Lecture 11
Categorical Data Analysis: Part 1

One sample z-test for a Population Proportion

Chi-Squared (χ^2) Goodness-of-Fit Test

In the Last Lecture….

- We found that, under certain conditions, the least squares regression model can be used to make predictions.

- The goodness-of-fit statistic (r^2) is a measure of how well our data fit the least squares regression model. It measures the proportion of variation in the dependent variable that can be explained by the independent variable.

- The correlation coefficient (r) is a measure of the strength and direction of the linear relation between the two numerical variables.

Quiz 1

You have a set of bivariate data of size 45 and test the hypotheses that $H_0\!: \beta = 0 \ vs \ H_1\!: \beta \neq 0$. You obtain a test statistic of 3.14. *State whether the following statements are True or False:*

a. There is a statistically significant linear relation between X and Y.

b. A confidence interval for β will contain zero.

c. X is a useful predictor of Y.

d. The slope of the line in the population will be positive.

Quiz 2

For a set of bivariate data which has a goodness-of-fit measure of $r^2 = 25\%$, *state whether the following statements are True or False:*

a. the null hypothesis $H_0\!: \beta = 0$ will be rejected

b. the correlation coefficient will be 0.5

c. 75% of the variation in Y cannot be explained by the model

Revision of Data Types: Numerical Data

Numerical Data:

- Each recording taken on the variable of interest is measured ie. continuous or discrete

  eg. IQ scores, age, height, income, rental costs, price

- We have already addressed various research questions involving numerical variables eg.

  Is the average rent for an apartment in Paris $4292 per month?

  Is there a difference between the average time spent foraging for penguins with metal tags and penguins with electronic tags?

  Is the price of a used Toyota Corolla related to its age?

Revision of Data Types: Categorical Data

Categorical Data:

- Each subject in the study is placed into one, and only one, of a number of categories or groups.

  eg. sex (binary/nominal)

  preferred car colour (nominal)

  grade for unit (ordinal)

Quiz 3

The variables on the following slide were recorded by a real estate agent on a sample of prospective home buyers. Decide whether each variable is numerical or categorical. For the numerical variables, decide whether the variable is continuous or discrete. For the categorical variables, decide whether the variables is nominal or ordinal.

Solution to Quiz 3

Variable Name	Variable type
a. Client name	
b. Occupation	
c. Annual income	
d. First home buyer: yes/no	
e. Dwelling type: house/apartment	
f. Number of bedrooms required	
g. Price range: <$500k, $500k - $750k, $750k - $1 million, $1 million+	

One Sample z-test for a Population Proportion

A one sample z-test for a Population Proportion

- In lecture 7 we used one sample tests to test hypotheses about population means. A test for a **population mean** is used when the research question concerns a variable of interest which is **numerical**, and the distribution of sample means can be assumed to be normal. The null hypothesis for this test is $\boldsymbol{H_0: \mu = \mu_0}$
- If the variable of interest is **categorical**, the research question will concern population proportions, rather than means. If we are hypothesising about the proportion in a particular group for a categorical variable, and the distribution of sample proportions can be assumed to be normal, we can use a one sample z-test for a **population proportion**. The null hypothesis for this test is $\boldsymbol{H_0: \pi = \pi_0}$

Steps in Hypothesis Testing: z-test for a Population Proportion

H ***Hypotheses***:

H_0: $\pi = \pi_0$ H_1: $\pi \neq \pi_0$

A ***Assumptions:***

We need to ensure that sample proportions follow a normal distribution ie. check that the Central Limit Theorem applies. The CLT applies to sample proportions only when **both $n\pi$ and $n(1-\pi)$ are ≥ 5**

T ***Test Statistic:***

For this test, we use $z = \dfrac{p-\pi}{\sqrt{\dfrac{\pi(1-\pi)}{n}}}$ where p is the sample proportion

P ***p-value:***

We will obtain the p-value from the two tails of the z distribution.

