Institute of Psychiatry, Psychology and Neuroscience

Dr Silia Vitoratou

Department: Biostatistics and Health Informatics

Topic materials:
Silia Vitoratou

Contributions:
Zahra Abdulla

Improvements:
Nick Beckley-Hoelscher
Kim Goldsmith
Professor Sabine Landau
Dr Silia Vitoratou

Module Title: Introduction to Statistics

Session Title: Equality of proportions: χ^2 tests

Topic title: Comparing groups I (parametric methods)

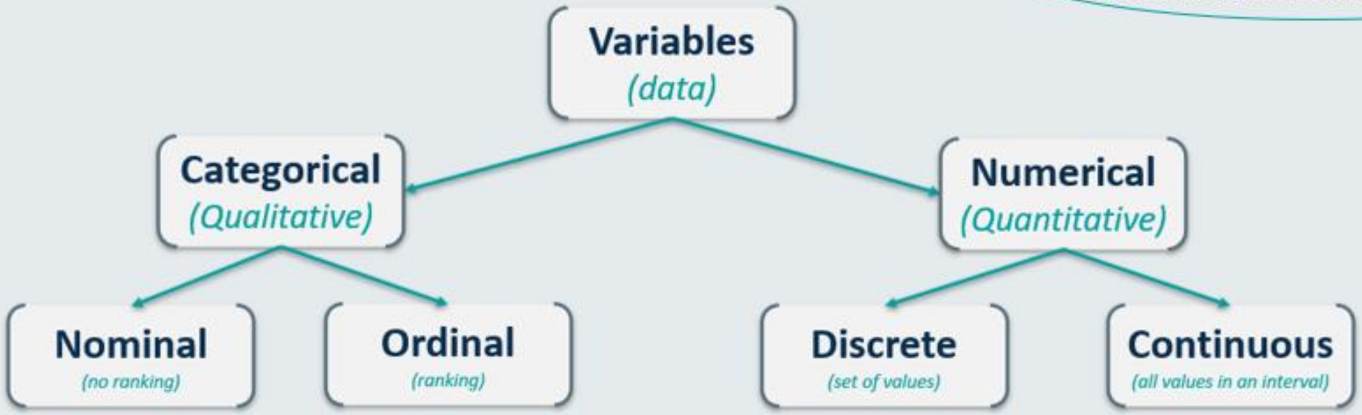
Learning Outcomes

- Learn when and how to use the χ^2 -tests for equality of proportions.
- Understand the assumptions of the various test of equality of proportions.
- Be able to conduct these tests in a statistical software.

Previously on ‘Introduction to Statistical Methods’…

Based on the type of each variable, we use different ways to describe the data.

Blue for symmetrical (normal)
Purple italics for skewed or discrete data



Ethnicity:
1= Asian
2= Black
3= White

Economic status:
1= Poor
2= Middle
3= Rich

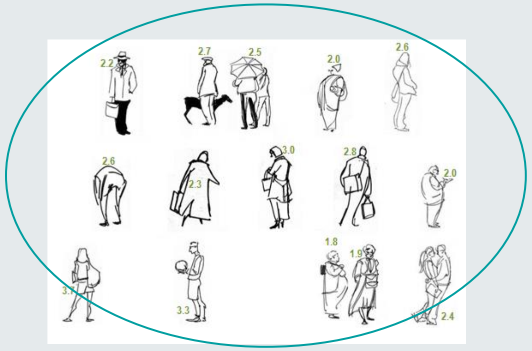
Number of children:
0, 1, 2, 3,…
(very specific values)

Height, weight, age:
(any value within range)

	Categorical	Numerical
• Descriptive indices	Frequencies (Percentages %)	location: mean, *median*, mode Dispersion: SD, *min-max*, range
• Charts/plots	Bar Chart	Histogram, Box plot

Equality of Means: The Three t-tests

<u>one sample t-test</u>



$H_0: \mu=\mu_0$

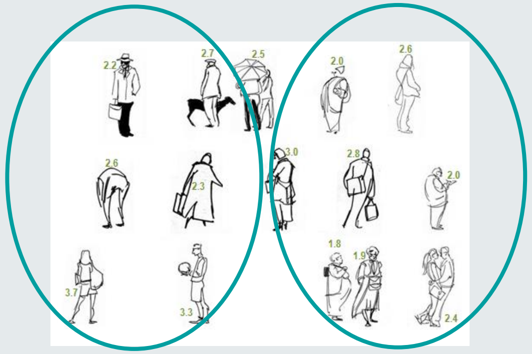
$H_a: \mu\neq\mu_0$

<u>Examples</u>

Difference from test value:

- age ≠ 25yo
- height ≠ 1.60cm
- weight ≠ 80kg

<u>independent samples t-test</u>



$H_0: \mu_A = \mu_B$

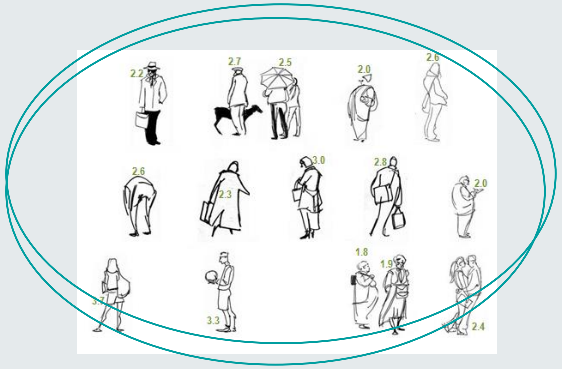
$H_a: \mu_A \neq \mu_B$

<u>Examples</u>

Difference in the means:

- young vs old
- males vs females
- City A vs City B

<u>paired samples t-test</u>



$H_0: \mu_1 = \mu_2$

$H_a: \mu_1 \neq \mu_2$

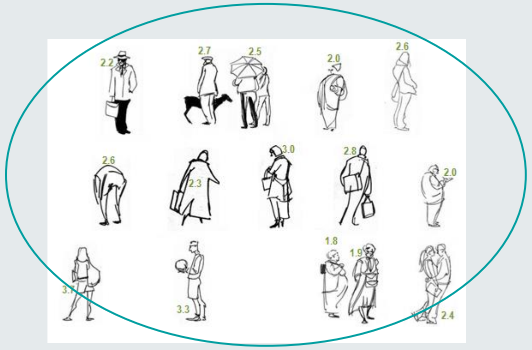
<u>Examples</u>

Difference in the means:

- before and after treatment
- twin studies
- matched cases vs controls

Equality of proportions: The three χ^2-tests

one sample χ^2-test



$H_0: \pi = \pi_0$

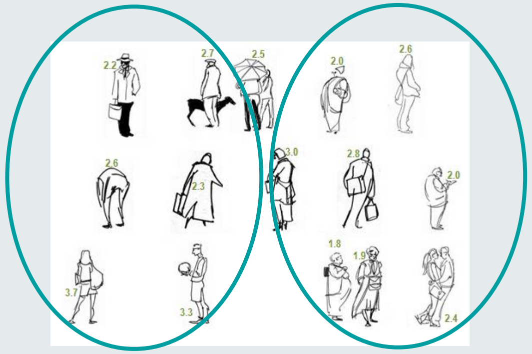
$H_a: \pi \neq \pi_0$

Examples

Difference from test value:

- % age 25yo ≠10%
- % taller ≠10%
- % gain weight ≠ 25%

independent samples χ^2-test



$H_0: \pi_A = \pi_B$

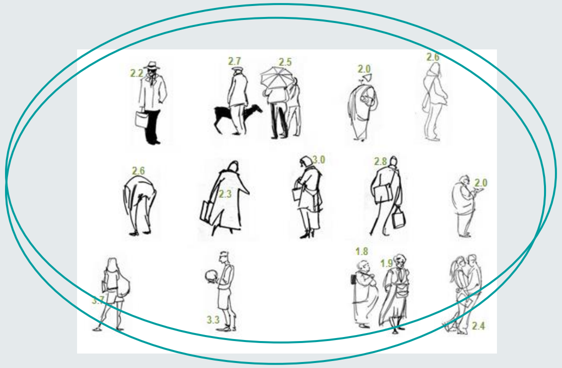
$H_a: \pi_A \neq \pi_B$

Examples

Difference in the proportions:

- % young vs %old
- % males vs %females
- %City A vs %City B

paired samples χ^2-test



$H_0: \pi_1 = \pi_2$

$H_a: \pi_1 \neq \pi_2$

Examples

Difference in the proportions:

- % before and %after treatment
- % twin studies
- % matched cases vs controls

Equality of Means: The Three t-tests

Hypotheses	**Suitable test**	**Decision**
H_0: is equal H_a: not equal	*test statistic*	p-value>0.05 do not reject the H_0 p-value≤0.05 reject the H_0

Hypotheses

H_0: $\mu=\mu_0$

H_a: $\mu\neq\mu_0$

One sample t-test

$$t = \frac{\bar{x}-\mu_0}{s.e.},\ \text{df=n-1}$$

$$s.e. = \sqrt{s^2/n}$$

In the population, is the population mean (μ) equal to a certain value (μ_0)?

Hypotheses

H_0: $\mu_A=\mu_B$

H_a: $\mu_A\neq\mu_B$

Independent samples t-test

$$t = \frac{\bar{x}_A-\bar{x}_B}{\sqrt{s_A^2/n_A+s_B^2/n_B}},\ \text{df=}\ n_A+n_B\text{-2}$$

In the population, is the mean of group A (μ_A) equal to the mean of group B (μ_B)?

Hypotheses

H_0: $\mu_1=\mu_2$

H_a: $\mu_1\neq\mu_2$

Paired samples t-test

$$t = \frac{\bar{x}_{diff}}{\sqrt{s_{diff}^2/n}},\ \text{df=n-1}$$

In the population, is the mean of a group in one condition (μ_1) equal to the mean of the same (or paired) group in another condition (μ_2)?

Equality of proportions: The three χ^2-tests

Hypotheses	**Suitable test**	**Decision**
H_0: % is equal H_a: % not equal	*test statistic*	p-value>0.05 do not reject the H_0 p-value≤0.05 reject the H_0

Hypotheses

H_0: $\pi=\pi_0$

H_a: $\pi\neq\pi_0$

One sample χ^2-test

$$\chi^2 = \sum \frac{(\mathrm{O}-\mathrm{E})^2}{\mathrm{E}}, \text{ df=c-1}$$

In the population, is the population proportion (π) equal to a certain value (π_0)?

Hypotheses

H_0: $\pi_A=\pi_B$

H_a: $\pi_A\neq\pi_B$

Pearson's χ^2 test

$$\chi^2 = \sum \frac{(\mathrm{O}-\mathrm{E})^2}{\mathrm{E}}, \text{ df=}(c_1\text{-1}) (c_2\text{-1})$$

In the population, is the proportion of group A (π_A) equal to the proportion of group B (π_B)?

