Exploring, Representing, and Interpreting Data

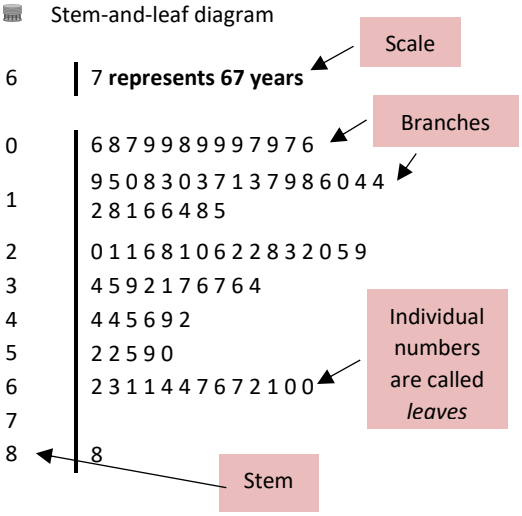
A. Exploring Data

Tally

Tallying is a quick, straightforward way of grouping data into suitable intervals.

Stated age (years)	Tally	Frequency
0 - 9	𝍸 𝍸 |||	13
10 – 19	𝍸 𝍸 𝍸 𝍸 𝍸 |	26
20 – 29	𝍸 𝍸 𝍸 |	16
30 – 39	𝍸 𝍸	10
40 – 49	𝍸 |	6
50 – 59	𝍸	5
60 – 69	𝍸 𝍸 ||||	14
70 - 79		
80 - 89	|	1
⋮		
130 - 139	|	1
Total		92

Stem-and-leaf diagram



HIGH 138

Extreme values are placed on a separate HIGH or LOW branch.

B. Measures of Central Tendency

When describing a typical value to represent a data set most people think of a value at the centre and use the word average. When using the word average they are often referring to arithmetic mean, which is usually just called mean, which is obtained by adding up the data values and divde by the total number of data values.

Σ notation and the mean, $\bar{x}$

A sample size n taken from a population can be identified as follows. The first item is called x_1, the second item x_2 and so on up to x_n. The sum of these n items of data is given by $x_1 + x_2 + x_3 + \ldots + x_n$. A shorthand for this is $\sum_{i=1}^{i=n} x_1$ or $\sum_{i=1}^{n} x_i$. This read as 'the sum of all the terms x_1 when i equals to n'. If there is no ambiguity about the number of items of data, the subscripts i can be dropped and $\sum_{i=1}^{n} x_i$ becomes $\sum x$.

The mean of these n items of data is written as

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

where $\bar{x}$ is the symbol for the mean. It is usual to write $\bar{x} = \frac{\sum x}{n}$.

The mean from a frequency table

Often data is presented in a frequency table. The notation for the mean is slightly different in such cases. The difference is that each item, x_n, has to be multiplied with it's frequency, f_n. Therefore it is written as $\bar{x} = \frac{\sum xf}{n}$ with n the equivalent of $\sum f$.

Median

The median is the value of the middle item when all the data items are ranked in order. If there are n items of data then the median is the value of the $\frac{n+1}{2}$th item.

- If n is odd then there is a middle value and this is the median.
- If n is even, the middle values are a and b. Therefore, the median is $\frac{a+b}{2}$

Mode

The **mode** is the value which occurs most frequently. If two non-adjacent values occurs more frequently than the rest, the distribution is said to be **bimodal**.

C. Grouped Data

Grouping means putting the data into a number of classes. The number of data items falling into any classes is called the frequency for that class.

When numerical data are grouped, each item of data falls within a class interval lying between class boundaries.

- Discrete data

Numbers of Cars, x	Frequency, f
0 - 9	5
10 – 19	8
20 – 29	13
30 – 39	20
40 – 49	22
50 – 59	21
60 - 69	11
Total	**100**

- Estimating the mean

To estimate the mean you first assume that all the values are equally spaced about a midpoint. The midpoints are taken as representative values of the intervals.