D ***Decision:***

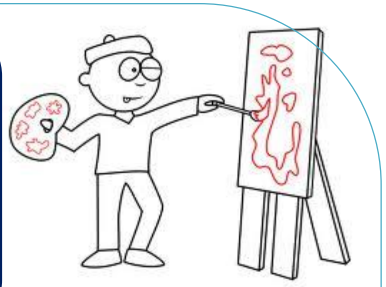
We will continue to use a significance level: $\alpha = 0.05$ ie. We will only reject H_0 when the p-value is < 0.05. When the p-value is ≥ 0.05, we do not reject H_0.

C ***Conclusion:***

We write a conclusion to the original research question in terms of the target population.

Example of a z-test for a Population Proportion

Research Question: Is the proportion of artists who are left handed any different to the proportion of left-handers in the general population?



Why are some people left-handed??? A researcher who is interested in left-handers wants to determine whether the proportion of left-handers is any different among artists than in the general population. The researcher samples 150 artists and finds that 18 of them are left handed.

It is generally claimed that the proportion of left handers in the general population is 10%. We will use a z-test of a population proportion to address the research question.

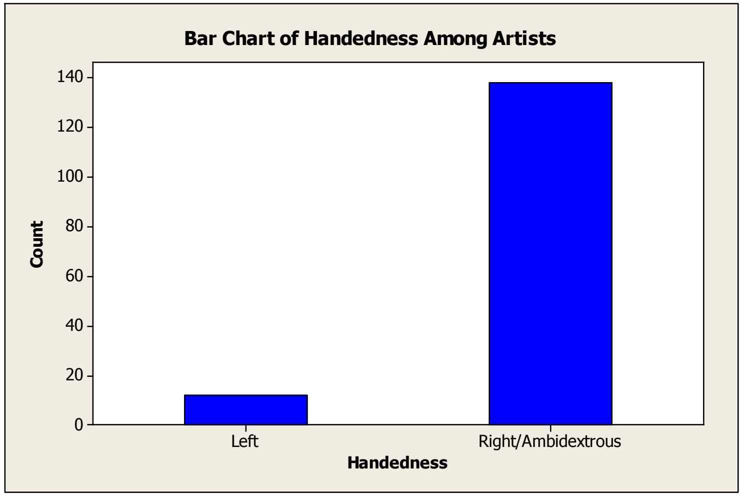
Example: Left Handers

Some Sample Information:

Variable of interest:

Y = Handedness of artists

$n = 150, \quad p = {}^{18}/_{150} = 0.12$



Left Handers: z-test for a Population Proportion

Hypothesis Test

H $H_0: \pi = 0.1 \quad H_1: \pi \neq 0.1$

A $n\pi = 150 \times 0.1 = 15$ and $n(1-\pi) = 150 \times 0.9 = 135$. Since both $n\pi$ and $n(1-\pi)$ are ≥ 5, the CLT applies and the test is valid.

T $$z = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.12-0.1}{\sqrt{\frac{0.1(1-0.1)}{150}}} = 0.82$$

P p-value $= 2 \times 0.2061 = 0.4122$

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

C 10% of all artists could be left handed. We do not have evidence to suggest that the proportion of left handers is different among artists than among the general population.

Quiz 4

Research Question: Had the proportion of Greens supporters changed five months after the 2013 federal election?

In the federal election held in September, 2013, 8.7% of votes in the House of Representatives went to the Greens.

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. 284 responded that they would have given their vote to the Greens.

Does this survey provide any evidence to indicate that support for the Greens had changed five months after the election?

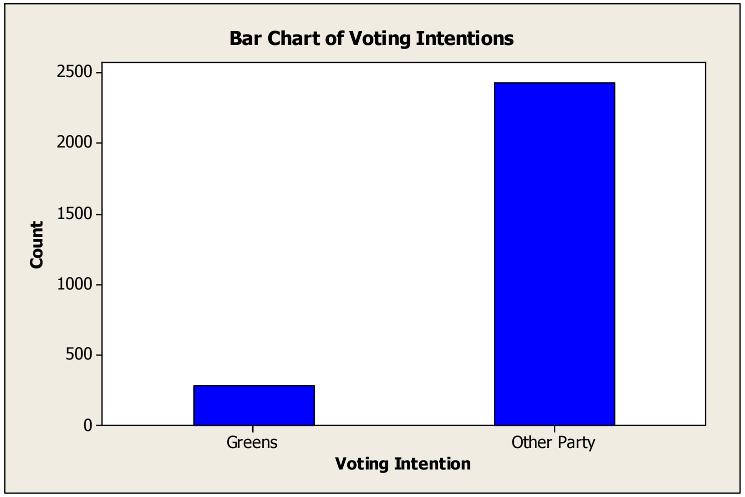
Source: http://www.roymorgan.com/findings

