

# Introductory Statistics for the Applied Sciences

# Chapter 1

## Introduction and Basic concepts

- **Statistics:** is a body of techniques and procedures dealing with the collection, organization, analysis, interpretation and presentation of information that can be stated numerically so that can be used in testing hypothesis or making decisions.
- Mainland (1963) defines Statistics as the “science and art of dealing with variation in such away as to obtain reliable results” .
- K. Person defines Statistics as the grammar of sciences since it is used in many branches of sciences as in Medicine, Public Health, Sociology, economics, Marketing, Management, Finance, Education, Agriculture,...

# What Do Statistician Do

- **To guide the design of an experiment or survey.**

A statistician ought to be consulted in the early planning stages so that investigations can be carried out efficiently

- **To analyse data.** Data analysis may take many forms, such as examining the relationships among several variables, describing and analysing the variation of certain characteristics, or determining whether a difference in some response is significant
- **To present and interpret results.** Results are best evaluated in terms of probability statements that will facilitate the decision making

# Types of Statistics

## 1. Descriptive statistics:

census in some countries, in which all the residents are requested to provide such information as age, sex, race, and marital status. The data obtained in such a census can then be compiled and arranged into tables and graphs that describe the characteristics of the population at a given time. Concern with enumeration, organization, and graphical representation of data. Descriptive statistics summarizes or describes the characteristics of a data set by numerical measures e.g. of descriptive statistics is the decennial

## 2. Inferential statistics:

A set of methods that use sample results to make a decision about the entire population.

Because a sample is typically only a part of the whole population, sample data provide only limited information about the population.

As a result, sample statistics are generally imperfect representatives of the corresponding population parameters.

Data Collection

Data Presentation

Tabulation  
Diagrams  
Graphs

Descriptive Statistics

Measures of Location  
Measures of Dispersion  
Measures of Skewness &  
Kurtosis

Inferential Statistics

Estimation Hypothesis  
Testing  
Point estimate  
Interval estimate

Univariate analysis

Multivariate analysis

# Why study statistics





- Statistics is important in evaluating, planning, problem solution, researching and developing.
- To discriminate between fact and fancy in everyday life in reading newspapers and watching television, and in making daily comparisons and evaluations.
- A knowledge of statistics is essential for persons who wish to keep their education up to date. To keep abreast of current developments in one's field, it is important to review and understand the writings in scientific journals, many of which use statistical terminology and methodology.
- No matter what your career, you will make professional decisions that involve data. An understanding of statistical methods will help you make these decisions effectively
- To help one know when and for what purposes a statistician should be consulted.

# Sources of Data

## 1. Historical Records

Old census, annual statistical records, textbooks and journals, annual and monthly reports of different companies, commercially available data bank and hospital medical records.

## 2. Survey Methods:

- **Census** :collection of data about every individual in the society.
- **Sample Survey**: collection of data about a certain portion of the population named the sample. It is the most commonly method for collection of data .

