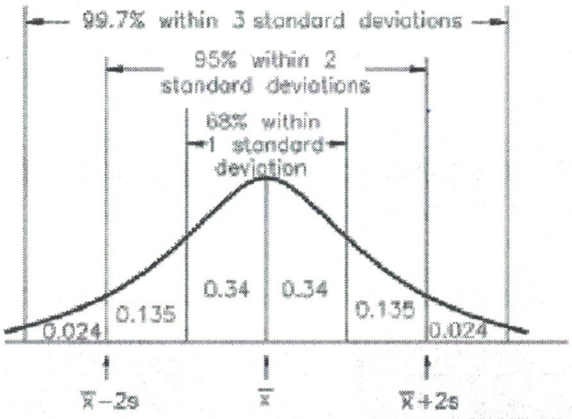
1 General properties of frequency (probability) distributions

1.1 The Empirical rule and standard deviation

For a distribution similar to the normal distribution (Bell curve), about 68% of data lies within 1 standard deviation, about 95% lies within 2 standard deviations and about 99.7% of all data lies within 3 standard deviations of the mean as shown:



Exercise 1: Estimate the percentage of the data points such that $x > \bar{x} + 2s$ or $x < \bar{x} - 2s$ for an approximately *normal* distribution. The occurence of such values of x is often considered "rare".

2 Measures of Position (relative standing)

2.1 Z scores

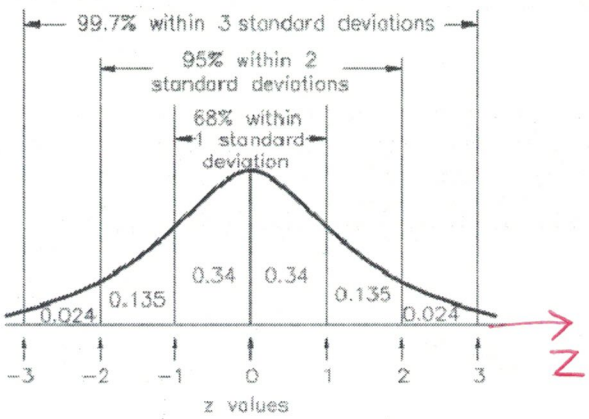
The standard score, or z score is the number of standard deviations that a given value of x is above or below the mean. It is found by using one of these formulas:

$$z = \frac{x - \bar{x}}{s} \quad \cdots \quad \text{for a sample}$$

$$z = \frac{x - \mu}{\sigma} \quad \cdots \quad \text{for a population}$$

$$Z = \frac{x - \text{mean}}{\text{sd}}$$

The meaning of the z score is that for data having a bell-shaped distribution, about 68% of the data has $-1 < z < 1$, about 95% of the data has $-2 < z < 2$, and about 99.7% of the data has $-3 < z < 3$, as shown below. Thus we can conclude that $-2 < z < 2$ is ordinary but $z > 2$ or $z < -2$ is unusual.

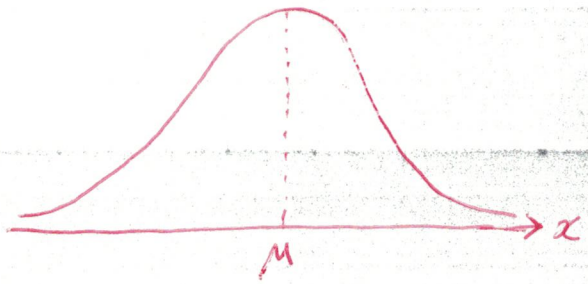


Exercise 2: What is the z score of the mean of any data set? **Exercise 3:** What is the z-score of a

$$Z = \frac{\text{mean} - \text{mean}}{sd} = 0$$

if $x > \text{mean} \iff Z > 0$

if $x < \text{mean} \iff Z < 0$



data value that is 1.5 standard deviation below the mean?

$$x = \text{mean} - 1.5\,sd$$

$$\Rightarrow Z = \frac{x - \text{mean}}{sd} = \frac{\text{mean} - 1.5\,sd - \text{mean}}{sd} = \frac{-1.5\,sd}{sd}$$

$$= -1.5$$

Exercise 4: Which is relatively better; a score of 83 on a psychology test (mean = 86, s.d. = 7) or a score of 55 on an economics test (mean = 57, s.d. = 8)?

