# DESCRIPTIVE STATISTICS

Adventist University of Central Africa (AUCA)

April 15, 2024

# 1. Introduction

What does Statistics mean?

"Statistics is a field of study (or science of numerical data) concerned with data collection, organization, summarization, and analysis that produces meaningful inferences about a body of data when only a part of the data is observed."

# Subdivisions of Statistics

Statistics has two main divisions:

- **Descriptive statistics**: Descriptive statistics is the science studying the methods and procedures used in collecting, organizing and presenting, summarizing, analyzing, interpreting of the numerical data.
- **Inferential statistics**: Inferential statistics is a field of statistics that uses analytical tools for drawing conclusions about a population by examining random samples.

# Cont.

Examples:

1. Based on sample information, the pollster predicted Demosthenes would be elected. **Inferential statistics**
2. The population of Rwanda in 1984 was 5 million. **Descriptive Statistics**
3. According to the pool forecasting, Demosthenes would get 54.3 percent of the votes cast. **Inferential statistics**
4. An engagement approach to learning work; the 29 students who generated summaries and inferences performed 20 percent better than those who just memorized. **Descriptive Statistics**

# Terms and Vocabulary

- **Population**: The universe (the set) of all potential observations having a common characteristic that is being studied and about which the experimenter wishes to make some general statements or inference.
- **Sample**: Census is practically impossible for an infinite population or for a population with large size. In such cases, the enumeration will be restrained to a limited number of individuals in the population called a sample.
- **Experimental unit**: Person, thing, event, or any item involved with a statistical study.

  Eg. Height of the people, duration of the exam, distance from the school to my house.

  Here people, exam, and road/venue from the school to my house are experimental units.

# Cont.

- **Sampling**: The different methods and rules to apply for selecting atypical sample of the population are called sampling. Sampling is often called sampling methods or sampling procedures. These are different sampling methods:
  1. **Simple random sample:** Draw a sample of size $n$ from the population of size $N$ in a such way that every sample of size $n$ has the same chance of being selected. A such sample may proceeds from sampling **with replacement** or **without replacement**;
  2. **Systematic Sampling :** Draw $n$ values from the population for a sample according to the initial values $x_i$ at the $i^{th}$ position, and other values are $x_{i+kh}$ at $(i + kh)^{th}$ positions, $k = 1, 2, \ldots, n - 1$ ,and $h$ being the sampling period;
  3. **Stratified Random Sampling:** Subdivide the population into $k$ strata according to different criteria; and then select from each strata a simple random sample of size $n^{'}$. The last are gathered in a final sample of size $n = kn^{'}$.

- **Population size**: The total number of items in a population. It is represented by $N$ but the population of the sample is noted by $n$
- **Census**: An enumeration or evaluation of every member of a population.
- **Parameter**: Descriptive measure obtained by calculations from numerical data of the population or Any constant value calculated from the population. For example the mean, proportion, and variance.
- **Statistic**: Descriptive Measure obtained by calculation from numerical data of a sample or any constant value calculated from the sample. For example the mean, proportion, and variance.

# Data Variable and Scale of Measurements

The variable is defined as **observable characteristic common** to each experimental unit concerned with a statistics study, This characteristic may take different values on different experimental units.

Examples:

- diastolic blood pressure of a patient
- heart rate of a cat
- the heights of adult males
- the weights of preschool children,
- color of the t-chart
- identification number of the student

# Cont. Variable

Statistics count two types of variables: Qualitative and Quantitative variables.

**Quantitative Variables**: A quantitative variable is one that can be assigned numerical values by the process of measuring using an instrument of measuring (**Continuous variable**), or by counting (**Discrete variable**).
Example:

- height of adult males
- weight of preschool children
- age of patients seen in a dental clinic

Note: Measurements made on quantitative variables convey information regarding the amount.

**Qualitative Variables**:

Some characteristics are not capable of being measured in the sense that height, weight, and age are measured. But they can be categorized or identified by a number only. Such characteristics are called qualitative variables.

Examples:

- health status of a patient
- color of the t-chart
- gender of a student
- nationality of the person
- academic performance (grand distinction, distinction, satisfaction, and failure)

Measurements made on qualitative variables convey information regarding attributes.

# Scale of measurements

Definition:

**Measurement** Measurement is defined as the assignment of a numerical value to different experimental units in conformity with a set of rules. For this reason, various scales result from the fact that measurement may be carried out under different sets of rules.

There exist 4 scales of measurement of variables in statistics:

- Nominal scale
- Ordinal scale
- Interval scale
- Ratio scale

**The Nominal Scale**: The lowest measurement scale is the nominal scale. As the name implies it consists of **"naming"** observations or classifying them into various mutually exclusive and collectively exhaustive categories.

Example:

- color of the t-chart (blue = 1, green =2, dark = 3, red = 4)
- gender of a student (male = 1, female = 2)
- nationality of the person (Rwandan = 1, Ugandan = 2, Kenyan = 3, Gabonese = 4, Chadian = 5, Zambian = 6)

**The Ordinal Scale**: Whenever observations are not only different from category to the category but can be ranked according to some criterion, they are said to be measured on an ordinal scale.

Example:

- Convalescing patients may be characterized as (unimproved = 0, improved = 1, much improved = 2)
- Individuals may be classified according to socioeconomic status as (low = 1, medium = 2, or high = 3)
- The intelligence of children may be (above average = 2, average = 1, or below average = 0)
- academic performance (grand distinction = 1, distinction = 2, satisfaction = 3, and failure = 4)

**Note**: The function of numbers assigned to ordinal data is to order (or rank) the observations from lowest to highest and, hence, the term ordinal.

**The Interval Scale**: The interval scale is a more sophisticated scale than the nominal or ordinal in that with this scale not only is it possible to order measurements but also the distance between any two measurements is known.

We know, say, that the difference between a measurement of 20 and a measurement of 30 is equal to the difference between measurements of 30 and 40.

The ability to do this implies choosing arbitrarily two points of reference for measuring:

- a zero point; 0
- a unit distance; 1

**Note**: The selected zero point is not necessarily a true zero in that it does not have to indicate a total absence of the quantity being measured.

Example: The following example are measured in interval scale:

- Temperature of the body ( zero degrees doesn't necessarily mean the absence of the heat of the experimental unit)
- price of the commodity on the market

**The Ratio Scale**: The highest level of measurement is the ratio scale. This scale is characterized by the fact that equality of ratios, as well as equality of intervals, maybe determined. Fundamental to the ratio scale is a true zero point. The measurement of such familiar traits as height, weight, and length makes use of the ratio scale.
Example:

- amount of money in the pocket ( zero degrees mean the total absence of money or empty pocket)
- number of the students per classroom ( zero students mean the total absence of students in the classroom)

# 2. Data Variable organization and Presentation.

## 2.1 Frequency Distribution.

They are terms that need to be known for a better understanding of the frequency distribution.

- **Array data:** is a set of the data variable sorted in ascending or descending order. For eg. For eg. 1, 2, 3, 4, 4, 5, 4, 2, 3, 6, 7, 7.
  Its correspondence array data is 1, 2, 2, 3, 3, 4, 4, 4, 5, 6, 7, 7.
- **Frequency**: The frequency of the observed value $x_i$ is the number of times it appears in the set of data. The frequency of the value $x_i$ is denoted by $f_i$. However, some books use $n_i$ to present this frequency.

  For eg. 1, 2, 3, 4, 4, 5, 4, 2, 3, 6, 7, 7. In this data variable, the frequency of 1 is 1, the frequency of 2 is 2, the frequency of 3 is 2, the frequency of 4 is 3, the frequency of 6 is 1, and the frequency of 7 is 2.

- **Frequency distribution** is a simple table of two rows for the different distinct observed values $x_i$ with their respective frequency $f_i$.

  For eg. the frequency distribution of the 7 different distinct values is the table here below:

| $x_i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $f_i$ | 1 | 2 | 2 | 3 | 1 | 1 | 2 |

  This frequency distribution is appropriate for only **discrete variable** when the number of the different distinct values $k$ is less than 12 i.e. $k \leq 12$. Otherwise, values are grouped into class intervals. The frequency distribution with class intervals is called "**grouped frequency distribution**"

- **Grouped Frequency Distribution**: It is a table in two rows; the 1st row for different class intervals in which fall observations and the 2nd exclusively for the frequency of the corresponding class interval.

  For eg. Find the grouped frequency distribution of the following data: 69 84 52 93 81 74 89 85 88 63 87 64 67 72 74 55 82 91 68 77

  Note that this set of values has more than 12 different distinct values. Values have to be grouped into class intervals.

  Let's group them into class intervals of length $c = 10$ starting from the value 50. These 20 observed values are put into 5 class intervals as follows:

| a-b | $50-60$ | $60-70$ | $70-80$ | $80-90$ | $90-100$ |
|---|---|---|---|---|---|
| $f_i$ | 2 | 5 | 4 | 7 | 2 |

**Rmrk**: Grouped frequency is recommended for a frequency distribution of continuous data variable

Some important elements should be known before the construction of the grouped frequency distribution. These are:

- The minimum number $k$ of class intervals required for the given data variable of size N. The following Sturges's rule should be used $k = 1 + 3.322 * log_{10}(N)$ for that purpose.

