Chapter summary

Statistics

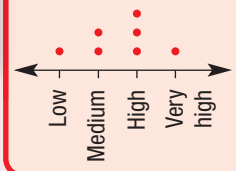
Data

Categorical
- Nominal (red, blue, ...)
- Ordinal (low, medium, ...)

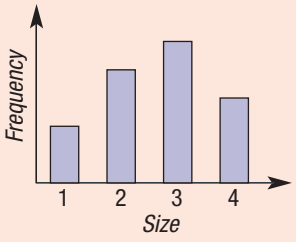
Numerical
- Discrete (1, 2, 3, ...)
- Continuous (0.31, 0.481, ...)

Graphs for a single set of categorical or discrete data

Dot plot



Column graphs



Stem-and-leaf plot

Stem	Leaf
0	1 6
1	2 7 9
2	3 8
3	4

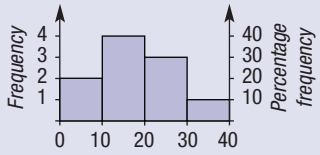
2|3 means 23

Frequency tables

Class interval	Frequency	Percentage frequency
0–	2	20%
10–	4	40%
20–	3	30%
30–40	1	10%
Total	10	100%

$$\text{Percentage frequency} = \frac{\text{frequency}}{\text{total}} \times 100$$

Histogram



Measures of centre

- Mean $(\bar{x}) = \frac{\text{sum of all values}}{\text{number of scores}}$
- Median (Q_2) = middle value of ordered data

odd number	even number
1 4 6 9 12	1 2 4 6 7 11
median	Median $= \frac{4+6}{2} = 5$

- Mode = most common value
- A data set with two modes is bimodal.

Measures of spread

- Range = max − min
- Interquartile range (IQR) = $Q_3 - Q_1$

Quartiles

Q_1: above 25% of the data (middle value of lower half)
Q_3: above 75% of the data (middle value of upper half)

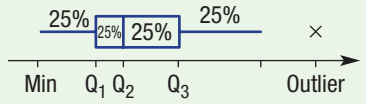
2 3 5 7 8 | 9 11 12 14 15

$Q_1 = 5$ $Q_2 = 8.5$ $Q_3 = 12$

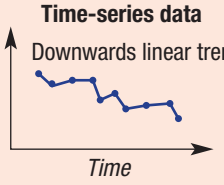
1 2 3 | 6 | 7 10 13

$Q_1 = 2$ Q_2 $Q_3 = 10$

Box plots

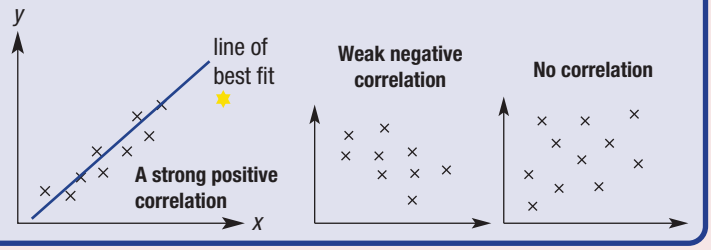


Time-series data



Bivariate data

- Two related variables
- Scatter plot



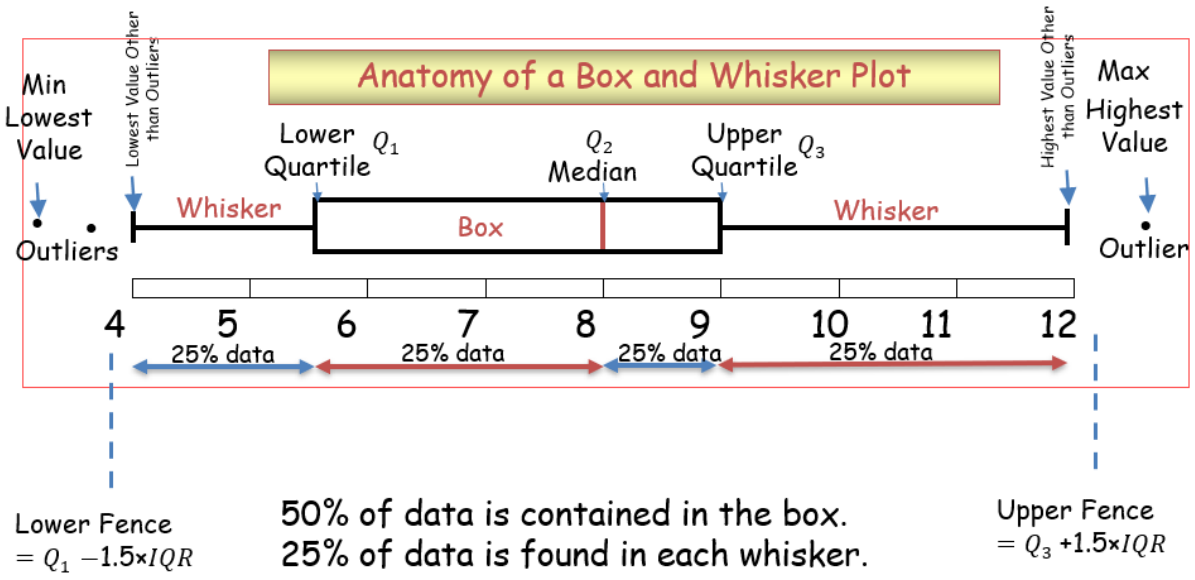
Outliers

- Single data set
  - less than $Q_1 - 1.5 \times \text{IQR}$ or more than $Q_3 + 1.5 \times \text{IQR}$
- Bivariate
  - not in the vicinity of the rest of the data

ISBN 978-1-108-87859-3
© Greenwood et al. 2021 Cambridge University Press
Photocopying is restricted under law and this material must not be transferred to another party.

Box and Whisker Plots

Box plots are graphical representations of 5 number summary.



Lower Fence $= Q_1 - 1.5 \times IQR$

50% of data is contained in the box.
25% of data is found in each whisker.

Upper Fence $= Q_3 + 1.5 \times IQR$

The standard deviation

$$s^2 = \frac{\text{sum of all (data value} - \text{mean)}^2}{\text{total freqency} - 1} = \frac{\sum(x-\bar{x})^2}{n-1}, \qquad s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

Example VCAA 2016 Sample Question 3 / VCAA 2014 Question 3

The mean, $\bar{x}$, and the standard deviation, s_x, of the variable, *distance*, for these seven drivers are closest to

***Distance* (km)**	$\boldsymbol{Distance - \bar{x}}$
4.2	$4.2 - 2.8 = 1.4$
0.8	$0.8 - 2.8 = -2$
3.9	$3.9 - 2.8 = 1.1$
5.6	$5.6 - 2.8 = 2.8$
0.9	$0.9 - 2.8 = -1.9$
1.7	$1.7 - 2.8 = -1.1$
2.5	$2.5 - 2.8 = -0.3$

$$\bar{x} = \frac{4.2 + 0.8 + 3.9 + 5.6 + 0.9 + 1.7 + 2.5}{7} = \frac{19.6}{7} = 2.8$$

$$s_x^2 = \frac{1.4^2 + (-2)^2 + 1.1^2 + 2.8^2 + (-1.9)^2 + (-1.1)^2 + (-0.3)^2}{7-1}$$

$$\Rightarrow s_x^2 = \frac{19.92}{6} = 3.32, \qquad \therefore s_x = \sqrt{3.32} = 1.82208671\ldots \approx 1.8$$

The mean distance travelled is 2.8 km and
the standard deviation from the mean is 1.8 km.