Lecture 12
Categorical Data Analysis: Part 2

Chi-Square Test of Independence

Some Summaries

In the Last Lecture….

- We analysed a categorical variable using a z distribution and using a χ^2 distribution.
- A categorical variable with two categories or groups (ie. a binary variable) can be analysed using either a z-test for a proportion or a χ^2 goodness-of-fit test.
- A categorical variable with more than two groups can be analysed using a goodness-of-fit test with degrees of freedom equal to (#categories – 1) = (c – 1).

Quiz 1

Research Question: Would the percentage of two party preferred votes going to the Labor Party have changed from those received in the 2013 federal election if the election had been held five months later?

In Australia's two party preferred electoral system, all voting preferences from minor candidates are distributed until a final percentage is obtained which designates all votes as Labor or Non-Labor.

In the 2013 Federal election, the final result in the House of Representatives after all preferences had been distributed was 46.51% to the Labor Party (ALP) and the remaining 53.49% to the Liberal-National Coalition (LNP).

Roy Morgan Research conducted a survey in February 2014. **An Australia wide cross section of 2709 Australian electors aged 18 years and over were asked what their voting intention would be if an election was held on that day. Under the two party preferred system, 1409 of these votes would have gone to the ALP and the remainder to the LNP.**

Use the output provided on the following slide to write up the appropriate hypothesis test addressing the research question.

Quiz 1

Research Question: Would the percentage of two party preferred votes going to the Labor Party have changed from those received in the 2013 federal election if the election had been held five months later?

Test and CI for One Proportion

Test of p = 0.4651 vs p not = 0.4651

Sample	X	N	Sample p	95% CI	Z-Value	P-Value
ALP	1409	2709	0.520118	(0.501305, 0.538931)	5.74	0.000

Using the normal approximation.

Solution to Quiz 1

Quiz 2

Research Question: Is absenteeism higher on some days of the week than on others?

It has been estimated that employee absenteeism costs Australian companies billions of dollars per year. As a first step in addressing the rising cost of absenteeism, the personnel department of a large corporation recorded the week days during which individuals in a sample of 362 absentees were away over the past several months. Do these data suggest that absenteeism is higher on some days of the week than on others?

Day	Mon	Tues	Wed	Thurs	Fri	Total
Number Absent	87	62	71	68	74	362

Use the output on the following page to write up an appropriate hypothesis test to address the question.

Source: Selvanathan et al, *Business Statistics,* 2011, Cengage

Quiz 2

Research Question: Is absenteeism higher on some days of the week than on others?

Chi-Square Goodness-of-Fit Test for Observed Counts in Variable: Absentees

Using category names in Day

Category	Observed	Test Proportion	Expected	Contribution to Chi-Sq
Monday	87	0.2	72.4	2.94420
Tuesday	62	0.2	72.4	1.49392
Wednesday	71	0.2	72.4	0.02707
Thursday	68	0.2	72.4	0.26740
Friday	74	0.2	72.4	0.03536

N	DF	Chi-Sq	P-Value
362	**	*****	*****

Solution to Quiz 2

Quiz 3

a. A sample is drawn from a target population. Which of the two measures, parameter or statistic, are you always able to calculate? Explain why.

b. You have a sample mean of 164 cm and, after testing H_0: $\mu = 170$, you do not reject H_0. What does the difference of 6cm represent in this instance?

c. You develop an exercise program aimed at reducing heart rate. You trial it on a sample and record heart rates before and after six weeks on the program. Write a conclusion if the test had produced:

   i. a p-value of 0.048

   ii. a p-value of 0.0048

Solution to Quiz 3

Chi-Square Test of Independence

Associations Between Two Categorical Variables

If we take a sample from some target population and we record information on each of two categorical variables for each subject in the sample, we can carry out a chi-square test of independence to determine whether the two variables are dependent or independent. We can this test to answer questions such as:

- Is occupation independent of education level among people between 40 and 50 years of age?
- Is the proportion of male students who own cars different to the proportion of female students who own cars?
- Among shift workers, is there an association between the day on which an employee is absent from work and the shift on which they are rostered?

