

# **An Introductory Data Analysis**

**08/12/2019**

**Dapo Olaitan, Biostatistician**

## Lecture One

- Introductory Descriptive Statistics
  - Statistical Concepts (Population, Sample, Parameter etc. )
  - Sampling Distribution
- Data Measurement
  - Measures of Central tendency (Location)
  - Measures of Variation
  - Measures of Association

# What is Statistics?

**Statistics** is the **science** of learning from **data**, and of measuring, controlling, and communicating uncertainty; and it thereby provides the navigation essential for controlling the course of scientific and societal advances (*Davidian, M. and Louis, T. A., 10.1126/science.1218685*).

**Statistics** is also an **ART** ... of conducting a study, analyzing the data, and derive useful conclusions from numerical outcomes about real life problems...

# What is Biostatistics?

**Biostatistics** is the application of statistics in medical research,

e.g.:

- Clinical trials
- Epidemiology
- Pharmacology
- Medical decision making
- Comparative Effectiveness Research etc..

# Terms in Biostatistics

## **Data :**

All the information we collect to answer the research question

## **Variables :**

Outcome, treatment, study population characteristics

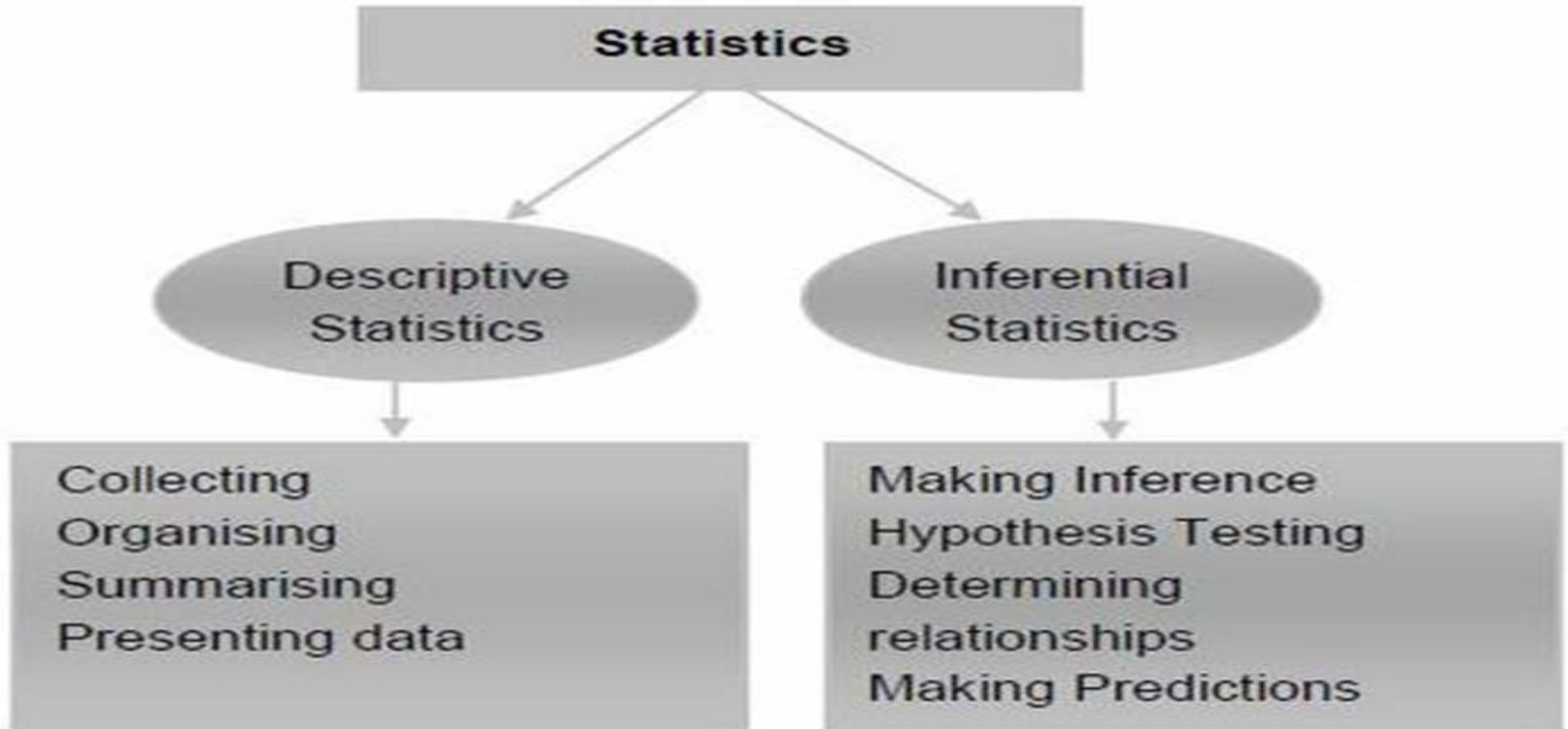
## **Subjects :**

Units on which characteristics are measured

## **Observations :**

Data elements

# Statistical Analysis

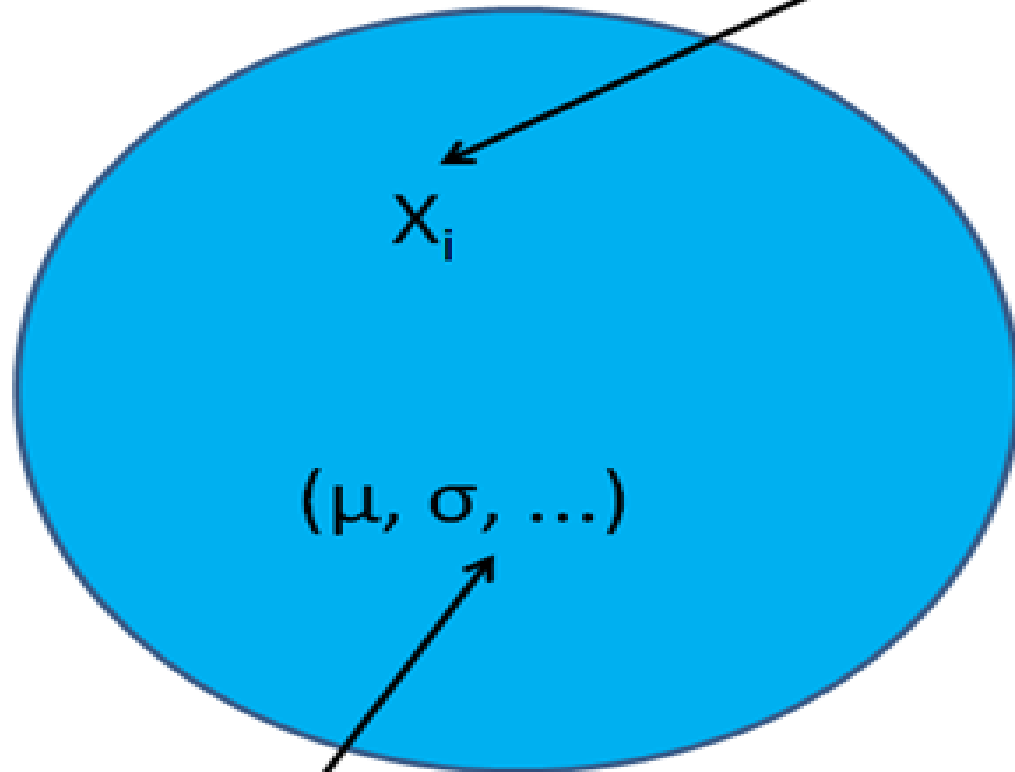


# Statistical Concepts

**Population**

(N)

Units



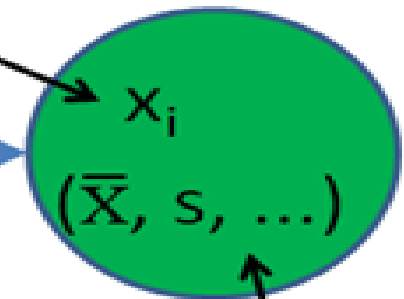
$X_i$

$(\mu, \sigma, \dots)$

random  
selection

**sample**

(n)



