%
No	40%	60%	65%	35%
Column totals	100%	100%	100%	100%
Number of participants	300	300	200	200

Reveal suppressed association

	Age	
Desire to travel abroad	Under **45**	**45** or older
Yes	50%	50%
No	50%	50%
Column totals	100%	100%
Number of participants	500	500

Cross-tabulation indicate no association.

Let's introduced gender as the third variable.

	Gender			
	Male age		Female age	
Desire to travel abroad	Under **45**	**45** or older	Under **45**	**45** or older
Yes	60%	40%	35%	65%
No	40%	60%	65%	35%
Column totals	100%	100%	100%	100%
Number of participants	300	300	200	200

Among men, 60% of those under 45 indicated a desire to travel abroad compared with 40% of those 45 or older. The pattern was reversed for women.

Reveal suppressed association

	Age	
Desire to travel abroad	Under **45**	**45** or older
Yes	50%	50%
No	50%	50%
Column totals	100%	100%
Number of participants	500	500

Cross-tabulation indicate no association.

Let's introduced gender as the third variable.

	Gender			
	Male age		Female age	
Desire to travel abroad	Under **45**	**45** or older	Under **45**	**45** or older
Yes	60%	40%	35%	65%
No	40%	60%	65%	35%
Column totals	100%	100%	100%	100%
Number of participants	300	300	200	200

Among men, 60% of those under 45 indicated a desire to travel abroad compared with 40% of those 45 or older. The pattern was reversed for women.

Since the association between desire to travel abroad and age runs in the *opposite direction* for males and females, the relationship between these two variables is masked when the data are aggregated across gender.

Reveal suppressed association

Desire to travel abroad	Age: Under **45**	Age: **45** or older
Yes	50%	50%
No	50%	50%
Column totals	100%	100%
Number of participants	500	500

Cross-tabulation indicate no association.

Let's introduced gender as the third variable.

	Gender			
	Male age		Female age	
Desire to travel abroad	Under **45**	**45** or older	Under **45**	**45** or older
Yes	60%	40%	35%	65%
No	40%	60%	65%	35%
Column totals	100%	100%	100%	100%
Number of participants	300	300	200	200

Among men, 60% of those under 45 indicated a desire to travel abroad compared with 40% of those 45 or older. The pattern was reversed for women.

Since the association between desire to travel abroad and age runs in the *opposite direction* for males and females, the relationship between these two variables is masked when the data are aggregated across gender.

But when the effect of gender is controlled, the **suppressed association** between preference and age is **revealed** for the separate categories of males and females.

No change in initial relationship

	Family size	
Eat frequently in fast-food restaurants	Small	Large
Yes	65%	65%
No	35%	35%
Column totals	100%	100%
Number of participants	500	500

No change in initial relationship

	Family size	
Eat frequently in fast-food restaurants	Small	Large
Yes	65%	65%
No	35%	35%
Column totals	100%	100%
Number of participants	500	500

No association is observed.

No change in initial relationship

Eat frequently in fast-food restaurants	Family size: Small	Family size: Large
Yes	65%	65%
No	35%	35%
Column totals	100%	100%
Number of participants	500	500

No association is observed.

The participants were further classified into high- or low-income groups based on a median split.

Eat frequently in fast-food restaurants	Income: Low-income family size		Income: High-income family size	
	Small	Large	Small	Large
Yes	65%	65%	65%	65%
No	35%	35%	35%	35%
Column total	100%	100%	100%	100%
Number of participants	250	250	250	250

No change in initial relationship

Eat frequently in fast-food restaurants	Family size: Small	Family size: Large
Yes	65%	65%
No	35%	35%
Column totals	100%	100%
Number of participants	500	500

No association is observed.

The participants were further classified into high- or low-income groups based on a median split.

Eat frequently in fast-food restaurants	Income: Low-income family size		Income: High-income family size	
	Small	Large	Small	Large
Yes	65%	65%	65%	65%
No	35%	35%	35%	35%
Column total	100%	100%	100%	100%
Number of participants	250	250	250	250

Again, no association was observed.

No change in initial relationship

	Family size	
Eat frequently in fast-food restaurants	Small	Large
Yes	65%	65%
No	35%	35%
Column totals	100%	100%
Number of participants	500	500

No association is observed.

The participants were further classified into high- or low-income groups based on a median split.