Hypotheses

H_0: $\pi_1=\pi_2$

H_a: $\pi_1\neq\pi_2$

McNemar χ^2 test

$$\chi^2 = \frac{(b-c)^2}{b+c}, \text{ df=1}$$

In the population, is the proportion of a group in one condition (π_1) equal to the proportion of the same (or paired) group in another condition (π_2)?

Equality of proportions: The three χ^2-tests

Hypotheses	Suitable test	Decision
H_0: % is equal H_a: % not equal	*test statistic*	p-value>0.05 do not reject the H_0 p-value≤0.05 reject the H_0

Hypotheses	One sample χ^2-test
H_0: $\pi=\pi_0$ H_a: $\pi\neq\pi_0$	$\chi^2=\sum\frac{(O-E)^2}{E}$, df=c-1

In the population, is the population proportion (π) equal to a certain value (π_0)?

One Sample Chi-Square Test

When to use

To test if according to the current data, the proportion in the population equals a certain, pre-specified, value.

Hypotheses:

H_0: the proportion in the population equals a certain pre-specified value
H_a: the proportion in the population is different than a certain pre-specified value

Assumptions:

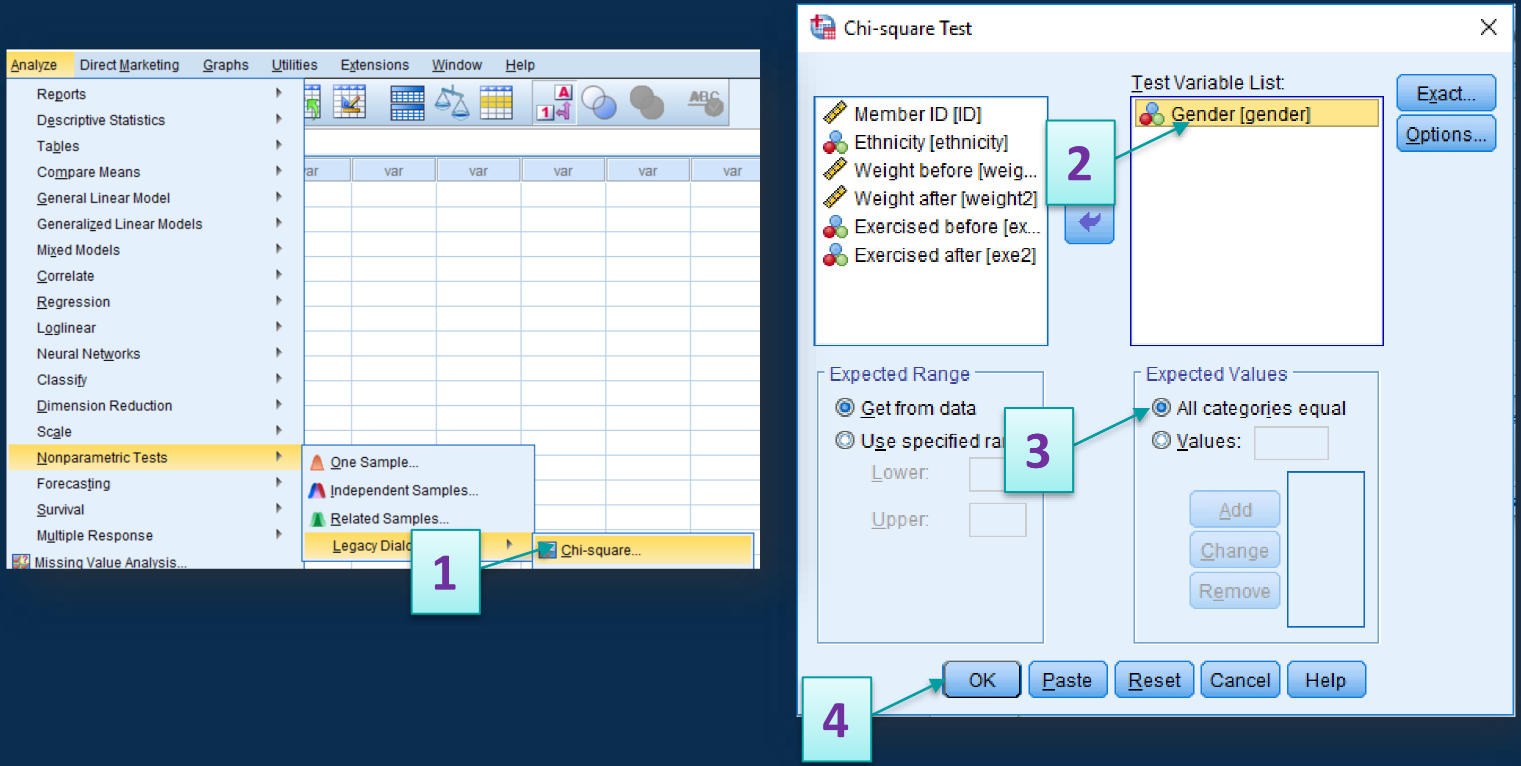
- The observations are randomly and independently drawn
- The number of cells with expected frequencies less than 5, are less than 20%
- The minimum expected frequency is at the very least 1.

SPSS Slide: ‘how to’

We start with testing whether men or women are more likely to come to the programme. In other words, is the proportion of men equal to the proportion of women?

Step 1: Use the appropriate test, here ‘one-sample chi-square test’.

