# Statistics

**( Probabilities and Statistics)**

***Introduction :***

Probability and statistics are concerned with events which occur by chance. Examples include occurrence of **accidents**, **errors of measurements**. In each case we may have some knowledge of the likelihood of various possible results, but we cannot predict with any certainty the outcome of any particular trial. Probability and statistics are used throughout engineering. In electrical engineering, signals and noise are analyzed by means of probability theory.

**Probability versus Statistics**

Probability and statistics are related areas of mathematics which concern themselves with analyzing the relative frequency of events. Still, there are fundamental differences in the way they see the world:

- ***Probability*** deals with predicting the likelihood of ***future events***, while ***statistics*** involves the analysis of the frequency of ***past events***.
- ***Probability*** is primarily a theoretical ***branch of mathematics***, which studies the consequences of mathematical definitions. ***Statistics*** is primarily an **applied** branch of mathematics, which tries to make sense of observations in the real world.

# ( Statistics )

## Describing a set of Data with numerical measures :

***Graphs*** can help you describe the basic shape of a data distribution; “a picture is worth a thousand words.” But there are ***limitations of using graph***. Therefore, we need to find another way to convey a mental picture of the data.

One way to overcome these problems is to use ***numerical measures***, which can be calculated for either a **sample** or a **population** of measurements. You can use the data to calculate a set of numbers that will convey a good mental picture of the frequency distribution. These measures are called **parameters** when associated with the **population**, and they are called ***statistics*** when calculated from ***sample measurements***.

Descriptive Statistics is the part of Statistics in charge of representing, analysing and summarizing the information contained in the sample.

After the sampling process, this is the next step in every statistical study and usually consists of:

1. **To classify, group and sort the data of the sample.**
2. **To tabulate and plot data according to their frequencies.**
3. **To calculate numerical measures that summarize the information contained in the sample (*sample statistics*).**

# Frequency distribution :

The study of a statistical variable starts by measuring the variable in the individuals of the sample and classifying the values. There are two ways of classifying data:

- **Non-grouping**: Sorting values from lowest to highest value (if there is an order). Used with qualitative variables and discrete variables with few distinct values.

  **1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, 0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2**

- **Grouping**: Grouping values into intervals (classes) and sort them from lowest to highest intervals. Used with continuous variables and discrete variables with many distinct values.

  ,185 ,111 ,111 ,172 ,171 ,158 ,171 ,181 ,173 ,171  ,171
  ,111 ,188 ,151 ,115 ,178 ,177 ,118 ,187 ,112  175, 182,
  167, 169, 172, 186, 172, 176, 168, 187.

**Sample classification :**

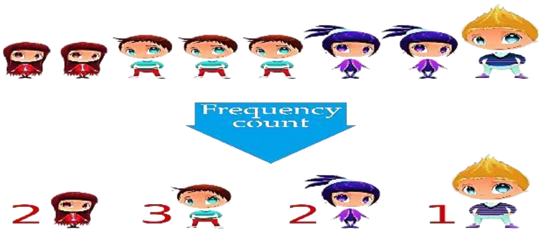
It consists in grouping the values that are the same and sorting them if there is an order among them.

**Example**. X = Height

**Frequency count :**

It consists in counting the number of times that every value appears in the sample.

**Example**. X=Height



## Sample frequencies :

**Definition - Sample frequencies**. Given a sample of $n$ values of a variable $X$, for every value $x_i$ of the variable we define

□ **Absolute Frequency** $n_i$: The number of times that value $x_i$ appears in the sample.

□ **Relative Frequency** $f_i$: The proportion of times that value $x_i$ appears in the sample.

$$f_i = n_i / n$$

□ **Cumulative Absolute Frequency** $N_i$: The number of values in the sample less than or equal to $x_i$.

$$N_i = n_1 + \dots + n_i = N_{i-1} + n_i$$

□ **Cumulative Relative Frequency** $F_i$ : The proportion of values in the sample less than or equal to $x_i$ .

$$F_i = N_i / n$$

## Frequency table :

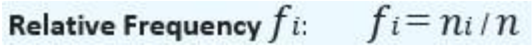
The set of values of a variable with their respective frequencies is called **frequency distribution** of the variable in the sample, and it is usually represented as a **frequency table**.

| *X* values | Absolute frequency | Relative frequency | Cumulative absolute frequency | Cumulative relative frequency |
|---|---|---|---|---|
| $x_1$ | $n_1$ | $f_1$ | $N_1$ | $F_1$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $x_i$ | $n_i$ | $f_i$ | $N_i$ | $F_i$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $x_k$ | $n_k$ | $f_k$f | $N_k$ | $F_k$ |

### Example - Quantitative variable and non-grouped data.

Find (fi, Ni, Fi) for the following of number of children in 25 families are:
1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, 0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2

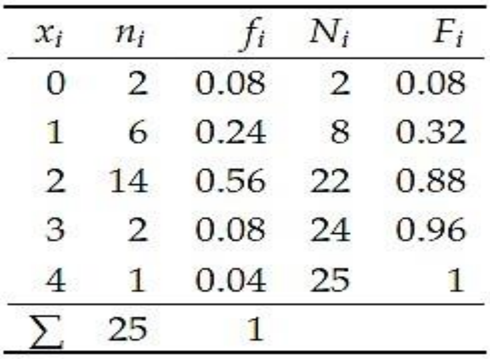
**Solution** :The frequency table for the number of children in this sample is