$x_{Psy} = 83$ mean = 86, sd = 7

$$Z_{Psy} = \frac{83 - 86}{7} = -\frac{3}{7} = -0.429$$

$x_{Eco} = 55$ mean = 57, sd = 8

$$Z_{Eco} = \frac{55 - 57}{8} = \frac{-2}{8} = -0.25$$

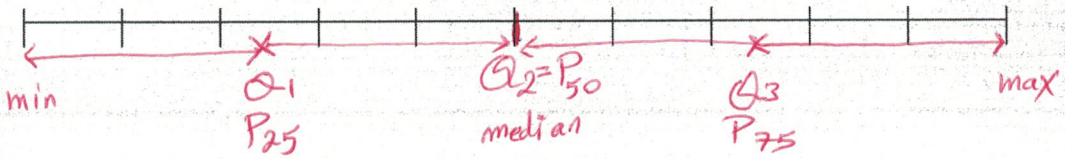
$\therefore Z_{Psy} < Z_{Eco} \Rightarrow$ 55 in Economic is better

2.2 Quartiles and Percentiles

Just as the median divides a ranked (ordered) data set into 2 groups of equal size,

- the 3 quartiles Q_1 , Q_2 and Q_3 divide a data set into 4 groups of equal size, and
- the rth percentile P_r divides data set into two sets, the first r% of the elements sorted in the order.

Exercise 5: Label the following with the median, the quartiles and typical percentiles.



To find the percentile corresponding to a data value x, use the following formula:

$$\text{percentile} = \frac{\text{number of data values less than or equal to } x}{\text{total number of data values}} \times 100^{\text{th}}$$

Exercise 6: Find the percentile of your test score which is 85, if the set of test scores were {50, 60, 75, 80, 85, 90}.

$\text{Percentile} = \frac{5}{6} \times 100 = 83.3$ i.e. $P_{83.3} = 85$

Find the test score corresponding to 50th percentile.

$\text{median} = P_{50} = Q_2 = \frac{75+80}{2} = 77.5$

Exercise 7: Find 80th, 45th and 50th percentile in $\{3, 4, 8, 11, 13, 16, 17, 19, 19\}$ $n = 9$

80th: $80\% \times 9 = 7.2$ round up 8 (8th in data is 80th Percentile)

45th: $45\% \times 9 = 4.05 \longrightarrow 5$ (5th " " 45th ")

50th: $50\% \times 9 = 4.5 \longrightarrow 5$ (5th " " 50th ")

$50^{\text{th}} = Q_2 = \text{median}$

Exercise 8: In $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ find Q_1, Q_2 and Q_3.

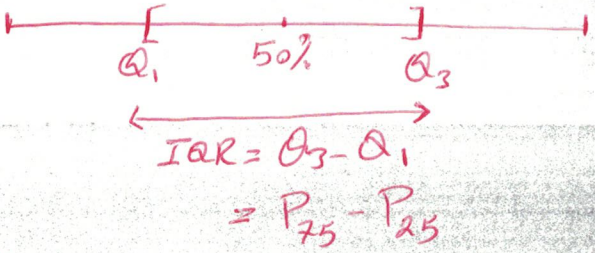
$Q_2 = \text{median} = \frac{5+6}{2} = 5.5$

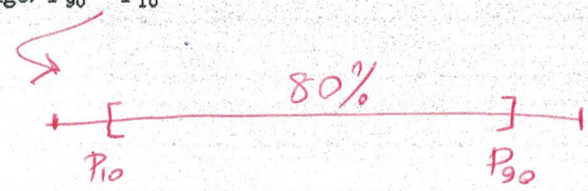
$Q_1 = \text{median of the 1st half} = 3$

$Q_3 = \text{median of the 2nd half} = 8$

2.3 Other useful quantities

- Interquartile Range (or IQR): $Q_3 - Q_1$
- Semi-interquartile Range: $\frac{Q_3 - Q_1}{2}$
- Midquartile: $\frac{Q_3 + Q_1}{2}$
- 10-90 Percentile Range: $P_{90} - P_{10}$