Eat frequently in fast-food restaurants	Income			
	Low-income family size		High-income family size	
	Small	Large	Small	Large
Yes	65%	65%	65%	65%
No	35%	35%	35%	35%
Column total	100%	100%	100%	100%
Number of participants	250	250	250	250

Again, no association was observed.

In some cases, the introduction of the third variable **does not change the initial relationship** observed, regardless of whether the original variables were associated.

This suggests that the third variable does not influence the relationship between the first two.

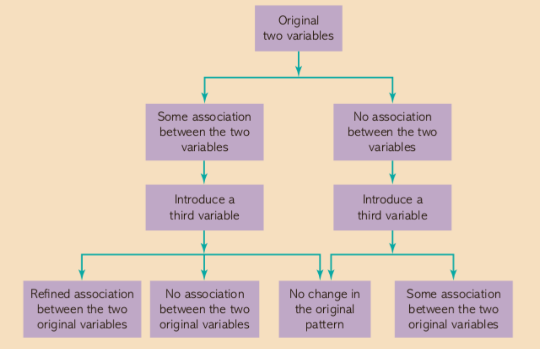


Figure: The introduction of a third variable in cross-tabulation

Spurious association – i.e. the introduction of a third variable in the cross-tabulation reveals that there is no association between the two variables despite the observed initial association.

Suppressed association – i.e. after introducing a third variable, the cross-tabulation reveals association between the two variables although no association was initially observed.

Chi-square test

Q: Is there a systematic association exists between the two variables?

Chi-square test

Q: Is there a systematic association exists between the two variables?

The chi-square statistic (χ^2) is used to test the statistical significance of the observed association in a cross-tabulation.

H_0 : there is no association between the variables.

H_a : there is association between the variables.

Chi-square test

Q: Is there a systematic association exists between the two variables?

The chi-square statistic (χ^2) is used to test the statistical significance of the observed association in a cross-tabulation.

H_0 : there is no association between the variables.

H_a : there is association between the variables.

Idea: compare the cell frequencies that would be expected if no association were present between the variables, given the existing row and column totals.

f_e – expected cell frequencies

f_o – actual observed frequencies.

The greater the discrepancies between the expected and observed frequencies, the larger the value of the statistic.

Chi-square test

The expected frequency for each cell can be calculated by using a simple formula:

$$f_e = \frac{n_r n_e}{n}$$

where

$$\begin{aligned} n_r = & \text{ total number in the row} \\ n_c = & \text{ total number in the column} \\ n = & \text{ total sample size.} \end{aligned}$$

The expected frequency for each cell can be calculated by using a simple formula:

$$f_e = \frac{n_r n_e}{n}$$

where

$$\begin{aligned} n_r = &\ \text{total number in the row} \\ n_c = &\ \text{total number in the column} \\ n = &\ \text{total sample size.} \end{aligned}$$

Then the value of χ^2 is calculated as follows:

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e} \sim \chi^2_{(r-1)\times(c-1)}$$

Chi-square test

The expected frequency for each cell can be calculated by using a simple formula:

$$f_e = \frac{n_r n_e}{n}$$

where

$$n_r = \text{ total number in the row}$$
$$n_c = \text{ total number in the column}$$
$$n = \text{ total sample size.}$$

Then the value of χ^2 is calculated as follows:

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e} \sim \chi^2_{(r-1)\times(c-1)}$$

The null hypothesis (H_0) of no association between the two variables will be rejected only when the calculated value of the test statistic is greater than the critical value of the chi-square distribution with the appropriate degrees of freedom.

Chi-square test

The expected frequency for each cell can be calculated by using a simple formula:

$$f_e = \frac{n_r n_e}{n}$$

where

$$n_r = \text{ total number in the row}$$
$$n_c = \text{ total number in the column}$$
$$n = \text{ total sample size.}$$

Then the value of χ^2 is calculated as follows:

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e} \sim \chi^2_{(r-1)\times(c-1)}$$

The null hypothesis (H_0) of no association between the two variables will be rejected only when the calculated value of the test statistic is greater than the critical value of the chi-square distribution with the appropriate degrees of freedom.

The chi-square distribution is a skewed distribution whose shape depends solely on the number of degrees of freedom.

As the number of degrees of freedom increases, the chi-square distribution becomes more symmetrical.