**Relative Frequency** $f_i$: $f_i = n_i / n$

**Cumulative Absolute Frequency** $N_i$: $N_i = n_1 + \dots + n_i$

**Cumulative Relative Frequency** $F_i$: $F_i = N_i / n$

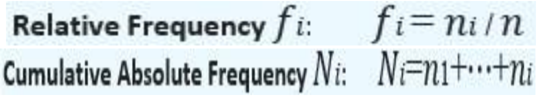


| $x_i$ | $n_i$ | $f_i$ | $N_i$ | $F_i$ |
|---|---|---|---|---|
| 0 | 2 | 0.08 | 2 | 0.08 |
| 1 | 6 | 0.24 | 8 | 0.32 |
| 2 | 14 | 0.56 | 22 | 0.88 |
| 3 | 2 | 0.08 | 24 | 0.96 |
| 4 | 1 | 0.04 | 25 | 1 |
| $\sum$ | 25 | 1 | | |

**Example - Quantitative variable and grouped data**. The heights (in cm) of 30 students are:

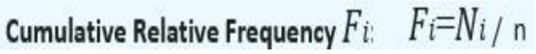
179, 173, 181, 170, 158, 174, 172, 166, 194, 185,
162, 187, 198, 177, 178, 165, 154, 188, 166, 171,
175, 182, 167, 169, 172, 186, 172, 176, 168, 187.

**Solution** :The frequency table for the height in this sample is :

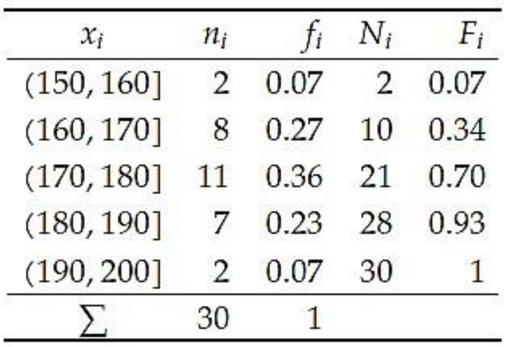


**Relative Frequency** $f_i$: $f_i = n_i / n$

Cumulative Absolute Frequency $N_i$: $N_i = n_1 + \cdots + n_i$



Cumulative Relative Frequency $F_i$: $F_i = N_i / n$



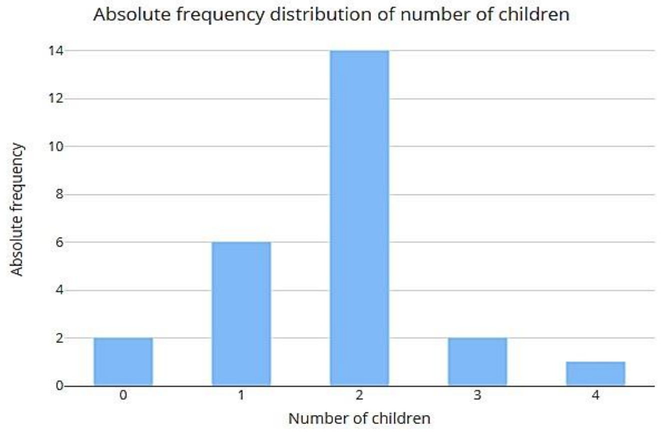
| $x_i$ | $n_i$ | $f_i$ | $N_i$ | $F_i$ |
|---|---|---|---|---|
| (150, 160] | 2 | 0.07 | 2 | 0.07 |
| (160, 170] | 8 | 0.27 | 10 | 0.34 |
| (170, 180] | 11 | 0.36 | 21 | 0.70 |
| (180, 190] | 7 | 0.23 | 28 | 0.93 |
| (190, 200] | 2 | 0.07 | 30 | 1 |
| $\sum$ | 30 | 1 | | |

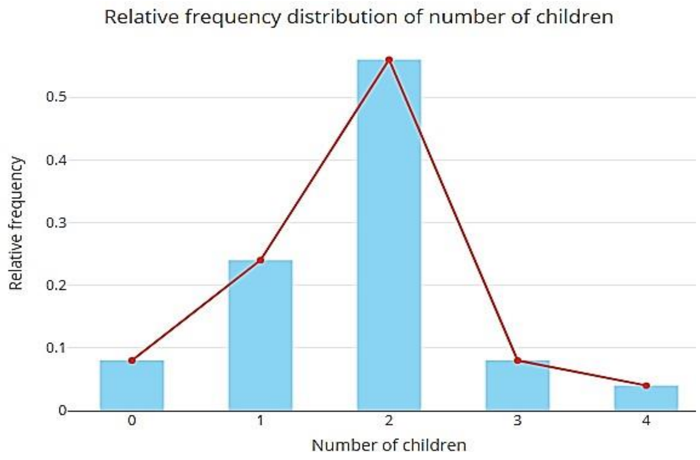
## Frequency distribution graphs :

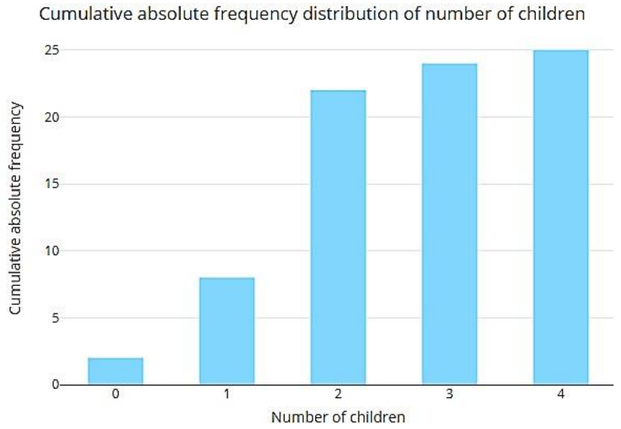
Usually the frequency distribution is also displayed graphically. Depending on the type of variable and whether data has been grouped or not, there are different types of charts:

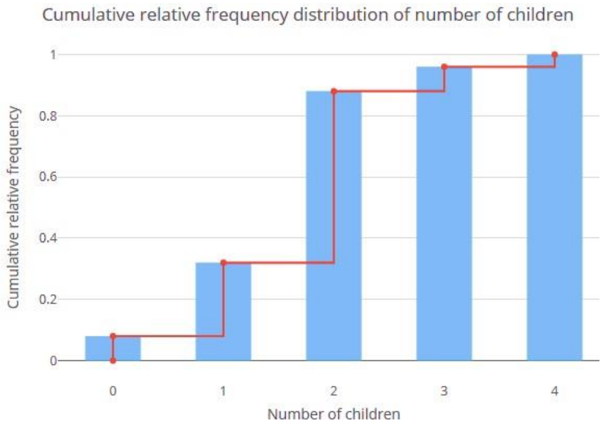
- Bar chart
- Histogram
- Line or polygon chart.
- Pie chart

## Bar chart :









## Sample statistics :

The frequency table and charts summarize and give an overview of the distribution of values of the studied variable in the sample, but it is difficult to describe some aspects of the distribution from it, as for example, which are the most **representative values of the distribution**, **how is the spread of data**, which **data could be considered outliers**, or **how is the symmetry of the distribution**.

To describe those aspects of the sample distribution more specific numerical measures, called **sample statistics**, are used.

According to the aspect of the distribution that they study, there are different types of statistics: **Location statistics and Measures of dispersion.**

## Location statistics :

There are two groups:

**Central location measures(Measures of center)**: They measure the values where data are concentrated, usually at the centre of the distribution. These values are the values that best represents the sample data. The most important are:

1. Arithmetic mean
2. Median
3. Mode

***Which central tendency statistic should I use?***

In general, when all the central tendency statistics can be calculated, is advisable to use them as representative values in the following order:

1. The mean. Mean takes more information from the sample than the others, as it takes into account the magnitude of data.
2. The median. Median takes less information than mean but more than mode, as it takes into account the order of data.
3. The mode. Mode is the measure that fewer information takes from the sample, as it only takes into account the absolute frequency of values.

But, *be careful with outliers*, as the mean can be distorted by them. In that case it is better to use the median as the value most representative.

**Measures of dispersion**: They measure the spread of data.

1- Range

2- Variance

3- Standard deviation

## Central location measures(Measures of center):

### 1- Arithmetic mean

**Sample arithmetic mean $\bar{X}$**. The *sample arithmetic mean* of a variable *X* is the sum of observed values in the sample divided by the sample size:

$$\boxed{\bar{x} = \frac{\sum x_i}{n}}$$

Also , it can be calculated from the ***frequency table*** with the formula :

$$\boxed{\bar{x} = \frac{\sum x_i n_i}{n} = \sum x_i f_i}$$

**Example - Non-grouped data.**

The number of children in 25 families are: 1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, 0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2

**Solution:**

$$\boxed{\bar{x} = \frac{\sum x_i}{n}}$$

$$\bar{x} = \frac{1+2+4+2+2+2+3+2+1+1+0+2+2}{25} + \\ + \frac{0+2+2+1+2+2+3+1+2+2+1+2}{25} = \frac{44}{25} = 1.76 \text{ children.}$$

or using the frequency table

| $x_i$ | $n_i$ | $f_i$ | $x_i n_i$ | $x_i f_i$ |
|---|---|---|---|---|
| 0 | 2 | 0.08 | 0 | 0 |
| 1 | 6 | 0.24 | 6 | 0.24 |
| 2 | 14 | 0.56 | 28 | 1.12 |
| 3 | 2 | 0.08 | 6 | 0.24 |
| 4 | 1 | 0.04 | 4 | 0.16 |
| $\sum$ | 25 | 1 | 44 | 1.76 |

$$\bar{x} = \frac{\sum x_i n_i}{n} = \frac{44}{25} = 1.76 \text{ children} \qquad \bar{x} = \sum x_i f_i = 1.76 \text{ children.}$$

That means that the value that best represent the number of children in the families of the sample is 1.76 children.

**Example – Grouped data.** Using the data of the sample of student heights, the arithmetic mean is

$$\boxed{\bar{x} = \frac{\sum x_i}{n}} \qquad \bar{x} = \frac{179 + 173 + \cdots + 187}{30} = 175.07 \text{ cm}.$$

or using the frequency table and taking the class marks as $x_i$,

| X | $x_i$ | $n_i$ | $f_i$ | $x_i n_i$ | $x_i f_i$ |
|---|---|---|---|---|---|
| (150, 160] | 155 | 2 | 0.07 | 310 | 10.33 |
| (160, 170] | 165 | 8 | 0.27 | 1320 | 44.00 |
| (170, 180] | 175 | 11 | 0.36 | 1925 | 64.17 |
| (180, 190] | 185 | 7 | 0.23 | 1295 | 43.17 |
| (190, 200] | 195 | 2 | 0.07 | 390 | 13 |
| $\sum$ | | 30 | 1 | 5240 | 174.67 |

$$\bar{x} = \frac{\sum x_i n_i}{n} = \frac{5240}{30} = 174.67 \text{ cm} \qquad \bar{x} = \sum x_i f_i = 174.67 \text{ cm}.$$

Observe that when the mean is calculated from the table the result differs a little from the real value, because the values used in the calculations are the class marks instead of the actual values.

## Weighted mean

In some cases the values of the sample have different importance. In that case the importance or *weight* of each value of the sample must be taken into account when calculating the mean.

*weighted mean* of variable $X$ is the sum of the product of each value by its weight, divided by sum of weights

$$\bar{x}_w = \frac{\sum x_i w_i}{\sum w_i}$$

From the frequency table can be calculated with the formula

$$\bar{x}_w = \frac{\sum x_i w_i n_i}{\sum w_i}$$

**Example.** Assume that a student wants to calculate a representative measure of his/her performance in a course. The grade and the credits of every subjects are

| Subject | Credits | Grade |
|---|---|---|
| Maths | 6 | 5 |
| Economics | 4 | 3 |
| Chemistry | 8 | 6 |

The arithmetic mean is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{5 + 3 + 6}{3} = 4.67 \text{ points}.$$

However, this measure does not represent well the performance of the student, as not all the subjects have the same importance and require the same effort to pass. Subjects with more credits require more work and must have more weight in the calculation of the mean.

