Descriptive Statistics: Measures of Dispersion (Measures of Variation):

A measure of dispersion: conveys information regarding the amount of variability present in a set of data.

مقياس التشتت: ينقل المعلومات المتعلقة بمقدار التباين الموجود في مجموعة من البيانات.

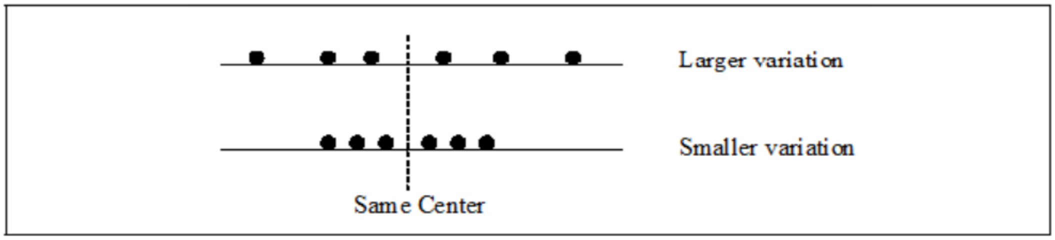
There are several measures of dispersion, some of which are:

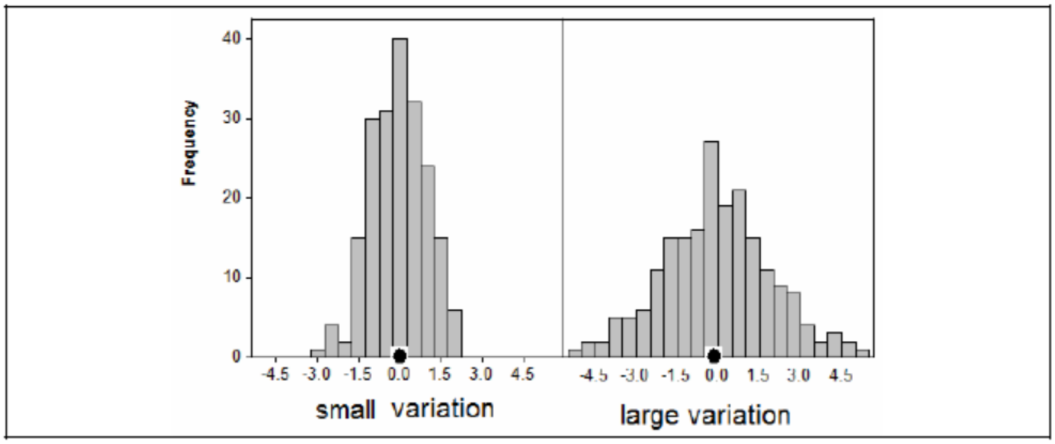
1) **Range (R).**
2) **The Sample Variance s^2.**
3) **The Sample Standard Deviation s.**
4) **The Population Variance (σ^2).**
5) **The Population Standard Deviation (σ).**
6) **Coefficient of variation ($C.V$).**
7) **Standard error (SE)**
8) **Quartiles and inter quartile range (QR)**

Variance.

The variation or dispersion in a set of values refers to how spread out (ينتشر) the values is from each other.

- The dispersion (variation) is small when the values are close together.
- There is no dispersion (no variation) if the values are the same.





1. The Range (*R*):

The Range is the difference between the largest value (Max) and the smallest value (Min).

Range (*R*) = Max – Min

Example:

Find the range for the sample values**:** 26, 25, 35, 27, 29, 29.

Solution:

$$max = 35 \qquad min = 25$$

$$Range\ (R) = 35 - 25 = 10\ (unit)$$

Notes:

1. The unit of the range is the same as the unit of the data
2. The usefulness of the range is limited.

The range is a poor measure of the dispersion because it only takes into account two of the values: however, it plays a significant role in many applications.

2. The sample Variance (s^2). (Variance computed from the sample)

The **variance** is one of the most important measures of **dispersion**.

