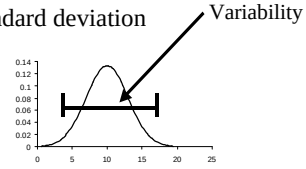
## Measure of Variability (Dispersion, Spread)

- Variance, standard deviation
- Range
- Inter-Quartile Range
- Pseudo-standard deviation



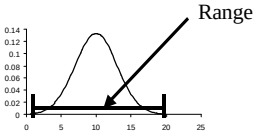

## Range

**Definition**

Let min = the smallest observation

Let max = the largest observation

Then Range =max - min




## Inter-Quartile Range (IQR)

**Definition**

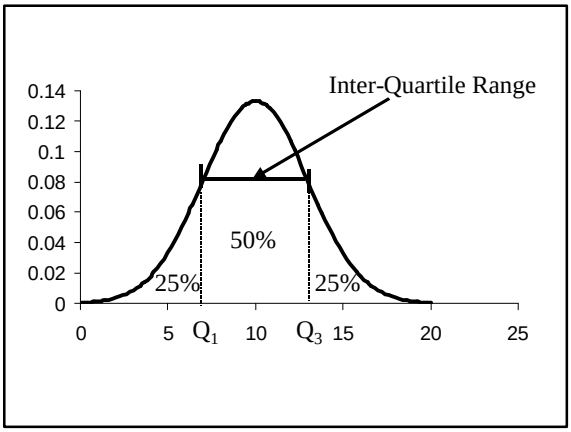
Let $Q_1$ = the first quartile,

$Q_3$ = the third quartile

Then the

Inter-Quartile Range

= IQR = $Q_3 - Q_1$



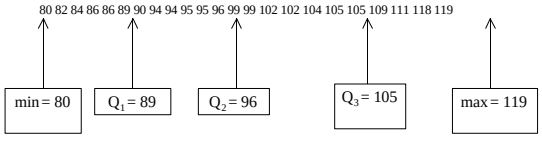

***Example***

The data Verbal IQ on n = 23 students arranged in increasing order is:

80 82 84 86 86 89 90 94
94 95 95 96 99 99 102 102
104 105 105 109 111 118 119

***Example***

The data Verbal IQ on n = 23 students arranged in increasing order is:



## Range

Range = max – min = 119 – 80 = 39

Inter-Quartile Range
= IQR = $Q_3$ - $Q_1$ = 105 – 89 = 16

## Some Comments

- Range and Inter-quartile range are relatively easy to compute.
- Range slightly easier to compute than the Inter-quartile range.
- Range is very sensitive to outliers (extreme observations)

## Sample Variance

Let $x_1, x_2, x_3, \ldots x_n$ denote a set of n numbers. Recall the mean of the n numbers is defined as:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + \ldots + x_{n-1} + x_n}{n}$$

The numbers

$$d_1 = x_1 - \bar{x}$$
$$d_2 = x_2 - \bar{x}$$
$$d_3 = x_3 - \bar{x}$$
$$\vdots$$
$$d_n = x_n - \bar{x}$$

are called deviations from the the mean

The sum

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

is called the sum of squares of deviations from the the mean.

Writing it out in full:

$$d_1^2 + d_2^2 + d_3^2 + \cdots + d_n^2$$

or

$$(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2$$

## The Sample Variance

Is defined as the quantity:

$$\frac{\sum_{i=1}^{n} d_i^2}{n-1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

and is denoted by the symbol $s^2$

**Example**

Let $x_1, x_2, x_3, x_3, x_4, x_5$ denote a set of 5 denote the set of numbers in the following table.

| $i$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x_i$ | 10 | 15 | 21 | 7 | 13 |

Then

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$$
$$= 10 + 15 + 21 + 7 + 13$$
$$= 66$$

and

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + \ldots + x_{n-1} + x_n}{n}$$
$$= \frac{66}{5} = 13.2$$

The deviations from the mean $d_1, d_2, d_3, d_4, d_5$ are given in the following table.

| i | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x_i$ | 10 | 15 | 21 | 7 | 13 |
| $d_i$ | -3.2 | 1.8 | 7.8 | -6.2 | -0.2 |

The sum

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$$
$$= (-3.2)^2 + (1.8)^2 + (7.8)^2 + (-6.2)^2 + (-0.2)^2$$
$$= 10.24 + 3.24 + 60.84 + 38.44 + 0.04$$
$$= 112.80$$

and

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{112.8}{4} = 28.2$$

## The Sample Standard Deviation s

**Definition:** The Sample Standard Deviation is defined by:

$$s = \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

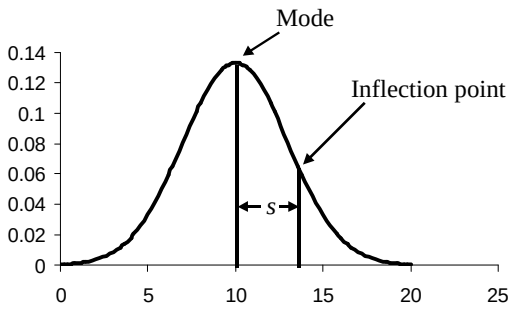
Hence the Sample Standard Deviation, s, is the square root of the sample variance.

