FIT2086 Self-Study/Revision
Summary Statistics, Basic Graphs and Revision

Daniel F. Schmidt

Faculty of Information Technology, Monash University

July 22, 2019

Outline

1. Descriptive Statistics
   - Descriptive Statistics
   - Associations Between Variables
2. Mathematics Revision

Outline

1. Descriptive Statistics
   - Descriptive Statistics
   - Associations Between Variables

2. Mathematics Revision

Basic Types of Data – Refresher

- Categorical-Nominal:
  - Discrete numbers of values, no inherent ordering
  - E.g., country of birth, sex
- Categorical-Ordinal:
  - Discrete number of states, but with an ordering
  - E.g., Education status, State of disease progression
- Numeric-Discrete:
  - Numeric, but the values are enumerable
  - e.g., Number of live births, Age (in whole years)
- Numeric-Continuous:
  - Numeric, not enumerable (i.e., real numbers)
  - E.g., Weight, Height, Distance from CBD
- Quantitative vs Qualitative:
  - Generally, categorical data is qualitative, numeric data is quantitative

Graphical representations – Refresher

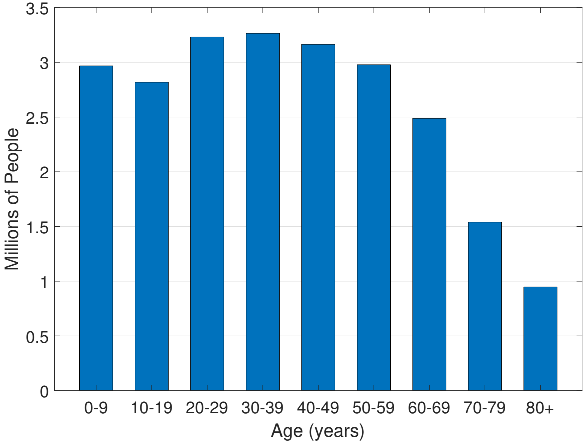
- It is often useful to visualise data
  - Can sometimes quickly reveal patterns
  - However, going beyond two dimensions is problematic
- For categorical data, standard visualisations include:
  - Frequency tables
  - Bar graphs
  - Pie charts (not recommended!)
- For numeric data (continuous and discrete), we can use:
  - Histograms
  - Box-and-whisker plots

Frequency Tables

Age (years)	Number of People
0-9	2,967,425
10-19	2,818,778
20-29	3,231,395
30-39	3,265,526
40-49	3,164,712
50-59	2,977,883
60-69	2,488,396
70-79	1,540,373
80+	947,411

Australian Population by Age (2016 Census)

Bar charts



Australian population by age (2016 Census)

Histograms

- Histograms are a special type of bar chart
  - Bar-charts only applicable to categorical data
- Group numeric data into categories by putting it bins
- If $\mathbf{y} = (y_1, \ldots, y_n)$ are our data points, we divide them between K equally spaced bins, i.e.,
  - The number of samples that fall in bin (category) k are

$$v_k = \#\{y_j \in (\min\{\mathbf{y}\} + (k-1)w, \;\; \min\{\mathbf{y}\} + kw)\}$$