## Methods of survey:

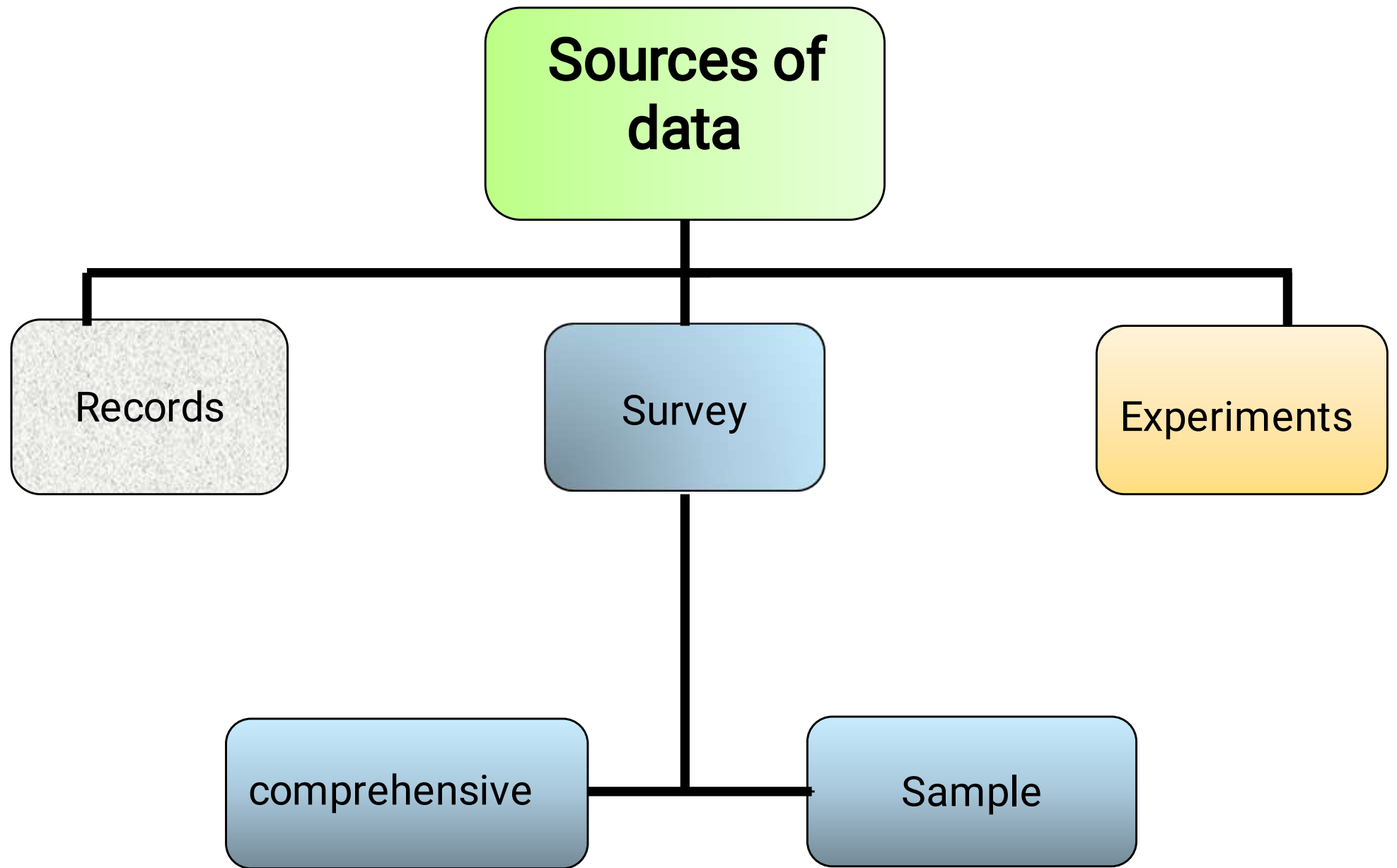
- Personal interview
- Telephone interview
- Questionnaire

## 3. Experiment:

The goal of an experiment is to demonstrate a cause and effect relationship between two variables; that is, to show that changing the value of one variable causes changes to occur in a second variable. e.g. to study the effect of

a new vaccine against the common cold.

•



# Population and Sample

The entire group of all elements (individuals, items, materials, or objects) whose characteristics are being studied is called the **population**.

Usually the populations in which we are interested are so large that a researcher cannot examine the entire group. In this case we select a portion of observations from a population and use it to infer something and to help answer questions about the characteristics of the population.

This portion is called **a sample**

# Sample

- The number of elements in the sample is called the **sample size**.
- The primary objective for selecting a sample from a population is to draw **inferences** about that population.
- The sample should be **representative**

## **Representative sample**

Is the sample that represents the characteristics of the population as closely as possible.