Analyse -> Non-parametric tests -> legacy dialogs -> ‘Chi-square’



Add the variable of interest Gender) in to the ‘Variables box’.
Choose to Test ‘All categories are equal’
(If you have specific proportions to test these would be entered in values in the order of the coding of the variable in the main dataset)
Click ‘OK’

Output & Interpretation Slide

SPSS prints a table with descriptive statistics and one with the one sample t-test

Gender

	Observed N	Expected N	Residual
Female	143	150.0	-7.0
Male	157	150.0	7.0
Total	300		

Test Statistics

	Gender
Chi-Square	.653[a]
df	1
Asymp. Sig.	.419

a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 150.0.

Test statistic $\sum \frac{(\mathbf{O}-\mathbf{E})^2}{\mathbf{E}}$

Degrees of freedom: c-1

P-value

$$\sum \frac{(O-E)^2}{E} = \frac{(143-150)^2}{150} + \frac{(157-150)^2}{150} = 0.653$$

H_0: π=50%

H_a: $\pi \neq$50%

Based on our sample, the expected proportion of men joining in the programme is not different than the expected proportion of women (χ^2=0.653, df=1, p=0.419).

Output & Interpretation

Step 2: Check the suitability of the data, do the assumptions of the chi-square test hold?

Test Statistics

	Gender
Chi-Square	.653[a]
df	1
Asymp. Sig.	.419

a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 150.0.

Gender

	Observed N	Expected N	Residual
Female	143	150.0	-7.0
Male	157	150.0	7.0
Total	300		

For the test to work properly we need to have enough data for all 'cells' (categories). Specifically:

- Only up to 20% of the cells are allowed to have expected frequencies less than 5.
- The minimum expected frequency needs to be larger than 1.

Equality of proportions: The three χ^2-tests

Hypotheses	Suitable test	Decision
H_0: % is equal H_a: % not equal	*test statistic*	p-value>0.05 do not reject the H_0 p-value≤0.05 reject the H_0

Hypotheses

H_0: $\pi=\pi_0$

H_a: $\pi\neq\pi_0$

One sample χ^2-test

$$\chi^2 = \sum \frac{(O-E)^2}{E}, \text{ df=c-1}$$

In the population, is the population proportion (π) equal to a certain value (π_0)?

Hypotheses

H_0: $\pi_A=\pi_B$

H_a: $\pi_A\neq\pi_B$

Pearson's χ^2 test

$$\chi^2 = \sum \frac{(O-E)^2}{E}, \text{ df=}(c_1\text{-}1)\ (c_2\text{-}1)$$

In the population, is the proportion of group A (π_A) equal to the proportion of group B (π_B)?

Pearson's Chi-Square Test

<u>When to use</u>

To test if, according to the current data, the proportions in the population of one variable change based on another variable.

<u>Hypotheses:</u>

H_0: there is no association between the two variables
H_a: there is an association between the two variables

<u>Assumptions:</u>

- The observations are randomly and independently drawn
- The number of cells with expected frequencies less than 5, are less than 20%
- The minimum expected frequency is at the very least 1.
- The observations are not paired

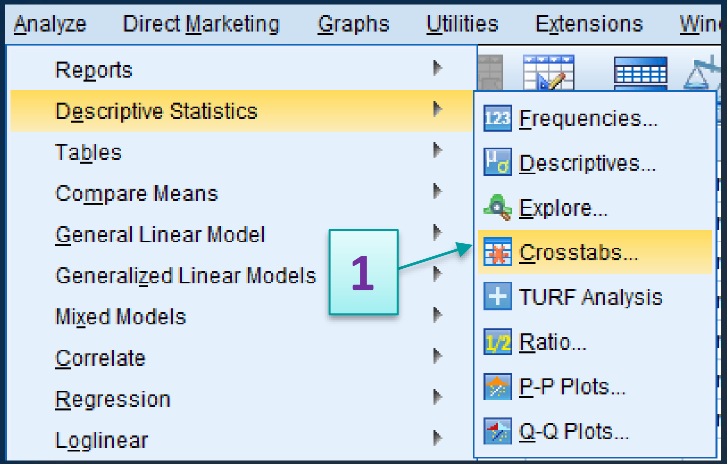
SPSS Slide: ‘how to’

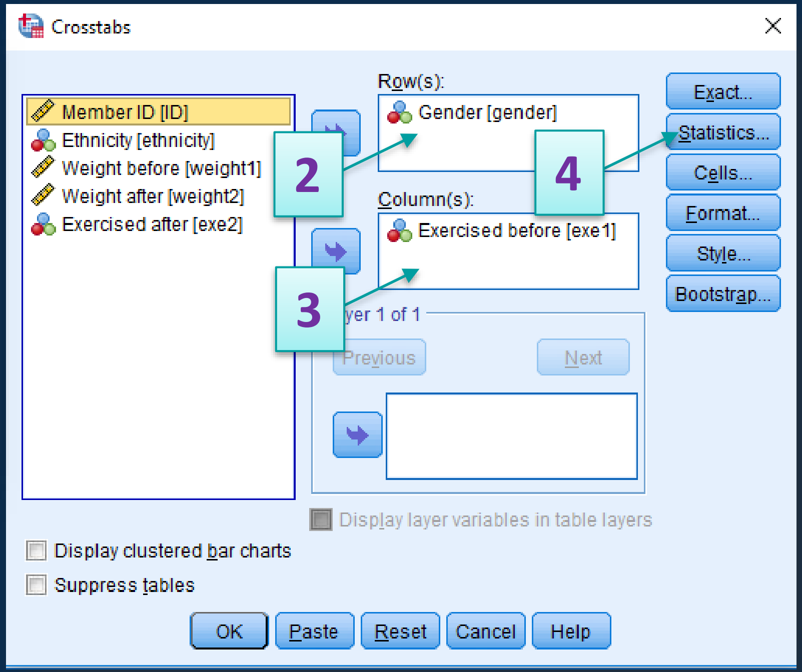
The next question is: do men exercise more than women prior to entering the programme?
Are the proportions of those exercised before the programme, different for men and women?