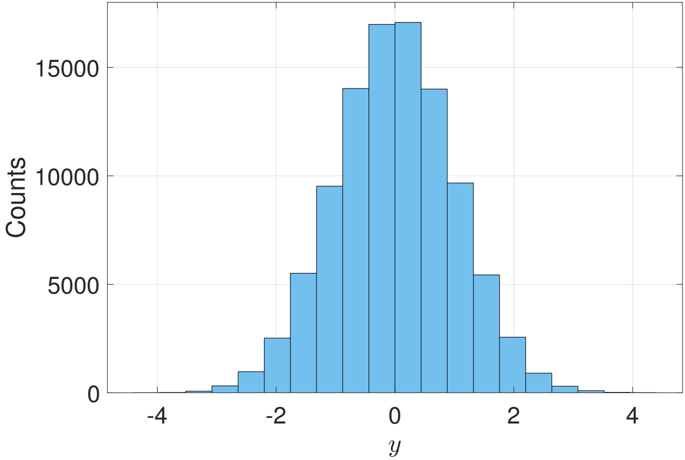
  where

$$w = \frac{\max\{\mathbf{y}\} - \min\{\mathbf{y}\}}{K}$$

  is the width of the bins

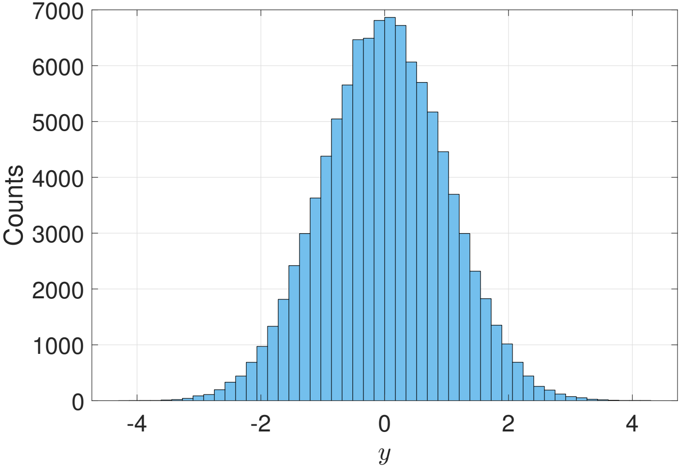
  $\Longrightarrow$ plot $v_1, \ldots, v_K$ using bar-chart

Histograms: Example



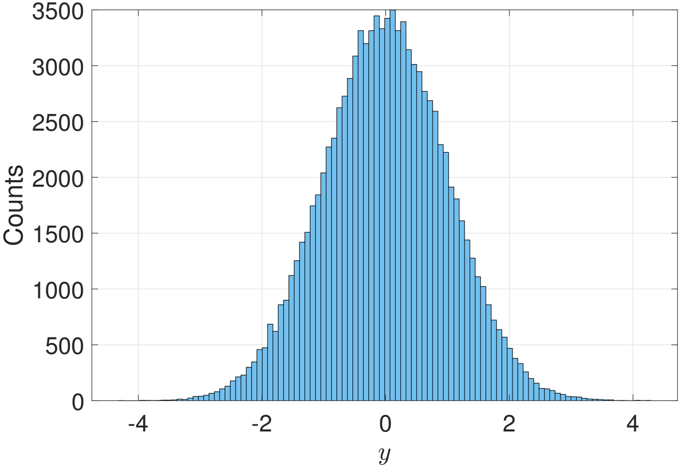
Histogram with $K = 20$ bins

Histograms: Example



Histogram with $K = 50$ bins; looks smoother

Histograms: Example



Histogram with $K = 100$; starting to look ragged

Descriptive Statistics

- Descriptive statistics summarise aspects of the data
- What is a "statistic"?
  - Let $\mathbf{y}$ denote a sample of data
  - Then a statistic is any function $s(\mathbf{y})$ of the data
- Some functions (statistics) more useful than others
  - But all describe properties of the data

- Let $\mathbf{y} = (y_1, \ldots, y_n)$ be a sample of n data points
- The the most common measure of centrality, or averageness, is the arithmetic mean

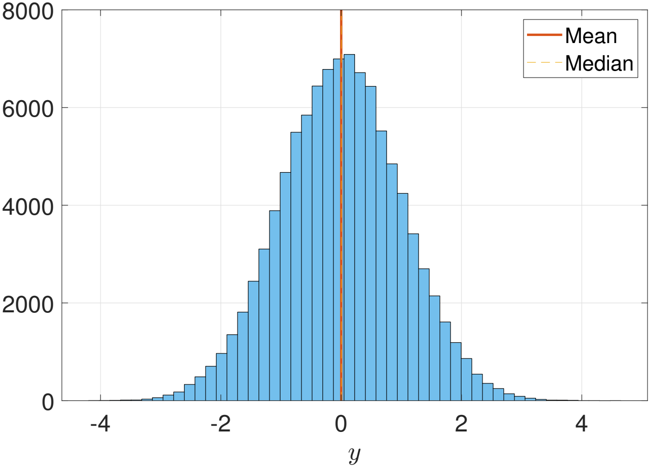
$$\bar{y} = \frac{1}{n} \sum_{j=1}^{n} y_j$$