Quiz 4

For each of these research questions, describe the target population, and write down the variables which you would record on each subject sampled:

a. Is occupation independent of education level among people between 40 and 50 years of age?

   i. Target Population:

   ii. Variables:

b. Is the proportion of male students who own cars any different to the proportion of females who own cars?

   i. Target Population:

   ii. Variables:

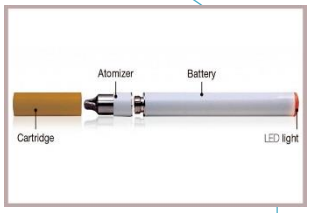
c. Among shift workers, is there an association between the days on which an employee is absent from work and the shift on which they are rostered?

   i. Target Population:

   ii. Variables:

χ^2 Test of Independence: e-Cigarettes vs Nicotine Patches

Research Question: Are e-cigarettes and nicotine patches equally effective in quitting smoking?



In 2013, researchers in New Zealand conducted a trial to compare the effectiveness of e-cigarettes and nicotine patches for smoking cessation. 584 smokers wanting to quit were randomised into groups with 289 allocated to the nicotine e-cigarette group and 295 allocated to the nicotine patch group. After six months of treatment, abstinence from smoking was verified for 21 of the nicotine e-cigarette group and 17 of the nicotine patch group.

Source: Bullen, Howe, Laugesen, McRobbie, Parag, Williman and Walker, *Electronic Cigarettes for Smoking Cessation: A Randomised Controlled Trial,* The Lancet, Volume 382 November 16, 2013

χ^2 Test of Independence

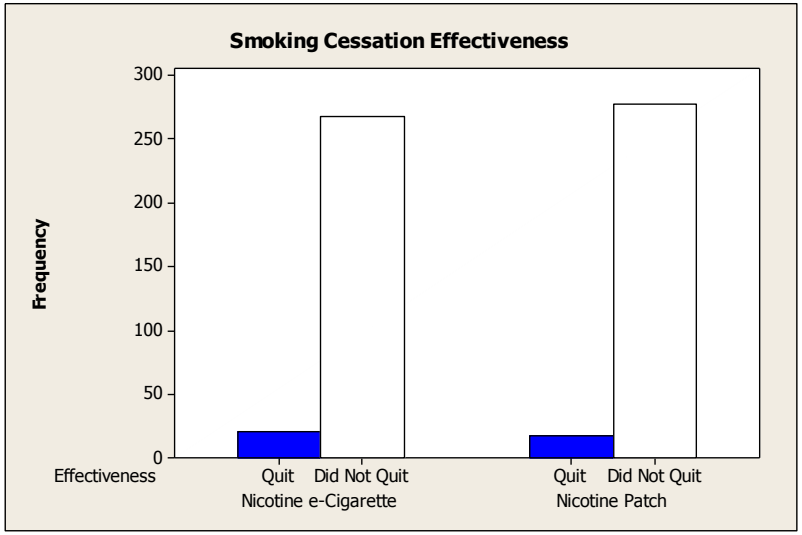
Some Sample Information:

Variable of interest:

X = Quit smoking method (nicotine e-cigarettes/nicotine patches)

Y = Effectiveness (quit/did not quit)

	Nicotine e-Cigarettes	Nicotine Patches	Total
Quit	21	17	38
Did not quit	268	278	546
Total	289	295	584



χ^2 Test of Independence: Hypotheses

H

H_0: There is no association between the method used to quit smoking and its effectiveness

(**or** method and effectiveness are independent

or the proportions of people who quit smoking is the same for those using nicotine e-cigarettes and those using nicotine patches)

H_1: There is an association between the method used to quit smoking and its effectiveness

(**or** method and effectiveness are dependent

or the proportions of people who quit smoking differs for those using nicotine e-cigarettes and those using nicotine patches)

χ^2 Test of Independence: Observed and Expected Values

We have the **observed** values from the sample already:

	Nicotine e-Cigarettes	Nicotine Patches	Total
Quit	21	17	38
Did not quit	268	278	546
Total	289	295	584

For the expected values, we need to calculate the values we would **expect** if the null hypothesis was true.

eg. if there is no association between method and effectiveness, the proportion of people who quit using e-cigarettes would be the same as the proportion who quit using patches. Since $^{38}/_{584} = 0.065$ is the estimated proportion of all people who quit we can use this to work out expected values for people who quit using each method:

$0.065 \times 289 = 18.8$ expected under H_0 for the e-cigarette users and

$0.065 \times 295 = 19.2$ expected under H_0 for the nicotine patch users

χ^2 Test of Independence: Observed and Expected Values

	Nicotine e-Cigarettes	Nicotine Patches	Total
Quit	21 (18.8)	17 (19.2)	38
Did not quit	268 (270.2)	278 (275.8)	546
Total	289	295	584

Since $^{546}/_{584}$ is the overall proportion who did not quit we can use this to work out expected values for people who did not quit using each method:

$^{546}/_{584} \times 289 = 270.2$ for the e-cigarette users and

$^{546}/_{584} \times 295 = 275.8$ for the nicotine patch

After finding the expected value for the first cell we could have just used subtraction to find the last row and the last column.

The easiest way to think of calculating the expected values is:

$$\boldsymbol{expected\ value} = \frac{\boldsymbol{Row\ Total} \times \boldsymbol{Column\ Total}}{\boldsymbol{Grand\ Total}}$$

χ^2 Test of Independence: Assumptions and Test Statistic

A We have calculated the expected value for each of the four cells in the table and since they are all at least 5, the test is valid

T
$$\chi^2 = \sum_{ij} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} = \frac{(21 - 18.8)^2}{18.8} + \frac{(17 - 19.2)^2}{19.2}$$
$$+ \frac{(268 - 270.2)^2}{270.2} + \frac{(278 - 275.8)^2}{275.8}$$
$$= 0.256 + 0.251 + 0.018 + 0.017$$
$$= 0.543$$

$$\boldsymbol{df = (rows - 1) \times (columns - 1) = (2 - 1) \times (2 - 1)}$$
$$\boldsymbol{= 1df}$$

χ^2 Test of Independence: Decision and Conclusion

P 0.2 < p-value < 0.5

D Since the p-value is ≥ 0.05 we do not reject H_0.

C **We have not found any evidence of an association between the method used to quit smoking and its effectiveness among people who want to quit smoking**.

The proportion of all people who quit using e-cigarettes could be the same as the proportion who quit using patches.

Another Example: χ^2 Test of Independence

Research Question: Is there an association between a voter's opinion on how to improve Australian economic growth and their political affiliation?

One of the issues argued during all recent election campaigns is how to improve Australian economic growth. A random sample of 1000 voters was each asked which option they supported (cut public spending/introduce tax reforms/job creation/increase educational funding) and their political affiliation (Labor/Liberal-National Coalition/Other). The output on the following slides was constructed from the results of the survey.

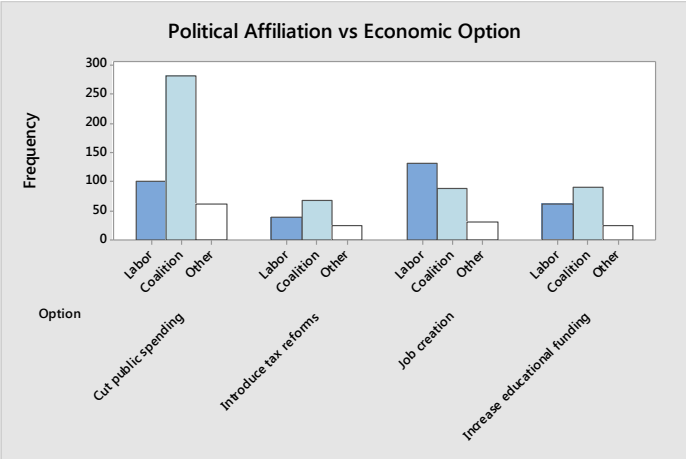
Source: Selvanathan et al, *Business Statistics,* 2014, Cengage

Minitab Output: χ^2 Test of Independence

Variables of interest:

X = Political Affiliation

Y = Economic Option



Chi-Square Test for Association: Option, Affiliation

Rows: Option Columns: Affiliation

	Labor	Coalition	Other	All
Cut public spending	101	282	61	444
	146.96	233.99	63.05	
	14.3756	9.8516	0.0665	
Introduce tax reforms	38	67	25	130
	43.03	68.51	18.46	
	0.5880	0.0333	2.3170	
Job creation	131	88	31	250
	82.75	131.75	35.50	
	28.1337	14.5280	0.5704	
Increase educational funding	61	90	25	176
	58.26	92.75	24.99	
	0.1292	0.0817	0.0000	
All	331	527	142	1000

Cell Contents: Count
Expected count
Contribution to Chi-square

Pearson Chi-Square = 70.675, DF = 6, P-Value = 0.000
Likelihood Ratio Chi-Square = 69.334, DF = 6, P-Value = 0.000

Example: χ^2 Test of Independence

Hypothesis Test

H H_0:There is **no** association between political affiliation and economic option

H_1:There **is** an association between political affiliation and economic option

A All expected values are at least 5, so the test is valid.

T $\chi^2 = 70.675$ with $6df$

P $\text{p} - \text{value} \approx 0$

D $\text{Since the p} - \text{value is} < 0.05, \text{we}$ **reject H_0**

C There ***is*** evidence of an association between political affiliation and the option chosen to improve economic growth. It appears that overall ***Labor voters are more likely to choose job creation and Coalition voters are more likely to choose to cut public spending.***

Quiz 5

Research Question: Do the proportions of right and left handers in different professions vary?

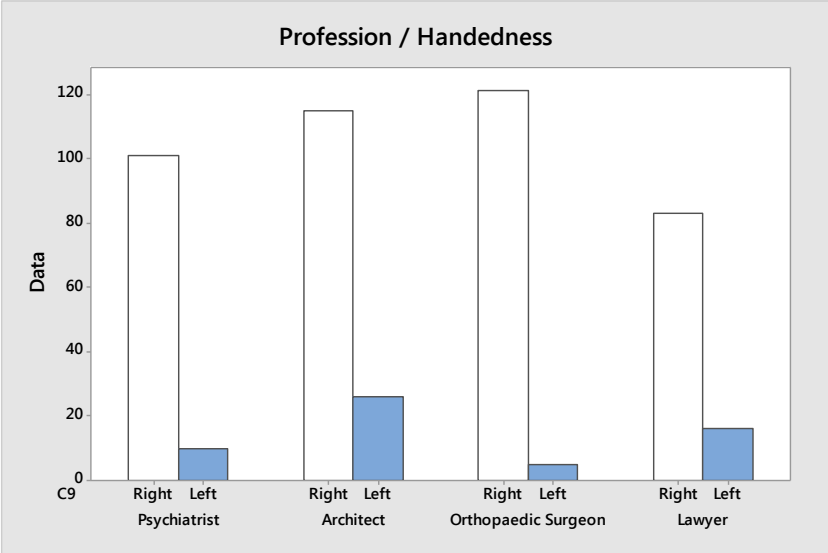
In Lecture 11 we used a goodness of fit test to investigate the proportion of left handers among artists. Another researcher is interested in comparing the proportions of left handers among professionals trained in various fields. He selected a sample of 477 professionals and recorded each person's profession as well as his/her handedness. Use the table below to address the research question.

	Psychiatrist	Architect	Orthopaedic Surgeon	Lawyer	Total
Right	101	115	121	83	420
Left	10	26	5	16	57
Total	111	141	126	99	477

Solution to Quiz 5

Some Sample Information:

Variables of interest:



	Psychiatrist	Architect	Orthopaedic Surgeon	Lawyer	Total
Right	101	115	121	83	420
Left	10	26	5	16	57
Total	111	141	126	99	477

Solution to Quiz 5:

What if any of the Expected Values are < 5?

A chi-square test will not give valid results if any of the expected values are less than 5. This is the case for both a chi square goodness-of-fit test and for a chi-square test of independence.

If we don't have enough data or we have some categories where there are very few values, we may not get large enough expected values.

Consider the following table constructed from information collected on a sample of 664 students. Preferred diet and smoking status was recorded for each student.