$X_i$

$(\bar{X}, s, \dots)$

inference

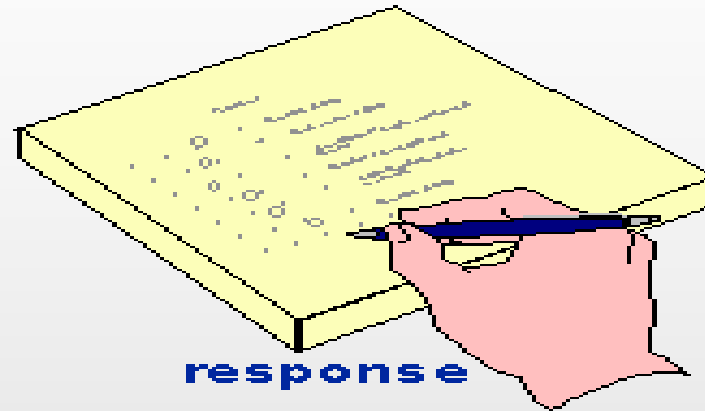
Statistics

Parameters



# Sampling

**Variable**



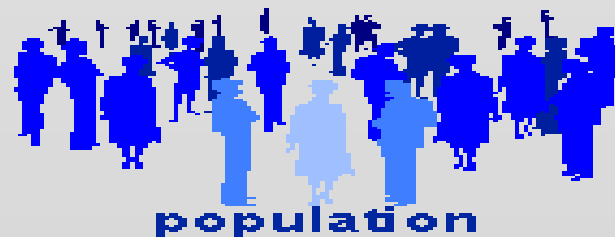
1 2 3 4 5

**Statistic**



**Average = 3.75**

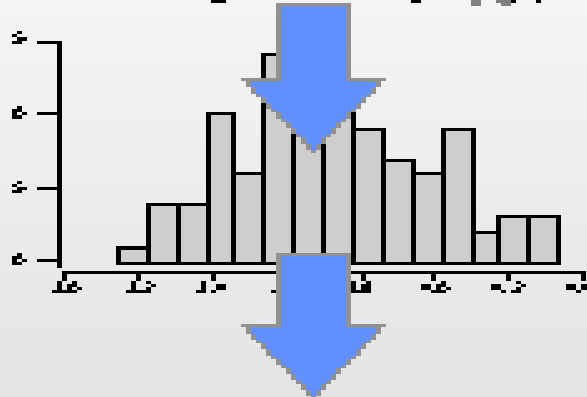
**Parameter**



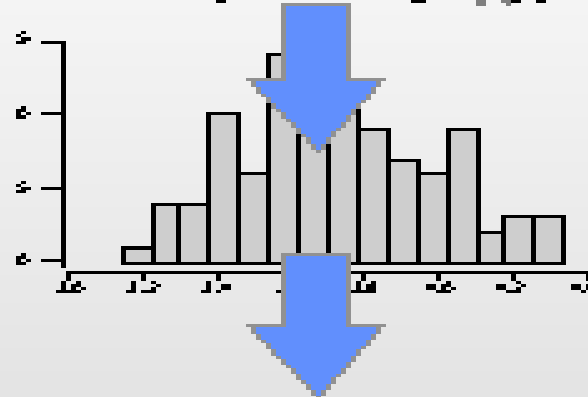
**Average = 3.72**



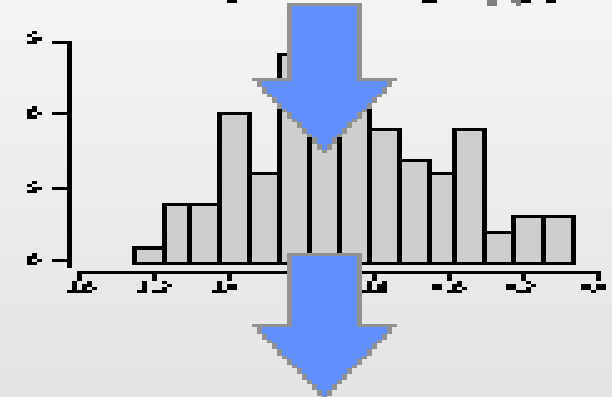
# Sampling Distribution



**Average**

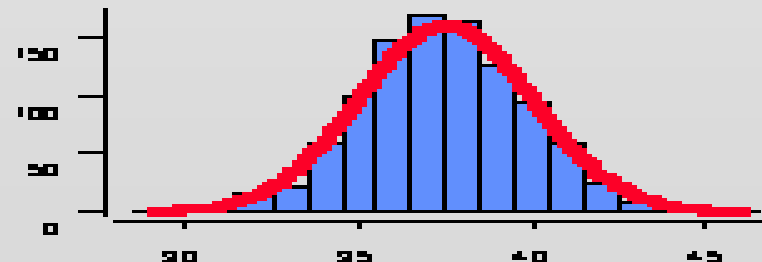


**Average**



**Average**

**The Sampling  
Distribution...**



**...is the distribution  
of a statistic across  
an infinite number  
of samples**

# Sampling Error

## **Sample :**

A sample is just one of a potentially infinite number of samples.

