MATH2801 Notes

Contents

Does the research question involve:					
	One variable		Two variables		
Data type:	Categorical	Quantitative	Both categorical	One of each	Both quantative
Numerics:	Table of frequencies	Mean/sd Median/quantiles	Two-way table	Mean/sd per group	Correlation
Graphs:	Bar chart	Dotplot Boxplot Histogram *etc.*	Clustered bar chart	Dotplot Boxplots Histograms *etc.*	Scatterplot

Figure 1: Summary of descriptive methods

Chapter 1: Descriptive statistics

2 Steps to Data Analysis:

1. What is the research question?
2. What properties of the variables of primary interest?

2 Types of variables:

- Categorical → Responses can be sorted into a finite set of unordered categories
- Quantitative → Responses are measured on some sort of scale

Categorical Data

Problems that summarise one categorical variable and the association between two categorical variables are extremely similar in scope so we'll cover both here.

Numerical summaries of cateogircal data

The main tool is a table of frequencies (both one way for a single variable and two way for two variables)

One way table:

Party	Liberal	Labor
	300	295

Two way table:

	Survived	Died
Male	142	709
Female	308	154

Graphical summaries of categorical data

2 types:

- Bar chart of frequencies → 1 var
- Clustered bar chart (of frequencies) → 2 vars

DON'T USE PIE CHARTS U DUMB FUCKS

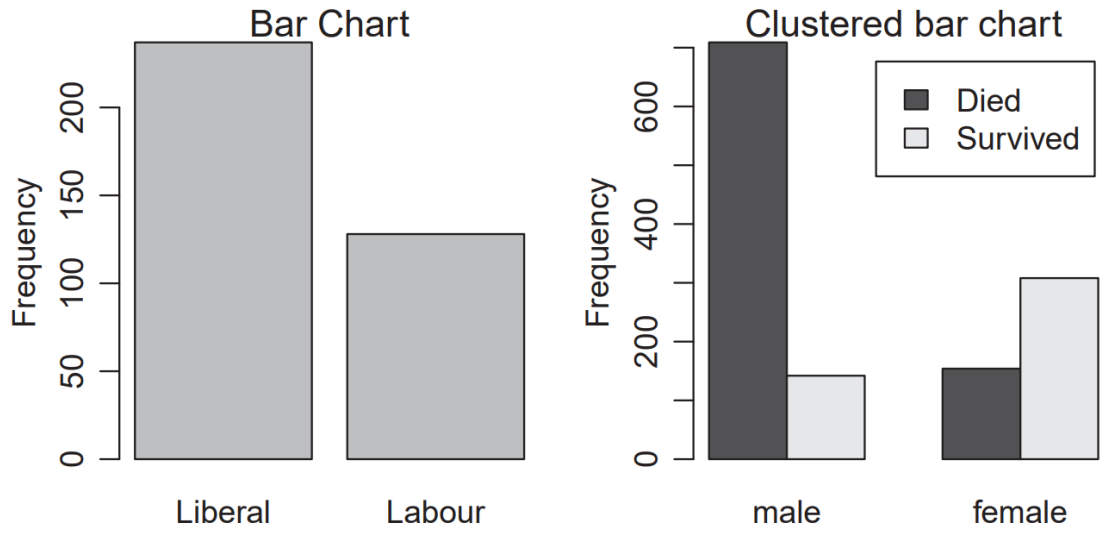


Figure 2: Barchart of frequencies and Clustered bar chart

Numerical summaries

Sample mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Sample deviation:

$$s = \sqrt{s^2}$$

Sample median

$$\tilde{x}_{0.5} = \begin{cases} x_{(\frac{n+1}{2})} \text{ if n is odd} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n+2}{2})}) \text{ if n is even} \end{cases}$$

pth sample quantile:

$$\tilde{x}_p = x_{(k)} \quad \text{where} \quad p = \frac{k-0.5}{n} \quad \text{for} \quad k \in \{1, 2, 3, \ldots, n\}$$

Inter-quartile Range:

$$IQR = \tilde{x}_{0.75} - \tilde{x}_{0.25}$$

Range based observations (IQR, median,) are much less sensitive to outliers than other measures (mean, variance, sd)