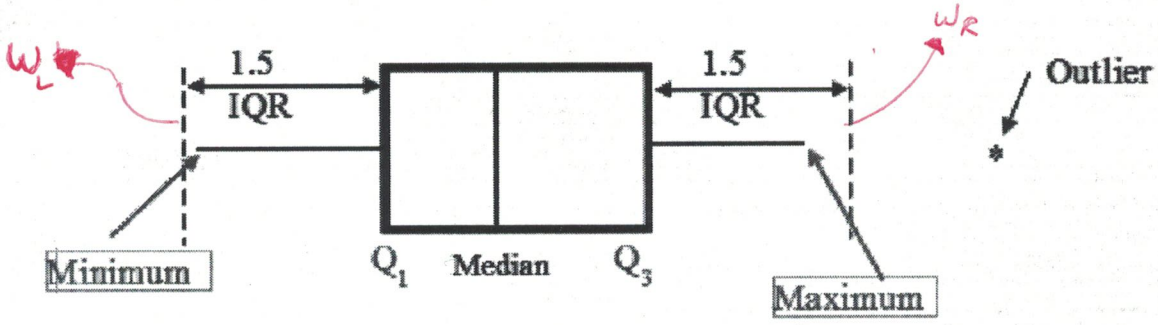
2.4 Outliers

A value that is located very far away from almost all the other values. An extreme value can have a dramatic effect on the mean, standard deviation, and on the scale of the histogram so that the true nature of the distribution is totally obscured.

Steps for Identifying Outliers

1. Arrange the data in order and find Q_1 and Q_3
2. Find the interquartile range: $\text{IQR} = Q_3 - Q_1$
3. Check the data set for any data value x such that $x < Q_1 - 1.5(\text{IQR})$ or $x > Q_3 + 1.5(\text{IQR})$.

2.5 Box plots



Exercise 9: Construct a boxplot using the following data.

$$\{1, 2, 3, \cdots, 97, 98, 99, 151\}$$

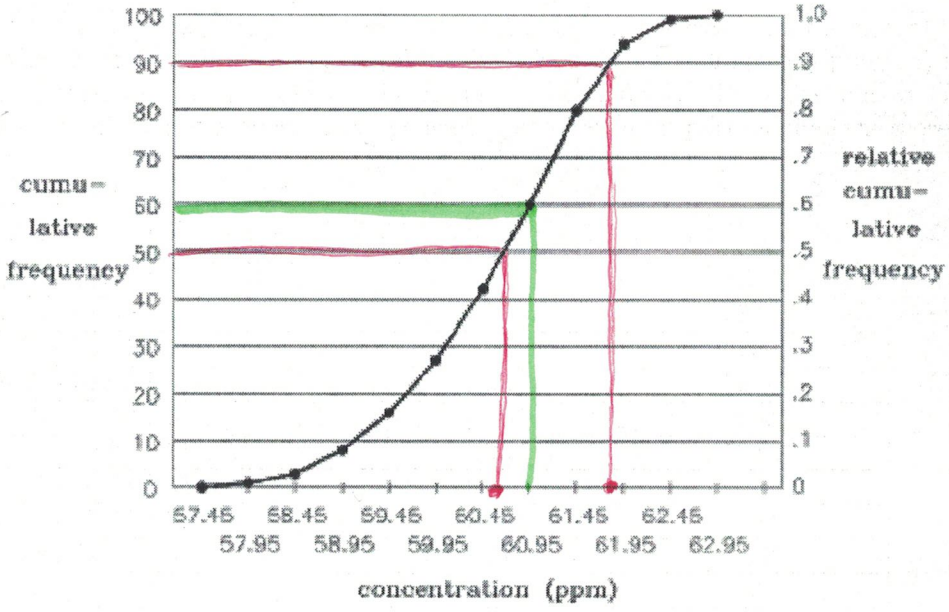
(all the integers from 1 to 99, and 151)

2.6 Cumulative distributions: The Ogive

The cumulative frequency of a class is the frequency of that class plus the sum of the frequencies of all the classes before it. An ogive (also called the cumulative frequency polygon) is a line graph of the cumulative frequency plotted versus the class boundaries. The ogive is used to show how many scores are below some value. It can be used to show what proportion of all the scores are below some value.

Exercise 10:

Class	Frequency	Cumulative Frequency	Relative Cumulative Freq.	Class label x
- 57.45	0	0	0	57.45
57.45 - 57.95	1	1 = 1+0	0.01	57.95
57.95 - 58.45	2	3 = 1+2	0.03	58.45
58.45 - 58.95	5	8 = 3+5	0.08	58.95
58.95 - 59.45	8	16 = 8+8	0.16	59.45
59.45 - 59.95	11	27 = 16+11	0.27	59.95
59.95 - 60.45	15	42	0.42	60.45
60.45 - 60.95	18	60	0.60	60.95
60.95 - 61.45	20	80	0.80	61.45
61.45 - 61.95	14	94	0.94	61.95
61.95 - 62.45	5	99	0.99	62.45
62.45 - 62.95	1	100	1.00	62.95



Exercise 11: Find the median and the 90th percentile.

$P_{50} = Q_2 = \text{median} = 60.6$

$P_{90} = 61.8$