Numbers of Cars, x (mid-values)	Frequency, f	$x \times f$
4.5	5	22.5
14.5	8	116.0
24.5	13	318.5
34.5	20	690.0
44.5	22	979.0
54.5	21	1144.5
64.5	11	715.0
Total	**100**	**3985.5**

The mean is given by:

$$\bar{x} = \frac{\sum xf}{\sum f}$$

$$= \frac{3985.5}{100} = 39.855$$

- Continous data

Height, h	Mid-value, x	Frequency, f	xf
157 < h ≤ 159	158	4	632
159 < h ≤ 161	160	11	1760
161 < h ≤ 163	162	19	3078
163 < h ≤ 165	164	8	1312
165 < h ≤ 167	166	5	830
167 < h ≤ 169	168	3	504
Totals		**100**	**8116**

$$\bar{x} = \frac{8116}{50}$$

$$= 162.32$$

D. Measures of Spread (variation)

- Range

The simplest measure of spread is range. This is just the difference between the largest value in the data set and the smallest value.

Range = largest - smallest

- The variance and standard deviation

An alternative to ignoring the signs is to square the differences or deviations. This gives rise to a measure of spread called the variance, which when square rooted gives the standard deviation.

$$Variance = \frac{\Sigma(x-\bar{x})}{n}$$

The square root of the varience is called the standard deviation.

$$sd = \sqrt{\frac{\Sigma(x-\bar{x})}{n}}$$

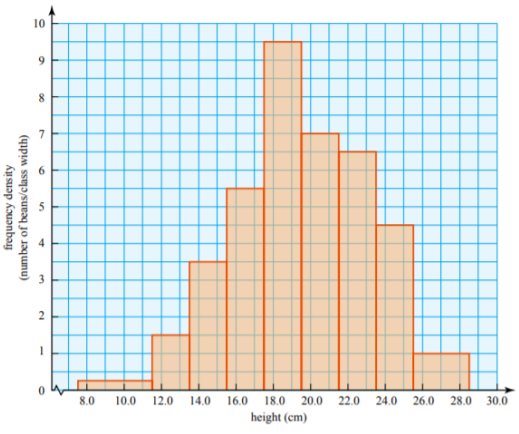
Though not as easy to calculate as the mean absolute deviation, the standard deviation has an important role in the study of more advanced statistics.

E. Representating and Interpreting Data

Histograms

Histograms are used to illustrate continous data. The collumns in a histogram may have different widhts and the area of each column is proportional to the frequency.

Height (cm)	Frequency	Class width (cm)	Frequency density
$7.5 \leqslant x < 11.5$	1	4	0.25
$11.5 \leqslant x < 13.5$	3	2	1.5
$13.5 \leqslant x < 15.5$	7	2	3.5
$15.5 \leqslant x < 17.5$	11	2	5.5
$17.5 \leqslant x < 19.5$	19	2	9.5
$19.5 \leqslant x < 21.5$	14	2	7
$21.5 \leqslant x < 23.5$	13	2	6.5
$23.5 \leqslant x < 25.5$	9	2	4.5
$25.5 \leqslant x < 28.5$	3	3	1

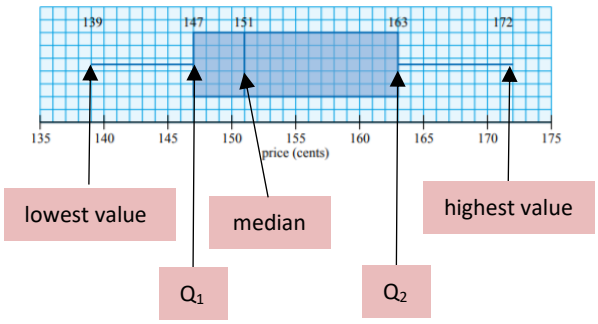


When the class widths are unequal you can use

$$frequency\ density = \frac{frequency}{class\ width}$$

Box-and-whiskers plots (boxplots)

Boxplots consists of three quartiles and two extreme values.