In this case it is better to use the weighted mean, using the credits as the weights of grades, as a representative measure of the student effort

$$\bar{x}_w = \frac{\sum x_i w_i}{\sum w_i} = \frac{5 \cdot 6 + 3 \cdot 4 + 6 \cdot 8}{6 + 4 + 8} = \frac{90}{18} = 5 \text{ points}.$$

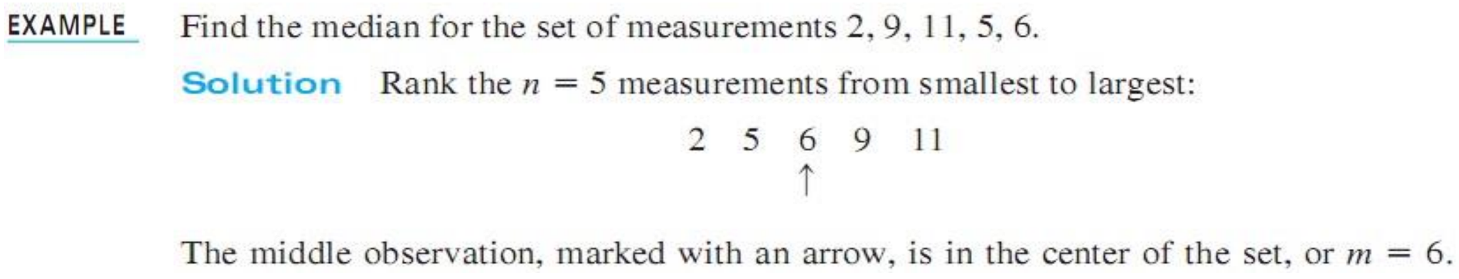
## 2- ***Median:***

A second measure of ***central tendency*** is the ***median***, which is the value in the ***middle position*** in the set of measurements ***ordered from smallest to largest***.

***Definition*** : The median ***m*** of a set of n measurements is the value of ***x*** that falls in the ***middle position*** when the measurements are ordered ***from smallest to largest***.

**We can know the order and the value of median by using the following :**

The value " **.5(n + 1)** " indicates the **position** of the median in the ordered data set. If the position of the median is a number that ends in the value .5, you need to average the two adjacent values.

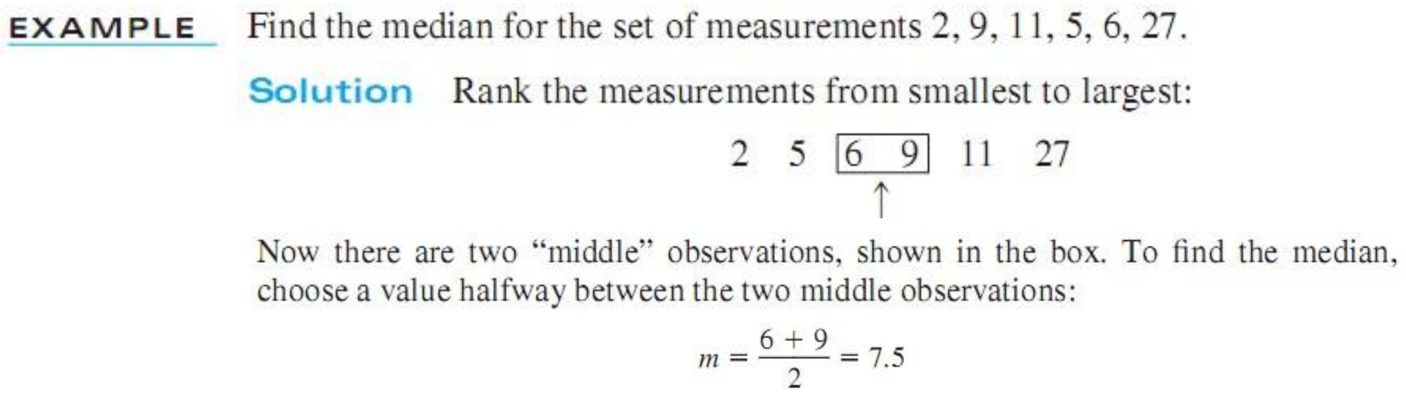


EXAMPLE Find the median for the set of measurements 2, 9, 11, 5, 6.

Solution Rank the $n = 5$ measurements from smallest to largest:

$$\begin{array}{ccccc} 2 & 5 & 6 & 9 & 11 \\ & & \uparrow & & \end{array}$$

The middle observation, marked with an arrow, is in the center of the set, or $m = 6$.



EXAMPLE Find the median for the set of measurements 2, 9, 11, 5, 6, 27.

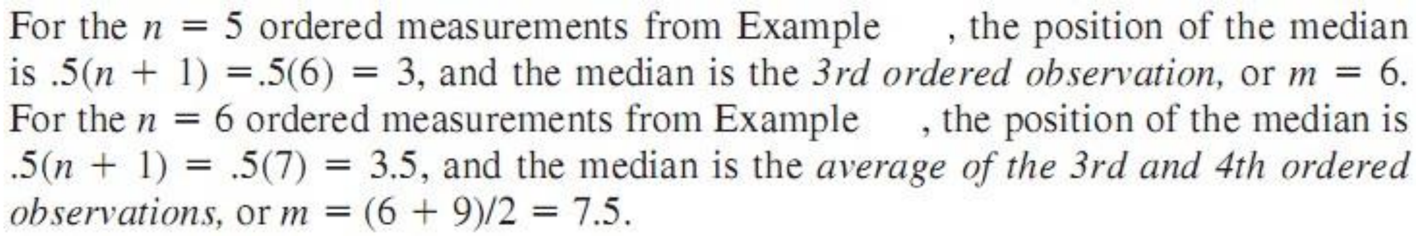
Solution Rank the measurements from smallest to largest:

$$\begin{array}{cccc} 2 & 5 & \boxed{6 \quad 9} & 11 \quad 27 \\ & & \uparrow & \end{array}$$