# Why a sample

1. With sampling, we have

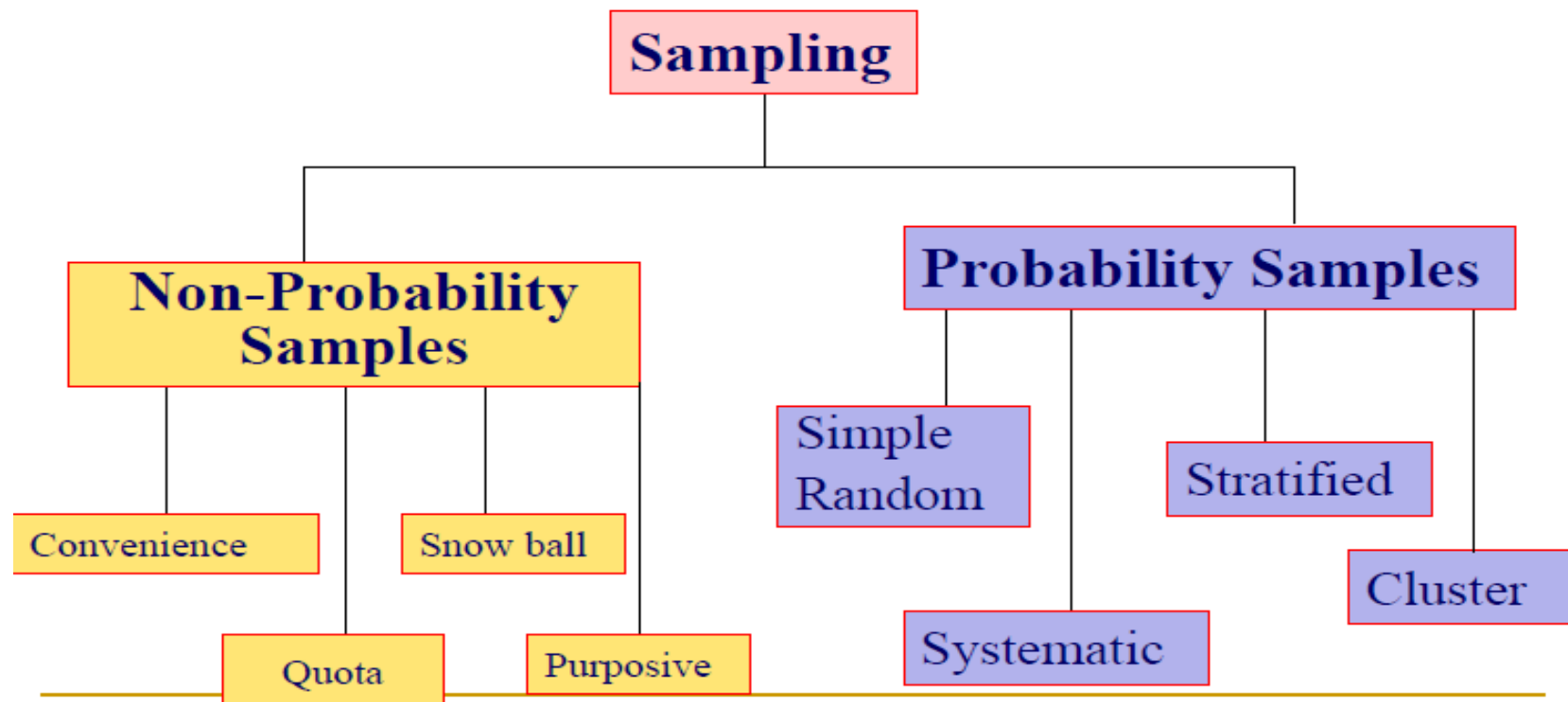
- Less costs
- Less field time
- Less efforts.

2. Sometimes it is impossible or destructive to study the entire population .e.g.

- It is impossible to obtain the weight of every tuna in the Pacific Ocean.
- We can't drain all the blood from a person and count every white cell.

**Sampling**: is a procedure by which some members of a given population are selected as representatives of the entire population