## **Standard Deviation :**

The standard deviation of the sampling distribution shows how different samples would be distributed.

## **Sampling (Standard) Error :**

The sampling error gives some idea of the precision of a statistical estimate.

# Sampling Distribution

## Population :

A group of persons, objects, or items from which samples are taken for measurement.

## Sample :

A measurable portion of a population whose properties are studied in order to gain information about the population.

	Population	Sample	
<b>Descriptive Measure</b>	Parameter	statistic	Summary of a characteristic
<b>Size</b>	N	n	Total number of subjects
<b>Mean</b>	$\mu$	$\bar{x}$	Average
<b>Variance</b>	$\sigma^2$	$s^2$	Variance

# Count

## Frequency (f) :

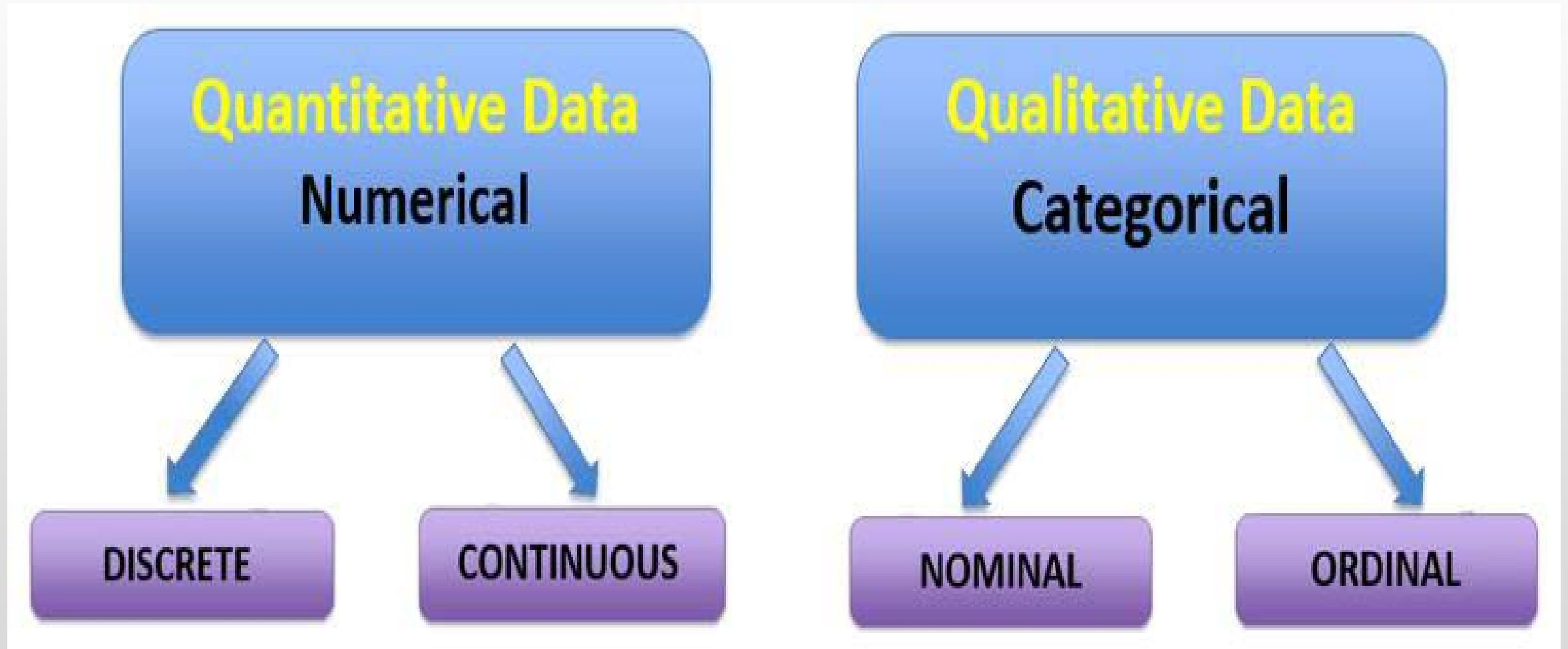
This is the number (#) of subjects in each category.