✓ Solution to Quiz 4

Some Sample Information:

Variable of interest:

n =   p =



Solution to Quiz 4

95% Confidence Interval for π

Recall from lecture 6 - to calculate a 95% confidence interval to estimate a population proportion:

$$95\%\ CI\ for\ \pi = p \pm z_{crit} \times est.se_p$$

where z_{crit} for a 95% confidence interval = 1.96 ie. the critical z-value that cuts off 5% in the two tails of the z-distribution and the estimated standard error for sample proportion,

$$est.se_p = \sqrt{\frac{p(1-p)}{n}}$$

Since the confidence interval uses an estimated standard error, a confidence interval for π will not always confirm the results of the corresponding one sample z-test of proportions, which uses the standard error for sample proportions,

$$se_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

95% Confidence Interval for π

To calculate a 95% confidence interval to estimate the proportion of all voters who would vote for the Greens if an election had been held in February 2014:

$CLT\ applies\ as\ np = 284\ and\ n(1-p) = 2425\ which\ are\ both\ at\ least\ 5$

$$95\%\ CI\ for\ \pi = p \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

$$= 0.1048 \pm 1.96 \times \sqrt{\frac{0.1048(1-0.1048)}{2709}}$$

$$= (0.093, 0.116)$$

We can be 95% confident that between 9.3% and 11.6% of voters would have voted for the Greens if an election had been held in February 2014.

Minitab Output

```
Test and CI for One Proportion: Voting Intention

Test of p = 0.087 vs p ≠ 0.087

Event = Greens


Variable             X     N  Sample p           95% CI          Z-Value  P-Value
Voting Intention   284  2709  0.104836  (0.093300, 0.116372)       3.29    0.001

Using the normal approximation.
```

Minitab uses the notation 'p' to describe the population proportion which we call 'π'

Minitab uses the notation 'sample p' to describe the sample proportion which we call 'p'

Chi Square (χ^2) Goodness-of-Fit Test

What if the Variable is Not Binary?

A one sample z-test of proportions can only be used to test a hypothesis about a single proportion. We have treated each of the variables used in the previous z-tests of proportions as **binary variables**:

Handedness = left handed/not left handed

Voting Intention = Greens/Other Parties

For each question we hypothesised about the proportion in a particular group and implicitly hypothesised that any other observations fell into a second group:

$$\boldsymbol{H_0: \pi_{left} = 0.1 \quad (\pi_{not\ left} = 0.9)}$$

$$\boldsymbol{H_0: \pi_{Greens} = 0.087 \ (\pi_{Others} = 0.913)}$$

What if we wish to test a hypothesis about proportions for a categorical variable where there are more than two groups of interest?

What if the Variable is Not Binary?

- A one sample z-test for proportions has no extension to three or more categories.
- When the hypothesis we wish to test involves proportions in three or more categories we need to use a different method of analysis.
- The chi square (χ^2) goodness-of-fit test can be used to test a hypothesis about any number of categories.

Example: χ^2 Goodness-of-Fit Test

Research Question: Are 80% of people ***right handed,*** *12%* ***left handed*** *and remainder* ***ambidextrous****?*

Another researcher interested in ‘handedness’ believes that 80% of people are right handed, 12% left handed and the remainder ambidextrous (able to use the right and left hands equally well).