  For eg. Find the minimum number $k$ of class intervals in order to represent the following data by a grouped frequency distribution: 69 84 52 93 81 74 89 85 88 63 87 64 67 72 74 55 82 91 68 77

  Apply Sturges's rule by taking $n = 20$, we have $k = 1 + 3.322 * log_{10}20 = 1 + 3.322 * 1.3010299957 = 5.322022$. We should round up the value of $k$ to the next natural number. i.e., $k = 6$. but $k \leq 12$, therefore $k = 6, 7, 8, 9, \ldots, 12$. We have to modify the previously grouped frequency distribution so as to have more than 5 class intervals.

Suppose now we take the number of class intervals $k = 6$, then the length of each class interval; called **class width** $c$ is equal to the ratio obtained from the division of the difference of largest and smallest values divided by the number of class intervals e.i., $c = (maxvalue - minvalue)/k$, and adjust its value to accommodate all values of the given data variable.

Find the value of "the class width" $c$ needed to group the following data 69 84 52 93 81 74 89 85 88 63 87 64 67 72 74 55 82 91 68 77 into 6 class intervals

Solution: The value of $c = \frac{(93-52)}{6} = 6.833333$ is rounded up a little to allow the last class interval to contain the maximum value. Choose $c = 7$ or $c = 8$

The modified grouped frequency distribution of the example to have six class intervals of class width $c = 8$ is:

| a-b | $52 - 60$ | $60 - 68$ | $68 - 76$ | $76 - 84$ | $84 - 92$ | $92 - 100$ |
|---|---|---|---|---|---|---|
| $f_i$ | 2 | 3 | 5 | 3 | 6 | 1 |

- **Extended Frequency Distribution Table**:

  Extended Frequency Distribution table is a table with $k$ rows and headings $i$: the counter of different distinct values, $x_i$: the $i^{th}$ observed values, $f_i$: the frequency of the $i^{th}$ observed value, and $cf_i$ the cumulative frequency of the $i^{th}$ observed value (or the position of the last $x_i$ value in an array of values sorted in ascending order)

  Find the extended frequency distribution that corresponds to the following table:

| $x_i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $f_i$ | 1 | 2 | 2 | 3 | 1 | 1 | 2 |

**Note**:

The cumulative frequency of the $x_i$ is given by the formula $cf_1 = f_1$ for the first value $x_1$ and $cf_i = cf_{i-1} + f_i$ for all $i > 1$

Example:

Find the cumulative frequencies of the values presented by the previous frequency distribution

$cf_1 = f_1 = 1,$
$cf_2 = cf_1 + f2 = 1 + 2 = 3,$
$cf_3 = cf_2 + f_3 = 3 + 2 = 5,$
$cf_4 = cf_3 + f_4 = 5 + 3 = 8,$
$\ldots$

- **Extended table for Simple Frequency Distribution Table:**

| i | $x_i$ | $f_i$ | $cf_i$ |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 2 | 2 | 2 | 3 |
| 3 | 3 | 2 | 5 |
| 4 | 4 | 3 | 8 |
| 5 | 5 | 1 | 9 |
| 6 | 6 | 1 | 10 |
| 7 | 7 | 2 | 12 |
| $\sum_{i=1}^{7}$ | | 12 | |

Note that the cumulative frequency distribution indicates the position of the observed value $x_i$ in an array sorted in ascending order. For eg. the only one value of 1 is in position 1, the 2nd value of 2 is in position 3, the 2nd value of 3 is in position 5, the 3rd value of 4 is in position 8, the only one values of 5 and 6 are in their positions 9 and 10, and the 2nd value of 7 is in position 12

- **Extended Grouped Frequency Distribution Table:**

  This table is different from the last one. Apart to contain the counter, different class intervals, frequency, and cumulative frequency, it has an inserted column for midpoints $m_i$ just after the column for different class intervals.

  The following table presents the extended grouped frequency distribution. Remark that the midpoint of the $i^{th}$ class interval, $a - b$, is $m_i = \frac{(b+a)}{2}$

| i | a-b | $m_i$ | $f_i$ | $cf_i$ |
|---|---|---|---|---|
| 1 | $52 - 60$ | 56 | 2 | 2 |
| 2 | $60 - 68$ | 64 | 3 | 5 |
| 3 | $68 - 76$ | 72 | 5 | 10 |
| 4 | $76 - 84$ | 80 | 3 | 13 |
| 5 | $84 - 92$ | 88 | 6 | 19 |
| 6 | $92 - 100$ | 96 | 1 | 20 |
| $\sum_{i=1}^{6}$ | | | 20 | |

# Activity II

1. Represent the following data of the ages of 62 people who live in a certain neighborhood by an appropriate frequency distribution. Construct its corresponding extended frequency distribution table:
   2, 5, 6, 12, 14, 15, 15, 16, 18, 19, 20, 22, 23, 25, 27, 28, 30, 32, 33, 35, 36, 36, 37, 38, 39, 40, 40, 41, 42, 43, 43, 44, 44, 45, 45, 46, 47, 47, 48, 49, 50, 51, 56, 57, 58, 59, 59, 60, 62, 63, 65, 65, 67, 69, 71, 75, 78, 80, 82, 84, 90, 96
2. Octane levels for various gasoline blends are given below: 87.9 84.2 86.9 87.7 91.7 88.8 95.3 93.5 94.3 88.1 90.2 91.4 91.3 93.9
   Represent these data by an appropriate extended frequency distribution table. Explain why you made a such choice.
3. The following are data on the number of students per classroom in AUCA, Faculty of Education. Represent them by an appropriate frequency distribution table
   14 11 10 8 12 13 11 10 16 11 11 9 9 7 14 12 9 10 11 6 13 8 11 11 9 8 13 16 10 11 9 8 12 11 10

# 2.2. Graphical Presentation of Data Variable

Statistical data variables are often presented by a graph or chart. The type of chart depends upon the nature of the variable it may represent.

- Qualitative variables are either represented by a pie chart or a bar chart.
- Quantitative discrete variables are represented by rod/spike chart, frequency polygon, and cumulative frequency chart (ogive in stairs form).
- Quantitative continuous variables are represented by a Histograms chart, polygon frequency chart, and ogives in a continuous curve form.

# 2.2.1. Graphical Presentation of Qualitative Data Variable

- **Pie Chart:**
  A pie chart is often used to indicate relative frequencies when the data are not numerical in nature (for categorical and nominal variable). A circle is constructed and then sliced into different sectors; one for each distinct type of data value. The relative frequency of a data value is indicated by the area of its sector, this area is described by an angular sector $\theta_i$ proportional to the relative frequency $F_i = f_i/N$ of the data value. $\theta_i = 360^0 * F_i$

**Example:**

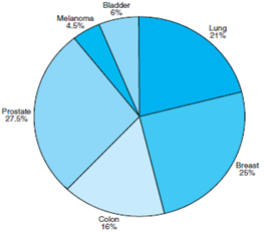
The following data relate to the different types of cancers affecting the 200 most recent patients to enroll at a clinic specializing in cancer.

| *Type* | *lung* | *Bearst* | *Colon* | *Prostate* | *Melanoma* | *Bladder* |
|---|---|---|---|---|---|---|
| *No cases* | 42 | 50 | 32 | 55 | 9 | 12 |

**Solution:**

Generate the extended frequency distribution table:

| $i$ | *category* | $f_i$ | $F_i = \frac{fi}{N}$ | $\theta_i = 360^0 * F_i$ |
|---|---|---|---|---|
| 1 | *lung* | 42 | 0.21 | 75.6 |
| 2 | *Breast* | 50 | 0.25 | 90 |
| 3 | *Colon* | 32 | 0.16 | 57.6 |
| 4 | *Prostate* | 55 | 0.275 | 99 |
| 5 | *Melanoma* | 9 | 0.045 | 16.2 |
| 6 | *Bladder* | 12 | 0.06 | 21.6 |
| $\sum_{i=1}^{6}$ | | 200 | 1 | 360 |

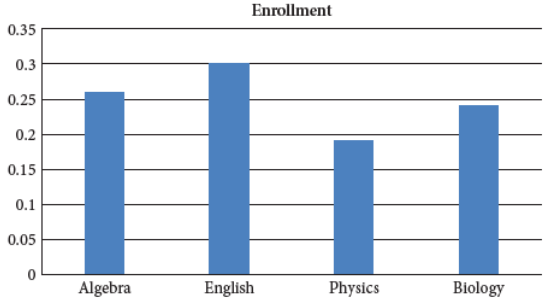

Bladder
6%
Melanoma
4.5%
Lung
21%
Prostate
27.5%
Breast
25%
Colon
16%

- **Bar Chart:**
  Bar chart is another alternative representation of qualitative data variable. It consists of a sequence of the equidistant vertical rectangles proportional to the frequency $f_i$ for each value $x_i$, drawn in the $XY-$plane

**Example:**

Frequency distribution of the enrollment of four classes in a high school is given in the following table.

| $i$ | *Class* | $f_i$ | $F_i = \frac{fi}{N}$ | $distance(d)(cm)$ |
|---|---|---|---|---|
| 1 | *Algebra* | 26 | 0.26 | 8.6 |
| 2 | *English* | 30 | 0.30 | 10 |
| 3 | *Physics* | 19 | 0.19 | 6.3 |
| 4 | *Biology* | 24 | 0.24 | 8 |
| $\sum_{i=1}^{4}$ | | 99 | 1 | |




**Note:** The bar chart is fitted within the available space by the scale defined by the distance at which the largest frequency is fixed. i.e at $x$ cm from the origin of the y-axis (axis of frequencies).
30 freq $\longrightarrow$ 10 cm
1 freq $\longrightarrow \frac{10}{30} = 0.3cm$ The positions are indicated in the last column of the above-extended frequency distribution table.