Example

Internet usage	Gender		Row total
	Male	Female	
Light	5	10	15
Heavy	10	5	15
Column total	15	15	

Example

Internet usage	Gender		Row total
	Male	Female	
Light	5	10	15
Heavy	10	5	15
Column total	15	15	

The expected frequencies for the cells, going from left to right and from top to bottom, are:

$$15 \times 15/30 = 7.50, 1515/30 = 7.50, 15 \times 15/30 = 7.50, 15 \times 15/30 = 7.50$$

Example

Internet usage	Gender		Row total
	Male	Female	
Light	5	10	15
Heavy	10	5	15
Column total	15	15	

The expected frequencies for the cells, going from left to right and from top to bottom, are:

$$15 \times 15/30 = 7.50, 1515/30 = 7.50, 15 \times 15/30 = 7.50, 15 \times 15/30 = 7.50$$

The value of χ^2 is calculated as:

$$\begin{aligned}
\chi^2 &= (5-7.5)^2/7.5 + (10-7.5)^2/7.5 + (10-7.5)^2/7.5 + (5-7.5)^2/7.5 \\
&= 0.833 + 0.833 + 0.833 + 0.833 \\
&= 3.333
\end{aligned}$$

Example

Internet usage	Gender		Row total
	Male	Female	
Light	5	10	15
Heavy	10	5	15
Column total	15	15	

The expected frequencies for the cells, going from left to right and from top to bottom, are:

$$15 \times 15/30 = 7.50, 1515/30 = 7.50, 15 \times 15/30 = 7.50, 15 \times 15/30 = 7.50$$

The value of χ^2 is calculated as:

$$\begin{aligned}\chi^2 &= (5-7.5)^2/7.5 + (10-7.5)^2/7.5 + (10-7.5)^2/7.5 + (5-7.5)^2/7.5 \\ &= 0.833 + 0.833 + 0.833 + 0.833 \\ &= 3.333\end{aligned}$$

Number of degree of freedom: $df = (2-1) \times (2-1) = 1$

Critical value at 5%: $\chi^2_{1,0.95} = 3.841$

Example

Internet usage	Gender		Row total
	Male	Female	
Light	5	10	15
Heavy	10	5	15
Column total	15	15	

The expected frequencies for the cells, going from left to right and from top to bottom, are:

$$15 \times 15/30 = 7.50, 1515/30 = 7.50, 15 \times 15/30 = 7.50, 15 \times 15/30 = 7.50$$

The value of χ^2 is calculated as:

$$\begin{aligned}\chi^2 &= (5-7.5)^2/7.5 + (10-7.5)^2/7.5 + (10-7.5)^2/7.5 + (5-7.5)^2/7.5 \\ &= 0.833 + 0.833 + 0.833 + 0.833 \\ &= 3.333\end{aligned}$$

Number of degree of freedom: $df = (2-1) \times (2-1) = 1$

Critical value at 5%: $\chi^2_{1,0.95} = 3.841$

The null hypothesis of no association cannot be rejected, indicating that the association is not statistically significant at the 0.05 level.

Example

Internet usage	Gender		Row total
	Male	Female	
Light	5	10	15
Heavy	10	5	15
Column total	15	15	

The expected frequencies for the cells, going from left to right and from top to bottom, are:

$$15 \times 15/30 = 7.50, 1515/30 = 7.50, 15 \times 15/30 = 7.50, 15 \times 15/30 = 7.50$$

The value of χ^2 is calculated as:

$$\begin{aligned}\chi^2 &= (5-7.5)^2/7.5 + (10-7.5)^2/7.5 + (10-7.5)^2/7.5 + (5-7.5)^2/7.5 \\ &= 0.833 + 0.833 + 0.833 + 0.833 \\ &= 3.333\end{aligned}$$

Number of degree of freedom: $df = (2-1) \times (2-1) = 1$

Critical value at 5%: $\chi^2_{1,0.95} = 3.841$

The null hypothesis of no association cannot be rejected, indicating that the association is not statistically significant at the 0.05 level.

Note that this lack of significance is mainly due to the small sample size (30).

If, instead, the sample size were 300 and each data entry were multiplied by 10, test statistics would be multiplied by 10 $\chi^2_{obs} = 33.33$, which is significant at the 0.05 level.

Test assumptions

– The chi-square statistic should be estimated only on counts of data. When the data are in percentage form, they should first be converted to absolute counts or numbers.