## Relative frequency (%) :

This is the proportion (%) of subjects in each category.

Period	Frequency (f)	Relative Frequency (%)	Cumulative Relative Frequency (%)
1	450	$(450 \div 1450) \times 100 = 31$	31
2	750	52	83
3	250	17	100
<b>Total</b>	<b>1450</b>	<b>100</b>	

## Types Of Data



# Data Measurement

There are three basic measurement methods primarily used in descriptive statistics.

## Types :

- Measures of Central Tendency or Location
- Measures of Dispersion
- Measures of Association

# Measures of Central Tendency or Location

The Measures of Central Tendency components are powerful tools when summarizing and comparing data.

- Mean or Arithmetic Average
- Median
- Mode

# Mean or Arithmetic Average

$$\text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$

Definition	Formula
<ul style="list-style-type: none"><li data-bbox="193 951 698 1011">• Average value.</li><li data-bbox="193 1036 1462 1096">• A typical value for the variable of interest.</li></ul>	$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$



# Measures of Central Tendency or Location

## Median (The middle item of a data)

Definition	Formula
<ul style="list-style-type: none"><li>The middle value of the variable of interest.</li><li>50% of the collected values are less and 50% are greater than the median.</li></ul>	<ul style="list-style-type: none"><li>If <math>n</math> odd: the <math>\frac{(n+1)^{th}}{2}</math> observation</li><li>If <math>n</math> even: mean of the <math>\frac{n}{2}^{th}</math> and the <math>(\frac{n}{2} + 1)^{th}</math> observations</li></ul> <p>in the ordered sample</p>

# Median

## Example :

If there is an odd number of values in the data set, then the median is the middle value

Data : 6, 9, 1, 2, 6, 5, 1

Arrange from lowest to highest : 1, 1, 2, 5, 6, 6, 9

Median = 5

# Median

## Example :

If there is an even number of values in the data set, then the median is the mean of the two middle values

Data : 6, 9, 1, 2, 6, 1

Arrange from lowest to highest : 1, 1, 2, 6, 6, 9

$$\text{Median} = \frac{2 + 6}{2} = 4$$

# Quartiles

## Definition

- First ( $Q_1$ ): 25% of the collected values are less than  $Q_1$ .
- Second ( $Q_2$ ): 50% of the collected values are less than  $Q_2$  (**median**).
- Third ( $Q_3$ ): 75% of the collected values are less than  $Q_3$ .

# Percentiles

## Definition

- $q_p$ :  $p\%$  of the collected values are less than  $q_p$ .
- E.g.,  $q_1$  is that value of the population (or sample) with 1% of the observed values being less and 99% being greater than it.

# Mode / Min / Max

## Definition

- **Min:** the minimum of the collected values ( $X_{(1)}$ ).
- **Max:** the maximum of the collected values ( $X_{(n)}$ ).
- **Mode:** the most frequent of the collected values.

Unordered	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$
	121	110	114	100	160	130	130
Ordered	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$	$X_{(7)}$
	100	110	114	121	130	130	160