Now there are two "middle" observations, shown in the box. To find the median, choose a value halfway between the two middle observations:

$$m = \frac{6+9}{2} = 7.5$$

Now if we use the value " **.5(n + 1)** :



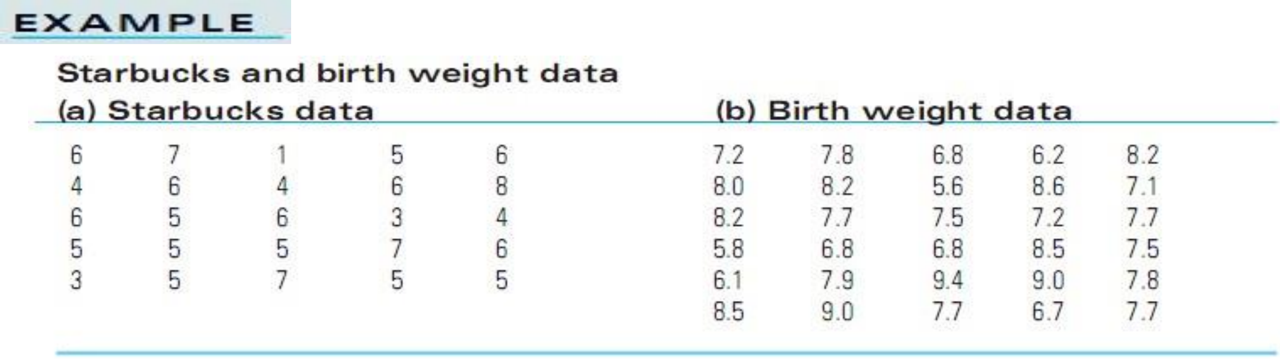
For the $n = 5$ ordered measurements from Example , the position of the median is $.5(n + 1) = .5(6) = 3$, and the median is the *3rd ordered observation*, or $m = 6$. For the $n = 6$ ordered measurements from Example , the position of the median is $.5(n + 1) = .5(7) = 3.5$, and the median is the *average of the 3rd and 4th ordered observations*, or $m = (6 + 9)/2 = 7.5$.

If a distribution is ***tilt*** to the right, the ***mean*** shifts to the right; if a distribution is skewed to the left, the ***mean*** shifts to the left. The median is not affected by these extreme values because the numerical values of the measurements are not used in its calculation. **When a distribution is symmetric**, the mean and the median are ***equal***. If a distribution is strongly skewed by one or more extreme values, you should use the median rather than the mean as a measure of center.

## 3- ***The Mode :***

Another way to ***locate the center of a distribution*** is to look for the value of ***x*** that occurs with the ***highest frequency***. This measure of the center is called the **mode**.

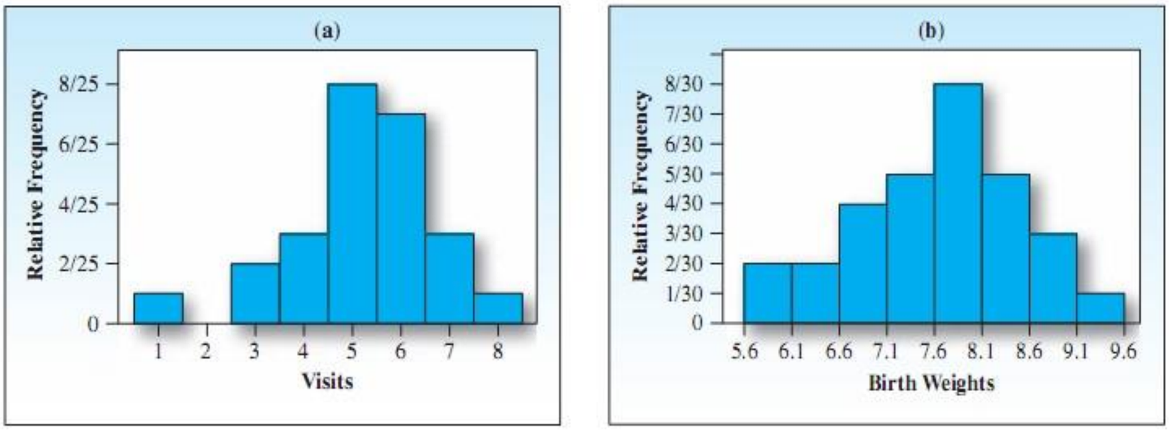
***Definition*** : The mode is the category that occurs ***most frequently***, or the ***most frequently occurring*** value of ***x***. ***When measurements on a continuous variable*** have been ***grouped as a frequency*** or relative frequency ***histogram***, the class with the highest peak or frequency is called the ***modal class***, and the midpoint of that class is taken to be the ***mode***. ***Note*** : The mode is generally used to describe ***large data sets***, whereas the mean and median are used for both ***large*** and ***small*** data sets.



EXAMPLE

Starbucks and birth weight data

| (a) Starbucks data | | | | | (b) Birth weight data | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 6 | 7 | 1 | 5 | 6 | 7.2 | 7.8 | 6.8 | 6.2 | 8.2 |
| 4 | 6 | 4 | 6 | 8 | 8.0 | 8.2 | 5.6 | 8.6 | 7.1 |
| 6 | 5 | 6 | 3 | 4 | 8.2 | 7.7 | 7.5 | 7.2 | 7.7 |
| 5 | 5 | 5 | 7 | 6 | 5.8 | 6.8 | 6.8 | 8.5 | 7.5 |
| 3 | 5 | 7 | 5 | 5 | 6.1 | 7.9 | 9.4 | 9.0 | 7.8 |
| | | | | | 8.5 | 9.0 | 7.7 | 6.7 | 7.7 |



***Solution :***

For ***The visits*** :

Table: From the data in Example reproduced in Table (a), the **mode** of the distribution of the number of reported weekly visits to Starbucks for 30 Starbucks customers is 5.