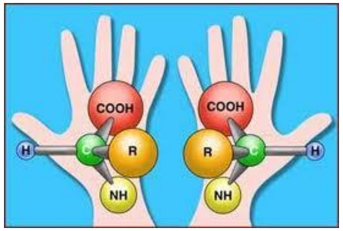
To test her belief, she selects a sample of 635 adults from various professions and records their ‘handedness’. She finds that 536 are predominantly right handed, 67 predominantly left handed and the remainder ambidextrous.

To test her claim we need to consider the proportion in all **three** groups so we cannot use a z-test of proportions. We can use a chi square goodness of fit test.

Source: Lock et al, *Unlocking the Power of Data,* 2013, Wiley

Example: χ^2 Goodness-of-Fit Test

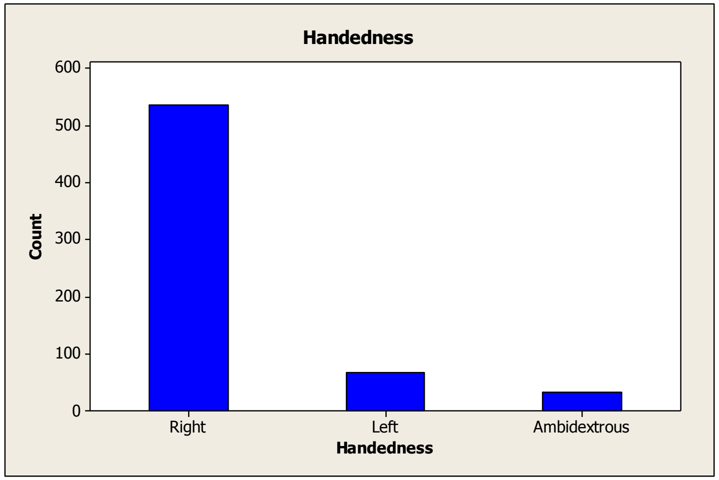
Some Sample Information:



Variable of interest:

Y = Handedness

Right	Left	Ambidextrous	Total
536	67	32	635



A Null Hypothesis for a χ^2 Goodness-of-Fit Test

- For the null hypothesis we need to define a parameter (π) for each group so we could use:

  π_R = proportion of right handers in the population

  π_L = proportion of left handers in the population

  π_A = proportion of ambidextrous people in the population

- The null hypothesis is a hypothesis of no effect/no change so here it will need to say that the claim made by the researcher could be correct ie. the proportions are not different to those claimed. The hypothesised proportions must add to one in total so each group must be represented in the null hypothesis. This gives us:

  H $\boldsymbol{H_0: \pi_R = 0.80, \pi_L = 0.12, \pi_A = 0.08}$

An Alternative Hypothesis for a χ^2 Goodness-of-Fit Test

- If the null hypothesis is rejected it may be because **all** the proportions are different to those hypothesised **but** it may be because **only some** of the proportions are different to those hypothesised.
- So the alternative hypothesis (which is believable if the null hypothesis is rejected) would have to be stated in the following format:

H $\boldsymbol{H_1}$**: The proportions are not as stated in the null hypothesis**

Observed Values and Expected Values

A χ^2 test compares observed counts to the expected counts under the null hypothesis. We know what the observed values are from the sample. We use the null hypothesis to calculate the expected counts:

E (right): $635 \times 0.80 = 508.0$

E (left): $635 \times 0.12 = 76.2$

E (ambidextrous): $635 \times 0.08 = 50.8$

Handedness				
	Right	Left	Ambidextrous	Total
Observed	536	67	32	635
(Expected)	(508)	(76.2)	(50.8)	(635)

A The test is only valid if all expected counts are ≥ 5

Discrepancies Between Observed Values and Expected Values

Discrepancies between **O**bserved and **E**xpected values for each group:

Right Handed: $\boldsymbol{O_R - E_R} = 536 - 508 = 28.0$

Left Handed: $\boldsymbol{O_L - E_L} = 67 - 76.2 = -9.2$

Ambidextrous: $\boldsymbol{O_A - E_A} = 32 - 50.8 = -18.8$

Total discrepancy between observed and expected = 0

Since the total discrepancy will always equal zero, the discrepancies are squared in order to calculate a test statistic which compares observed and expected values.