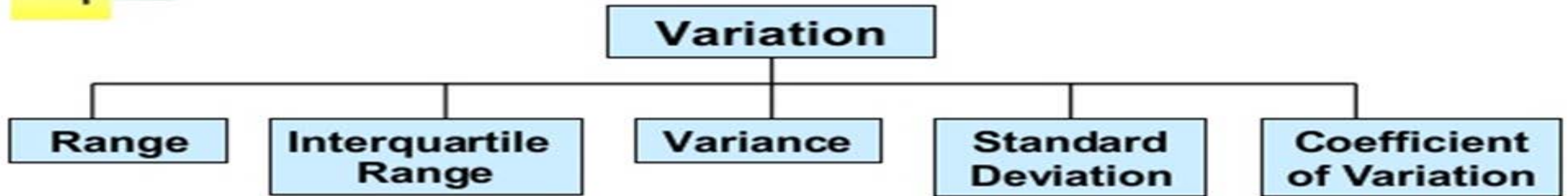
**Min** → 100

**Max** ← 160

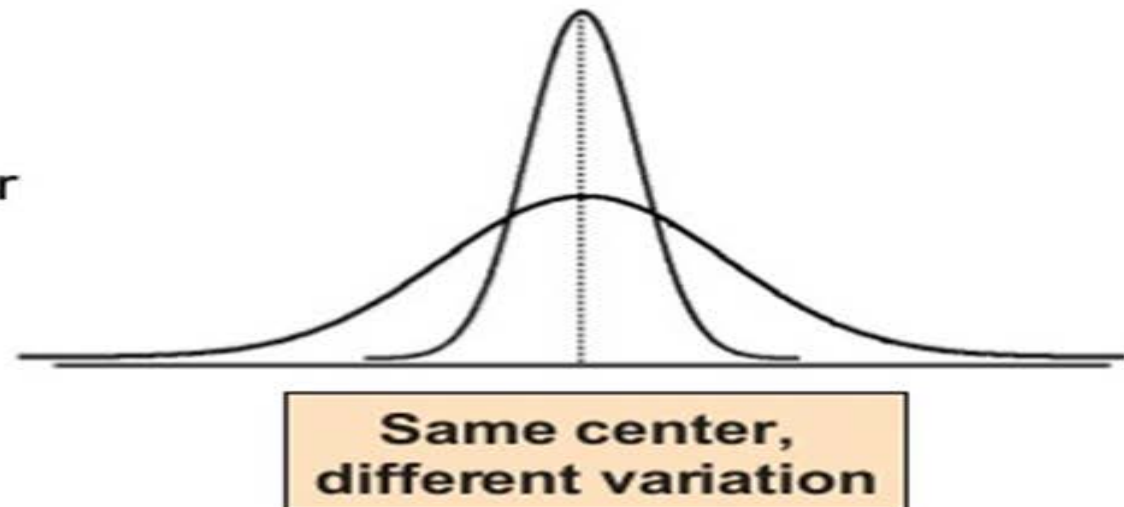
**Mode = 130**

# Measures of Dispersion (Variation)

## Measures of Variation



- Measures of variation give information on the **spread** or **variability** of the data values.



# Measures of Dispersion (Variation)

## Range

largest value minus smallest value

## Interquartile Range (IQR)

difference between 75<sup>th</sup> percentile and  
25<sup>th</sup> percentile values

## Variance ( $s^2$ )

$$\frac{1}{n-1} \sum (x - \bar{x})^2$$

## Standard deviation ( $s$ )

$$\sqrt{\text{variance}}$$



## Variance ( $s^2$ )

Definition	Formula
<ul style="list-style-type: none"> <li>Average squared deviation from the mean.</li> </ul>	$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$

$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$
121	110	114	100	160	130	130