Test assumptions

– The chi-square statistic should be estimated only on counts of data. When the data are in percentage form, they should first be converted to absolute counts or numbers.

– The observations are drawn independently.

Test assumptions

– The chi-square statistic should be estimated only on counts of data. When the data are in percentage form, they should first be converted to absolute counts or numbers.

– The observations are drawn independently.

– Chi-square analysis should not be conducted when the expected or theoretical frequency in any of the cells is less than five.

Test assumptions

– The chi-square statistic should be estimated only on counts of data. When the data are in percentage form, they should first be converted to absolute counts or numbers.

– The observations are drawn independently.

– Chi-square analysis should not be conducted when the expected or theoretical frequency in any of the cells is less than five.

– If the number of observations in any cell is less than 10, or if the table has two rows and two columns (a 2×2 table), a correction factor should be applied (Yates's chi-squared test).

Test assumptions

– The chi-square statistic should be estimated only on counts of data. When the data are in percentage form, they should first be converted to absolute counts or numbers.

– The observations are drawn independently.

– Chi-square analysis should not be conducted when the expected or theoretical frequency in any of the cells is less than five.

– If the number of observations in any cell is less than 10, or if the table has two rows and two columns (a 2×2 table), a correction factor should be applied (Yates's chi-squared test).

Yates's correction for continuity :

Test assumptions

Yates's correction for continuity :

Idea: correcting the error introduced by assuming that the discrete probabilities of frequencies in the table can be approximated by a continuous distribution (chi-squared).

Test assumptions

Yates's correction for continuity :

Idea: correcting the error introduced by assuming that the discrete probabilities of frequencies in the table can be approximated by a continuous distribution (chi-squared).

Hence, the effect of Yates's correction is to prevent overestimation of statistical significance for small data.

Yates's correction for continuity :

Idea: correcting the error introduced by assuming that the discrete probabilities of frequencies in the table can be approximated by a continuous distribution (chi-squared).

Hence, the effect of Yates's correction is to prevent overestimation of statistical significance for small data.

The following is Yates's corrected version of Pearson's chi-squared statistics:

$$\chi^2_{\text{Yates}} = \sum_{i=1}^{N} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

where: O_i = an observed frequency E_i = an expected (theoretical) frequency, asserted by the null hypothesis N = number of distinct events

Test assumptions

Yates's correction for continuity :

Idea: correcting the error introduced by assuming that the discrete probabilities of frequencies in the table can be approximated by a continuous distribution (chi-squared).

Hence, the effect of Yates's correction is to prevent overestimation of statistical significance for small data.

The following is Yates's corrected version of Pearson's chi-squared statistics:

$$\chi^2_{\text{Yates}} = \sum_{i=1}^{N} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

where: O_i = an observed frequency E_i = an expected (theoretical) frequency, asserted by the null hypothesis N = number of distinct events

In some cases, Yates's correction may adjust too far, and so its current use is limited.

Example

Internet usage	Gender		Row total
	Male	Female	
Light	5	10	15
Heavy	10	5	15
Column total	15	15	

The value of Yates χ^2 is calculated as:

$$\chi^2 = (\mid 5 - 7.5 \mid -0.5)^2/7.5 + (\mid 10 - 7.5 \mid -0.5)^2/7.5 + (\mid 10 - 7.5 \mid -0.5)^2/7.5 + (\mid 5 -$$
$$= 2.133$$

Critical value at 5%: $\chi^2_{1,0.95} = 3.841$

Result: H_0 is not rejected

Test assumptions

Note, that Yates's correction is related to **continuity correction**

Test assumptions

Note, that Yates's correction is related to **continuity correction**

If a random variable X has a binomial distribution with parameters n and p, i.e., X is distributed as the number of "successes" in n independent Bernoulli trials with probability p of success on each trial, then

$$P(X \leq x) = P(X < x + 1)$$

for any $x \in \{0, 1, 2, \dots n\}$.

Test assumptions

Note, that Yates's correction is related to **continuity correction**

If a random variable X has a binomial distribution with parameters n and p, i.e., X is distributed as the number of "successes" in n independent Bernoulli trials with probability p of success on each trial, then

$$P(X \leq x) = P(X < x + 1)$$

for any $x \in \{0, 1, 2, \ldots n\}$.