# 2.2.2. Graphical Representation of Quantitative Discrete Data Variable

- Rod/Spike Chart:
  This graphic is drawn in XY-plane where different distinct values $x_i$ are found on X-axis, and their frequency $f_i$ on Y-axis. The graphic consists of vertical line segments starting from X-axis at the point $x_i$, and of height proportional to the frequency $f_i$.

**Example:**

Represent the following data variable by Rod / Spike Chart:

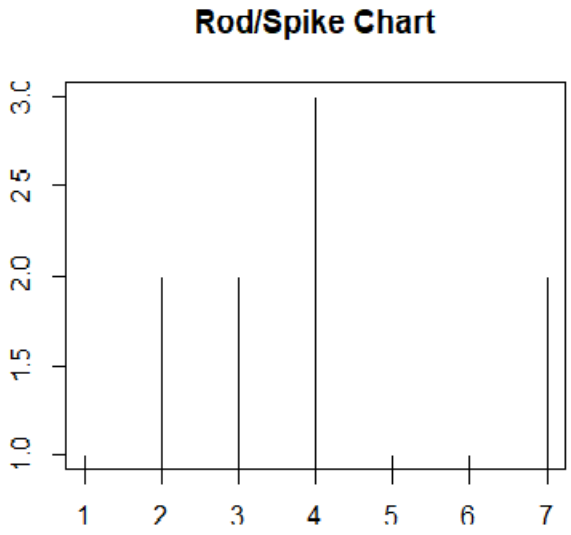
| $x_i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $f_i$ | 1 | 2 | 2 | 3 | 1 | 1 | 2 |

The Extended frequency distribution table of the given example is:

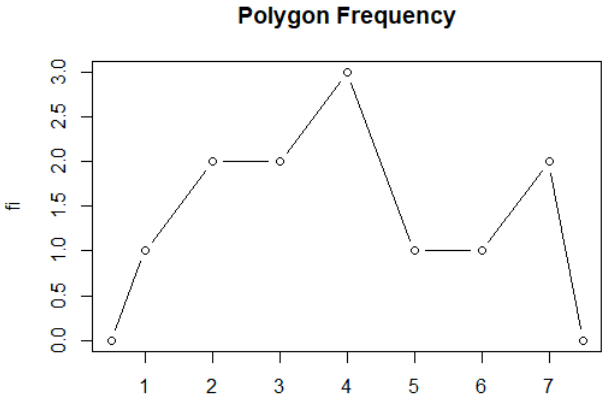
| i | $x_i$ | $f_i$ | $cf_i$ | $location(cm)$ |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 3 |
| 2 | 2 | 2 | 3 | 6 |
| 3 | 3 | 2 | 5 | 6 |
| 4 | 4 | 3 | 8 | 9 |
| 5 | 5 | 1 | 9 | 3 |
| 6 | 6 | 1 | 10 | 3 |
| 7 | 7 | 2 | 12 | 6 |
| $\sum_{i=1}^{7}$ | | 12 | | |

Let's locate the largest frequency 3 on Y-axis (axis of frequencies) at 9 cm. The location of other frequencies follows from the correspondence 3 freq $\longrightarrow$ 9 cm. We have the correspondence: 1 freq $\longrightarrow \frac{9}{3} = 3cm$ (one unit of frequency must be marked at 3 cm from the origin 0 on Y-axis.)

Finally, we have the following Rod/Spike Chart:

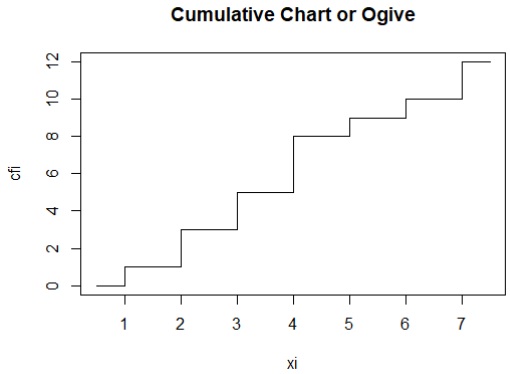


- **Polygon Frequency:**
  Polygon frequency is obtained by joining every two consecutive upper points of the rod/spike chart by a line segment. Here below is the polygon chart generated from the previous rod/spike chart.



- **Cumulative frequency chart (Ogive):**
  Ogive is the chart drawn in XY-plane defined by the step function.

$$f(x) = \begin{cases} 0 & , x < x_1, \\ cf_i & , x_i \le x < x_{i+1}, \text{ for all } i = 1, 2, \ldots, (k-1) \\ 1 & , x \ge x_k \text{ the last observed value} \end{cases}$$



# 2.2.3. Graphical Representation of Quantitative Continuous Data Variable

In this part, we shall use the following extended grouped frequency distribution table that presents the lifetimes of 200 incandescent lamps.

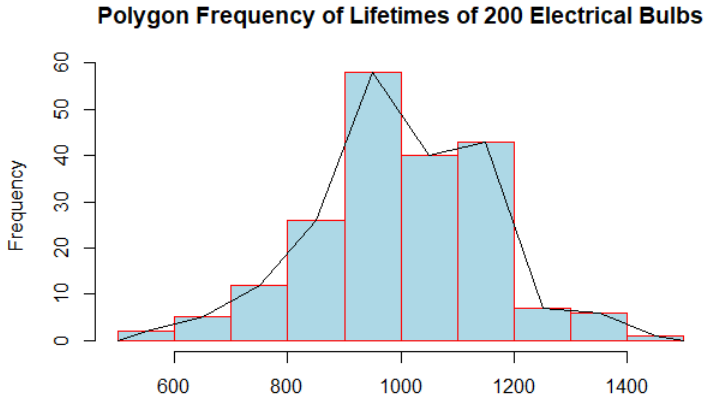
| **Number** (i) | **Class** $(a_i - b_i)$ | **Midpoint** $(m_i = \frac{a_i + b_i}{2})$ | **Frequency** $(f_i)$ | **Cumulative frequency** $(cf_i)$ |
|---|---|---|---|---|
| 1 | **500 – 600** | 550 | 2 | 2 |
| 2 | ***600–700*** | 650 | 5 | 7 |
| 3 | ***700–800*** | 750 | 12 | 19 |
| 4 | **800 – 900** | 850 | 25 | 44 |
| 5 | 900 – 1000 | 950 | 58 | 102 |
| 6 | **1000 – 1100** | 1050 | 41 | 143 |
| 7 | **1100 – 1200** | 1150 | 43 | 186 |
| 8 | **1200 – 1300** | 1250 | 7 | 193 |
| 9 | **1300 – 1400** | 1350 | 6 | 199 |
| 10 | **1400 – 1500** | 1450 | 1 | 200 |
| $\sum_{i=1}^{7}$ | | | 200 | |

- **Histogram:**
  Histogram is a series of contingent rectangles of equal breadth, drawn in XY-plane, whose heights are proportional to the frequency of each class interval.

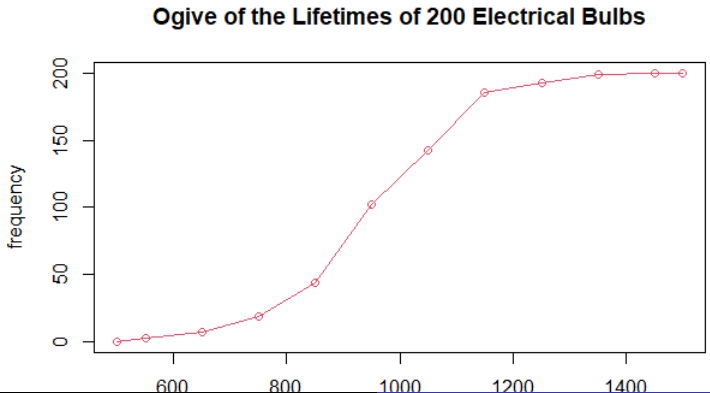
Remember that to draw the histogram, it is necessary to fix first the largest frequency at a specified distance $x$ (units of distance) from the origin 0 of the axis of frequencies.