For ***the birth weight*** :

Table: For the birth weight data in Table (b), a birth weight of 7.7 occurs four times, and therefore the **mode** for the distribution of birth weights is 7.7

**Using the histogram(a)** for ***The visits***: The ***modal class*** and the value of x occurring with the highest frequency are the same, as shown in Figure (a).

**Using the histogram(b)** for ***the birth weight***: Using the histogram to find the **modal class**, you find that the class with the highest peak is the fifth class, from 7.6 to 8.1. Our choice for the mode would be the midpoint of this class, or (**7.6 + 8.1**) 7.85. See Figure (b).

11

## ***Dispersion statistics(Measure of variability):***

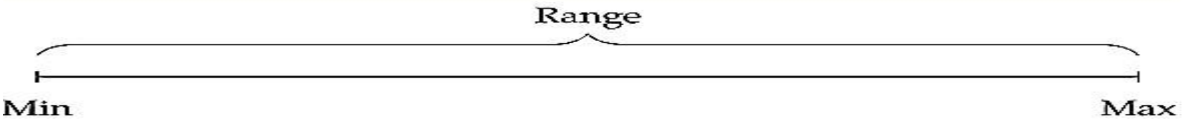
*Dispersion* or *spread* refers to the ***variability of data***. So, dispersion statistics measure how the data values are scattered in general, or with respect to a central location measure. For quantitative variables, the most important are:

1- Range
2- Variance
3- Standard deviation

### 1- Range :

Definition - Sample range. The sample range of a variable *X* is the difference between the the maximum and the minimum values in the sample.

$$\text{Range} = \max_{x_i} - \min_{x_i}$$



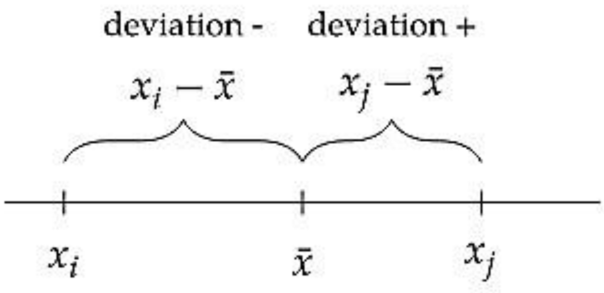
The range measures the largest variation among the sample data. However, it is very sensitive to ***outliers***, as they appear at the ends of the distribution, and ***for that reason is rarely used***.

**Example**: the measurements " 5, 7, 1, 2, 4 " vary from 1 to 7. Hence, the range is ( 7 - 1 = 6) . The range is easy to calculate, easy to interpret, and is an adequate measure of variation for small sets of data. ***For large data sets, the range is not an adequate measure of variability***.

 ***Deviations from the mean:***

Another way of measuring spread of data is with respect to a central tendency measure, as for example the mean.

In that case, it is measured the distance from every value in the sample to the mean, that is called **deviation from the mean**·



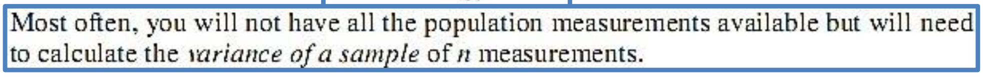
***If deviations are big, the mean is less representative than when they are small.***

## 2- Variance and standard deviation

**Variance of a population**:

The ***variance of a population*** of *N* measurements is the average of the squares of the deviations of the measurements about their mean "μ". The population variance is denoted by " $\boldsymbol{\sigma^2}$ " and is given by the formula.

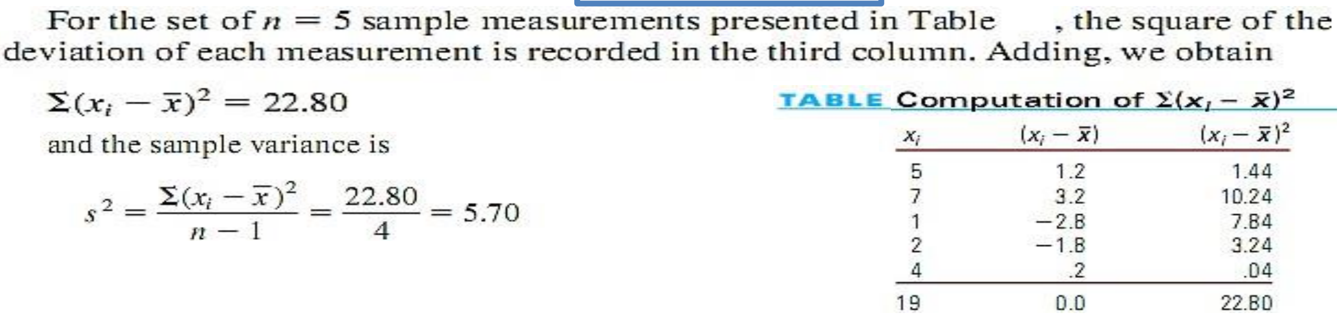
$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N}$$


Most often, you will not have all the population measurements available but will need to calculate the *variance of a sample* of *n* measurements.

**The variance of a sample :**

The ***variance of a sample*** of **"n"** measurements is the sum of the squared deviations of the measurements about their mean " $\bar{\mathbf{x}}$ " divided by (n - 1). The sample variance is denoted by $\mathbf{s^2}$ and is given by the formula.