<u>Step 1</u>: Use the appropriate test, here: ‘Pearson’s chi-square test’.

Analyse -> Descriptive Statistics -> Crosstabs

$$H_0: \pi_M = \pi_F$$
$$H_a: \pi_M \neq \pi_F$$





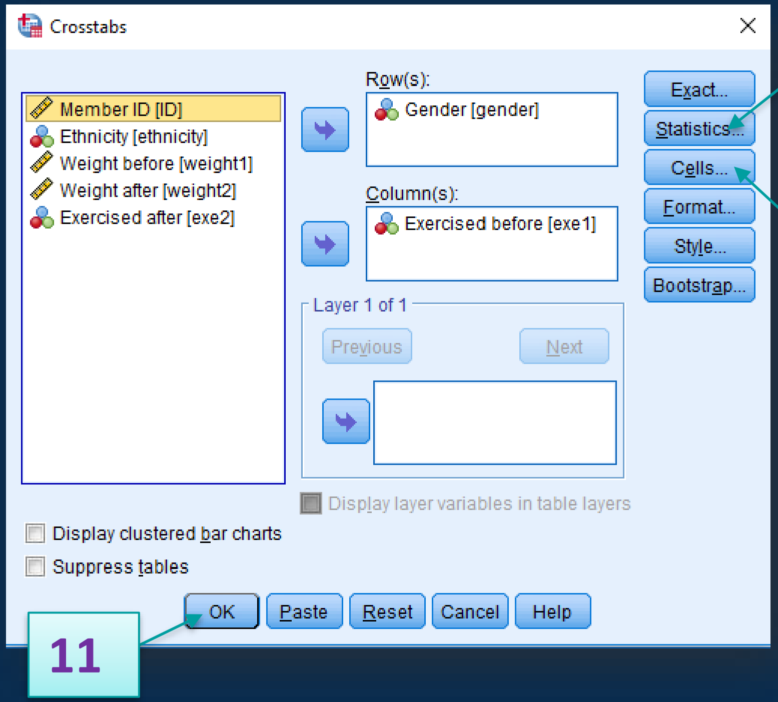
Add the variable of interest (Gender) in to the ‘Rows box’.
Add the second variable interest (Exe before) in the ‘columns box’.

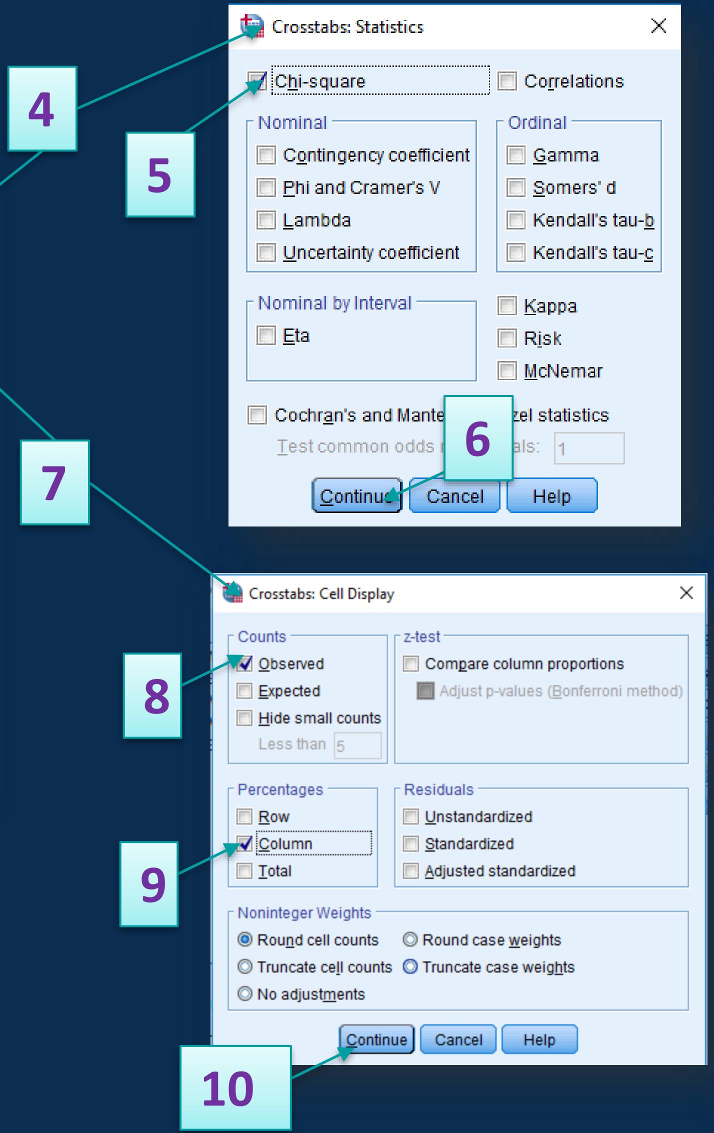
SPSS Slide: ‘how to’

Step 1: Use the appropriate test, here: ‘Pearson’s chi-square test’.

