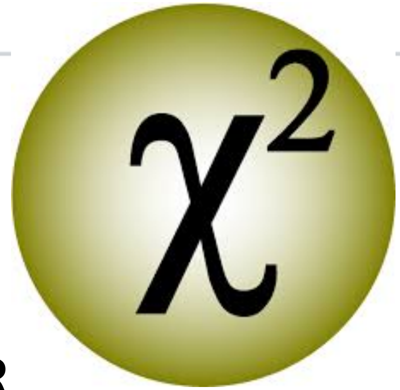
# Associations between categorical variables

| |  |  |
|---|---|---|
| ♂ |  104 |  156 |
| ♀ |  177 |  83 |

Dr. Margriet A. Groen

# Outline: Aims and Objectives



- Introduction to Pearson's $\chi^2$ test
  What it is and how/when it is applied

- Carrying out Pearson's $\chi^2$ test by hand and with R
  And measuring the strength of the association (effect size)

- Interpreting the source of difference
- Fisher's exact test
- Reporting Pearson's $\chi^2$ test in APA style

- Frequencies and percentages
- More than 2x2 – partitioning and combining

# What is Chi-Square?

Chi-square tests are ***non-parametric*** tests of inference for ***categorical*** data.


## ➢ Test of independence / Pearson's chi-square (2x2)

- Measures the relationship between two (or more) nominal variables: Are observations contingent upon another categorical variable?
- Tests whether the frequency counts could be expected by chance or whether there is a relationship between the categorical variables

# Chi-square: Test of independence

Tests whether two nominal variables are associated

**Examples**

- *Is gender associated with preferred subject?*
- *Is ownership of a dog associated with residence (country/city)?*
- *Is smoking associated with drinking?*

**Null hypothesis**

There is no association between the two variables

**Alternative hypothesis**

The two variables are associated

# Calculating *chi-square* by hand

# So how do we conduct the test?

- **Construct a contingency table** representing frequencies of both nominal variables (example data sourced from Howitt & Cramer, 2017)

*Researchers are interested is assessing the relationship between children's record of fighting in school and their preference for a violent or non-violent TV programme*

| | | Fight in school? | |
|---|---|---|---|
| | | **Yes** | **No** |
| **TV programme preference** | **Violent** | 40 | 15 |
| | **Non-violent** | 30 | 70 |

# Calculating $\chi^2$ by hand

| | | Fight in school? | |
|---|---|---|---|
| | | Yes | No |
| TV programme preference | Violent | 40 | 15 |
| | Non-violent | 30 | 70 |

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Σ = Sum of

O = Observed frequencies (40, 15, 30, 70)

E = Expected Frequencies = $\frac{Column\ Total\ (CT) \times Row\ Total\ (RT)}{N}$

# Calculating *expected frequencies* by hand

| | Fighters | Non-fighters | TOTAL |
|---|---|---|---|
| Violent TV | 40 | 15 | 55 |
| Non-violent TV | 30 | 70 | 100 |
| TOTAL | 70 | 85 | 155 |

$$E = \frac{CT \times RT}{N}$$

***Fighters/Violent TV*** = (**70** fighters X **55** violent) / **155** = **24.84**

# Calculating *expected frequencies* by hand

| Observed *(Expected)* | **Fighters** | **Non-fighters** | **TOTAL** |
|---|---|---|---|
| **Violent TV** | 40 ***(24.84)*** | 15 ***(30.16)*** | **55** |
| **Non-violent TV** | 30 ***(45.16)*** | 70 ***(54.84)*** | **100** |
| **TOTAL** | **70** | **85** | **155** |

$$E = \frac{CT \times RT}{N}$$

***Fighters/Violent TV*** = (**70** fighters X **55** violent) / **155** = ***24.84***

***Fighters/Non-violent TV*** = (**70** fighters X **100** non-violent) / **155** = ***45.16***

***Non-fighters/Violent TV*** = (**85** non-fighters X **55** violent) / **155** = ***30.16***

***Non-fighters/Non-violent TV*** = (**85** non-fighters X **100** non-violent) / **155 =** ***54.84***

# Calculating $\chi^2$ by hand

- $\chi^2 = \sum \frac{(O-E)^2}{E}$

| Cell | Observed | Expected | $(O-E)^2$ | $(O-E)^2 / E$ |
|---|---|---|---|---|
| Fighters/Violent | 40 | 24.84 | 229.83 | 9.25 |
| Fighters/Non-violent | 30 | 45.16 | 229.83 | 5.09 |
| Non-fighters/Violent | 15 | 30.16 | 229.83 | 7.62 |
| Non-fighters/Non-violent | 70 | 54.84 | 229.83 | 4.19 |

| Observed *(Expected)* | **Fighters** | **Non-fighters** | **TOTAL** |
|---|---|---|---|
| **Violent TV** | 40 ***(24.84)*** | 15 ***(30.16)*** | **55** |
| **Non-violent TV** | 30 ***(45.16)*** | 70 ***(54.84)*** | **100** |
| **TOTAL** | **70** | **85** | **155** |

# Calculating $\chi^2$ by hand

***Degrees of freedom***

$$df = (number\ of\ rows - 1)\times(numbers\ of\ columns - 1)$$

$$df = (2-1)\times(2-1)$$

***df = 1***

# What are ‘degrees of freedom’?

… the number of independent pieces of information that went into calculating the estimate

… or the number of values that are free to vary

Example:

Question: Pick a set of numbers that have a mean of 10.