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$$


For the set of $n = 5$ sample measurements presented in Table , the square of the deviation of each measurement is recorded in the third column. Adding, we obtain

$$\Sigma(x_i - \bar{x})^2 = 22.80$$

and the sample variance is

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{22.80}{4} = 5.70$$

TABLE Computation of $\Sigma(x_i - \bar{x})^2$

| $x_i$ | $(x_i - \bar{x})$ | $(x_i - \bar{x})^2$ |
|---|---|---|
| 5 | 1.2 | 1.44 |
| 7 | 3.2 | 10.24 |
| 1 | −2.8 | 7.84 |
| 2 | −1.8 | 3.24 |
| 4 | .2 | .04 |
| 19 | 0.0 | 22.80 |

The **variance** ( $\mathbf{s^2}$) is measured in terms of the square of the original units of measurement. If the original measurements are in inches, the variance is expressed in square inches. Taking the square root of the variance, we obtain the ***standard deviation***, which returns the measure of variability to the original units of measurement.

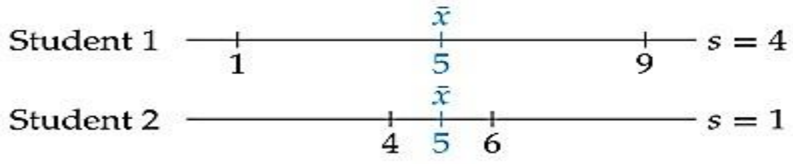
**Definition** : The ***standard deviation(S)*** of a set of measurements is equal to the positive square root of the variance

$$s = +\sqrt{s^2}$$

Both variance and standard deviation measure the spread of data around the mean. When the variance or the standard deviation are small, the sample data are concentrated around the mean, and the mean is a good representative measure. In contrast, when variance or the standard deviation are high, the sample data are far from the mean, and the mean does not represent so well.

| | | |
|---|---|---|
| Standard deviation small | ⇒ | Mean is representative |
| Standard deviation big | ⇒ | Mean is unrepresentative |

**Example.** The following samples contains the grades of 2 students in 2 subjects



*Which mean is more representative?*

## NOTATION

$n$: number of measurements in the sample

$s^2$: sample variance

$s = \sqrt{s^2}$: sample standard deviation

$N$: number of measurements in the population

$\sigma^2$: population variance

$\sigma = \sqrt{\sigma^2}$: population standard deviation

For the set of $n = 5$ sample measurements in Table , the sample variance is $s^2 = 5.70$, so the sample standard deviation is $s = \sqrt{s^2} = \sqrt{5.70} = 2.39$. The more variable the data set is, the larger the value of $s$.

## ***Shortcut method for calculating ( $s^2$).***

### THE COMPUTING FORMULA FOR CALCULATING $s^2$

$$\boxed{s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}} \quad \Longleftrightarrow \quad s^2 = \frac{\Sigma x_i^2 - \dfrac{(\Sigma x_i)^2}{n}}{n-1}$$

$\Sigma x_i^2$ = Sum of the squares of the individual measurements

$(\Sigma x_i)^2$ = Square of the sum of the individual measurements

**EXAMPLE** Calculate the variance and standard deviation for the five measurements in Table , which are 5, 7, 1, 2, 4. Use the computing formula for $s^2$ and compare your results with those obtained using the original definition of $s^2$.

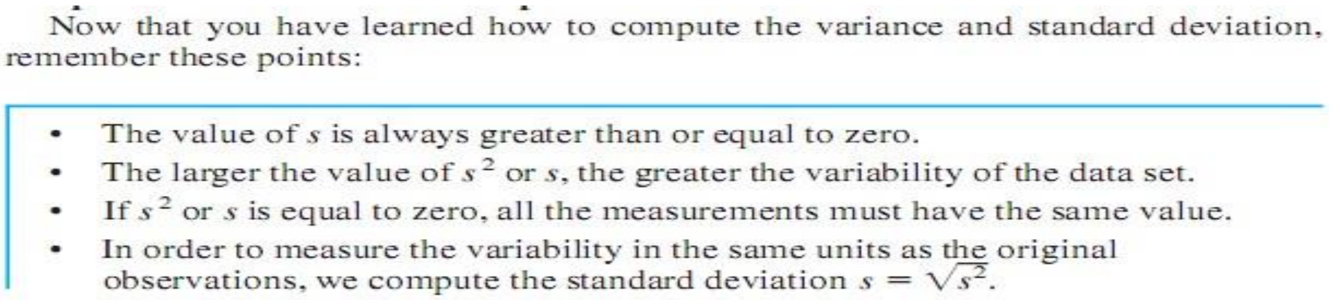
**Table for Simplified Calculation of $s^2$ and $s$**

| $x_i$ | $x_i^2$ |
|---|---|
| 5 | 25 |
| 7 | 49 |
| 1 | 1 |
| 2 | 4 |
| 4 | 16 |
| 19 | 95 |

**Solution** The entries in Table are the individual measurements, $x_i$, and their squares, $x_i^2$, together with their sums. Using the computing formula for $s^2$, you have

$$s^2 = \frac{\Sigma x_i^2 - \dfrac{(\Sigma x_i)^2}{n}}{n-1} = \frac{95 - \dfrac{(19)^2}{5}}{4} = \frac{22.80}{4} = 5.70$$

and $s = \sqrt{s^2} = \sqrt{5.70} = 2.39$, as before.



Now that you have learned how to compute the variance and standard deviation, remember these points:

- The value of $s$ is always greater than or equal to zero.
- The larger the value of $s^2$ or $s$, the greater the variability of the data set.
- If $s^2$ or $s$ is equal to zero, all the measurements must have the same value.
- In order to measure the variability in the same units as the original observations, we compute the standard deviation $s = \sqrt{s^2}$.