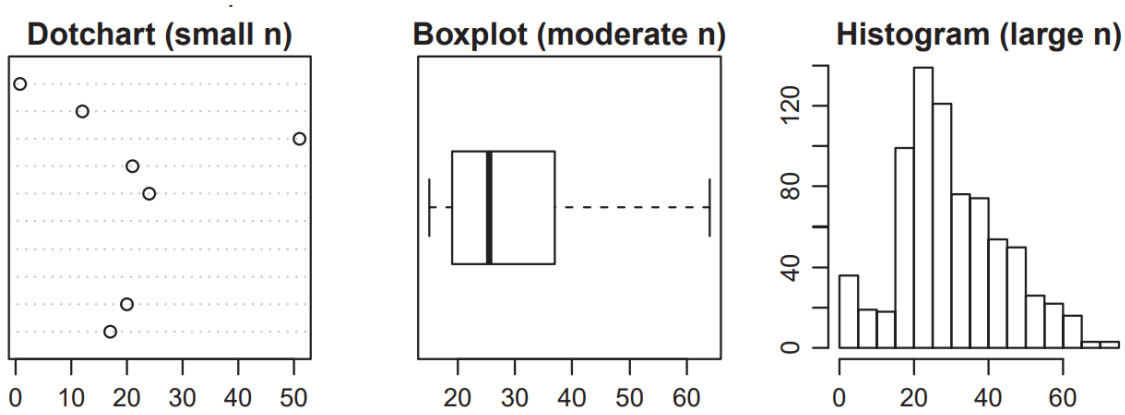


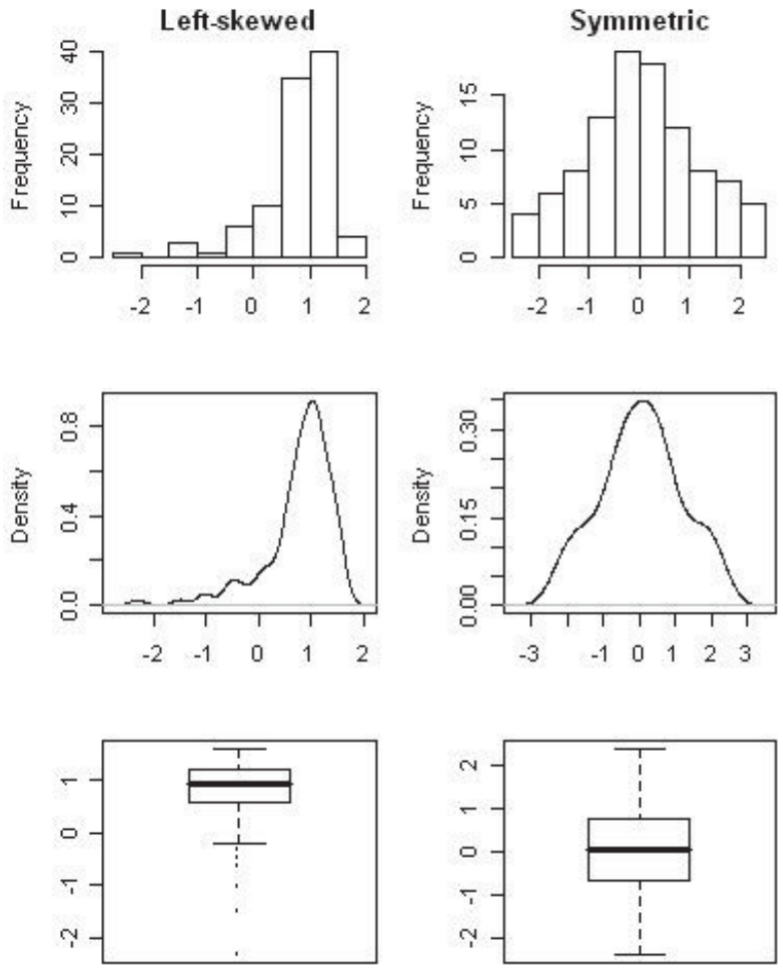
Figure 3: Dotchart Boxplot, Histogram

Graphical summaries of quantitative data

Kernel density estimator:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} w_h(x - x_i)h \rightarrow \text{bandwidth parameter}$$

Shape of a distribution



Here are some sample distributions in 3 different skews:

It's also worth checking for outliers that can influence the shape of the data

Summarising Associations Between Variables

correlation coefficeint (2 quant vars):

$$r = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{x_i - \bar{x}}{s_x})(\frac{y_i - \bar{y}}{s_y})$$

where $\bar{x}$ and s_x are the sample mean and standard deviation of x,similarly for y.

3 Types of result:

- $|r| \leq 1$
- $r = -1$
- $r = 1$

Where the second and third results are linear relationships between the two varibles (negative and postive gradient)

2 measures:

- Relationship strength → how close r is to -1 or 1
- Direction of assosiation → values less than one suggest a decreasing relationship, values greater than one suggest an increasing relationship

Associations between categorical and quantitative variables

Just use a comparative boxplot smh

Transforming Data

Linear transformations

Linear Transformations take the shape of

$$y_i = a + bx_i$$

for each i and b $\neq 0$

It doesn't affect the shape of the distribution → only the location and spread.

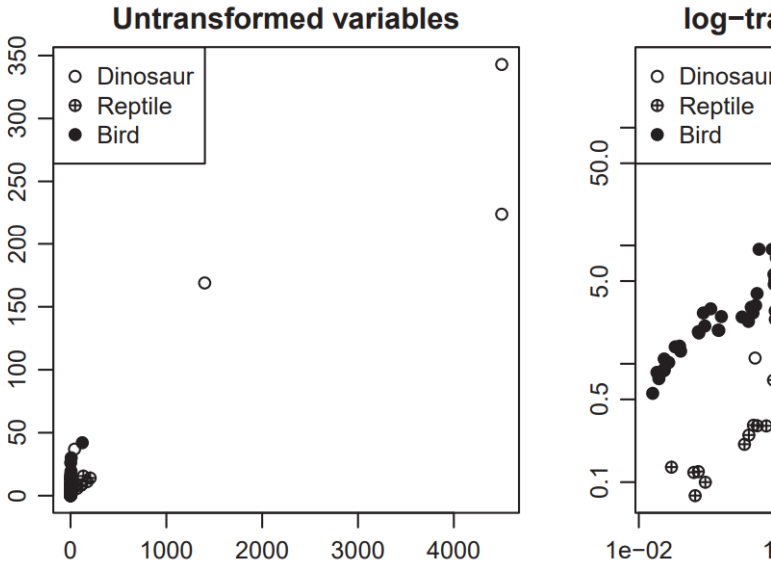
A common Linear transformation is the z-score or standardised score:

$$z = \frac{x - \bar{x}}{s_x}$$

It measures how many standard deciations above/below the value is from the mean (ie as $|z| \to 1$) the more unusal it is.

Nonlinear transformations

The most common Nonlinear transformation is a log-transformation, it can reveal intresting relationships and



structures for values that may seem too close together

Important Note: Let (y = h(x)) be some on linear transformation of real values x. In most cases:

$$\bar{y} \neq h(\bar{x})$$

ie: the mean of the transform won't be equal to the mean of the original data

Chapter 2