Topic: Exploratory Data Analysis (EDA)

Presentation of Bivariate Data

Part A: Two categorical variables

School of Mathematics and Applied Statistics

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Bivariate data: Two Variables

Different tables / plots for different data types . . .

For two qualitative variables:

- two-way tables
- stacked bar graphs
- clustered bar graphs

For one quantitative and one qualitative variable:

- side-by-side box plots
- back-to-back stem & leaf plots

For two quantitative variable/s:

- scatterplots
- line plots (against time)

Contingency Table

A two-way table or contingency table summarises bivariate data of two categorical variables.

Example: Titanic data

Survived	1st Class	2nd Class	3rd Class	Crew	Total
No	122	167	528	673	1490
Yes	203	118	178	212	711
Total	325	285	706	885	2201

Is there any association between the two variables?

Is the proportion of *Survived* the same for *Class of passenger*?

Contingency Table - Conditional probability

Example: Titanic: Observed data

Survived	1st Class	2nd Class	3rd Class	Crew	Total
No	122	167	528	673	1490
Yes	203	118	178	212	711
Total	325	285	706	885	2201

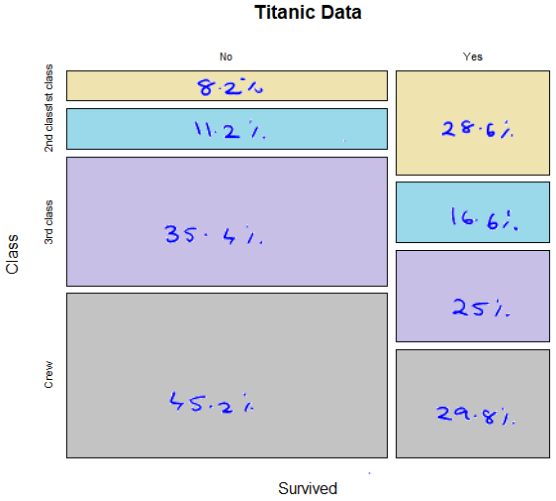
$P(1st \mid No) = \frac{122}{1490} = 0.082 = 8.2\%$

Exercise: Determine the row percentages:

Survived	1st Class	2nd Class	3rd Class	Crew	Total
No	122/1490=	167/1490=	528/1490 =	673/1490=	100%
	8.2%	11.2%	35.4%	45.2%	
Yes	203/711 =	118/711=	178/711=	212/711 =	100%
	28.6%	16.6%	25.0%	29.8%	

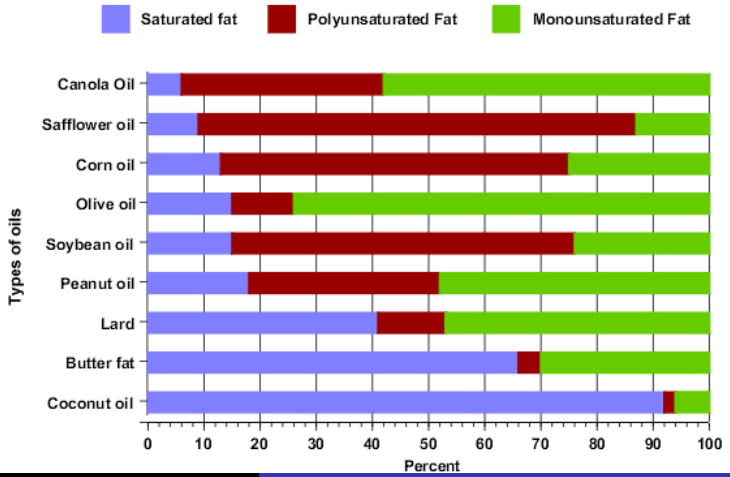
Mosaic Plot

This mosaic plot displays counts of two-way table, as areas proportional to frequencies within a row.



Example of a stacked bar chart

Stacked bar graphs are often used to represent parts of a whole.



9 different types cooking oil.

Least % of sat fat

≈92% Coc. oil Sat. fat.