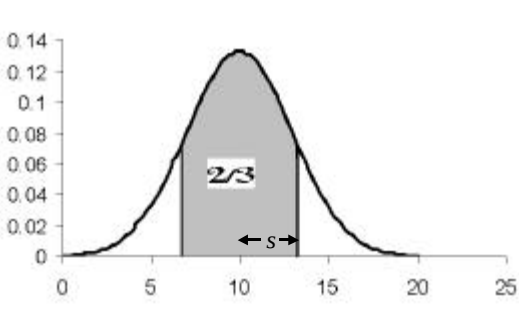
In the last example

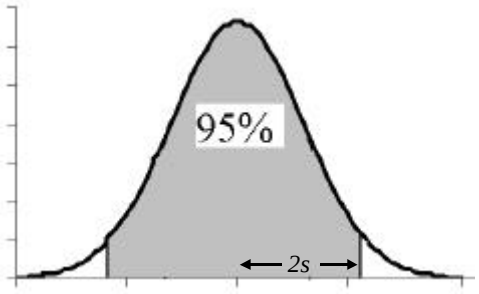
$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{112.8}{4}} = \sqrt{28.2} = 5.31$$

## Interpretations of s

- In Normal distributions
  - Approximately 2/3 of the observations will lie within one standard deviation of the mean
  - Approximately 95% of the observations lie within two standard deviations of the mean
  - In a histogram of the Normal distribution, the standard deviation is approximately the distance from the mode to the inflection point






## Example

A researcher collected data on 1500 males aged 60-65.

The variable measured was cholesterol and blood pressure.

- The mean blood pressure was 155 with a standard deviation of 12.
- The mean cholesterol level was 230 with a standard deviation of 15
- In both cases the data was normally distributed

## Interpretation of these numbers

- Blood pressure levels vary about the value 155 in males aged 60-65.
- Cholesterol levels vary about the value 230 in males aged 60-65.

- 2/3 of males aged 60-65 have blood pressure within 12 of 155. Ii.e. between 155-12 =143 and 155+12 = 167.
- 2/3 of males aged 60-65 have Cholesterol within 15 of 230. i.e. between 230-15 =215 and 230+15 = 245.

- 95% of males aged 60-65 have blood pressure within 2(12) = 24 of 155. Ii.e. between 155-24 =131 and 155+24 = 179.
- 95% of males aged 60-65 have Cholesterol within 2(15) = 30 of 230. i.e. between 230-30 =200 and 230+30 = 260.

## A Computing formula for:

Sum of squares of deviations from the the mean :

$$\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$$

The difficulty with this formula is that $\bar{x}$ will have many decimals.

The result will be that each term in the above sum will also have many decimals.

The sum of squares of deviations from the the mean can also be computed using the following identity:

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^{\,2}-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}$$

To use this identity we need to compute:

$$\sum_{i=1}^{n}x_i=x_1+x_2+\cdots+x_n \quad \text{and}$$

$$\sum_{i=1}^{n}x_i^2=x_1^2+x_2^2+\cdots+x_n^2$$

Then:

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^{\,2}-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}$$

$$\text{and } s^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}=\frac{\sum_{i=1}^{n}x_i^{\,2}-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}}{n-1}$$

and

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}}$$

***Example***

The data Verbal IQ on n = 23 students arranged in increasing order is:

80 82 84 86 86 89 90 94
94 95 95 96 99 99 102 102
104 105 105 109 111 118 119

$$\sum_{i=1}^{n} x_i = 80 + 82 + 84 + 86 + 86 + 89 + 90 + 94 + 94 + 95 + 95 + 96 + 99 + 99 + 102 + 102 + 104 + 105 + 105 + 109 + 111 + 118 + 119 = 2244$$

$$\sum_{i=1}^{n} x_i^2 = 80^2 + 82^2 + 84^2 + 86^2 + 86^2 + 89^2 + 90^2 + 94^2 + 94^2 + 95^2 + 95^2 + 96^2 + 99^2 + 99^2 + 102^2 + 102^2 + 104^2 + 105^2 + 105^2 + 109^2 + 111^2 + 118^2 + 119^2 = 221494$$

Then:

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^{\,2}-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}$$

$$=221494-\frac{(2244)^2}{23}=2557.652$$

and $s^2=\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}=\dfrac{\sum_{i=1}^{n}x_i^{\,2}-\dfrac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}}{n-1}$

$$=\frac{221494-\dfrac{(2244)^2}{23}}{22}=\frac{2557.652}{22}=116.26$$

Also $s=\sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}=\sqrt{\dfrac{\sum_{i=1}^{n}x_i^{\,2}-\dfrac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}}{n-1}}$

$$=\sqrt{\frac{221494-\dfrac{(2244)^2}{23}}{22}}=\sqrt{\frac{2557.652}{22}}=\sqrt{116.26}$$

$$=10.782$$

## A quick (rough) calculation of s

$$s \approx \frac{\text{Range}}{4}$$

The reason for this is that approximately all (95%) of the observations are between $\bar{x} - 2s$ and $\bar{x} + 2s$.