The χ^2 Test Statistic

The chi square test statistic is a standardised measure of the total discrepancy between observed and expected values:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where ***O**$_i$* = *observed value for group i*

***E**$_i$* = *expected value for group i*

The χ^2 statistic is valid only if all the expected values are $\geq$ 5.

Calculating the χ^2 Test Statistic

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{(536-508)^2}{508} + \frac{(67-76.2)^2}{76.2} + \frac{(32-50.8)^2}{50.8}$$

$$= 1.54 + 1.11 + 6.96 = 9.61$$

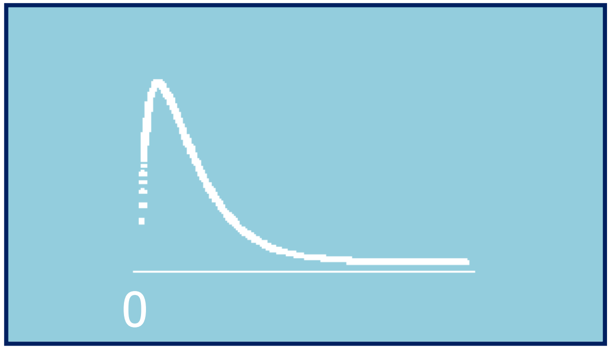
The distribution of this statistic is chi-square (χ^2) with degrees of freedom equal to:

number of categories – 1

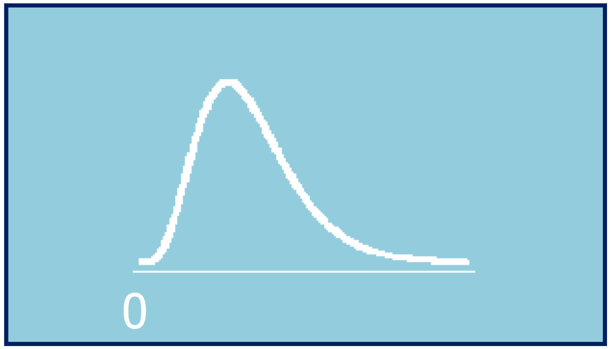
T So χ^2 = 9.61 with (3 – 1) = 2 df

The χ^2 Distribution

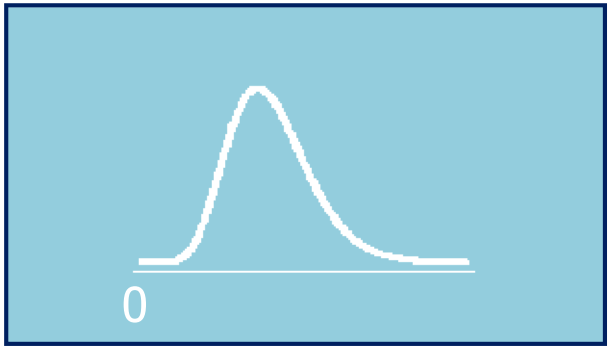
degrees of freedom = 3



degrees of freedom = 10



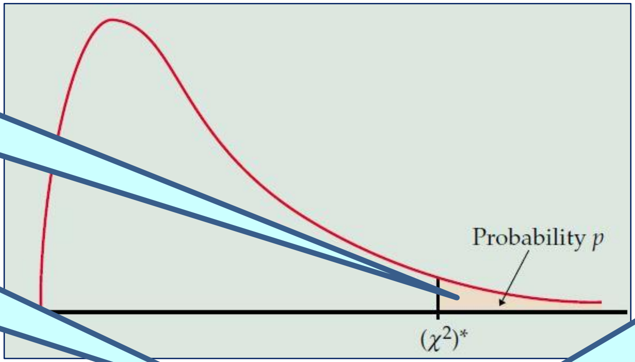
degrees of freedom = 20



Finding a p-value from a χ^2 Table

The area in the right hand tail is the p-value for a chi square test

So the p-value is somewhere between 0.005 and 0.01



We had a test statistic of 9.61 with 2df

p	0.0005	0.001	0.002	0.005	0.01	0.02	0.05	0.1	0.2	0.5
ν										
1	12.12	10.83	9.55	7.879	6.635	5.412	3.842	2.706	1.642	0.455
2	15.20	13.82	12.43	10.600	9.210	7.824	5.992	4.605	3.219	1.386
3	17.73	16.27	14.80	12.840	11.340	9.837	7.815	6.251	4.642	2.366
4	...									