- The mode is the most frequently occuring value in the sample
  - Of limited use for continuous numeric data
- Another common measure is the median, $\mathrm{med}(\mathbf{y})$
  - Value such that 50% of samples have values less than $\mathrm{med}(\mathbf{y})$
  - Easily found by sorting samples and finding middle sample

Mean vs Median

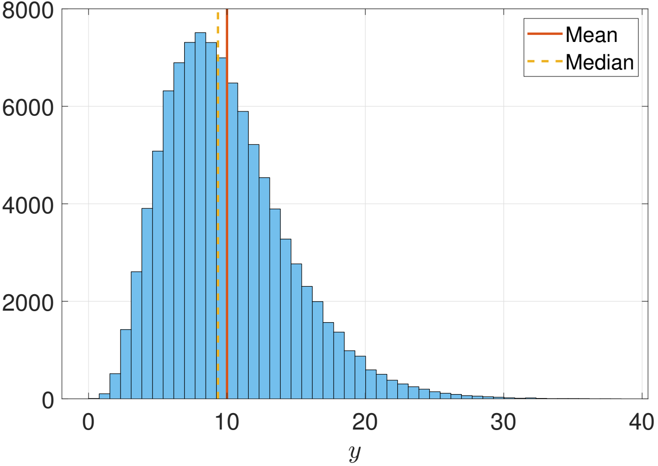
- The mean uses *all* the values of the sample
  - Any change to any sample changes the mean
  - The mean can be changed as much as desired by changing just one sample by a large enough amount
- The median uses at most two of the values of the sample
  - Is very resistant to changes to the samples not in the middle
- Example:
  - $\mathbf{y} = (1, 2, 3, 4, 5) \Rightarrow \bar{y} = 3, \quad \mathrm{med}(\mathbf{y}) = 3$
  - $\mathbf{y} = (1, 2, 3, 4, 50) \Rightarrow \bar{y} = 12, \quad \mathrm{med}(\mathbf{y}) = 3$

Mean vs Median: Symmetric Distributions



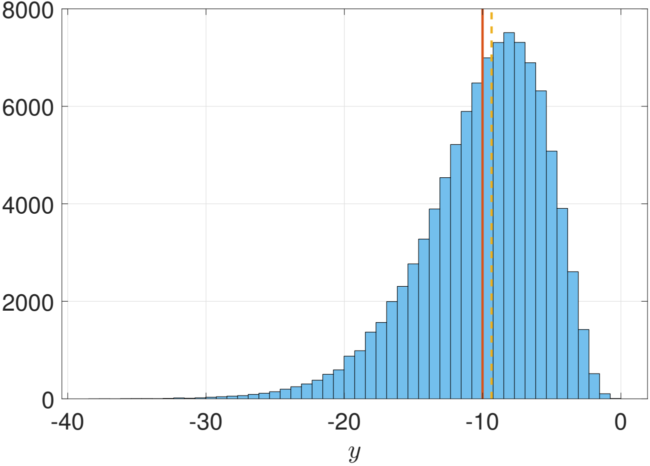
Symmetric distribution of data; mean and median (nearly) the same