The next ppt shows a histogram that corresponds to the lifetimes of 200 incandescent lamps. We should fit the chart within 10 cm i.e. fix the frequency 58 at 10 cm from the origin of the axis of frequencies

- **Polygon frequency:**
  Polygon frequency of the grouped frequency distribution is obtained by joining the consecutive upper midpoint of each class interval (midpoint of the upper bases of the rectangles) by line segments.



- **Cumulative frequency (Ogive):**
  The Cumulative frequency (ogive) that represents the grouped frequency distribution is a simple smooth curve, described by the points $(m_i, cf_i)$, through the upper points of midpoints drawn from the first and then through all other midpoints in their successive orders.



# 2.2.4. Data Presentation by Stem-and-Leaf Graphic (Chart)

One simple graph, stem-and-leaf or stemplot, comes from exploratory data analysis. It is a good choice when the data sets are small or for grouped data.

To create the chart, divide each data observation into a stem and a leaf. The **leaf** consists of the **last significant digit** of the observed value, and the **stem is the remaining** part of that value.

For example:

- 23 has stem 2 and leaf 3 (3=the last significant digit)
- 432 has stem 43 and leaf 2 (2=the last significant digit)
- 5,432 has stem 543 and leaf 2 (2=the last significant digit)
- 9.3 has stem 9 and leaf 3

**A stemplot is a table with two columns in which the first column contains all stems in ascending order, and their respective leaves sorted in ascending order, in the second column**

**Example:**

Generate the stemplot corresponding to the following scores for the final exam of descriptive statistics.

33 42 49 49 53 55 55 61 63 67 68 68 69 69 72 73 74 78 80 83 88 88 88 90 92 94 94 94 94 96 100

The following is the stem-and-leaf graphic of the above data

```
 3 | 3
 4 | 299
 5 | 355
 6 | 1378899
 7 | 2348
 8 | 03888
 9 | 0244446
10 | 0
```

# 3. Summarizing Data Variable

In statistics, data variables are summarized by some values computed (or determined) from the values of that data variable. These values are often referred to as statistical descriptor measures of the statistical data variable.

The statistical descriptor values are classified into two categories: **measures of central tendency** and **measures of spread**. The measures of central tendency comprise **the mean, the median, the mode, and the quantiles** while the tree **measures; the ranges, variance, standard deviation, and coefficient of variation** are expressing the spread or volatility of the data variable.

# 3.1. Measures of Central Tendency

One of the important objectives of statistical analysis is to determine various numerical measures that describe the inherent characteristics of a frequency distribution. These measures are called averages. They condense the set of numerical data into a single numerical value to represent the entire distribution. Averages are the values that lie between the smallest and the largest observations of the distribution. They reflect the pattern of concentration of the values in the central part of the distribution. Averages are helpful for the following reasons:

1. concisely describing the distribution;
2. Comparative study of different distributions;
3. Computing various other statistical measures such as correlation, regression, dispersion, skewness, and other various basic characteristics of a mass of data.

# 3.1.1. The mean

The mean in statistics stands for the representative of other values. It is involved in good number of calculations needed for statistical analysis of a data variable(s). In statistics, there are different kinds of means which are **arithmetic mean, geometric mean, harmonic mean, quadratic mean, and weighed mean**

- **Arithmetic mean:**
  The arithmetic mean of N observed values equals the sum of the resulting values divided by the total number of these values (i.e. N).

  The arithmetic mean of the statistical variable x, is the value denoted by $\mu$ and $\bar{x}$ (for the sample) and is given by the direct formula by:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

  This formula is given when the data are given in form of the raw data

# Cont.

**Example**:
Find the arithmetic mean of the following data of the variable x: 24 39 7 48 16 29 34 20 43 18

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i$$

$$\bar{x} = \frac{24+39+7+48+16+29+34+20+43+18}{10} = \frac{278}{10} = 27.8$$

- **Arithmetic mean for data in a simple frequency distribution:**
  If data are given in frequency distribution, the previous formula must include the frequency for each distinct observed value $x_i$. The arithmetic mean is given by

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} x_i f_i$$

**Example**:

Find the arithmetic mean of the data variable $x$ represented by the following frequency distribution:

| $x_i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $f_i$ | 1 | 2 | 2 | 3 | 1 | 1 | 2 |

**Solution**:

The problem must be solved by adding column for $x_i f_i$ to the extended frequency distribution table. This looks like:

| i | $x_i$ | $f_i$ | $cf_i$ | $x_i f_i$ |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 |
| 2 | 2 | 2 | 3 | 4 |
| 3 | 3 | 2 | 5 | 6 |
| 4 | 4 | 3 | 8 | 12 |
| 5 | 5 | 1 | 9 | 5 |
| 6 | 6 | 1 | 10 | 6 |
| 7 | 7 | 2 | 12 | 14 |
| $\sum$ | | | | |

$$\bar{x} = \frac{1}{12}\sum_{i=1}^{7} x_i f_i$$

$$\bar{x} = \frac{48}{12} = 4$$

**Note:**

When the different distinct values $x_i$ and $f_i$ have many digits, the product $x_i f_i$ may take time to be computed to fill in the columns of the extended frequency distribution table. To simplify the task, we have to use the following formula called **the short-cut formula** for the calculation of the arithmetic mean

$$\bar{x} = A + \frac{1}{N}\sum_{i=1}^{k} d_i f_i$$

Where, $A$: the assumed mean (One of the $x_i$ value taken in the central part of the column of $x_i$). For the previous example $A = 4$, the $4^{the}$ observed value.

$d_i$: the deviation of the value $x_i$ from the assumed mean $A$. e.i, $d_i = x_i - A$

Find the arithmetic mean of the previous example by using the short-cut formula:

| i | $x_i$ | $f_i$ | $cf_i$ | $d_i = x_i - 4$ | $d_i f_i$ |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | $-3$ | $-3$ |
| 2 | 2 | 2 | 3 | $-2$ | $-4$ |
| 3 | 3 | 2 | 5 | $-1$ | $-2$ |
| 4 | 4 | 3 | 8 | 0 | 0 |
| 5 | 5 | 1 | 9 | 1 | 1 |
| 6 | 6 | 1 | 10 | 2 | 2 |
| 7 | 7 | 2 | 12 | 3 | 6 |
| $\sum_{i=1}^{7}$ | | 12 | 48 | | 0 |

$$\bar{x} = 4 + \frac{1}{12} \sum_{i=1}^{7} d_i f_i$$

$$\bar{x} = 4 + \frac{0}{12} = 4$$

- **Arithmetic mean for data in a grouped frequency distribution**
  When data variable $x$ is represented par a grouped frequency distribution, the arithmetic mean $\bar{x}$ is obtained by the formula,

$$\bar{x} = A + \frac{c}{N}\sum_{i=1}^{k} D_i f_i$$

where
$k$: the number of different class intervals
$A$: the assumed mean which is one of the mid-points $m_i$ picked from the central part of the column of the mid-points.
$D_i$: the reduced deviation of the mid-point $m_i$ from the assumed mean $A$ i.e., $D_i = \frac{m_i - A}{c}$
$c$: Class width or the length of each class interval

**Note:** The above formula is called **step-deviation formula** and gives an approximate value of the arithmetic mean.

Find the arithmetic mean of the data variable here below represented by the following grouped frequency distribution with the first class interval: $0 - 10$

| *class* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $f_i$ | 5 | 10 | 25 | 30 | 20 | 10 | 5 | 5 |

**Solution:**

| i | a-b | $m_i$ | $f_i$ | $cf_i$ | $D_i = \frac{m_i - 35}{10}$ | $D_i f_i$— |
|---|---|---|---|---|---|---|
| 1 | $0 - 10$ | 5 | 5 | 5 | $-3$ | $-15$ |
| 2 | $10 - 20$ | 15 | 10 | 15 | $-2$ | $-20$ |
| 3 | $20 - 30$ | 25 | 25 | 40 | $-1$ | $-25$ |
| 4 | $30 - 40$ | 35 | 30 | 70 | 0 | 0 |
| 5 | $40 - 50$ | 45 | 20 | 90 | 1 | 20 |
| 6 | $50 - 60$ | 55 | 10 | 100 | 2 | 20 |
| 7 | $60 - 70$ | 65 | 5 | 105 | 3 | 15 |
| 8 | $70 - 80$ | 75 | 5 | 110 | 4 | 20 |
| $\sum_{i=1}^{7}$ | | | 110 | | | 15 |

Finally, we obtain the approximate value of the arithmetic mean equal to:

$$\bar{x} = A + \frac{c}{N} \sum_{i=1}^{k} D_i f_i$$

$$\bar{x} = 35 + \frac{10x15}{110}$$

$$\bar{x} = 35 + \frac{15}{11}$$

$$\bar{x} = 36.36$$

- **Properties of the arithmetic mean:**

The arithmetic mean of the data variable $x$ satisfies the following properties:

P1: The sum of deviations of observed values from the actual arithmetic mean $\bar{x}$ is zero. That is $\sum(x_i - \bar{x}) = 0$

P2: The sum of squared deviations of the observed values from the arithmetic mean is minimum compared to other sum of squared deviations of the observed values from any value A different form the arithmetic mean. That is $\sum(x_i - \bar{x})^2 \leq \sum(x_i - A)^2$, for any value A.