	Meat	Vegetarian	Vegan	Total
Smokers	65 (65.4)	14 (13.0)	1 (1.6)	80
Non-Smokers	478 (477.6)	94 (95.0)	12 (11.4)	584
Total	543	108	13	664

Expected Counts < 5

Not all of the expected counts are at least 5. The problem here is that there are not many students who are smokers or vegans. To overcome this problem we could:

- collect more data
- design a study such that we deliberately select more vegans or more smokers

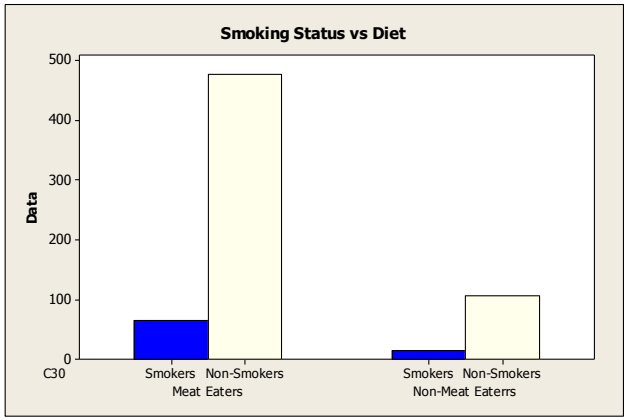
Since that isn't possible here, we will reorganise the table by collapsing the two categories of vegetarian and vegan since they are all non-meat eaters. This is shown on the next slide.

Quiz 7

Research Question: Is there as association between a student's diet and his/her smoking status?

	Meat Eaters	Non-Meat Eaters	Total
Smokers	65 (65.4)	15 (14.6)	80
Non-Smokers	478 (477.6)	106 (106.4)	584
Total	543	121	664

Variables of interest:



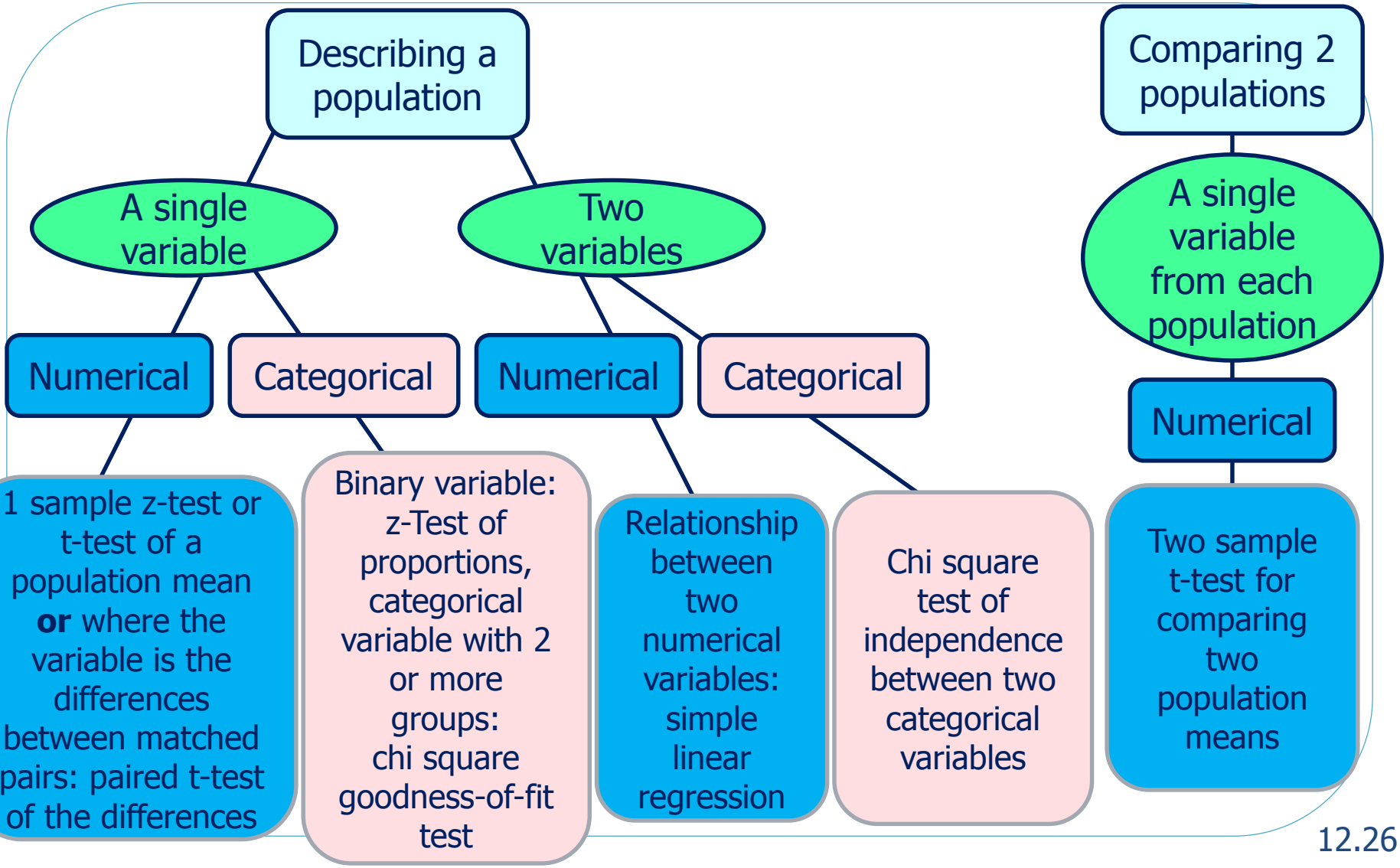
Solution to Quiz 7

Some Summaries

Graphical Summaries and Analyses

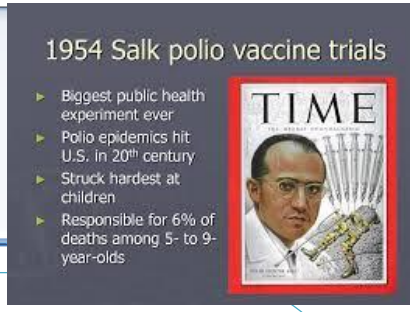
DATA	Categorical	Numerical	
Categorical	○ Clustered Bar Chart ○ Chi Square Test of Independence Between Two Categorical Variables	○ Comparative Box Plots ○ Two Sample t-Test Comparing Two Population Means	○ Simple Bar Chart ○ One Sample z-Test of Proportions/Chi Square Goodness-of-Fit Test
Numerical		○ Scatter Plot ○ Simple Linear Regression for a Linear Relation Between Two Numerical Variables	○ Histogram ○ One sample z/t-Test for a Population Mean/Paired t-Test on the Differences Between Population Means

Another Summary



Homework Questions

Homework Quiz 1



Research Question: Was the Salk vaccine effective against poliomyelitis?

In 1954 a large study was undertaken in the USA and Canada involving more than 400,000 children. The parents of these children all agreed to them participating in an experiment to test a vaccine against polio developed by Jonas Salk. Each child was randomly assigned into one of two groups. One group was given the Salk vaccine whilst the other group was given a placebo. The experiment was double blind. Of the 201,229 children who were given the placebo, 115 contracted polio, whereas only 33 of the 200,745 children given the Salk vaccine contracted polio. Minitab output for this analysis is given on the following two slides. Use this output to write up an appropriate test to address the research question.

Solution to Homework Quiz 1

Some Sample Information:

Variables of interest:

	Polio	No Polio	Total
Placebo	115	201114	201229
Salk	33	200712	200745
Total	148	401826	401974

Solution to Homework Quiz 1

Chi-Square Test: Polio, No Polio

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

	Polio	No Polio	Total
Placebo	115	201114	201229
	74.09	201154.91	
	22.590	0.008	
Salk	33	200712	200745
	73.91	200671.09	
	22.645	0.008	
Total	148	401826	401974

Chi-Sq = ******, DF = ******, P-Value = ******

Solution to Homework Quiz 1

Lecture 12 Summary

- The association between two categorical variables can be analysed using a contingency table and a χ^2 test of independence with degrees of freedom equal to (#columns – 1) $\times$ (#rows – 1) = (r – 1)(c – 1).

Textbook References

Further information on the topics discussed in this lecture can be found in:

Modern Statistics: An Introduction
by Don McNeil and Jenny Middledorp
(ISBN 9781486007011).

- Chapter 10: Pages 226 – 235.

Note: The section on Odds Ratio is not covered in STAT170.