If np and $np(1-p)$ are large (sometimes taken as both ≥ 5), then the probability above is fairly well approximated by

$$P(Y \leq x + 1/2)$$

where Y is a normally distributed random variable with the same expected value and the same variance as X, i.e., $\mathrm{E}(Y) = np$ and $\mathrm{var}(Y) = np(1-p)$.

This addition of $1/2$ to x is a continuity correction.

Goodness-of-fit

The chi-square statistic can also be used in goodness-of-fit tests to determine whether certain models fit the observed data.

These tests are conducted by calculating the significance of sample deviations from assumed theoretical (expected) distributions and can be performed on cross-tabulations as well as on frequencies (one-way tabulations).

Goodness-of-fit

The chi-square statistic can also be used in goodness-of-fit tests to determine whether certain models fit the observed data.

These tests are conducted by calculating the significance of sample deviations from assumed theoretical (expected) distributions and can be performed on cross-tabulations as well as on frequencies (one-way tabulations).

(i) Equal proportion hypothesis:

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5$$

H_a : at least one p_i not equal

The alternative hypothesis is that at least one of the proportions is different from the others.

Goodness-of-fit

The chi-square statistic can also be used in goodness-of-fit tests to determine whether certain models fit the observed data.

These tests are conducted by calculating the significance of sample deviations from assumed theoretical (expected) distributions and can be performed on cross-tabulations as well as on frequencies (one-way tabulations).

(i) Equal proportion hypothesis:

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5$$

H_a : at least one p_i not equal

The alternative hypothesis is that at least one of the proportions is different from the others.

(ii) Unequal proportions.

$$H_0 : p_1 = 0.2, p_2 = 0.65, p_3 = 0.15$$

H_a : at least one p_i not equal to expected value

The alternative hypothesis is that at least one of the proportions is different from the others.

Goodness-of-fit

The chi-square statistic can also be used in goodness-of-fit tests to determine whether certain models fit the observed data.

These tests are conducted by calculating the significance of sample deviations from assumed theoretical (expected) distributions and can be performed on cross-tabulations as well as on frequencies (one-way tabulations).

(i) Equal proportion hypothesis:

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5$$

H_a : at least one p_i not equal

The alternative hypothesis is that at least one of the proportions is different from the others.

(ii) Unequal proportions.

$$H_0 : p_1 = 0.2, p_2 = 0.65, p_3 = 0.15$$

H_a : at least one p_i not equal to expected value

The alternative hypothesis is that at least one of the proportions is different from the others.

We calculate the test statistic using the formula below:

$$\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{n-1}$$

Example

Consider a categorical variable which is the flavors of candy.

We collect a random sample of ten bags. Each bag has 100 pieces of candy and five flavors. Our hypothesis is that the proportions of the five flavors in each bag are the same.

Each bag has 100 pieces of candy. Each bag has five flavors of candy. We expect to have equal numbers for each flavor: $100 / 5 = 20$ pieces of candy in each flavor from each bag.

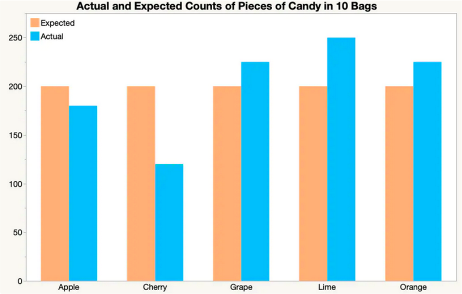
Example

Consider a categorical variable which is the flavors of candy.

We collect a random sample of ten bags. Each bag has 100 pieces of candy and five flavors. Our hypothesis is that the proportions of the five flavors in each bag are the same.

Each bag has 100 pieces of candy. Each bag has five flavors of candy. We expect to have equal numbers for each flavor: 100 / 5 = 20 pieces of candy in each flavor from each bag.

For 10 bags in our sample, we expect 10 x 20 = 200 pieces of candy in each flavor.



Are the number of pieces “close enough” for us to conclude that across many bags there are the same number of pieces for each flavor?

Example

Let's start by listing what we expect if each bag has the same number of pieces for each flavor. Above, we calculated this as 200 for 10 bags of candy.