The **variance** is a measure that uses the **mean** as a point of reference.

- The **variance** of the data is **small** when the observations are **close** to the **mean**.
- The **variance** of the data is **large** when the observations are **spread out** from the **mean**.
- The **variance** of the data is **zero** (no variation) when all observations have the **same value** (concentrated at the mean).

Deviations of sample values from the sample mean:

$$let\ \boldsymbol{x_1, x_2, \ldots, x_n}\ be\ \boldsymbol{the\ sample\ values}, and$$
$$\boldsymbol{\overline{x}}\ be\ \boldsymbol{the\ sample\ mean}.$$

The deviation of the **value** $\boldsymbol{x_i}$ from the **sample mean** $\boldsymbol{\overline{x}}$

$$\boldsymbol{x_i - \overline{x}}$$

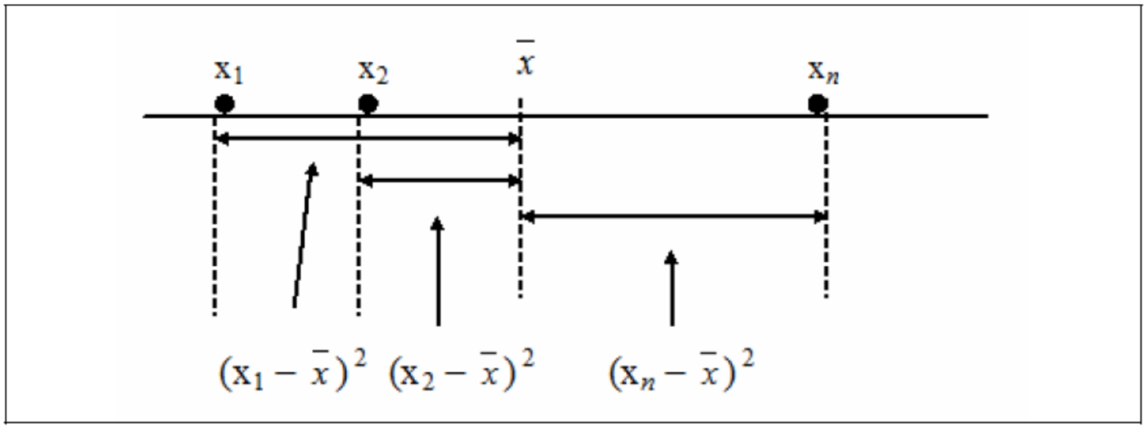
The squared deviation is:

$$(\boldsymbol{x_i} - \boldsymbol{\overline{x}})^2$$

The sum of squared deviations is:

$$\sum_{i-1}^{n}(\boldsymbol{x_i} - \boldsymbol{\overline{x}})^2$$

The following graph shows the squared deviations of the values from the mean.



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

$$s^2 = \frac{(x_1 - \overline{x})^2 + \ (x_2 - \overline{x})^2 + \ . \ . \ . \ \ (x_n - \overline{x})^2}{n-1} \qquad (unit)^2$$

Where $\overline{x} = \frac{\sum_{\mathrm{i-1}}^{\mathrm{n}} \mathrm{x_i}}{\mathrm{n}}$ is the sample mea, and
(n) is the sample size.

- S^2 is a statistic because it is obtained from the sample values (it is known).
- S^2 is used to approximate (estimate) σ^2.
- $S^2 \geq 0$
- $S^2 = 0$ ⇔ all observation have the same value

  ⇔ there is no dispersion (no variation)

Example:

We want to compute the sample variance of the following sample values: 10, 21, 33, 53, 54.

Solution:

N=5

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{10+21+33+53+54}{5} = \frac{171}{5} = 34.2$$

$$S^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1} = \frac{\sum_{i=1}^{5} (x_i - 34.2)^2}{5-1}$$

$$S^2 = \frac{(10-34.2)^2 + (21-34.2)^2 + (33-34.2)^2 + (53-34.2)^2 + (54-34.2)^2}{4}$$

$$= \frac{1506.8}{4} = 376.7 \quad (\text{unit})^2$$

3. The Sample Standard Deviation (*s*).

The **variance** represents **squared units**, therefore, is not appropriate measure of **dispersion** when we wish to express the concept of dispersion in terms of **the original unit**.