Click on ‘Statistics’ and choose ‘Chi square’
Click ‘continue’

Click on ‘cells’ choose ‘Observed’ and ‘column’ percentages.
Click ‘Continue’
Click ‘OK’





Choosing row or column percentages depends on the interpretations you are hoping to make.

Output and Interpretation

SPSS prints a double entry table with descriptive statistics and one with the χ^2-test

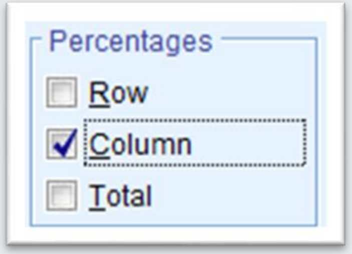
Gender * Exercised before Crosstabulation

			Exercised before		
			No	Yes	Total
Gender	Female	Count	119	24	143
		% within Exercised before	53.6%	30.8%	47.7%
	Male	Count	103	54	157
		% within Exercised before	46.4%	69.2%	52.3%
Total		Count	222	78	300
		% within Exercised before	100.0%	100.0%	100.0%

In our sample, we have

- 119 women who did not exercise before the programme
- 24 women who did exercise before the programme
- 103 men who did not exercise before the programme
- 54 men who did exercise before the programme.

Output and Interpretation



Gender * Exercised before Crosstabulation

			Exercised before		
			No	Yes	Total
Gender	Female	Count	119	24	143
		% within Exercised before	53.6%	30.8%	47.7%
	Male	Count	103	54	157
		% within Exercised before	46.4%	69.2%	52.3%
Total		Count	222	78	300
		% within Exercised before	100.0%	100.0%	100.0%

Add by column -> interpret by row ! **NEVER compare percentages which add up to 100%!**

Among the 'females', the proportion of those who did 'not exercise before' was higher than the proportion of those who 'exercise before' (53.6% versus 30.8%, respectively).

(Alternatively, we could have used the row proportions for men)

Output and Interpretation

Percentages
- [x] Row
- [ ] Column
- [ ] Total

Gender * Exercised before Crosstabulation

			Exercised before		
			No	Yes	Total
Gender	Female	Count	119	24	143
		% within Gender	83.2%	16.8%	100.0%
	Male	Count	103	54	157
		% within Gender	65.6%	34.4%	100.0%
Total		Count	222	78	300
		% within Gender	74.0%	26.0%	100.0%

Add by row -> interpret by column! **NEVER compare percentages which add up to 100%!**

Among those who did not ‘exercise before’, the proportion of ‘females’ was higher than the proportion of ‘males’ (83.2% versus 65.6%, respectively).

(Alternatively, we could have used the column proportions for those who do exercise)

Output and Interpretation

What essentially happens here, is that when you compare proportions which add up to 100%, then you essentially work with one of the variables. Therefore the association is not described.

On the contrary, when you compare proportions which do not add up to 100, you use information by both variables, thus you highlight the association!

Gender * Exercised before Crosstabulation

			Exercised before		
			No	Yes	Total
Gender	Female	Count	119	24	143
		% within Gender	83.2%	16.8%	100.0%
	Male	Count	103	54	157
		% within Gender	65.6%	34.4%	100.0%
Total		Count	222	78	300
		% within Gender	74.0%	26.0%	100.0%

Output and Interpretation

Recent example: people's wrong read of the benefit of the COVID vaccines against loss of life.

Let us assume a city has 1000 people and 80% are vaccinated against a virus.

Sadly over the course of a year, 30 people die due to the virus. Among them, 15 were vaccinated and 15 were not.

Headline: half of the people who died were vaccinated

FAKE NEWS

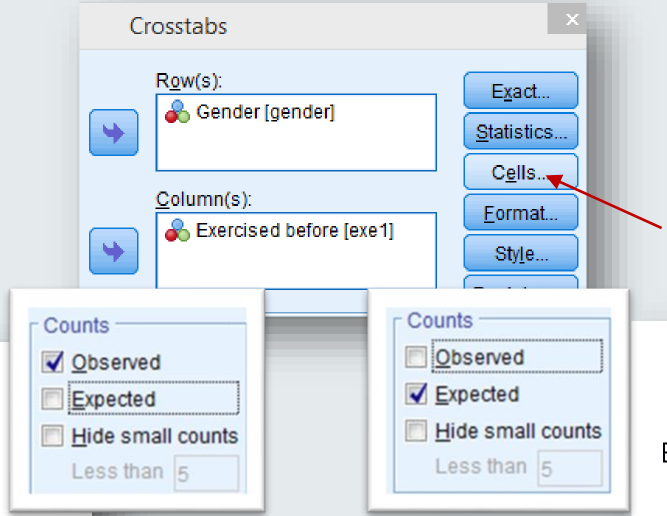
People naturally think: among those who died, 50% were vaccinated and 50% not. Thus equal proportions!

But we know, we do not compare proportions that add up to 100%

What we need to compare here, among those who died, it was the 15/800=2.25% of the vaccinated people, and the 7.5% of the unvaccinated people.

So we expect almost 2 out of 100 people who are vaccinated to be in danger, but almost 8 out of 100 people who are not vaccinated (4 times up!).