Flavor	Number of Pieces of Candy (10 bags)	Expected Number of Pieces of Candy	Observed-Expected	Squared Difference	Squared Difference / Expected Number
Apple	180	200	$180 - 200 = -20$	400	$400/200 = 2$
Lime	250	200	$250 - 200 = 50$	2500	$2500/200 = 12.5$
Cherry	120	200	$120 - 200 = -80$	6400	$6400/200 = 32$
Orange	225	200	$225 - 200 = 25$	625	$625/200 = 3.125$
Grape	225	200	$225 - 200 = 25$	625	$625/200 = 3.125$

Example

Let's start by listing what we expect if each bag has the same number of pieces for each flavor. Above, we calculated this as 200 for 10 bags of candy.

Flavor	Number of Pieces of Candy (10 bags)	Expected Number of Pieces of Candy	Observed-Expected	Squared Difference	Squared Difference / Expected Number
Apple	180	200	$180 - 200 = -20$	400	$400/200 = 2$
Lime	250	200	$250 - 200 = 50$	2500	$2500/200 = 12.5$
Cherry	120	200	$120 - 200 = -80$	6400	$6400/200 = 32$
Orange	225	200	$225 - 200 = 25$	625	$625/200 = 3.125$
Grape	225	200	$225 - 200 = 25$	625	$625/200 = 3.125$

Finally, we add the numbers in the final column to calculate our test statistic:

2+12.5+32+3.125+3.125=52.75

Critical value: $\chi^2_{5-1=4,0.95} = is 9.488$

Since $52.75 > 9.488$, we reject the null hypothesis that the proportions of flavors of candy are equal.

1) **Phi coefficient** (ϕ)

– works for the special case of a table with two rows and two columns (a 2×2 table)

For a sample of size n, this statistic is calculated as:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

Strength of association

1) **Phi coefficient** (ϕ)

– works for the special case of a table with two rows and two columns (a 2×2 table)

For a sample of size n, this statistic is calculated as:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

It takes the value of 0 when there is no association, which would be indicated by a chi-square value of 0 as well.

When the variables are perfectly associated, phi assumes the value of 1 and all the observations fall just on the main or minor diagonal.

1) **Phi coefficient** (ϕ)

– works for the special case of a table with two rows and two columns (a 2×2 table)

For a sample of size n, this statistic is calculated as:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

It takes the value of 0 when there is no association, which would be indicated by a chi-square value of 0 as well.

When the variables are perfectly associated, phi assumes the value of 1 and all the observations fall just on the main or minor diagonal.

Example: in our case, because the association was not significant at the 0.05 level, we would not normally compute the phi value. However, for the purpose of illustration, the value of phi is:

$$\phi = \sqrt{3.333/30} = 0.333$$

Thus, the association is not very strong.

Strength of association

2) **Contingency coefficient** (C)

– a more general case involving a table of any size

This index is also related to chi-square, as follows:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Strength of association

2) **Contingency coefficient** (C)

– a more general case involving a table of any size

This index is also related to chi-square, as follows:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

The contingency coefficient varies between 0 and 1.

The value of 0 occurs in the case of no association (i.e. the variables are statistically independent), but the maximum value of 1 is never achieved.

The maximum value of the contingency coefficient *depends on the size of the table* (number of rows and number of columns), hence, it should be used only to compare tables of the *same size*.

2) **Contingency coefficient** (C)

– a more general case involving a table of any size

This index is also related to chi-square, as follows:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

The contingency coefficient varies between 0 and 1.

The value of 0 occurs in the case of no association (i.e. the variables are statistically independent), but the maximum value of 1 is never achieved.

The maximum value of the contingency coefficient *depends on the size of the table* (number of rows and number of columns), hence, it should be used only to compare tables of the *same size*.

Example: The value of the contingency coefficient:

$$C = \sqrt{3.333/(3.333 + 30)} = 0.31$$

This value of C indicates that the association is not very strong.

3) **Cramer's** V

– a modified version of the phi correlation coefficient used in tables larger than 2×2.

For a table with r rows and c columns:

$$V = \sqrt{\frac{\phi^2}{\min(r-1),(c-1)}}$$

or

$$V = \sqrt{\frac{\chi^2/n}{\min(r-1),(c-1)}}$$

3) **Cramer's** V

– a modified version of the phi correlation coefficient used in tables larger than 2×2.

For a table with r rows and c columns:

$$V = \sqrt{\frac{\phi^2}{\min(r-1),(c-1)}}$$

or

$$V = \sqrt{\frac{\chi^2/n}{\min(r-1),(c-1)}}$$

When phi is calculated for a table larger than 2×2, it has no upper limit. Cramer's V is obtained by adjusting phi for either the number of rows or the number of columns in the table, based on which of the two is smaller. Hence, V will range from 0 to 1.