Mean vs Median: Positively Skewed Data



Positively skewed data; mean greater than median

Mean vs Median: Negatively Skewed Data

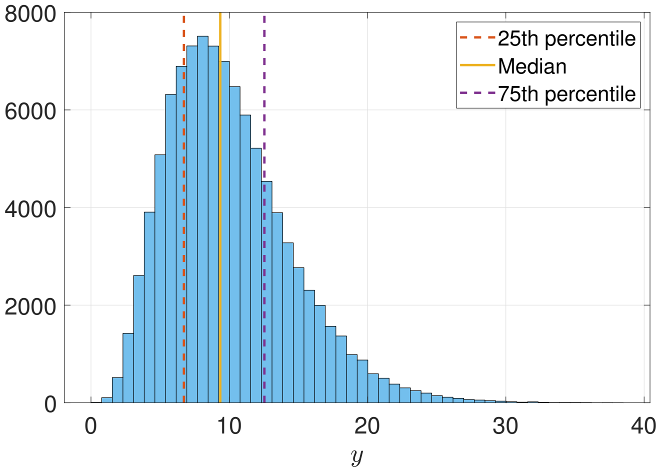


Negatively skewed data; mean less than median

Percentiles

- More generally, we can define the percentiles
  - The p-th percentile is the value, $Q(\mathbf{y}, p)$ such that $p\%$ of the values of the sample are lower than $Q(\mathbf{y}, p)$
- The median is simply the 50th percentile, $Q(\mathbf{y}, 50)$
- Other important percentiles are the 1st and 3rd quartiles
  - i.e., the 25th and 75th percentiles

Percentiles



Measures of Spread (1)

- Measures of centrality tell us about the typical value of the sample
- Measures of spread tell us how much the samples differ, on average, from the typical value
- The most straightforward is the range

$$\mathrm{rng}(\mathbf{y}) = \max\{\mathbf{y}\} - \min\{\mathbf{y}\}$$

where

- $\min\{\mathbf{y}\}$ denotes the minimum value in the sample;
- $\max\{\mathbf{y}\}$ denotes the maximum value in the sample.

Measures of Spread (2)

- The most common measure of spread used is the sample standard deviation

$$s(\mathbf{y}) = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

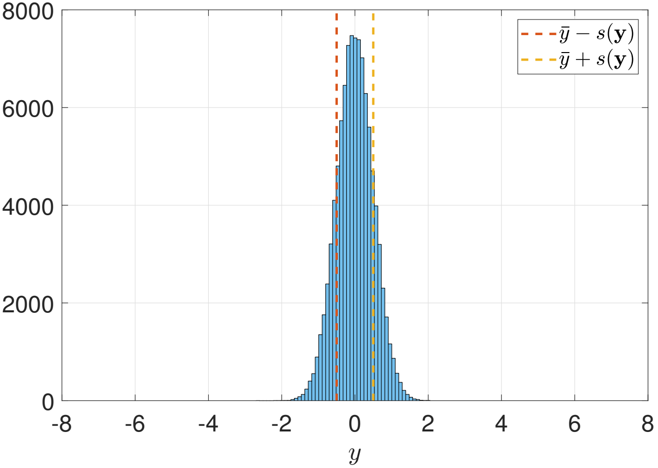
- The sample standard deviation is the arithmetic mean of the squared deviations from the sample mean
  $\Longrightarrow$ has the same unit as the data

- Like the mean, is sensitive to changes in the sample
- Often, the sample variance

$$v(\mathbf{y}) = s^2(\mathbf{y})$$

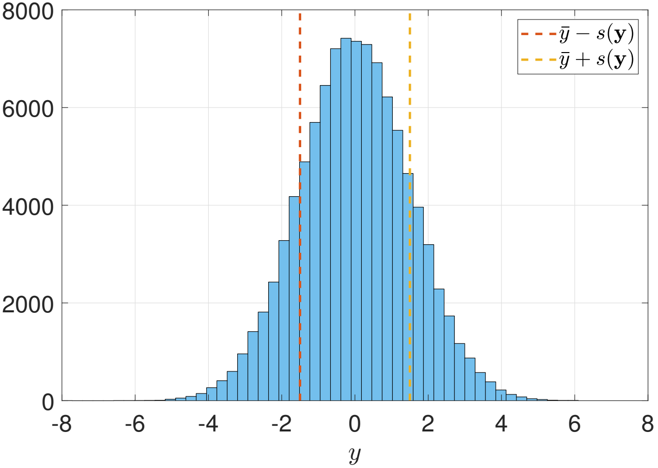
  is used, as it can be easier to work with