$$\bar{X} = 123.6$$

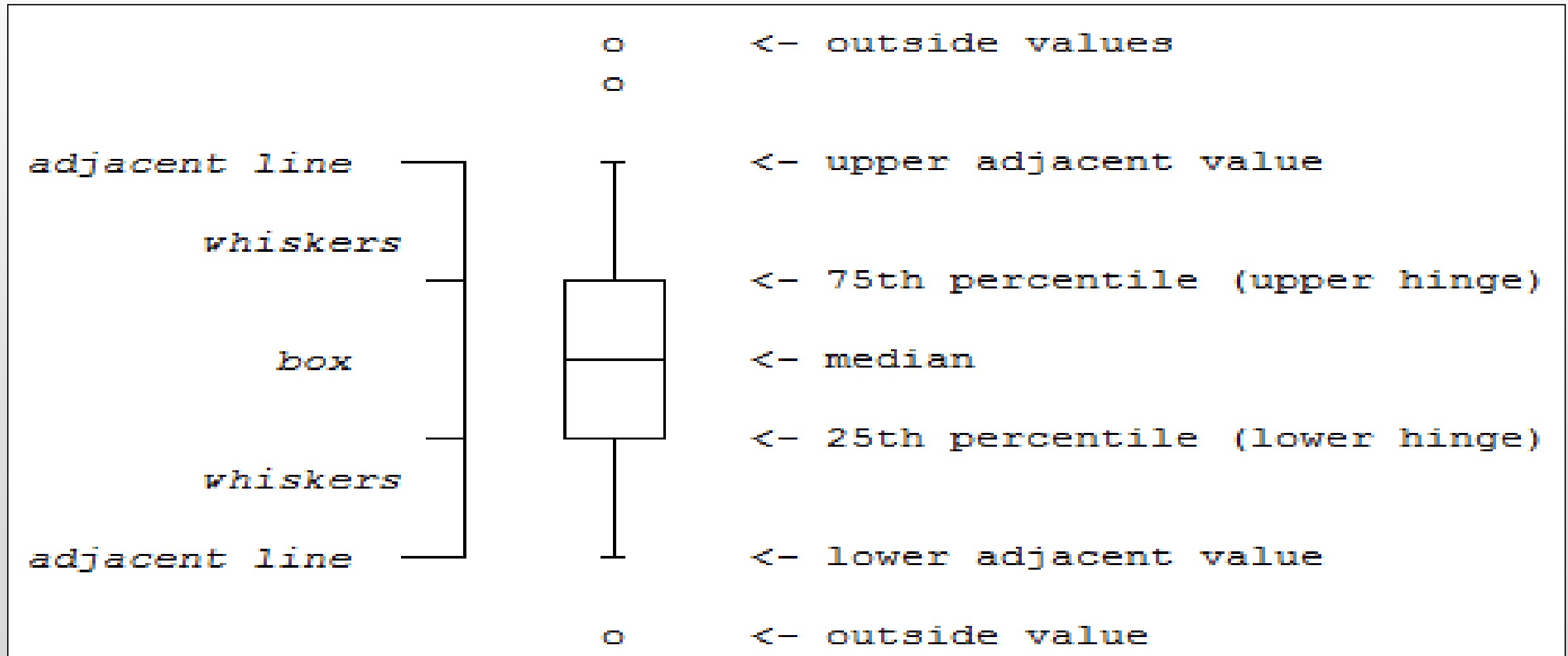
$$\begin{aligned}
 s^2 &= \frac{(X_1 - \bar{X})^2 + \dots + (X_7 - \bar{X})^2}{n-1} = \frac{(121 - 123.6)^2 + \dots + (130 - 123.6)^2}{7-1} \\
 &= \frac{2247.72}{6} = 374.62 \approx 374.6
 \end{aligned}$$

## Other Measures of Dispersion

Definition	Formula
<ul style="list-style-type: none"> <li>Standard deviation</li> </ul>	$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$
<ul style="list-style-type: none"> <li>Mean Absolute Deviation (MAD)</li> </ul>	$\text{MAD} = \frac{\sum_{i=1}^n  X_i - \bar{X} }{n}$
<ul style="list-style-type: none"> <li>Range</li> </ul>	<p>Max – Min</p>
<ul style="list-style-type: none"> <li>Interquartile Range (IQR)</li> </ul>	<p><math>Q_3 - Q_1</math></p>
<ul style="list-style-type: none"> <li>Coefficient of variation</li> </ul>	$\frac{s}{\bar{X}}$

# Graphical Methods for Continuous variables

## Box - Plot (Continuous Data) :



# Outliers

- Observations above  $Q_3 + 1.5IQR$  or  
below  $Q_1 - 1.5IQR$   
are called, “**outliers**”, in the box plot.
- Outliers are not caused by typo or errors.
- Outliers are simply part of data, which can not be ignored.
- Outliers explain how many extreme values are located at tails of a  
distribution.

# Measures of Association

The measures of association helps to determine the outcomes of epidemiological studies, and also explains the strength of association between two categorical research variable indicators.

## Types of Measures

1. Relative Risk (RR) or Risk Ratio,
2. Rate Ratio,
3. Odds Ratio (OR).

## 1. Relative Risk (RR) or Risk Ratio

The relative risk (risk ratio) quantifies a population's risk of disease from a particular exposure.