P3: The Combined average computed from the different arithmetic means of p groups of data is equal to the arithmetic mean $\bar{x}$ of the n groups put together and it is given by

$$\bar{x}_{1,2,3,\ldots,n} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_p\bar{x}_p}{n_1 + n_2 + \cdots + n_p}$$

Where

$\bar{x}_{1,2,3,\ldots,p}$ : The combined average (mean) of p groups;

$\bar{x}_i$: the arithmetic mean of the group i ;

$n_i$ : The number of items in the group i.

P4: The sum of all observed values in the set of statistical series is equal to the product of total number of observed values and the arithmetic mean of these observed values. Mathematically we write $\sum x_i f_i = N\bar{x}$.

**Example:**

The arithmetic mean age of 80 boys is 10 years old and that of the another group of 20 boys is 15 years old. Find the arithmetic mean of the two groups.

**Solution:**

The problem is concerned with two groups of boys of numbers $n_1 = 80$ and $n_2 = 20$ with their respective mean age old $x_1 = 10$ and $x_2 = 15$. Find the arithmetic mean of the two groups of boys put together.

$$\bar{x} = \bar{x}_{1,2}$$

$$\bar{x}_{1,2} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

$$\bar{x}_{1,2} = \frac{80(10) + 20(15)}{80 + 20}$$

$$\bar{x}_{1,2} = \frac{1100}{100}$$

**ACTIVITY**

1. In a class of 50 students, 10 have failed and their average mark is 2.5. The total marks secured by the entire class was 281. Find the average marks of those who have passed.

2. The mean weight of 100 students in a class is 50Kg. The mean weight of boys in the class is 55kg and that of girls is 45kg. Find the number of boys and girls in the class.

3. The combined mean height of 60 children is 60.8cms. The mean height of the first 30 boys is 62cms, and that of the last 20 boys is 61.2cms. Find the mean height of the remaining 10 girls.

4. 100 students appeared for an examination. The results of those who failed are given below

| Marks | 5 | 10 | 15 | 20 | 25 | 30 |
|---|---|---|---|---|---|---|
| No of students | 4 | 6 | 8 | 7 | 3 | 2 |

If the average marks of all the 100 students were 68.6, find out the average marks of those who passed.

5. In a city, 30 members were surveyed as to know how many domestic appliances they have and the following were the result of the survey:**(Group 1 to 5)**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 5 | 1 | 5 | 2 | 1 | 4 | 2 | 3 |
| 4 | 2 | 4 | 3 | 2 | 6 | 3 | 2 | 4 | 3 |
| 6 | 2 | 2 | 3 | 3 | 7 | 2 | 3 | 0 | 2 |

Prepare the frequency distribution and draw the bar chart and cumulative frequency graph..

6. The marks scored by 50 students in an English examination are given below:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 30 | 45 | 48 | 55 | 39 | 25 | 31 | 12 | 18 | 21 |
| 54 | 59 | 51 | 33 | 43 | 44 | 10 | 38 | 19 | 26 |
| 47 | 35 | 37 | 41 | 46 | 33 | 51 | 37 | 58 | 58 |
| 17 | 19 | 23 | 26 | 29 | 38 | 57 | 36 | 35 | 44 |
| 43 | 27 | 31 | 43 | 22 | 31 | 47 | 34 | 18 | 15 |

Prepare a frequency distribution taking class interval or class width 10, and draw the histograms and ogives

# 3.1.2.Other Types of Mean

Apart from the arithmetic mean, as we say, the central value tendency knows other measure under the global name of **mean**:

- **Harmonic mean**

Harmonic mean of N observed values is denoted by H. The following formula is applied to find the harmonic mean

$$\frac{1}{H} = \frac{1}{N}\sum_{i=1}^{p}\frac{f_i}{x_i}$$

Or

$$H = \frac{N}{\sum_{i=1}^{p}\frac{f_i}{x_i}}$$

Let a and b be two real numbers $\frac{1}{H} = \frac{1}{2}\left(\frac{1}{a} + \frac{1}{b}\right) = \frac{1}{2}\left(\frac{b+a}{ba}\right) = \frac{b+a}{2ba}$

$$H = \frac{2ab}{a+b}$$

Example:

Consider the following data 2 and 3. Evaluate the arithmetic mean and the harmonic mean and prove that $2 < H < \bar{X} < 3$

$$AM\,(2,3) = \frac{2+3}{2} = \frac{5}{2} = 2.5$$

$$\frac{1}{H} = \frac{1}{2}\left(\frac{1}{2} + \frac{1}{3}\right)$$

$$= \frac{1}{4} + \frac{1}{6} = \frac{5}{12}$$

$$H = \frac{12}{5} = 2.4$$

$$2 < 2.4 < 2.5 < 3$$

$$2 < H < \bar{X} < 3$$

- **Geometric mean**

Geometric mean of N statistical values is G and it is given by the following formula:

$$G = \sqrt[N]{\prod_{i=1}^{p} x_i^{f_i}}$$

Example:

$$G = \left(1, \frac{1}{2}, \frac{1}{4}\right) = \sqrt[3]{1 * \frac{1}{2} * \frac{1}{4}} = \sqrt[3]{\frac{1}{8}} = \frac{1}{2}$$

Note: This mean is evaluated when the values of X are in geometric sequence.

- **Weighted mean**

In the evaluation of the arithmetic mean, each observed value in the statistical series is considered as of equal importance or equal straight or equal power. However, in some cases, one value can differ from another by their consideration or by the power. In this case, the weight ($w_i$) of each observed value must be considered. For this reason, the arithmetic mean computed considering the power of each value will be called ***weighted arithmetic mean*** and it is denoted by $\bar{x}_w$. The following is the formula use to evaluate the weighted arithmetic mean:

$$\overline{X}_w = \frac{\sum_{i=1}^{n} w_i \; X_i}{\sum_{i=1}^{n} w_i}$$

**Example:**

The household expenses for four households was found to be 2400Rwf in A, 3600Rwf in B, 5800Rwf in C and 6800Rwf in D. If the number of members in the family A, B, C and D is 2, 3, 4 and 6 persons respectively, then find the average household expenses of these four households.

**Solution:**

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4}{w_1 + w_2 + w_3 + w_4}$$

$$\bar{x}_w = \frac{2400 \times 2 + 3600 \times 3 + 5800 \times 4 \quad + 6800 \times 6}{2 + 3 + 4 + 6}$$

$$\bar{x}_w = 530.67Rwf$$

- **Quadratic mean**

The quadratic mean of the $n$ observed values $x_1, x_2, \cdots, x_n$ is given by:

$$Q. = \sqrt{\frac{1}{N}\sum_{i=1}^{n} x_i^{\,2} f_i}$$

This kind of mean is used to evaluate the mean of surface. In other words, the square of the quadratic mean of the $n$ observed values $x_1, x_2, \cdots, x_n$ is equivalent to the arithmetic mean of the squared values $x_{1,}^2\ x_{2,}^2\ x_{3,}^2\ \cdots, x_{n,}^2$

$$Q^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 f_i$$

Example: The quadratic mean of the value 1, 2, 3, 4 is $Q = \sqrt{\frac{1}{4}(1+4+9+16)} = \sqrt{\frac{30}{4}}$

Solve the problems of the activity here below using the theory of weighted mean.

**ACTIVITY**

1. A student gets the following percentage of marks in B.A Examination. English 80%, Kinyarwanda 70%, Accountancy 60%, Costing 70% and statistics 50%. Find out the average marks of this student knowing that time load of each course is 3, 4, 2, 2 and 3 hours per week respectively.
2. An amount of \$6800 must be shared to 5 directors, 30 teachers and 6 secretaries in the company. The respective strength of the positions is 1, 0.8 and 0.3. Find the part in share of each person.

# 3.1.2. The Mode of the Data Variable

**The Mode.**

The mode is a positional average and it is defined as the most frequent observed value in the statistical series. In other words, it is the value repeated maximum times in the series. However, the definition is subject to failures when the statistical series has a large number of observed values.

To adjust the definition, the mode would be defined as the value about which the items are most closely concentrated or it is the value which has the greatest density in its immediate neighborhood.

**Example:**

From the following table, find the mode of size of shoes

| Size of shoes | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|
| No of persons | 10 | 20 | 25 | 40 | 22 | 15 | 6 |

**Solution:**

Here, the mode is 7 or Mo = 7, because that is the value with the higher frequency.