Measures of Spread: Example



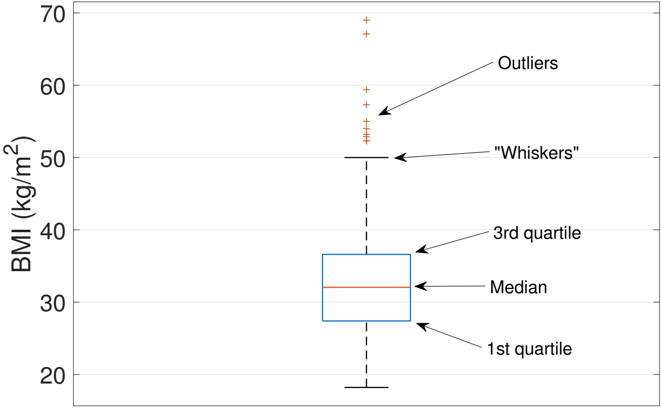
$$\operatorname{rng}(\mathbf{y}) = 4.63 \ (\min\{\mathbf{y}\} = -2.61, \ \max\{\mathbf{y}\} = 2.01), \ s(\mathbf{y}) = 0.5$$

Measures of Spread: Example



$$\text{rng}(\mathbf{y}) = 13.89 \ (\min\{\mathbf{y}\} = -7.84, \ \max\{\mathbf{y}\} = 6.05), \ s(\mathbf{y}) = 1.5$$

Visualising Continuous Data: Boxplots



Boxplot graphically captures centrality, spread and skewness in one plot

Association Between Two Continuous Variables

- Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ be two numeric variables measured on the same objects
  - We might ask if there is an association between $\mathbf{x}$ and $\mathbf{y}$
- Pearson correlation measures linear association

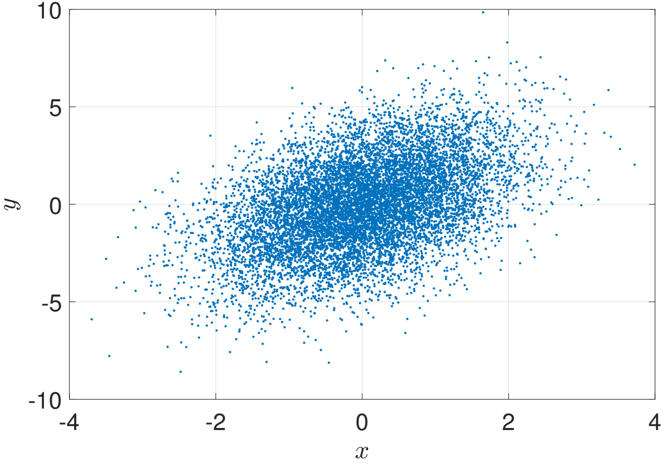
$$R(\mathbf{x}, \mathbf{y}) = \frac{\sum_{j=1}^{n} (x_j - \bar{x})(y_j - \bar{y})}{n\, s(\mathbf{x}) s(\mathbf{y})}$$

  - Correlation is always between -1 (completely negatively correlated) and 1 (completely positively correlated)
  - A correlation of zero implies there is no linear association $\implies$ does not imply no non-linear association
- Remember: correlation $\neq$ causation!