- The **standard deviation** is another measure of **dispersion**
- The **standard deviation** is the square root of the **variance**.
- The **standard deviation** is expressed in the **original unit** of the **data**.

) Sample standard deviation is: $S = \sqrt{S^2}$ (unit)

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

Example:

For the previous example, the sample standard deviation is

$$S = \sqrt{S^2} = \sqrt{376.7} = 19.41 \qquad \text{(unit)}$$

4. The Population Variance (σ^2).

(Variance computed from the population)

$$let\ \boldsymbol{X_1, X_2, ..., X_n}\ be\ \boldsymbol{the\ population\ values}, and$$

$$\boldsymbol{\mu}\ be\ \boldsymbol{the\ population\ mean}$$

$$\boldsymbol{\sigma^2} = \frac{\sum_{i=1}^{N} \boldsymbol{(X_i - \mu)^2}}{N}$$

$$\boldsymbol{\sigma^2} = \frac{(\boldsymbol{X_1} - \boldsymbol{\mu})^2 + (\boldsymbol{X_2} - \boldsymbol{\mu})^2 + ...\ (\boldsymbol{X_N} - \boldsymbol{\mu})^2}{N} \qquad (unit)^2$$

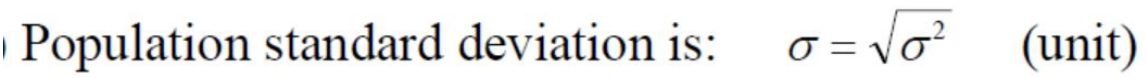
Where $\mu = \frac{\sum_{i-1}^{N} X_i}{N}$ is the population mea, and

(N) is the population size.

Notes:

- $\boldsymbol{\sigma^2}$ is a parameter because it is obtained from the population values (it is unknown in general).
- $\boldsymbol{\sigma^2} \geq \mathbf{0}$

5. The Population Standard Deviation (σ).



Population standard deviation is: $\sigma = \sqrt{\sigma^2}$ (unit)

6. Coefficient of variation ($C.V$).

Is a measure use to **compare the dispersion in two sets of data** which is **independent** of the **unit** of the measurement.

The **coefficient of variation** is the **ratio** of the **sample standard deviation** to the **sample mean of a distribution**.

- **It is a measure of the spread of the distribution relative to the mean of the distribution.**

$$C.V = \frac{S}{\bar{x}} 100\%$$

• where **S**: Sample standard deviation, and

• $\bar{x}$: Sample mean.

Example:

• Suppose two samples of human male yield the following data:

	Sampe1	Sample2
Age	**25-year-olds**	**11year-olds**
Mean weight (**X̅**)_	**145 pounds**	**80 pounds**
Standard deviation (**S**)	**10 pounds**	**10 pounds**

• We wish to know which is more variable.

Solution:

$$for\ \textbf{sample 1}\ \ C.V = \frac{10}{145} * 100\% = 6.9\%$$

$$for\ \textbf{sample 2}\ \ C.V = \frac{10}{80} * 100\% = 12.5\%$$

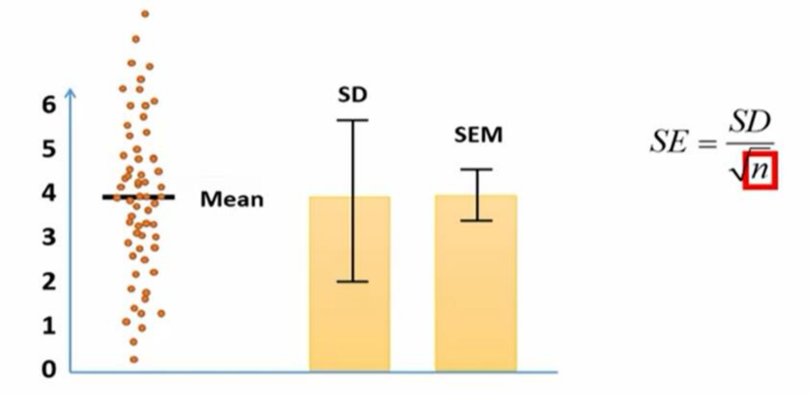
Then age of 11 − years old(sample2)is more variation than the (**sample 1**)of 25 − years

7. Standard error (SE)

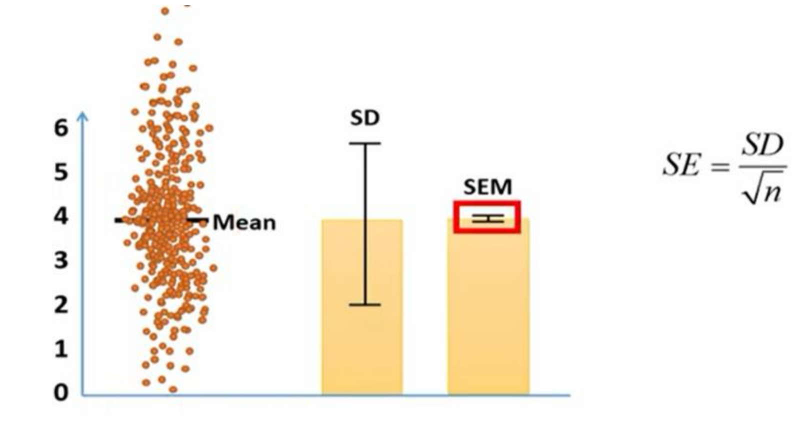
What is the value of the standard error of the mean?

The standard error of the mean, or simply standard error, **indicates how different the population mean is likely to be from a sample mean**. It tells you how much the sample mean would vary if you were to repeat a study using new samples from within a single population.

The standard error (SE) is used to describe the variability among separate sample means.



Another important difference between the standard deviation and the standard error is that the value of standard error is reduced if the sample is increased.



The standard error will be reduced because we are then more certain that our estimate mean is very close to the true population mean.

8. Quartiles and inter quartile range (QR)

Five-number summaries and box plots.

The five-number summary of a dataset consists of the numbers in the following list (and in this order):

Minimum, Q1, **Median**, **Q3**, **Maximum**

A box plot of a dataset is a graphical version of the five-number summary, with a few extras. Generic box plot:

Step-by-step procedure for constructing a box plot

1. Draw a horizontal (or vertical) reference scale based on the extent of the data.

2. Draw a box whose sides (or top and bottom) are located at Q1 and Q3.

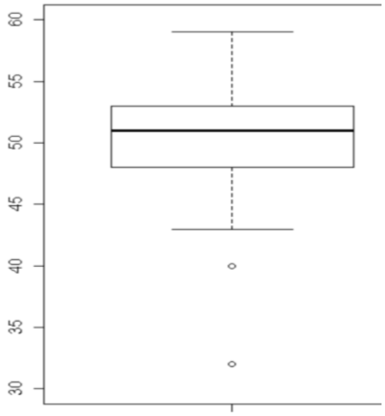
3. Draw a vertical (horizontal) line segment at the median.

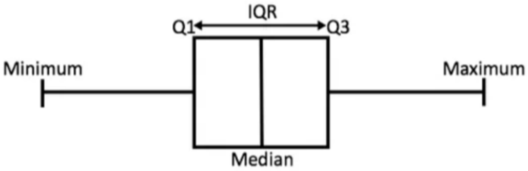
4. Compute the fences, $f_1 = Q1 - 1.5 * IQR$ and $f_3 = Q3 + 1.5 * IQR$.