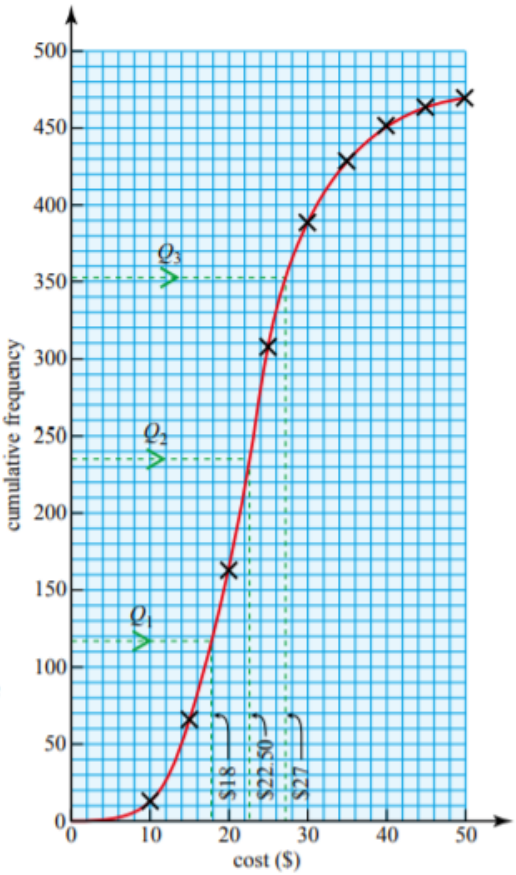


Cumulative frequency graphs

When working with large data sets or grouped data, percentiles and quartiles can be found from cumulative frequency graphs.

Cost, C ($)	Frequency	Cost	Cumulative frequency
$0 \leqslant C < 10$	13	$C < 10$	13
$10 \leqslant C < 15$	53	$C < 15$	66
$15 \leqslant C < 20$	97	$C < 20$	163
$20 \leqslant C < 25$	145	$C < 25$	308
$25 \leqslant C < 30$	81	$C < 30$	389
$30 \leqslant C < 35$	40	$C < 35$	429
$35 \leqslant C < 40$	23	$C < 40$	452
$40 \leqslant C < 45$	12	$C < 45$	464
$45 \leqslant C < 50$	6	$C < 50$	470

A cumulative frequency graph is obtained by plotting the upper boundary of each class against the cumulative frequency. The points are joined by a smooth curve.



In this example the actual values are unknown and the median must therefore be an estimate. It is usual in such cases to find the estimated value of the $\frac{n}{2}$th item.

F. Measure of Central Tendency and of Spread Using Quartiles

Quartiles for small data sets

For small data sets, calculation of Q_2, the median, is straightforward. However, there are no standard formulae for the calculation of lower quartile (Q_1) and upper quartile (Q_2). The quartiles depend on whether the number of items, n, is even or odd

- If n is even, there will be an equal number of items in the lower half and upper half of the data. To calculate Q_1, find the median of the lower half. To calculate Q_3, find the median of the upper half.
  Example:
  { 1, 3, 9, 14, 19, 20, 22, 40}
  Lower half – Upper half
  $Q_1 = \frac{3+9}{2} = 6$
  $Q_2 = \frac{20+22}{2} = 21$
- If n is odd, the lower half of the data are all the items below the median meanwhile the upper half are all the items above the median.
  Example:
  (1, 2, 10, 16, 19, 23, 30}
  Lower half – Median – Upper half
  $Q_1 = 2$
  $Q_2 = 23$

Interquartile range or quartile spread

The difference between the lower and upper quartiles is known as the interquartile range or quartile spread.

Interquartile range (IQR) = $Q_3 - Q_1$

Outliers

Data that are more than 1.5 × IQR beyond the lower or upper quartiles are regarded as outliers.

The corresponding boundary values beyond which outliers may be found are:

$Q_1 - 1.5 \times (Q_3 - Q_1)$ and $Q_3 - 1.5 \times (Q_3 - Q_1)$