Scatter Plots

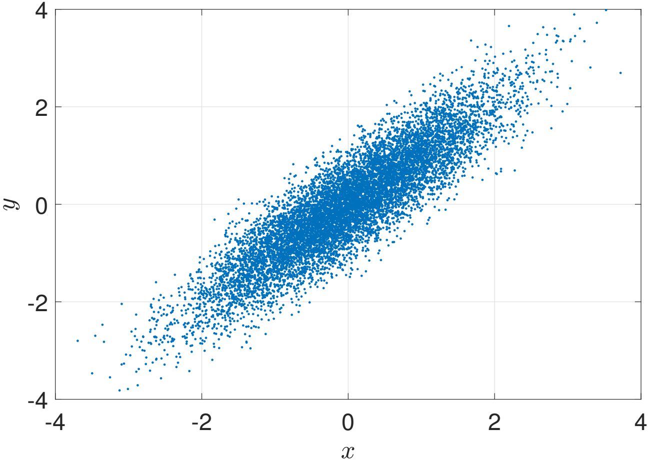
- Scatter plots help us visualise relationships between two (usually) numeric variables
  - Plot points, with one variable on x-axis and one on y-axis
- Can be used to visually look for association
- Correlation coefficients are statistics that quantatatively measure the strength of the association between two variables
  - The two can be combined for more information
- Three-variable scatter plots, like almost all three-dimensional plots, should be avoided

Correlation/Scatter Plot Example (1)



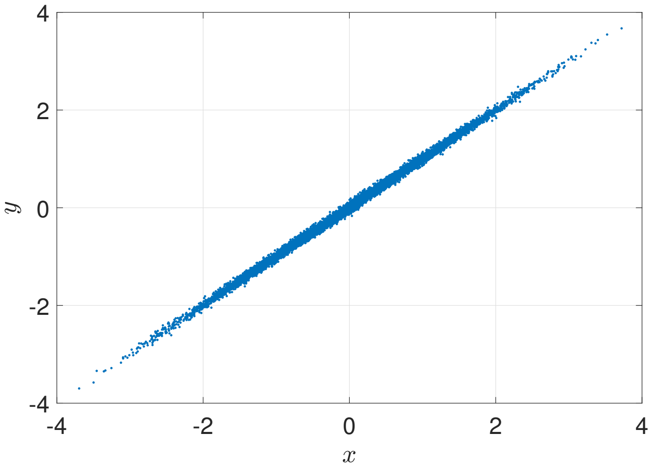
$$R \approx 0.44$$

Correlation/Scatter Plot Example (2)



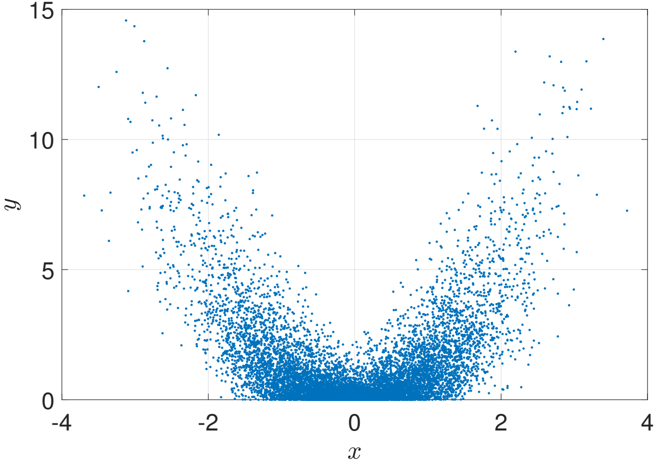
$$R = 0.9$$

Correlation/Scatter Plot Example (3)



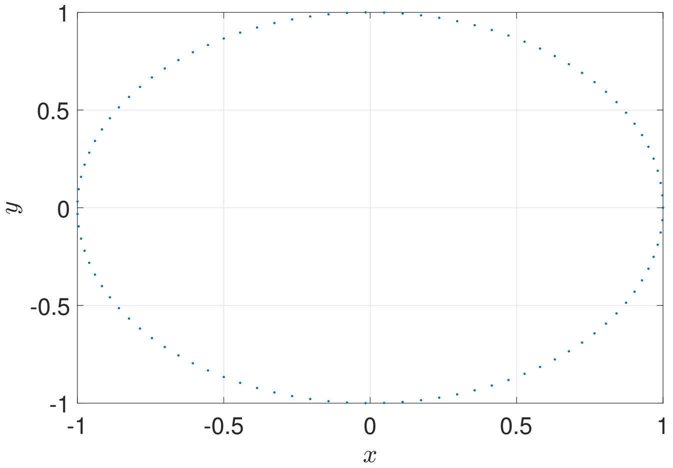
$$R \approx 0.999$$

Correlation/Scatter Plot Example (4)



