Chebyshev's Inequality

For <u>any</u> data set (no need to be symmetric or bell-shaped), for any value of $k \geq 1$, <u>more than</u> $100(1 - 1/k^2)\%$ of the data lie within the interval $(\bar{x} - ks,\ \bar{x} + ks)$.

For $k = 2$ (say), $100(1 - 1/k^2) = 75$. Therefore, we can say that more than 75% of the data lie within the interval $(\bar{x} - 2s,\ \bar{x} + 2s)$.

Example

For a particular data set, let $\bar{x} = 40$ and $s = 3$. Here,

$\bar{x} - 2s = 40 - 2 \times 3 = 34$
$\bar{x} + 2s = 40 + 2 \times 3 = 46$

Therefore, more than $100(1 - 1/2^2)\% = 75\%$ of the data lie between 34 and 46.

Again,

$\bar{x} - 3s = 40 - 3 \times 3 = 31$
$\bar{x} + 3s = 40 + 3 \times 3 = 49$

Therefore, more than $100(1 - 1/3^2)\% = 88.9\%$ of the data lie between 31 and 49.

* k can be a decimal number.

IQR

Interquartile range (IQR) is a measure of dispersion (variability) of a data set.

IQR$= Q_3 - Q_1$

Box Plot

A box plot is drawn based on the 5-number summary of the data. It has a 'box' and two 'whiskers' (lines). The box shows the three quartiles. A distance of 1.5 times IQR is measured out below the first quartile and a whisker is drawn down to the lowest observed data point that falls within this distance. Similarly, a distance of 1.5 times IQR is measured out above the third quartile and a whisker is drawn up to the highest observed data point that falls within this distance.

Box plot helps us detect outliers.

Example:

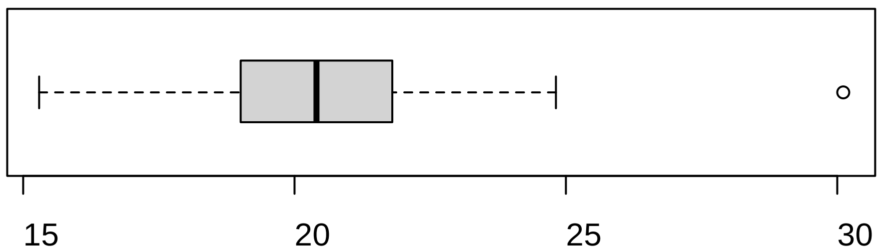
5-number summary of a data set: 15.3, 19.0, 20.4, 21.8, 30.1

IQR $= 2.8$

1.5 IQR$= 1.5 \times 2.8 = 4.2$

$19 - 4.2 = 14.8$

$21.8 + 4.2 = 26.0$



Paired data (bivariate data)

Sometimes we collect data on two related variables. That is, from each object or individual, we collect a pair of values (one value for each variable). When both variables are numerical, we often calculate 'correlation coefficient' which is discussed below.

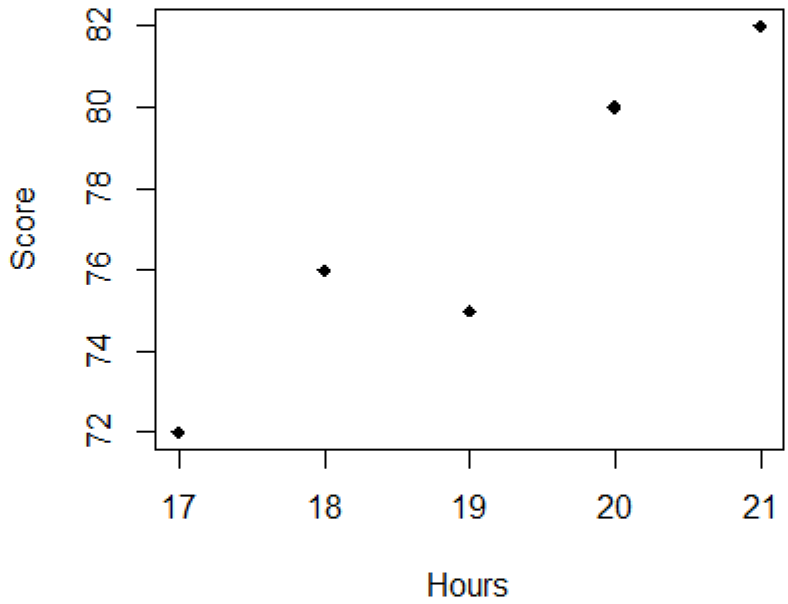
Correlation

It means association (more specifically, linear relationship) between 2 numerical variables.

Example

You want to study the relationship between hours of study and exam score. Listed below are data from a random sample of 5 students.

Hours (X):	17	18	19	20	21
Score (Y):	72	76	75	80	82



First, we draw a ‘scatter plot’ as above.

We observe that ‘Score’ increases as ‘Hours’ increases. Also, the relationship is approximately linear. To measure the linear relationship numerically, we calculate ‘correlation coefficient’ (r) by using the following formula:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

The calculation is shown in the following table:

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
17	72	-2	-5	10
18	76	-1	-1	1
19	75	0	-2	0
20	80	1	3	3
21	82	2	5	10

$$\sum(x-\bar{x})(y-\bar{y}) = 24$$

$$\sqrt{\sum(x-\bar{x})^2} = \sqrt{10} = 3.162$$

$$\sqrt{\sum(y-\bar{y})^2} = \sqrt{64} = 8$$

$$r = \frac{24}{3.162 \times 8} = 0.949$$

Properties of r

1. $-1 \leq r \leq 1$.

2. When r is positive: If X increases, then Y increases.
   When r is negative: If X increases, then Y decreases.

3. When r is close to -1 or $+1$, the linear relationship is strong.
   When r is close to zero, the linear relationship is weak.

Exercise (Do it yourself)

You want to study the relationship between the age of a car and its selling price. Listed below is a random sample of 5 used cars during the last year. Determine the correlation coefficient and comment.

Age (years)	7	8	9	11	12
Selling Price (lac)	8.0	7.0	6.1	4.6	4.0