- **The mode of the data in grouped frequency distribution**

When data are presented by a grouped frequency distribution, the mode of the data variable is the value $M_o$ defined by the formula:

$$M_0 = L + \frac{(f_1 - f_0)}{(f_1 - f_0) + (f_1 - f_2)} xc$$

, Where

$L$ :is the lower limit of the modal class interval (The modal class interval is the one that has the highest frequency).

$f_0$ : is the frequency of the class interval that **comes just before** the modal class interval

$f_1$ : is the frequency of the modal class interval

$f_2$ : is the frequency of the class interval that comes **just after** the modal class interval

$c$ : is the class width

**Example:**

Calculate the Mode of the following data:

| Marks | 10 – 20 | 20 -30 | 30 - 40 | 40 – 50 | 50 – 60 | 60 – 70 | 70 – 80 | 80 – 90 |
|---|---|---|---|---|---|---|---|---|
| No of the students | 8 | 10 | 15 | 25 | 20 | 18 | 9 | 5 |

**Solution:**

The calculations for the Mode are carried out from the following table:

| **Number** (i) | **Class** $(a_i - b_i)$ | **Midpoint** $(m_i = \frac{a_i + b_i}{2})$ | **Frequency** $(f_i)$ |
|---|---|---|---|
| *1* | 10 – 20 | 15 | 8 |
| *2* | 20 – 30 | 25 | 10 |
| 3 | 30 – 40 | 35 | $15 \leftarrow f_0$ |
| **4** | **40 – 50** | **45** | $\mathbf{25} \leftarrow f_1$ |
| 5 | 50 – 60 | 55 | $20 \leftarrow f_2$ |
| 6 | 60 – 70 | 65 | 18 |
| 7 | 70 – 80 | 75 | 9 |
| 8 | 80 – 90 | 85 | 5 |
| $\sum_{i=1}^{8}$ | | | 110 |

The higher frequency occurs in the fourth class **40 – 50**with frequency $f_1 = 25$

The frequency of the preceding class $f_0 = 15$

The frequency of the succeeding class $f_2 = 20$

The Lower class limit $L = 40$

The class width $C = 10$

$$M_0 = L + \frac{|f_1 - f_0|}{|f_1 - f_0| + |f_1 - f_2|} \times c$$

$$M_0 = 40 + \frac{|25 - 15|}{|25 - 15| + |25 - 20|} \times 10$$

$$M_0 = 40 + \frac{10}{10 + 5} \times 10$$

$$M_0 = 46.67$$

## ACTIVITY

i. The following table gives the wages of 150 workers in a factory. Find the mode

| Wages in Rwf | Number of workers |
|---|---|
| 0 to 400 | 4 |
| 400 to 800 | 12 |
| 800 to 1200 | 40 |
| 1200 to 1600 | 41 |
| 1600 to 2000 | 27 |
| 2000 to 2400 | 13 |
| 2400 to 2800 | 9 |
| 2800 to 3200 | 4 |

ii. Calculate the mode form the following details:

| Wages in Rwf | Number of Workers |
|---|---|
| Below 10 | 4 |
| Below 20 | 6 |
| Below 30 | 24 |
| Below 40 | 46 |
| Below 50 | 68 |
| Below 60 | 85 |
| Below 70 | 95 |
| Below 80 | 98 |
| Below 90 | 100 |

iv. From the following data of the wages of 122 workers, determine **the modal wages**, **the median** and **the mean**.

| Wages in Rwf | No of workers |
|---|---|
| 100 – 110 | 4 |
| 110 – 120 | 6 |
| 120 – 130 | 20 |
| 130 – 140 | 32 |
| 140 – 150 | 33 |
| 150 – 160 | 17 |
| 160 – 170 | 8 |
| 170 – 180 | 2 |

# 3.1.3. The Percentiles and Quartiles.

Given a set of numerical observations, we may transform it into an array of data (order the data in ascending order). In statistics, it is very important to understand the role of percentiles. The percentiles are positional values. They are describing the number in percent of the data less or equal to the value in a specific position within the set of the whole values.

**Example:**
If the grade of the student is in the 90th percentile, this does mean that 90% of his or her classmates have got grades less than or equal to his (or her) grades.

**Definition**: The $p^{th}$ percentiles of a group of observed values is that value below which lie p% of the numbers in the group. The position of the $p^{th}$ percentiles, $X_{p\%}$ is given by the formula (N+1)p%, where N is the number of data points arranged or ordered in ascending order of their magnitudes.

**Example:**

A large department store collects data on sales made by each of its salespeople. The data, number of sales made on a given day by each of 20 salespeople, are as follows:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 9 | 6 | 12 | 10 | 13 | 15 | 16 | 14 | 14 | 16 | 17 | 16 |
| 17 | 16 | 21 | 18 | 19 | 18 | 20 | 17 | | | | |

Find the $40^{th}$ percentiles.

**Solution**

The 40 percentiles, $X_{40\%}$ is the value occupies the position $(N+1) \times 40\%$. That is the position $(20+1) \times 40\% = 8.4$ . As the position is a positive integer, $X_{40\%}$ is the arithmetic mean of the values in the position 8 and 9 when these twenty sales made are arranged in ascending order of their magnitudes.

6 9 10 12 13 14 14 15 16 16 16 16

17 17 17 18 18 19 20 21

Values in the positions 8 and 9 are 15 and 16 respectively. This means that the 40 percentiles is

$$X_{40\%} = \frac{15+16}{2}$$

$$X_{40\%} = 15.5$$

**Remark:**

[1] The same solution can be read through the frequency distribution by looking for the position in the column of the cumulative frequency i.e. the values at which the cumulative frequency of 8 and 9 fall. (See the previous example)

| Counter (i) | Value ($x_i$) | Frequency ($f_i$) | Cumulative frequency($cf_i$) |
|---|---|---|---|
| 1 | 6 | 1 | 1 |
| 2 | 9 | 1 | 2 |
| 3 | 10 | 1 | 3 |
| 4 | 12 | 1 | 4 |
| 5 | 13 | 1 | 5 |
| 6 | 14 | 2 | 7 |
| 7 | 15 | 1 | 8← |
| 8 | 16 | 4 | 12← |
| 9 | 17 | 3 | 15 |
| 10 | 18 | 2 | 17 |
| 11 | 19 | 1 | 18 |
| 12 | 20 | 1 | 19 |
| 13 | 21 | 1 | 20 |
| $\sum_{i=1}^{15}$ | | 20 | |

The values in the position 8 and 9 are 15 and the first 16 because there are 4 values equal to 16.

$$X_{40\%} = \frac{15 + 16}{2}$$

$$X_{40\%} = 15.5$$

[2] When data are grouped in class intervals, the $p^{th}$ percentiles is defined by the following formula:

$$X_{p\%} = L + \frac{N \times p\% - cf}{f} \times C$$

Where L: Lower limit of the class interval in which the $p^{th}$ percentiles is contained.

$cf$: The cumulative frequency of the class preceding the class interval in which the $p^{th}$ percentiles is contained

$f$: The frequency of the class interval in which the $p^{th}$ percentiles is contained

**Note:**

The following are important $p^{th}$ percentiles:

- First quartile $Q_1$ is the $25^{th}$ percentiles
- Second quartile $Q_2$ is the $50^{th}$ percentiles. It is also the median (Me = $Q_2$)
- Third quartile $Q_3$ is the $75^{th}$ percentiles.

**Example:**

From the following data of the wages of 122 workers, the median, the first and third quartiles as well as $80^{th}$ -percentiles.

| Wages in Rwf | No of workers |
|---|---|
| 100 – 110 | 4 |
| 110 – 120 | 6 |
| 120 – 130 | 20 |
| 130 – 140 | 32 |
| 140 – 150 | 33 |
| 150 – 160 | 17 |
| 160 – 170 | 8 |
| 170 – 180 | 2 |

**Solution:**

To solve this problem use the following table

| **Number** (i) | **Class** $(a_i - b_i)$ | **Mid-point** $(m_i)$ | **Frequency** $(f_i)$ | $(cf)$ | $A = 145$ $D_i = \frac{m_i - A}{10}$ | $D_i f_i$ |
|---|---|---|---|---|---|---|
| *1* | 100 – 110 | 105 | 4 | 4 | -4 | -16 |
| 2 | 110 – 120 | 115 | 6 | 10 | -3 | -18 |
| 3 | 120 – 130 | 125 | 20 | 30 | -2 | -40 |
| **4** | 130 – 140 | 135 | 32 | 62 | -1 | -32 |
| 5 | 140 – 150 | 145 | 33 | 95 | 0 | 0 |
| 6 | 150 – 160 | 155 | 17 | 112 | 1 | 17 |
| 7 | 160 – 170 | 165 | 8 | 120 | 2 | 16 |
| 8 | 170 – 180 | 175 | 2 | 122 | 3 | 6 |
| $\sum_{i=1}^{8}$ | | | 122 | | | -67 |

The Median Me.