3) **Cramer's** V

– a modified version of the phi correlation coefficient used in tables larger than 2×2.

For a table with r rows and c columns:

$$V = \sqrt{\frac{\phi^2}{\min(r-1),(c-1)}}$$

or

$$V = \sqrt{\frac{\chi^2/n}{\min(r-1),(c-1)}}$$

When phi is calculated for a table larger than 2×2, it has no upper limit. Cramer's V is obtained by adjusting phi for either the number of rows or the number of columns in the table, based on which of the two is smaller. Hence, V will range from 0 to 1.

Example: The value of Cramer's V:

$$V = \sqrt{(3.333/30)/1 = 0.333}$$

Thus, the association is not very strong.

As can be seen, in this case $V = \phi$, which is always the case for a 2×2 table.

Strength of association

3) **Cramer's** V

– a modified version of the phi correlation coefficient used in tables larger than 2×2.

For a table with r rows and c columns:

$$V = \sqrt{\frac{\phi^2}{\min(r-1),(c-1)}}$$

or

$$V = \sqrt{\frac{\chi^2/n}{\min(r-1),(c-1)}}$$

Strength of association

3) **Cramer's** V

– a modified version of the phi correlation coefficient used in tables larger than 2×2.

For a table with r rows and c columns:

$$V = \sqrt{\frac{\phi^2}{\min(r-1),(c-1)}}$$

or

$$V = \sqrt{\frac{\chi^2/n}{\min(r-1),(c-1)}}$$

When phi is calculated for a table larger than 2×2, it has no upper limit. Cramer's V is obtained by adjusting phi for either the number of rows or the number of columns in the table, based on which of the two is smaller. Hence, V will range from 0 to 1.

3) **Cramer's** V

– a modified version of the phi correlation coefficient used in tables larger than 2×2.

For a table with r rows and c columns:

$$V = \sqrt{\frac{\phi^2}{\min(r-1),(c-1)}}$$

or

$$V = \sqrt{\frac{\chi^2/n}{\min(r-1),(c-1)}}$$

When phi is calculated for a table larger than 2×2, it has no upper limit. Cramer's V is obtained by adjusting phi for either the number of rows or the number of columns in the table, based on which of the two is smaller. Hence, V will range from 0 to 1.

Example: The value of Cramer's V:

$$V = \sqrt{(3.333/30)/1 = 0.333}$$

Thus, the association is not very strong.

As can be seen, in this case $V = \phi$, which is always the case for a 2×2 table.

Strength of association

4) **Lambda coefficient**

The lambda coefficient assumes that the variables are measured on a nominal scale.

Asymmetric lambda – measures the percentage improvement in predicting the value of the dependent variable, given the value of the independent variable.

Lambda also varies between 0 and 1. A value of 1 indicates that the prediction can be made without error.

Asymmetric lambda can be computed for each of the variables (treating it as the dependent variable). In general, the two asymmetric lambdas are likely to be different, since the marginal distributions are not usually the same.

Symmetric lambda – does not make an assumption about which variable is dependent. It measures the overall improvement when prediction is done in both directions.

Example: The value of asymmetric lambda, with usage as the dependent variable, is 0.333. This indicates that knowledge of gender increases our predictive ability by the proportion of 0.333 , i.e. a 33% improvement.

The symmetric lambda is also 0.333 .

Strength of association

5) **Non-parametric coefficients of association**

Note that in the calculation of the chi-square statistic, the variables are treated as being meaured only on a nominal scale.

Other statistics, such as tau *b*, tau *c* and gamma, are available to measure association between two *ordinal-level variables*.

5) **Non-parametric coefficients of association**

Note that in the calculation of the chi-square statistic, the variables are treated as being meaured only on a nominal scale.

Other statistics, such as tau *b*, tau *c* and gamma, are available to measure association between two *ordinal-level variables*.

All these statistics use information about the ordering of categories of variables by considering every possible pair of cases in the table.

Each pair is examined to determine whether its relative ordering on the first variable is the same as its relative ordering on the second variable (concordant), the ordering is reversed (discordant), or the pair is tied.