$R \approx 0.01$ – though clearly associated, as $y = x^2 + \text{noise}$

Correlation/Scatter Plot Example (5)

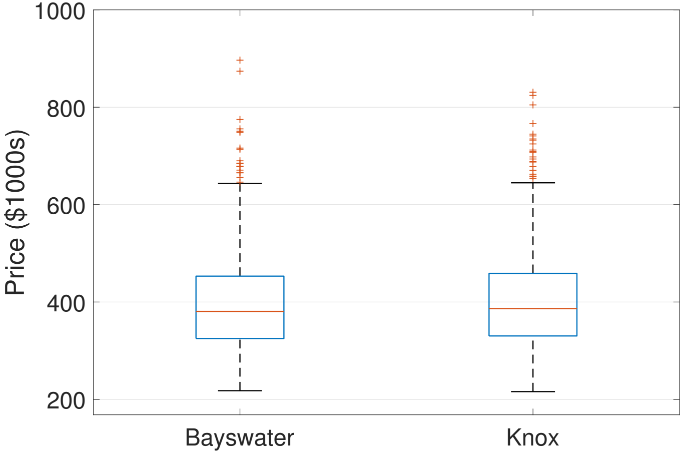


$R = 0$, though there is a deterministic association between x and y

Association Between Categorical and Numeric Variables

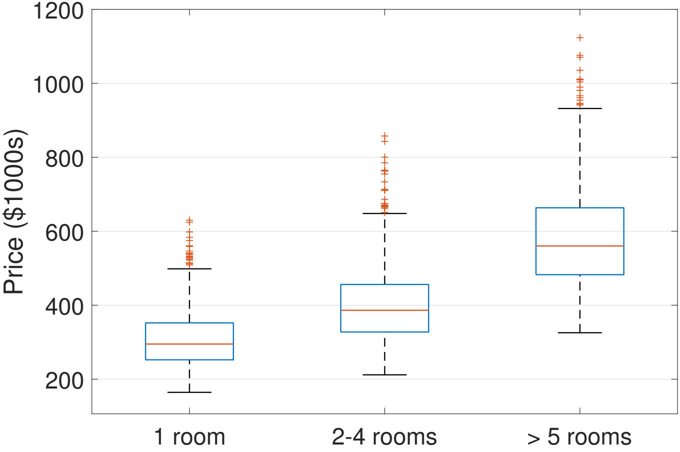
- If $\mathbf{x}$ is categorical, and $\mathbf{y}$ is numeric, how to visualise?
- A standard approach is the side-by-side boxplot
  - Divide the data between categories, then plot boxplots for each group
  - Do the boxplots look different?
- If $\mathbf{x}$ and $\mathbf{y}$ are both categorical, we can use a side-by-side bargraph instead
  - Are the distributions/bargraphs different between categories?
  - If so, there is a possible association

Example: Categorical and Numeric Variables (1)



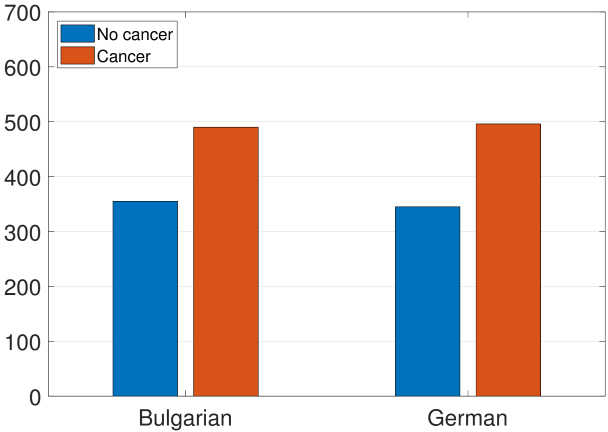
Distribution of price similar between suburbs