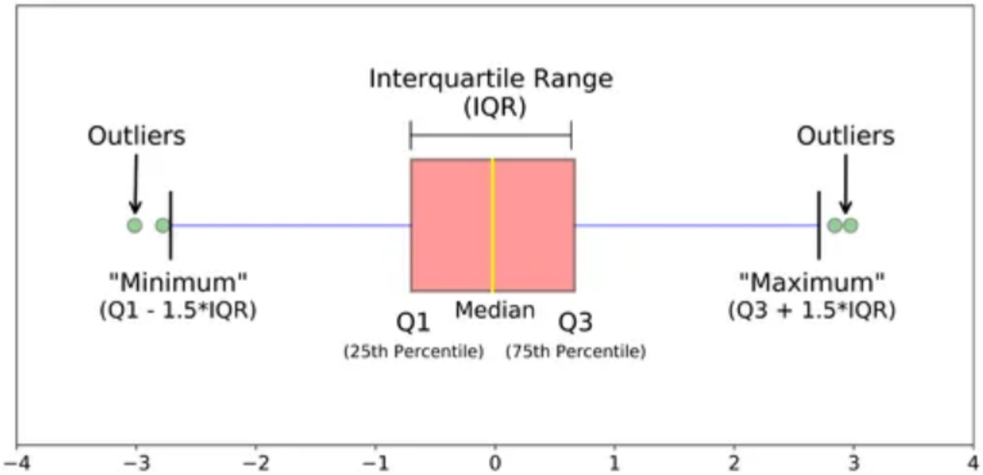
5. Extend a line segment (so-called whisker) from Q1 out to the most extreme observation that is at or inside the fence, i.e., $\geq f_1$). Repeat on the other side, i.e., from Q3 to the most extreme observation that is $\leq f_3$. Mark the end of these line segments with a ×.

6. Mark any observations beyond the fences with an open circle, ◦; these are regarded as outliers.

7. If you are constructing more than one box plot for comparison purposes, use the same scale for all of them and put them side by-side (or one on top of another)







Example: from the following data draw box-whisker plot:

10, 4, 11, 7, 3, 12, 11, 9, 5, 5, 8

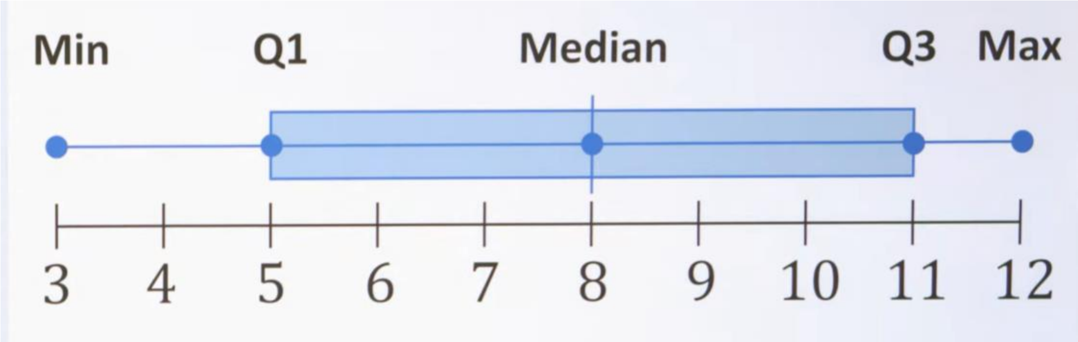
To calculate the quartiles, arrange of the data.

Data	**3**	**4**	**5**	**5**	**7**	**8**	**9**	**10**	**11**	**11**	**12**
Rank	1	2	3	4	5	6	7	8	9	10	11
			Q1			**Q2**			Q3		

a) Min=3
b) Max=12
c) Q2=median= 8

$$\left[rank = \frac{n+1}{2} = \frac{11+1}{2} = \frac{12}{2} = 6\right]$$

d) Q1=5
e) Q3=11


Min
Q1
Median
Q3
Max