It is primarily associated with the;

- Cohort study
- Cumulative Incidence (disease frequency by comparing ratio)
- Probabilities comparison in terms of their ratio (**p1 / p2**)
- Strength of association of categorical indicators

**Relative Risk :**

**Formula :**

<b>Disease</b>			
<b>Exposure</b>	<b>Yes</b>	<b>No</b>	<b>TOTAL</b>
<b>Yes</b>	a	b	<b>a + d</b>
<b>No</b>	c	d	<b>c + d</b>
<b>TOTAL</b>	<b>a + c</b>	<b>b + d</b>	<b>n</b>

$$\text{Risk in exposed group} = a / (a + b)$$

$$\text{Risk in un-exposed group} = c / (c + d)$$

$$\text{Relative Risk (RR)} = [a / (a + b)] / [c / (c + d)]$$

## Rules

### Relative Risk (Risk Ratio) :

- **RR = 1.0** (suggests no difference or little difference in risk : incidence in each group is the same).
- **RR > 1.0** (suggests an increased risk of that outcome in the exposed group).
- **RR < 1.0** (suggests a reduced risk in the exposed group).



## Prevalence Ratio (PR) :

- Measures prevalence rather than incidence (measured by risk ratio)
- Usually from a cross-sectional study

<b>Disease</b>			
<b>Exposure</b>	<b>Yes</b>	<b>No</b>	<b>TOTAL</b>
<b>Yes</b>	a	b	<b>a + d</b>
<b>No</b>	c	d	<b>c + d</b>
<b>TOTAL</b>	<b>a + c</b>	<b>b + d</b>	<b>n</b>

$$\text{Prevalence Ratio (PR)} = [a / (a + b)] / [c / (c + d)]$$

## 2. Rate Ratio :

The rate ratio Compares the rates of disease in two groups that differ by demographic characteristics or exposure history.

Formula ;

$$\text{Rate Ratio} = \frac{\text{Rate for group of primary interest}}{\text{Rate for comparison group}}$$

### 3. Odds Ratio (OR) :

The odds ratio helps in the study of rare diseases and the multiple exposures that may be related to an healthcare outcome.

The odds ratio is associated with ;

- the case - control studies,
- the ‘Odds’ , and helps to indicate how **much higher**

the odds of exposure among cases of a disease compared with controls.

## Odds Ratio (OR):

### Formula:

Odds of being exposed among the cases =  $a / c$

Odds of being exposed among the control =  $b / d$

Exposure odds ratio = odds of exposure (cases) / odds of exposure (con)

$$= (a/c) / (b/d)$$

$$= (a*d) / (b*c) \quad (\text{Cross-product ratio})$$

Exposed	Case	Control
Yes	a	b
No	c	d

# Rules

## Odds Ratio (OR):

**OR = 1** (Odds of exposure among cases '**Equals**' Odds of exposure among controls : **Exposure is not associated with the disease**).

**OR > 1** (Odds of exposure among cases '**Is Greater**' than Odds of exposure among controls : **Exposure may be a risk factor for the disease**).

**OR < 1** (Odds of exposure among cases '**Is Lower**' than the Odds of exposure among controls : **Exposure may be protective against the disease**).

## Prevalence Odds Ratio (POR):

- Similar to odds ratio from case control study
- Usually from a cross-sectional study

<b>Disease</b>			
<b>Exposure</b>	<b>Yes</b>	<b>No</b>	<b>TOTAL</b>
<b>Yes</b>	a	b	a + d
<b>No</b>	c	d	c + d
<b>TOTAL</b>	a + c	b + d	n

$$\text{Prevalence Odds Ratio (POR)} = (a*d) / (c*b)$$

# Chi - square Test for Independence

The Chi-squared test is applied when ;

- A single population has two categorical variables.
- We are about to determine whether there is a significant

association between the two categorical variables.

- We are about to determine whether results are statistically

significant.

## Chi Square ( $\chi^2$ ) Test - Assumptions

The following assumptions or fundamentals must be met prior to carrying out a Chi-square test ;

- The expected value of a cell count must be equal to, or greater than 5 in a contingency table.
- The observations within a sample must be independent of one another.
- The sample is a random sample of categorical observations from a large population.



## Chi-square – Formula :

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}.$$

where  $O_{i,j}$  is the observed frequency count at level  $i$  of Variable **A** and level  $j$  of Variable **B**,

$E_{i,j}$  is the expected frequency count at level  $i$  of Variable **A** and level  $j$  of Variable **B**.

## Chi-square – Hypothesis :

Chi-square hypothesis states that ;

The Null  $H_0$  : variable 1 and variable 2 are independent

(Assumes that there is no association between the two variables).

The Alternative  $H_a$  : variable 1 and variable 2 are not independent

(Assumes that there is an association between the two variables).

**Thank you**