Output and Interpretation



Gender * Exercised before Crosstabulation

Count

		Exercised before No	Exercised before Yes	Total
Gender	Female	119	24	143
	Male	103	54	157
Total		222	78	300

Gender * Exercised before Crosstabulation

Expected Count

		Exercised before No	Exercised before Yes	Total
Gender	Female	105.8	37.2	143.0
	Male	116.2	40.8	157.0
Total		222.0	78.0	300.0

$$\sum \frac{(O-E)^2}{E} = \frac{(119-105.8)^2}{105.8} + \frac{(103-116.2)^2}{116.2} + \frac{(24-37.2)^2}{37.2} + \frac{(54-40.8)^2}{40.8}$$

$$= 12.07$$

$df=(c_1-1)\,(c_2-1)=1*1=1$

Output and Interpretation Slide

Gender * Exercised before Crosstabulation

			Exercised before No	Exercised before Yes	Total
Gender	Female	Count	119	24	143
		% within Gender	83.2%	16.8%	100.0%
	Male	Count	103	54	157
		% within Gender	65.6%	34.4%	100.0%
Total		Count	222	78	300
		% within Gender	74.0%	26.0%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	12.065[a]	1	.001		
Continuity Correction[b]	11.167	1	.001		
Likelihood Ratio	12.342	1	.000		
Fisher's Exact Test				.001	.000
Linear-by-Linear Association	12.024	1	.001		
N of Valid Cases	300				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 37.18.

b. Computed only for a 2x2 table

Among those who ‘exercised before’ the programme, the proportion of ‘males’ was higher than the proportion of ‘females’ (34.4% versus 16.8%, respectively). This difference was statistically significant (Pearson χ^2=12.065, df=1, p<0.001).

Therefore, we conclude that men tend to exercise (before the programme) more often than women, in the population. The variables ‘gender’ and ‘exe1’ are related.

Output and Interpretation

<u>Step 2</u>: Check the suitability of the data, do the assumptions of the chi-square test hold?

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	12.065[a]	1	.001		
Continuity Correction[b]	11.167	1	.001		
Likelihood Ratio	12.342	1	.000		
Fisher's Exact Test				.001	.000
Linear-by-Linear Association	12.024	1	.001		
N of Valid Cases	300				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 37.18.

b. Computed only for a 2x2 table

For the test to work properly we need to have enough data for all ‘cells’ (categories). Specifically:

- Only up to 20% of the cells are allowed to have expected frequencies less than 5.
- The minimum expected frequency needs to be larger than 1.

Equality of proportions: The three χ^2-tests

Hypotheses	Suitable test	Decision
H_0: % is equal H_a: % not equal	*test statistic*	p-value>0.05 do not reject the H_0 p-value≤0.05 reject the H_0

Hypotheses

H_0: $\pi=\pi_0$

H_a: $\pi\neq\pi_0$

One sample χ^2-test

$$\chi^2 = \sum \frac{(O-E)^2}{E}, \text{ df=c-1}$$

In the population, is the population proportion (π) equal to a certain value (π_0)?

Hypotheses

H_0: $\pi_A=\pi_B$

H_a: $\pi_A\neq\pi_B$

Pearson's χ^2 test

$$\chi^2 = \sum \frac{(O-E)^2}{E}, \text{ df=}(c_1\text{-1})\ (c_2\text{-1})$$

In the population, is the proportion of group A (π_A) equal to the proportion of group B (π_B)?

Hypotheses

H_0: $\pi_1=\pi_2$

H_a: $\pi_1\neq\pi_2$

Mc Nemar χ^2 test

$$\chi^2 = \frac{(b-c)^2}{b+c}, \text{ df=1}$$

In the population, is the proportion of a group in one condition (π_1) equal to the proportion of the same (or paired) group in another condition (π_2)?

McNemar χ^2-test

When to use:

To test if, according to the current data, the proportions in the population of a variable change based on another matched variable.

Hypotheses:

H_0: there is no association between the two (paired) variables
H_a: there is an association between the two (paired) variables

Assumptions:

- The observations are randomly and independently drawn
- There are at least 25 observations in the discordant cells
- The data are paired