**Example - Non-grouped data.** Using the data of the sample with the number of children of families, with mean $\bar{x} = 1.76$ children, and adding a new column to the frequency table with the squared values,

$$\boxed{s^2 = \frac{\sum x_i^2 n_i}{n} - \bar{x}^2}$$

$$\boxed{\bar{x} = \frac{\sum x_i n_i}{n} = \frac{44}{25} = 1.76 \text{ children}}$$

| $x_i$ | $n_i$ | $x_i^2 n_i$ |
|---|---|---|
| 0 | 2 | 0 |
| 1 | 6 | 6 |
| 2 | 14 | 56 |
| 3 | 2 | 18 |
| 4 | 1 | 16 |
| $\sum$ | 25 | 96 |

$$s^2 = \frac{\sum x_i^2 n_i}{n} - \bar{x}^2 = \frac{96}{25} - 1.76^2 = 0.7424 \text{ children}^2.$$

and the standard deviation is $s = \sqrt{0.7424} = 0.8616$ children.

Compared to the range, that is 4 children, the standard deviation is not very large, so we can conclude that the dispersion of the distribution is small and consequently the mean, $\bar{x} = 1.76$ children, represents quite well the number of children of families of the sample.

**Example - Grouped data.** Using the data of the sample with the heights of students and grouping heights in classes, we got a mean $\bar{x} = 174.67$ cm. The calculation of variance is the same than for non-grouped data but using the class marks.

$$\boxed{s^2 = \frac{\sum x_i^2 n_i}{n} - \bar{x}^2 :}$$

$$\boxed{\bar{x} = \frac{\sum x_i n_i}{n} = \frac{5240}{30} = 174.67 \text{ cm}}$$

| $X$ | $x_i$ | $n_i$ | $x_i^2 n_i$ |
|---|---|---|---|
| (150, 160] | 155 | 2 | 48050 |
| (160, 170] | 165 | 8 | 217800 |
| (170, 180] | 175 | 11 | 336875 |
| (180, 190] | 185 | 7 | 239575 |
| (190, 200] | 195 | 2 | 76050 |
| $\sum$ | | 30 | 918350 |

$$s^2 = \frac{\sum x_i^2 n_i}{n} - \bar{x}^2 = \frac{918350}{30} - 174.67^2 = 102.06 \text{ cm}^2,$$

and the standard deviation is $s = \sqrt{102.06} = 10.1$ cm.

This value is quite small compared to the range of the variable, that goes from 150 to 200 cm, therefore the distribution of heights has little dispersion and the mean is very representative.

11

## PERMUTATIONS

**Permutations**:
The total number of ways of **arranging** $n$ objects, taking $r$ at a time is given by

$$\frac{n!}{(n-r)!}$$

Notation: We use the notation $^nP_r$ (read as "n–p–r') to denote $\frac{n!}{(n-r)!}$.

That is, $^nP_r = \frac{n!}{(n-r)!}$.

example
the total number of arrangements of 8 books on a bookshelf if only 5 are used

**solution**

$$^8P_5 = \frac{8!}{(8-5)!} = \frac{8!}{3!} = 6720\ .$$

example In how many ways can 5 boys be arranged in a row

(a) using three boys at a time?
(b) using 5 boys at a time?

We have 5 boys to be arranged in a row with certain constraints.

(a) The constraint is that we can only use 3 boys at a time. In other words, we want the number of arrangements (permutations) of 5 objects taken 3 at a time.
From rule 4: $n = 5, r = 3,$

Therefore, number of arrangements = $^5P_3 = \frac{5!}{(5-3)!} = \frac{120}{2} = 60$

(b) This time we want the number of arrangements of 5 boys taking all 5 at a time.
From rule 4: $n = 5, r = 5,$

Therefore, number of arrangements = $^5P_5 = \frac{5!}{(5-5)!} = \frac{120}{0!} = 120$

## permutations with repetitions:

The number of permutations of $n$ objects of which $n_1$ are identical, $n_2$ are identical, . . . , $n_k$ are identical is given by $\frac{n!}{n_1! \times n_2! \times \ldots \times n_k!}$.

**Example :** How many different arrangements of the letters of the word HIPPOPOTAMUS are there?

**Solution :** $\frac{12!}{3! \times 2!} = 39916800$ arrangements.

# COMBINATIONS

On the otherhand, combinations represent a counting process where the order has no importance. For example, the number of combinations of the letters A, B, C and D, if only two are taken at a time, can be enumerated as:

AB, AC, AD, BC,BD, CD,

That is, the combination of the letters A and B, whether written as AB or BA, is considered as being the same.

Instead of combination the term selection is often used.

**Combinations:**

The total number of ways of **selecting** $n$ objects, taking $r$ at a time is given by

$$\Longrightarrow \boxed{\frac{n!}{(n-r)!r!}}$$

Notation: We use the notation $\binom{n}{r}$ (read as "n–c–r') to denote $\frac{n!}{(n-r)!r!}$.

That is, $\binom{n}{r} = \frac{n!}{(n-r)!r!}$. Note:Sometimes ${}^{n}C_{r}$ is used instead of $\binom{n}{r}$.

**Example :** in how many ways can 5 books be selected from 8 different books?

**Solution** In this instance, we are talking about selections and therefore, we are looking at combinations. .Therefore we have. the selection of 8 books taking 5 at a time is equal to

$$\binom{8}{5} = \frac{8!}{(8-5)!5!} = \frac{8!}{3!5!} = 56$$

**Example :** A sports committee at the local hospital consists of 5 members. A new committee is to be elected, of which 3 members must be women and 2 members must be men. How many different committees can be formed if there were originally 5 women and 4 men to select from?

First we look at the number of ways we can select the women members (using Rule 6):

We have to select 3 from a possible 5, therefore, this can be done in ${}^{5}C_{3} = 10$ ways.

Similarly, the men can be selected in ${}^{4}C_{2} = 6$ ways.

Using Rule 2, we have that the total number of possible committees = ${}^{5}C_{3} \times {}^{4}C_{2} = 60$.