Example: Categorical and Numeric Variables (2)



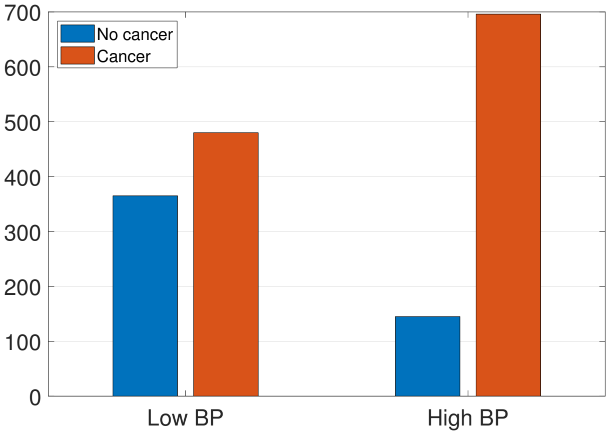
Distribution of price varies greatly with number of rooms

Example: Two Categorical Variables (1)



Frequency of cancer does not seem to change with ethnicity; unlikely to be associated

Example: Two Categorical Variables (2)



Frequency of cancer changes substantially with blood pressure; likely to be strong association

Outline

Useful Identities (1)

- The following logarithmic identities will be useful

$$
\begin{aligned}
\log 1 &= 0 \\
\log a\, b &= \log a + \log b \\
\log a/b &= \log a - \log b \\
\log a^b &= a \log b
\end{aligned}
$$

- The following exponential identities will be useful

$$
\begin{aligned}
e^a e^b &= e^{a+b} \\
e^{-a} &= 1/e^a \\
(e^a)^b &= e^{ab}
\end{aligned}
$$

Useful Identities (2)

- The following calculus identities will be useful

$$
\begin{aligned}
\frac{d}{dx}\{x^n\} &= nx^{n-1} \\
\frac{d}{dx}\{\log x\} &= \frac{1}{x} \\
\frac{d}{dx}\{e^x\} &= e^x \\
\text{Linearity}: \frac{d}{dx}\{a\,f(x)+b\} &= a\frac{d}{dx}\{f(x)\}+b \\
\text{Product Rule}: \frac{d}{dx}\{f(x)g(x)\} &= g(x)\frac{d}{dx}\{f(x)\}+f(x)\frac{d}{dx}\{g(x)\} \\
\text{Chain Rule}: \frac{d}{dx}\{f(g(x))\} &= \frac{d}{dg(x)}\{f(g(x))\}\cdot\frac{d}{dx}\{g(x)\}
\end{aligned}
$$

Useful Identities (3)

- If we have a function, $f(x, y)$, of two variables, the partial derivative

$$\frac{\partial f(x,y)}{\partial x}$$

is found by differentiating $f(x, y)$ w.r.t. x treating y as a constant.

- Example:

$$\begin{aligned}
\frac{\partial}{\partial x}\left\{y \log\left(x^2 y + 1\right)\right\} &= y\,\frac{\partial}{\partial x}\left\{\log\left(x^2 y + 1\right)\right\} \text{ (linearity)} \\
&= y \cdot \frac{1}{x^2 y + 1} \cdot \frac{\partial}{\partial x}\left\{x^2 y + 1\right\} \text{ (chain rule)} \\
&= \frac{2xy^2}{x^2 y + 1}
\end{aligned}$$

Reading/Terms to Revise

- Reading for this week: Chapter 2 of Ross.
- Terms you should know:
  - Histogram;
  - Measures of central tendency: mean, median, mode;
  - Measures of dispersion: standard deviation, variance, range;
  - Percentiles and quartiles;
  - Scatter plot;
  - Correlation coefficient