P $0.005 < \text{p- value} < 0.01$

D Since the p-value is below the 0.05 significance level, we reject H_0.

Conclusion for χ^2 Test

- ○ **If we hadn't rejected the null hypothesis** we would not have had any evidence against the null hypothesis. We would have concluded that the researcher's claim could have been correct.
- ○ **But… we did reject the null hypothesis** so we have evidence that the proportions are not as claimed by the researcher. **So which proportions have changed? All of them or only some of them?** We can't say for sure without further analysis but we can see that the standardised discrepancy between observed and expected values is much larger for the ambidextrous group than for either the right handers or the left handers **and…** when we compare observed and expected values for the ambidextrous group we see that there were less people observed (32) in this group than expected (50.8) under the null hypothesis.

C

The proportions are not as claimed by the researcher. There appears to be a lower proportion of ambidextrous people than the researcher has claimed.

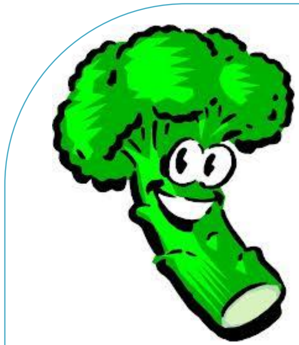
Minitab Output for χ^2 Goodness-of-Fit Test

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

Using category names in Handedness


                                Test                Contribution
Category        Observed  Proportion  Expected   to Chi-Sq
Right                536        0.80     508.0     1.54331
Left                  67        0.12      76.2     1.11076
Ambidextrous          32        0.08      50.8     6.95748