The median Me is the 50-percentiles or $X_{50\%}$ and it is contained in the class of cumulative frequency $(N+1) \times 50\% = 123 \times 0.5 = 61.5$. It is the class interval 130 – 140

The median is given by the following formula:

$$X_{p\%} = M_e = L + \frac{N \times p\% - cf}{f} \times C$$

$$X_{p\%} = M_e = 130 + \frac{122 \times 50\% - 30}{32} \times 10$$

$$M_e = 139.69$$

The First quartile $Q_1$.

The first quartile $Q_1$ is the 25-percentiles or $X_{25\%}$ and it is contained in the class of cumulative frequency $(N+1) \times 25\% = 123 \times 0.25 = 30.75$. It is the class interval 130 – 140

The first quartile is given by the following formula:

$$X_{p\%} = Q_1 = L + \frac{N \times p\% - cf}{f} \times C$$

$$X_{p\%} = Q_1 = 130 + \frac{122 \times 25\% - 30}{32} \times 10$$

$$Q_1 = 130.16$$

The Third quartile $Q_3$.

The Third quartile $Q_3$ is the 75-percentiles or $X_{75\%}$ and it is contained in the class of cumulative frequency $(N+1) \times 75\% = 123 \times 0.75 = 92.25$. It is the class interval 140 – 150

The Third quartile is given by the following formula:

$$X_{p\%} = Q_3 = L + \frac{N \times p\% - cf}{f} \times C$$

$$X_{p\%} = Q_3 = 140 + \frac{122 \times 75\% - 62}{33} \times 10$$

$$Q_3 = 148.94$$

# DESCRIPTIVE STATISTICS

Dr. Hategekimana Pascal

Adventist University of Central Africa (AUCA)

April 17, 2025

# 4. Measures of Dispersion

## 4.1. Definition

Let $x_1, x_2, \ldots, x_N$ be a data set of $N$ values of the variable $X$. The measures of dispersion describe the extent to which these $N$ values deviate (or vary) from one of the averages of this data set; the arithmetic mean $\bar{x}$, the median $M_e$ or the mode $M_o$.

The measures of dispersion of the $N$ data set are:

- Range, $R$;
- Inter-quantile range, $IQR$;
- Variance; $\sigma^2$ (population) and $s^2$ (sample);
- Standard deviation; $\sigma$ (population) and $s$ (sample);
- Coefficient of Variation; $CV$.

## 4.2. The Range

The range, $R$, of the N values $x_1, x_2, \ldots, x_N$ assigned by the variable $X$ is defined as the difference between the largest and smallest values

$$R = x_l - x_s$$

$x_l$: Largest value
$x_s$: smallest value

**Example:**
Find the range of the following data:
67.2 , 65.0 , 72.5 , 71.1, 69.1, 69.0, 70.2, 68.2, 68.5, 71.3, 67.5, 68.6, 73.1, 71.3, 69.4, 65.5, 69.5, 70.8, 70.0, 69.2
The range $R$ is:

**Answer:**
$R = 73.1 - 65 = 8.1$

## 4.3. The inter-quantile range

The inter-quantile range, $IQR$, of the N values $x_1, x_2, \ldots, x_N$ of the variable $X$, is equal to the difference between the largest and smallest quantiles

$$IQR = Q_3 - Q_1$$

$Q_1$: the first quantile,
$Q_3$: the third quantile.

Find the inter-quantile range of the following 20 values:
67.2 , 65.0 , 72.5 , 71.1, 69.1, 69.0, 70.2, 68.2, 68.5, 71.3, 67.5, 68.6, 73.1, 71.3, 69.4, 65.5, 69.5, 70.8, 70.0, 69.2

**Answer:**

- $Q_1$ is the $(20+1)*0.25 = 5.25^{th}$ value, therefore $Q_1$ is the midpoint of the values in highlighted positions 5 and 6 of the array: 65.0, 65.5, 67.2, 67.5, **68.2**, **68.5**, 68.6, 69.0, 69.1, 69.2, 69.4, 69.5, 70.0, 70.2, 70.8, 71.1, 71.3, 71.3, 72.5, 73.1.

  $Q_1 = \frac{(68.2+68.5)}{2} = 68.35$
- $Q_3$ is the $(20+1)*0.75 = 15.75^{th}$,therefore $Q_3$ is the midpoint of the values in highlighted positions 15 and 16 of the array: 65.0, 65.5, 67.2, 67.5, 68.2, 68.5, 68.6, 69.0, 69.1, 69.2, 69.4, 69.5, 70.0, 70.2, **70.8**, **71.1**, 71.3, 71.3, 72.5, 73.1.

  $Q_3 = \frac{(70.8+71.1)}{2} = 70.95$

The inter-quantile range $IQR = 70.95 - 68.35 = 2.6$

## 4.4. The Variance

The variance, $\sigma^2$ or $s^2$(for the sample), of the N values $x_1, x_2, \ldots, x_N$ of the variable $X$ is the arithmetic mean of the squared deviations $d_i^2$ from the arithmetic mean $\bar{x}$ Calculation of the variance follows from different formulae; depending whether the values belong exclusively to the population or sample.

The variance of the population data

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

or

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}x^2 - (\bar{x})^2$$

(formulae 1)

When data are presented by the frequency distribution, the formula must contain the frequency, $f_i$, the $k$ different distinct values

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i$$

or

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}x^2 f_i - (\bar{x})^2$$

(formulae 2)

The variance of the sample data

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

or

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

(formulae 1)

When data are presented by the frequency distribution, the formula must contain the frequency, $f_i$, the $k$ different distinct values

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

or

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} x^2 f_i - n(\bar{x})^2\right]$$

(formulae 2)

# Short-cut formulae

When values are presented by using the frequency distribution, the following are the appropriate formulae to use. We are advising to use them to avoid heavy calculations and their time consuming $\sigma^2$.

Formula for a simple frequency distribution:

$$\sigma^2 = \frac{\sum_{i=1}^{k} d_i^2 f_i}{N} - \left( \frac{\sum_{i=1}^{k} d_i f_i}{N} \right)^2$$

Here, $d_i = m_i - A$, where $A$ is one of the midpoints taken in the central part of the column for the midpoints.

# Cont...

Formula for a grouped frequency distribution:

$$\sigma^2 = \left[ \frac{\sum_{i=1}^{k} D_i^2 f_i}{N} - \left( \frac{\sum_{i=1}^{k} D_i f_i}{N} \right)^2 \right] x\ c^2$$

Remember that

$$D_i = \frac{m_i - A}{c}$$

where $A$ stands for the assumed mean, and $c$, the class width.

## 4.5. Relation between population and sample Variance

Suppose that $\sigma^2$ and $s^2$ denote the population and sample variances respectively of the same set of N values $x_1, x_2, \ldots, x_N$.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 f_i$$

and

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

From the above two formulae, the ratio of $\sigma^2$ and $s$ is:

$$\frac{\sigma^2}{s^2} = \frac{N-1}{N} \Longleftrightarrow s^2 = \frac{N}{N-1}\sigma^2$$

$\sigma^2$ is less than or equal to $s^2$

**Example 1.:** Find the variance, $\sigma^2$, of the following 20 values of the variable $x$:
67.2 , 65.0 , 72.5 , 71.1, 69.1, 69.0, 70.2, 68.2, 68.5, 71.3, 67.5, 68.6, 73.1, 71.3, 69.4, 65.5, 69.5, 70.8, 70.0, 69.2

**Answer:**
Using the formulea 1, we have: $x_i =$

65.0 65.5 67.2 67.5 68.2 68.5 68.6 69.0 69.1 69.2 69.4 69.5
70.0 70.2 70.8 71.1 71.3 71.3 72.5 73.1

$\bar{x} =$
69.35

$(x_i - \bar{x})^2 =$
18.9225 14.8225 4.6225 3.4225 1.3225 0.7225 0.5625 0.1225 0.0625 0.0225 0.0025 0
0.4225 0.7225 2.1025 3.0625 3.8025 3.8025 9.9225 14.0625

$$\sigma^2 = \frac{1}{20}\sum_{i=1}^{20}(x_i - \bar{x})^2 = 4.1265$$

**Example 2.**:
Find the variance of the data variable $x$ represented by the following frequency distribution:

| $x_i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $f_i$ | 1 | 2 | 2 | 3 | 1 | 1 | 2 |

**Solution**:
we have to use (formulae 2) especially

$$\sigma^2 = \frac{1}{12}\sum_{i=1}^{7} x^2 f_i - (\bar{x})^2$$

because, the values $x_i$ are decimal.