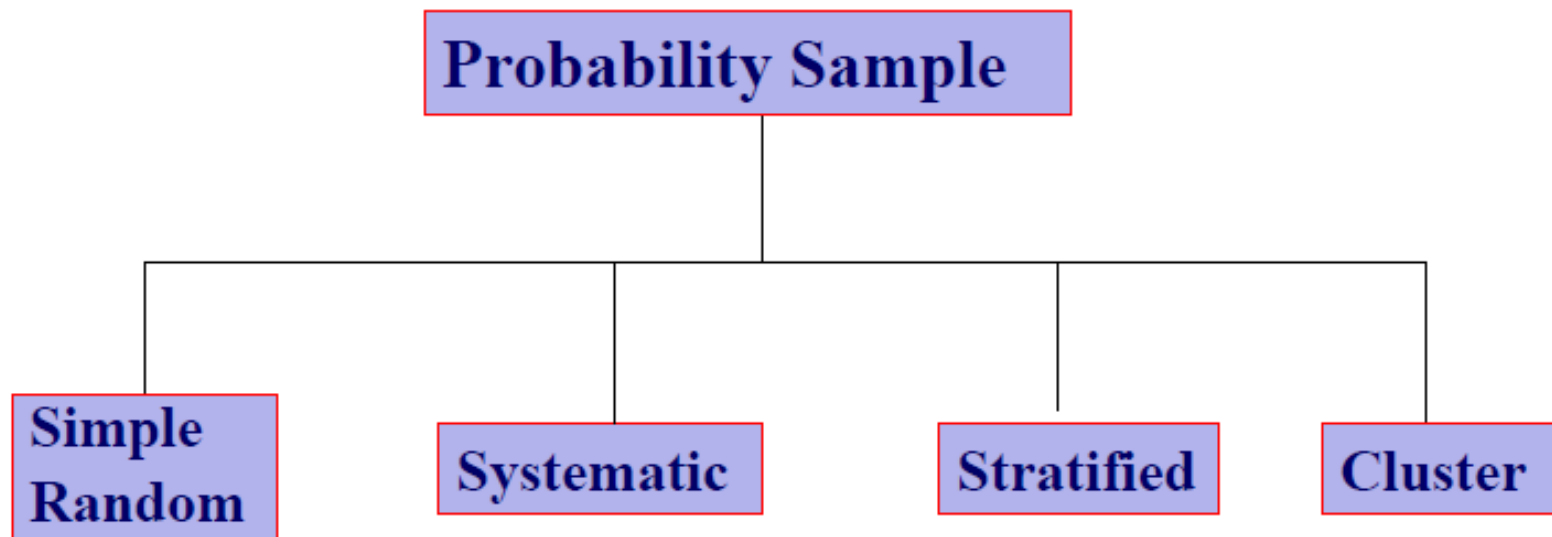
## Types of Sampling Methods





# Probability Sampling

- This is one in which each person in the population has a chance/probability of being selected



# Simple random sample

It is a sample drawn so that every element in the population has an equal chance of being selected. It requires a homogeneous population.

Two ways for selecting simple random sample

- **Lottery method**

Procedure

- Number all units
- Randomly draw units

- **Random number table**

# How to select a Random Sample? Using random number table

Selecting a random sample involves three steps:

1. Define the population.
2. Enumerate it.
3. Use a random number table to select the sample.

# Random number table

13962	70992	65172	28053	02190	83634	66012	70305	66761	88344
43905	46941	72300	11641	43548	30455	07686	31840	03261	89139
00504	48858	38051	59408	16508	82979	92002	63606	41078	86326
61274	57238	47267	35303	29066	02140	60867	39847	50968	96719
43753	21159	16239	50595	62509	61207	86816	29902	23395	72640
83503	51662	21636	68192	84294	38754	84755	34053	94582	29215
36807	71420	35804	44862	23577	79551	42003	58684	09271	68396
19110	55680	18792	41487	16614	83053	00812	16749	45347	88199
82615	86984	93290	87971	60022	35415	20852	02909	99476	45568
05621	26584	36493	63013	68181	57702	49510	75304	38724	15712
06836	37293	55875	71213	83025	46063	74665	12178	10741	58362
84981	60458	16194	92403	80951	80068	47076	23310	74899	87929
66354	88441	96191	04794	14714	64749	43097	83976	83281	72038
49602	94109	36460	62353	00721	68960	82554	90270	12312	56299
78430	72391	96973	70437	97803	78683	04670	70667	58912	21883
33331	51803	15934	75807	46561	80188	73440	29317	27971	16440
62843	84445	56652	91797	45284	25842	40938	73504	21631	81223
19528	15445	77764	33446	41204	70067	16926	70680	66664	75486
16737	01887	50934	43306	75190	86997	24057	79018	34273	25196
99389	06885	45945	62000	76228	60645	66318	46329	46544	95665
36160	38196	77705	28891	12106	56281	29579	66116	39626	06080
05505	45420	44016	79662	92069	27628	07440	32540	19848	27319
85962	19758	92795	00458	71289	05884	97407	23322	73243	98185
28763	04900	54460	22083	89279	43492	82470	40857	86568	49336
42222	40446	82240	79159	44168	38213	43797	26598	29983	67645
43626	40039	51492	36488	70280	24218	53872	04744	89336	35630
97761	43444	95895	24102	07006	71923	17132	32062	41425	66862
49275	44270	52512	03951	21651	53867	36136	70073	45542	22831
15797	75134	39856	73527	78417	36208	11283	76913	22499	68467
04497	24853	43879	07613	26400	17180	93285	66083	02196	10638

# Random Number Table

- Many calculators and computers also generate random numbers
- The digits 0 through 9 occur randomly throughout a random number table with each digit having an equal chance of occurring.
- To use a random number table, first randomly select a **starting position** and then **move in any direction** to select the numbers.