  N  DF    Chi-Sq  P-Value
635   2   9.61155    0.008
```

Standardised discrepancies

Minitab will give the exact p-value

χ^2 Goodness-of-Fit Test with 1 df



Consider again the example from Quiz 4. You used a z-test of proportions to determine whether the proportion of Greens supporters had changed between the 2013 federal election (when 0.087 of all voters voted for the Greens) and five months later when a Morgan Gallup Poll surveyed 2709 Australian electors and found that 284 would have given their vote to the Greens if an election had been held on that day. Let us repeat the analysis, this time using a χ^2 goodness-of-fit test and compare the results to those you obtained from Quiz 4.

Let π_G = population proportion who would vote for the Greens
Let π_O = population proportion who would vote for other parties

χ^2 Goodness-of-Fit Test with 1 df

	Greens	Other Parties	Total
observed (expected)	284 (235.7)	2425 (2473.3)	2709 (2709)

Hypothesis Test

H

H_0: $\pi_G = 0.087, \pi_O = 0.913$

H_1: $\mathit{The\ proportions\ are\ not\ as\ stated}$

A

The expected values are all at least 5 so the test is valid

T

$$\chi^2 = \frac{(284 - 235.7)^2}{235.7} + \frac{(2425 - 2473.3)^2}{2473.3}$$

$$= 9.91 + 0.94 = 10.85 \ \mathit{with\ 1df}$$

P

From the chi square table 0.0005 < p-value < 0.001

If we use Minitab, we find the p-value is almost exactly 0.001

D

Since the p-value is < 0.05, we reject H_0

C

As in the solution to Quiz 4 we would conclude that the evidence indicated that support for the Greens party appeared to have increased.

Comparing Results

$\chi^2 = 10.85$, p-value = 0.001

$z = 3.29$, p-value = 0.001

$z^2 = 3.29^2 = 10.85 = \chi^2$

These are consistent!

So for a binary variable the z test of proportions must give the same result as for a χ^2 test with one degree of freedom

Note that, for the chi square goodness-of-fit test, ALL categories must be included in the analysis.

Also note that we use counts for the chi-squared test, whereas we use proportions for the z-test.

Quiz 5

Bankcard was the first widely available credit card issued by Australian banks. It was launched in 1974 and used up until 2006. Since this time the use of bank cards as a source of consumer credit has become increasingly prevalent in Australia. VISA and Mastercard are currently the two most widely used credit cards in Australia.

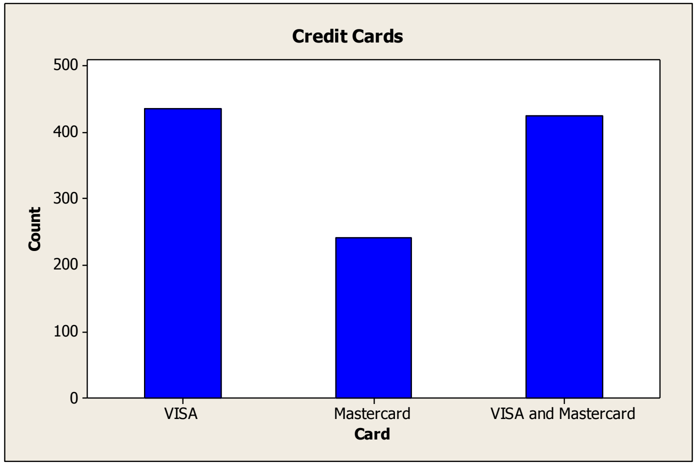
A recent report on credit card spending in Australia claimed that 40% of credit card users use a VISA card, 25% use Mastercard and the remainder use **both** VISA and Mastercard.
A researcher has been able to select a random sample of 1100 credit card users from bank records and finds that 435 of these use VISA cards only, 241 use Mastercards only and the remainder use both VISA and Mastercards.
Use an appropriate hypothesis test to test the claim made in the report on credit card spending.

Solution to Quiz 5

Some Sample Information:

Variable of interest:

Card	Count
VISA	435
Mastercard	241
VISA and Mastercard	424
Total	1100



Solution to Quiz 5

Homework Questions

Homework Question 1

Do consumer preferences for Brand A cola and Brand B cola differ?

A sample of 75 people take part in an experiment to compare two different brands of cola flavoured soft drink. Each person is blindfolded and given a small drink of one brand then a small drink of the other brand and asked to indicate which they prefer. The order of the drinks are randomised to avoid any bias related to which brand is consumed first and which is consumed second. 40 of those sampled indicate they prefer Brand A and the rest indicate that they prefer Brand B. Carry out an appropriate analysis to determine whether there is a significant difference between preferences for Brands A and B.

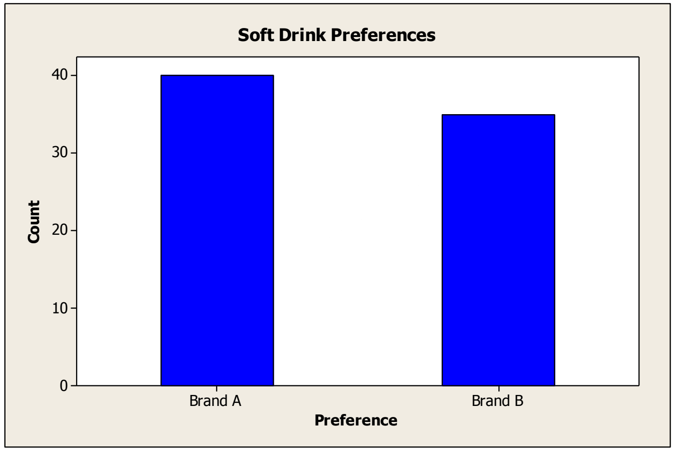
11.38Q

Homework Question 1 Solution

Some Sample Information:

Variable of interest:

Preference	Count
Brand A	
Brand B	
Total	



Lecture 11 Summary

- Categorical data can be analysed using a z or 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).

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 7: Pages 149 – 153
- Chapter 10: Pages 226 - 230