Conditional probability - another look

Example: Phone carriers and Gender

	Optus	Telstra	Vodafone	Total
Female	19	9	4	32
Male	36	17	20	73
Total	55	26	24	105

Q1: Given a female, what is the probability that they

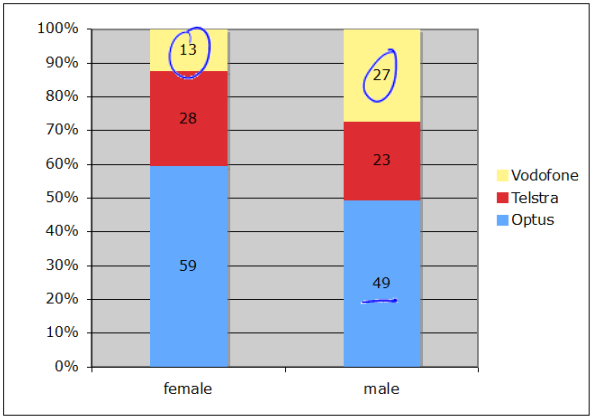
- Use Optus? $P(O \mid F) = 19/32 = 0.594$
- Use Telstra? $P(T \mid F) = 9/32 = 0.281$
- Use Vodafone? $P(V \mid F) = 4/32 = 0.125$

Q2: Write down

- P(Optus|Male)= $36/73 = 0.493$
- P(Telstra|Male)= $17/73 = 0.233$
- P(Vodafone|Male)= $20/73 = 0.274$

Q3: Is the pattern of usage the same for males and females?

Stacked Bar Charts & Independence



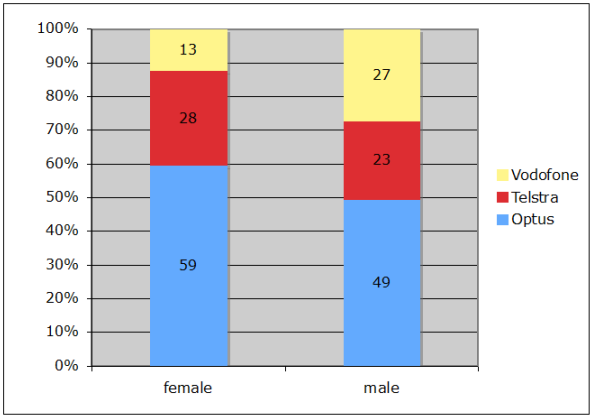
The proportion of Males using Vodafone is greater than that for females.

For Optus, it is less than than that for females..

Is this just the sample or is it a pattern evident in the population?

Stacked Bar Charts & Independence

When the pattern of use by males and females is the same then we have independence in the population



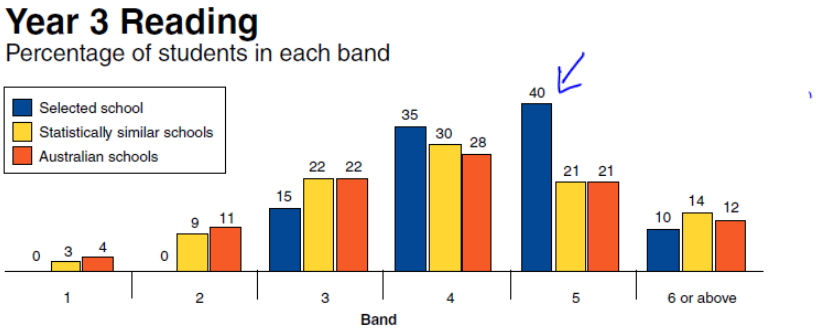
When we use a sample to infer something about the population then similar patterns imply independence.

BUT

How similar is similar?

Bar Charts can be: . . . clustered

For two qualitative variables:



Source: Lantites - Sample Questions p15, ACER

This graph shows the percentage of Year 3 students in six achievement bands for reading, for a selected school. It also shows comparable percentages for statistically similar schools and for all Australian schools.