The problem must be solved by adding an additional column for $x_i^2 f_i$ to the extended frequency distribution table. This looks like:

| i | $x_i$ | $f_i$ | $cf_i$ | $x_i f_i$ | $x_i^2 f_i$ | $(x_i - 4)^2 f_i$ |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 9 |
| 2 | 2 | 2 | 3 | 4 | 8 | 8 |
| 3 | 3 | 2 | 5 | 6 | 18 | 2 |
| 4 | 4 | 3 | 8 | 12 | 48 | 0 |
| 5 | 5 | 1 | 9 | 5 | 25 | 1 |
| 6 | 6 | 1 | 10 | 6 | 36 | 4 |
| 7 | 7 | 2 | 12 | 14 | 98 | 18 |
| $\sum_{i=1}^{7}$ | | 12 | | 48 | 234 | 42 |

The mean $\bar{x}$ is given by

$$\bar{x} = \frac{1}{12}\sum_{i=1}^{7} xf_i = \frac{48}{12} = 4$$

The variance,

$$\sigma^2 = \frac{1}{12}\sum_{i=1}^{7} x^2 f_i - (\bar{x})^2 = \frac{234}{12} - 4^2 = \frac{7}{2} = 3.5$$

or

$$\sigma^2 = \frac{1}{12}\sum_{i=1}^{7} (x_i - \bar{x})^2 f_i = \frac{42}{12} = 3.5$$

If we consider the data of the example 2 as data of the sample,

$$s^2 = \frac{12}{12-1} * 3.5$$

$$s^2 = 3.8181$$

## 4.6. The standard deviation and Coefficient of Variation

The standard deviation is equal to the square root of the variance $\sqrt{\sigma^2} = \sigma$

The coefficient of variation, $CV$, expresses the degree of homogeneity of values and it is equal to the percentage of the ratio of standard deviation and the arithmetic mean

$$CV = (\frac{\sigma}{\bar{x}}) * 100\%$$

**Example:**
The standard deviation and coefficient of variation of the two previous examples 1 and 2 are:
$\sigma = \sqrt{4.1265}$=2.031379 and $\sigma = \sqrt{3.5} = 1.870829$ for example 1 and 2 respectively.

$CV = \frac{2.031379}{69.35} * 100\% = 2.929169\%$ for the example 1

$CV = \frac{1.870829}{4} * 100\% = 46.77072\%$ for the example 2

## Scales of the variability or spread

CV indicates the homogeneity or the non-homogeneity of the values according to the following two scales:

- if $CV < 15\%$ values are said to be homogeneous (small spread)
- if $CV \geq 15\%$ values are said to be non-homogeneous or heterogeneous (high spread)

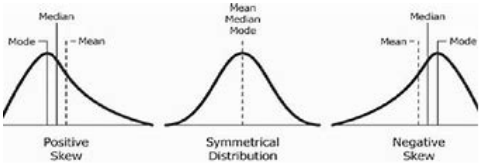
Based on the two scales of variability of data, you can realize that the data of example 2 are more spread that the values of the data of example 1. We say that the data of example 1 are more homogeneous than the data of example 2 (here they are even not homogeneous. They are rather heterogeneous $CV \geq 15\%$)

# 5. The shape of the Distribution

**Definition:**

The shape of the distribution is considered as the form described by the chart or the function curve that relates observation values with their frequencies ( or relative frequencies). In statistics, we distinguish 3 elementary shapes of the distribution:

- **Normal distribution**
- **Left (negative) skewed distribution**
- **Right (positive) skewed distribution**



The pattern of the data variable's shape can be predicted by comparing the mean , the median and the mode. For this reason, the distribution is

- normally distributed as long as $\bar{x} = M_e = M_0$
- left or negatively skewed distributed as long as $\bar{x} < M_e < M_o$
- right or positively skewed distributed when $M_0 < M_e < \bar{x}$

We can know the shape of the distribution from:

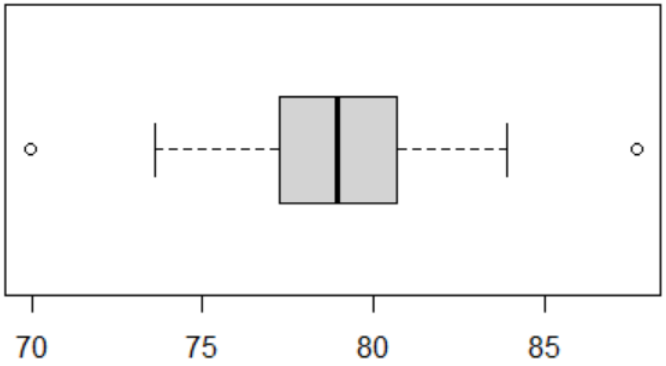
- The shape of the histogram chart;
- The shape of the stem-and-leaf chart;
- The shape of the polygon frequency;
- The shape of the box-whisker-plot or box-plot;
- the shape of the qq-normal plot;
- the sign of the skewness coefficient.

# 5.1. The Box-and-Whisker Plot (Boxplot)

A useful visual device for communicating the information contained in a data set is the box-and-whisker plot. It is generated as follows:

1. Find the quartiles $Q_1, Q_2, Q_3$ of the given data;
2. Represent values of variable of interest, say, $X$ on the horizontal axis;
3. Represent the lower limit of the inner fence $Q1 - 1.5 * IQR$ and largest limit of the inner $Q3 + 1.5 * IQR$;
4. Represents by $x_s$, the **smallest value** within the inner fence, and ,$x_l$, the **largest value** within the same inner fence;
5. Draw a box above the horizontal axis where the left-hand side is at $Q1$ and right-hand side at $Q_3$. At $Q2$, draw inside the box a vertical line segment that joins the two sides.
6. draw the left horizontal whisker by line segment from $x_s$ to the left side of the box plot, and another right horizontal whisker from $x_l$ to the right side of the box.

Note that all values less than $Q1 - 1.5 * IQR$ or greater than $Q3 + 1.5 * IQR$ are called the **outliers** and are extreme values considered as if they would not be part of the given set of values.
Outliers are indicated by either stars or dot-point above the horizontal axis of values of the variable, say $X$



**Example:**

Represent the following data by a boxplot:

87.7, 80.01, 77.28, 78.76, 81.52, 74.2, 80.71, 79.5, 77.87, 81.94, 80.7, 82.32, 75.78, 80.19, 83.91, 79.4, 77.52, 77.62, 81.4, 74.89, 82.95, 73.59, 77.92, 77.18 , 79.8 3, 81. 23, 79 .28, 7 8.44, 79.01, 80.47, 76.23, 78.89, 77.14, 69.94, 78.54, 79.7, 82.45, 77.29, 75.52, 77.21, 75.99, 81.94, 80.41, 77.7

**Solution:**

Found the values of $Q_1, Q_2, Q_3$ in their respective positions $(N+1) * p\%$ i.e., 11.25, 22.5 and 33.75 on these given data sorted in ascending order.

$Q_1 = 77.245$

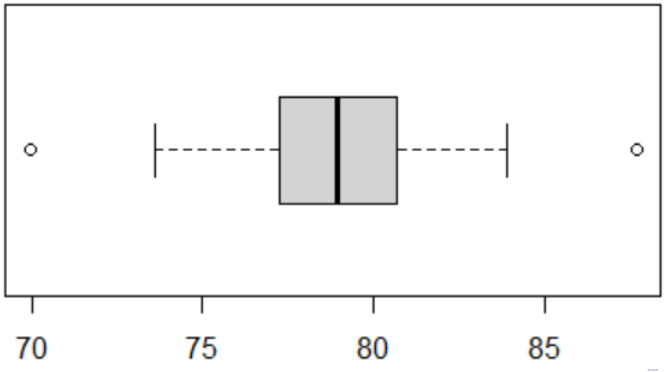
$Q_2 = 78.95$

$Q_3 = 80.705$, $IQR = 80.705 - 77.245 = 3.46$

the lower limit of the inner fence: $Q_1 - 1.5 * IQR = 77.245 - 1.5 * 3.46 = 72.055$, and upper limit of the inner fence: $Q_3 + 1.5 * IQR = 80.705 + 1.5 * 3.46 = 85.895$

Notice that 69.94 is the unique value less than 72.055, the limit of the lower fence, and 87.70 is the unique value greater than 85.895, the upper limit of the inner fence.
We should conclude that the data set has two outliers: 69.94 and 87.70.

Drawing the boxplot making reference to the above information we obtain the following chart (**or use R command "boxplot()"**)



# Activity and Exercise

Consider the following row data of a certain sample drawn from the population:
65 69 70 71 71 74 76 76 76 77 77 77 77 77
78 78 78 78 78 79 80 80 80 81 81 81 81 81
81 81 83 83 83 84 84 84 84 85 86 86 87 88
89 90 91 93 93 94 95 95

a. Represent these data by an appropriate frequency distribution table
b. Compute the sample's smallest quartile and upper quartile;
c. Compute the sample's mean;
d. Compute the sample standard deviation;
e. What proportion of the measurements lie in the interval $\bar{x} \pm 2s$? ($s$= the sample's standard deviation);
f. Find the sample's range and the interquartile range;
g. compute the $90^{th}$ percentile?
h. Construct and Interpret the information of the boxplot.