Thus $\max \approx \bar{x} + 2s$ and $\min \approx \bar{x} - 2s$.

and $Range = \max - \min \approx (\bar{x} + 2s) - (\bar{x} - 2s)$.

$= 4s$

Hence $s \approx \frac{\text{Range}}{4}$

***Example***

Verbal IQ on n = 23 students

min = 80 and max = 119

$$s \approx \frac{119 - 80}{4} = \frac{39}{4} = 9.75$$

This compares with the exact value of *s* which is 10.782.

The rough method is useful for checking your calculation of s.

## The Pseudo Standard Deviation (PSD)

**Definition:** The **Pseudo Standard Deviation (PSD)** is defined by:

$$\text{PSD} = \frac{\text{IQR}}{1.35} = \frac{\text{InterQuartile Range}}{1.35}$$

## Properties

- For Normal distributions the magnitude of the pseudo standard deviation (*PSD*) and the standard deviation (*s*) will be approximately the same value
- For leptokurtic distributions the standard deviation (*s*) will be larger than the pseudo standard deviation (*PSD*)
- For platykurtic distributions the standard deviation (*s*) will be smaller than the pseudo standard deviation (*PSD*)

***Example***

Verbal IQ on n = 23 students

Inter-Quartile Range

$= IQR = Q_3 - Q_1 = 105 - 89 = 16$

Pseudo standard deviation

$$= PSD = \frac{IQR}{1.35} = \frac{16}{1.35} = 11.85$$

This compares with the standard deviation

$$s = 10.782$$

- An ***outlier*** is a "wild" observation in the data
- Outliers occur because
  - of errors (typographical and computational)
  - Extreme cases in the population
- We will now consider the drawing of box-plots where outliers are identified

## To Draw a Box Plot we need to:

- Compute the Hinge (Median, $Q_2$) and the Mid-hinges (first & third quartiles – $Q_1$ and $Q_3$ )
- To identify outliers we will compute the inner and outer ***fences***
- ***Lower inner fence***
  - $f_1 = Q_1 - (1.5)IQR$

***Lower outer fence***

$F_1 = Q_1 - (3)IQR$

***Upper outer fence***

$F_2 = Q_3 + (3)IQR$

***Lower inner fence***

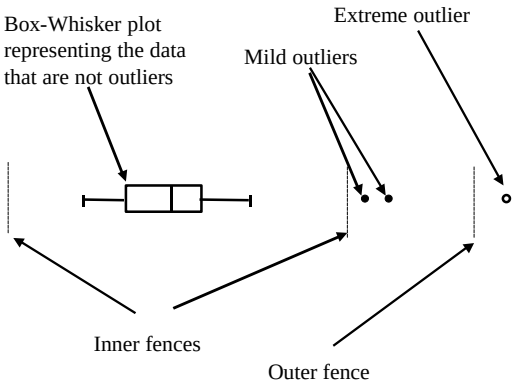
$f_1 = Q_1 - (1.5)IQR$

***Upper inner fence***

$f_2 = Q_3 + (1.5)IQR$

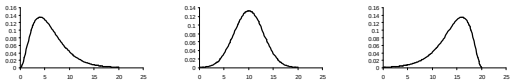
- Observations that are between the lower and upper fences are considered to be non-outliers.
- Observations that are outside the inner fences but not outside the outer fences are considered to be ***mild*** outliers.
- Observations that are outside outer fences are considered to be ***extreme*** outliers.

- ***mild*** outliers are plotted individually in a box-plot using the symbol ●
- ***extreme*** outliers are plotted individually in a box-plot using the symbol ○
- non-outliers are represented with the box and whiskers with
  - Max = largest observation within the fences
  - Min = smallest observation within the fences

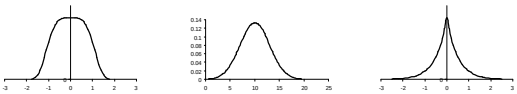


## Measures of Shape

- Skewness



- Kurtosis



- Skewness – based on the sum of cubes

$$\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^3$$

- Kurtosis – based on the sum of 4th powers

$$\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^4$$