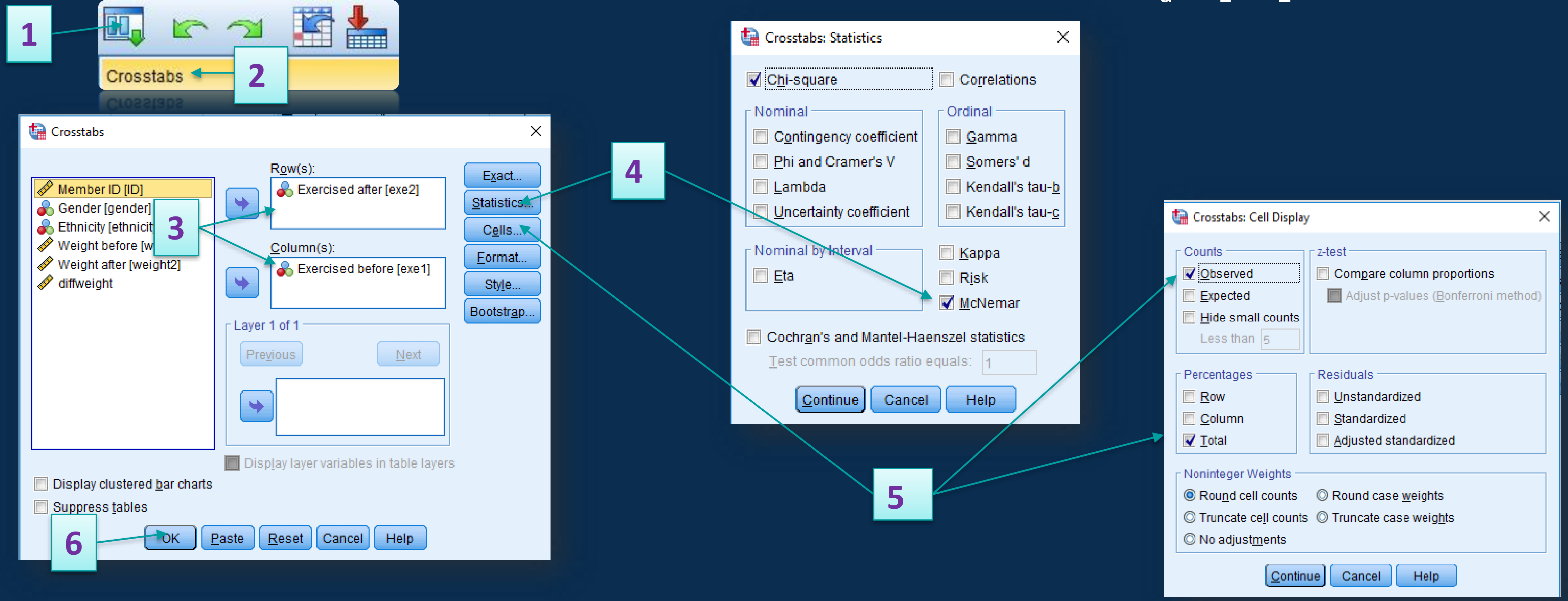
SPSS Slide: ‘how to’

Are the proportions of those exercised before the programme, different of those exercised after the programme?

H_0: $\pi_1=\pi_2$

H_a: $\pi_1 \neq \pi_2$

Step 1: Use the appropriate test, here: ‘McNemar chi-square test’.



Output and Interpretation

Exercised after * Exercised before Crosstabulation

Count

		Exercised before		
		No	Yes	Total
Exercised after	No	a 119	b 48	167
	Yes	c 103	d 30	133
Total		222	78	300

discordant cells

concordant *cells*

only ***discordant*** *cells play a role in McNemar test!*

The test statistic:

$$\chi^2 = \frac{(\boldsymbol{b}-\boldsymbol{c})^2}{\boldsymbol{b}+\boldsymbol{c}} \boxed{= \frac{(\mathbf{48}-\mathbf{103})^2}{\mathbf{151}}}$$

- The assumptions:
  - Paired data (here before-after)
  - b+c no less than 25 (here 151)

Output and Interpretation slide

Exercised before * Exercised after Crosstabulation

			Exercised after: No	Exercised after: Yes	Total
Exercised before	No	Count	119	103	222
		% of Total	39.7%	34.3%	74.0%
	Yes	Count	48	30	78
		% of Total	16.0%	10.0%	26.0%
Total		Count	167	133	300
		% of Total	55.7%	44.3%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	1.473[a]	1	.225	.236	.140	
Continuity Correction[b]	1.169	1	.280			
Likelihood Ratio	1.484	1	.223	.236	.140	
Fisher's Exact Test				.236	.140	
Linear-by-Linear Association	1.468[d]	1	.226	.236	.140	.051
McNemar Test				.000[c]	.000[c]	.000[c]
N of Valid Cases	300					

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 34.58.

b. Computed only for a 2x2 table

c. Binomial distribution used.

d. The standardized statistic is -1.211.

Note, we would have concluded the wrong result if we had used Pearson's chi-square. Pearson's chi-square is not valid for paired data.

The percentage of those who 'exercised after' (44.3%) is higher than the percentage of those who 'exercised before' (26.0%). This difference was statistically significant according to the McNemar test ($p<0.001$).

Therefore, we conclude that the proportions of people who exercise before and after the programme are not the same.

Knowledge Test Solution

Match the scenario with the correct test.

Tom wants to test if boys' proportions in ADHD high/low classification groups are different than those of girls.

Tom wants to test if mothers' reported ADHD high/low classification for children are different than those reported by fathers.

Tom wants to test if children's high classification ADHD proportion is higher than 50%.

One-sample χ^2-test

Pearson's χ^2-test

McNemar test

Reflection

Write down three examples from your research that would require the use of each of the three chi square tests.

Reference List

- **Agresti and Finlay (2009) Statistical Methods for the Social Sciences, 4th Edn, Pearson Hall, Upper Saddle River, NJ.**
  - Comparison of Two Groups, Ch 7, pages 183-209
  - Analyzing Association between Categorical Variables, Ch 8, pages 221-239

- **Field (2005) Discovering Statistics using SPSS, 2nd Edn, Sage, London.**
  - Comparing Two Means, Ch 7
  - Categorical Data, Ch 16

Thank you

Please contact your module leader or the course lecturer of your programme, or visit the module's forum for any questions you may have.

If you have comments on the materials (spotted typos or missing points) please contact Dr Vitoratou:

Silia Vitoratou, PhD
Psychometrics & Measurement Lab,
Department of Biostatistics and Health Informatics
IoPPN, King's College London, SE5 8AF, London, UK
silia.vitoratou@kcl.ac.uk

For any other comments or remarks on the module structure, please contact one of the three module leaders of the Biostatistics and Health Informatics department:

Zahra Abdula: **zahra.abdulla@kcl.ac.uk**

Raquel Iniesta: **raquel.iniesta@kcl.ac.uk**

Silia Vitoratou: **silia.vitoratou@kcl.ac.uk**