Answer:

| | | |
|---|---|---|
| 9 | 10 | … |
| 8 | 10 | … |
| 5 | 10 | … |

# Significance

| Degrees of freedom | 5% | 1% |
|---|---|---|
| 1 (1-tailed)[a] | 2.705 | 5.412 |
| 1 (2-tailed) | 3.841 | 6.635 |
| 2 (2-tailed) | 5.992 | 9.210 |
| 3 (2-tailed) | 7.815 | 11.345 |
| 4 (2-tailed) | 9.488 | 13.277 |
| 5 (2-tailed) | 11.070 | 15.086 |
| 6 (2-tailed) | 12.592 | 16.812 |
| 7 (2-tailed) | 14.067 | 18.475 |
| 8 (2-tailed) | 15.507 | 20.090 |
| 9 (2-tailed) | 16.919 | 21.666 |
| 10 (2-tailed) | 18.307 | 23.209 |
| 11 (2-tailed) | 19.675 | 24.725 |
| 12 (2-tailed) | 21.026 | 26.217 |

$\chi^2 = 26.15$

df = 1

***p* < .01**

# Assumptions

- Independence
  Data cannot be related (must use distinct nominal categories) – *cannot fall into both categories. Between subject designs: 1 response from each participant*
- Raw frequencies
  Chi-square should be conducted on raw frequencies, not percentages (more on this later)
- Sample/cell size
  No expected cell frequencies should be less than 1 and no more than 20% of the cells should be less than 5 (Cochran, 1954).
  - *Can collapse categories (Meat eaters/vegetarians/vegans)*
  - *Report Fisher's Exact test*

# Descriptive statistics, effect size and variance accounted for

- Percentages
  Chi-square should be conducted on raw frequencies, not percentages. However, percentages are useful to report in addition to raw frequencies
- Cramer's V
  Measure of effect size
- Variance accounted for
  We can square the effect size to see how much variance in one variable can be accounted for by the other variable

# Standardised residuals

- Help determine which cells are contributing to the 'significant association'.
- They are *z*-scores indicating how many SD's above or below the expected count, an observed count is (thus indicating how much they differ).

$\pm 1.96$ $p < .05$

$\pm 2.58$ $p < .01$

$\pm 3.29$ $p < .001$

| Observed *(Expected)* | **Fighters** | **Non-fighters** | **TOTAL** |
|---|---|---|---|
| **Violent TV** | 40 ***(24.84)*** | 15 ***(30.16)*** | **55** |
| **Non-violent TV** | 30 ***(45.16)*** | 70 ***(54.84)*** | **100** |
| **TOTAL** | **70** | **85** | **155** |

# Reporting chi-square

# Reporting results

There was a significant association between school fighting and TV programme preference (violent versus non-violent), $\chi^2$ (1, *N* = 155) = 26.16, *p* < .001, *Cramers V* = .41. Above expectations, 73% (40 out of 55) of children who preferred violent TV programmes, had also fought in school (*z* = 3.00, *p* < .01), and significantly less than expected had not fought *(z* = -2.76, *p* < .01). Conversely, 70% of children preferring non-violent TV programmes (70 out of 100), had not engaged in fighting, more than expected (*z* = 2.05, *p* < .05) and those who had fought were below expectations *(z* = -2.26, *p* < .05). Analysis showed 17% of the variance in school fighting could be accounted for by TV programme preference.

# Issues with chi-square

# Frequencies and percentages

*Raw frequencies/counts* should always be used, **not** *percentages*

- Pearson's $\chi^2$: using proportions/percentages can drastically change $\chi^2$ and significance value.

| | Male | Female | Total |
|---|---|---|---|
| Science | 70 | 46 | **116** |
| Literacy | 34 | 50 | **84** |
| **Total** | **104** | **96** | **200** |

| | Male | Female | Total |
|---|---|---|---|
| Science | 35 | 23 | **58** |
| Literacy | 17 | 25 | **42** |
| **Total** | **52** | **48** | **100** |

$\chi^2$ (1, $N$ = 200) = 7.71, $p$ = .006

$\chi^2$ (1, $N$ = 100) = 3.85, $p$ = .05

# Partitioning and combining categories

Larger contingency tables (categories have more than 2 levels) can be difficult to interpret. We can help understand the associations in a few different ways

- **Use standardized residuals** to determine main contributors
- **Partitioning:** Carry out multiple 2x2 chi-squares
  - Example: TV programme preferences (soap opera, crime drama, other) in male and female students

| | Soap | Crime |
|---|---|---|
| Males | | |
| Females | | |

| | Soap | Other |
|---|---|---|
| Males | | |
| Females | | |

| | Crime | Other |
|---|---|---|
| Males | | |
| Females | | |

- **Combine categories:** Alternatively, if it makes logical & theoretical sense, you can combine categories.
  - Example: Combine 'Soap opera' and 'Other' or combine 'Crime drama' and 'Other'.

# Summary

- Pearson's / test of independence chi-square investigates the association between two nominal variables.
- Assumptions of chi-square
- To understand the association we look at $\chi^2$, *p*, *Cramer's V* and standardized residuals
- Fisher's exact test should be used when assumptions re. minimum frequencies are not met
- We should always use raw frequencies and not percentages to avoid Type 1 or 2 errors
- When dealing with larger contingency tables, we can better understand our results using standardized residuals or partitioning/combining categories