Strength of association

The manner in which the ties are treated is the basic difference between these statistics:

tau b – makes an adjustment for ties and is the most appropriate when the table of variables is square.

tau c – makes an adjustment for ties and is most appropriate when the table of variables is rectangle.

Gamma – does not make an adjustment for ties.

Strength of association

The manner in which the ties are treated is the basic difference between these statistics:

tau *b* – makes an adjustment for ties and is the most appropriate when the table of variables is square.

tau *c* – makes an adjustment for ties and is most appropriate when the table of variables is rectangle.

Gamma – does not make an adjustment for ties.

tau *b* value varies between +1 and -1 . Thus the direction (positive or negative) as well as the strength (how close the value is to 1) of the relationship can be determined.

Gamma also varies between +1 and -1 and generally has a higher numerical value than tan *b* or tau *c*.

Example: for our case as gender is a nominal variable, it is not appropriate to calculate ordinal statistics.

Cross-tabulation in practice

While conducting cross-tabulation analysis in practice, it is useful to proceed through the following steps:

Cross-tabulation in practice

While conducting cross-tabulation analysis in practice, it is useful to proceed through the following steps:

1. Test the null hypothesis that there is no association between the variables using the chisquare statistic. If you fail to reject the null hypothesis, then there is no relationship.

Cross-tabulation in practice

While conducting cross-tabulation analysis in practice, it is useful to proceed through the following steps:

1. Test the null hypothesis that there is no association between the variables using the chisquare statistic. If you fail to reject the null hypothesis, then there is no relationship.

2. If H_0 is rejected, then determine the strength of the association using an appropriate statistic (phi coefficient, contingency coefficient, Cramer's V, lambda coefficient, or other statistics).

Cross-tabulation in practice

While conducting cross-tabulation analysis in practice, it is useful to proceed through the following steps:

1. Test the null hypothesis that there is no association between the variables using the chisquare statistic. If you fail to reject the null hypothesis, then there is no relationship.

2. If H_0 is rejected, then determine the strength of the association using an appropriate statistic (phi coefficient, contingency coefficient, Cramer's V, lambda coefficient, or other statistics).

3. If H_0 is rejected, interpret the pattern of the relationship by computing the percentages in the direction of the independent variable, across the dependent variable.

Cross-tabulation in practice

While conducting cross-tabulation analysis in practice, it is useful to proceed through the following steps:

1. Test the null hypothesis that there is no association between the variables using the chisquare statistic. If you fail to reject the null hypothesis, then there is no relationship.

2. If H_0 is rejected, then determine the strength of the association using an appropriate statistic (phi coefficient, contingency coefficient, Cramer's V, lambda coefficient, or other statistics).

3. If H_0 is rejected, interpret the pattern of the relationship by computing the percentages in the direction of the independent variable, across the dependent variable.

4. If the variables are treated as ordinal rather than nominal, use tau b, tau c or gamma as the test statistic. If H_0 is rejected, then determine the strength of the association using the magnitude, and the direction of the relationship using the sign of the test statistic.

Cross-tabulation in practice

While conducting cross-tabulation analysis in practice, it is useful to proceed through the following steps:

1. Test the null hypothesis that there is no association between the variables using the chisquare statistic. If you fail to reject the null hypothesis, then there is no relationship.

2. If H_0 is rejected, then determine the strength of the association using an appropriate statistic (phi coefficient, contingency coefficient, Cramer's V, lambda coefficient, or other statistics).

3. If H_0 is rejected, interpret the pattern of the relationship by computing the percentages in the direction of the independent variable, across the dependent variable.

4. If the variables are treated as ordinal rather than nominal, use tau b, tau c or gamma as the test statistic. If H_0 is rejected, then determine the strength of the association using the magnitude, and the direction of the relationship using the sign of the test statistic.

5. Translate the results of hypothesis testing, strength of association and pattern of association into managerial implications and recommendations.

Summary

Summary:

1) Contingency tables with three variables

– refine the association

– spurious association

– reveal suppressed association

– no change in the initial association.

Summary

Summary:

1) Contingency tables with three variables

– refine the association

– spurious association

– reveal suppressed association

– no change in the initial association.

2) Chi-square statistic

– should be estimated only on counts of data

– the observations are drawn independently

– should not be conducted when the expected or theoretical frequency in any of the cells is less than five

– if the number of observations in any cell is less than 10, or for 2×2 table, Yates's correction can be applied