# Simple Random Sampling

**Example:** evaluate the prevalence of tooth decay among the 850 children attending a school

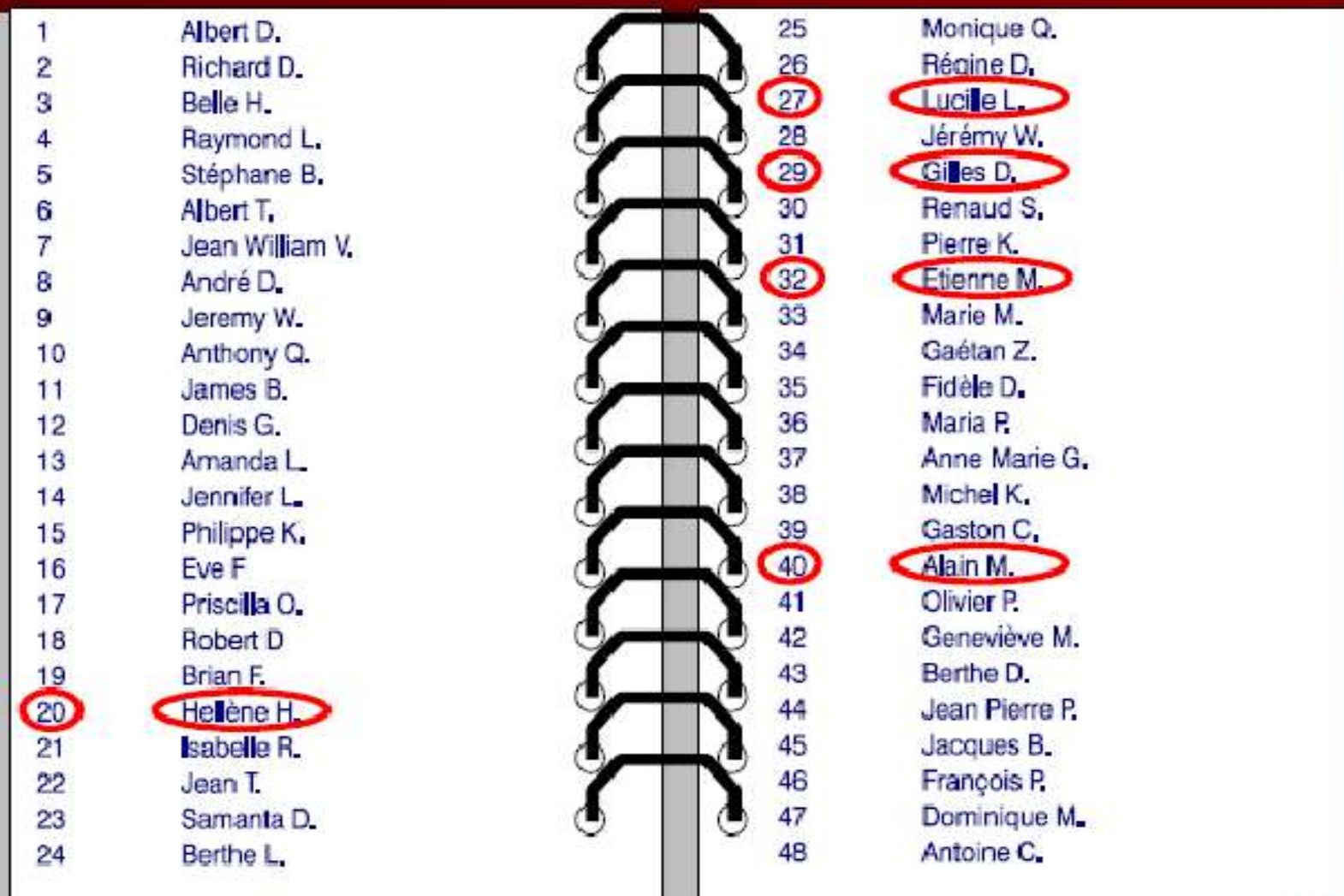
- List of children attending the school
- Children numerated from 1 to 850
- Sample size = 30 children
- Random sampling of 30 numbers between 1 and 850

# Random number table

13962	70992	65172	28053	02190	83634	66012	70305	66761	88344
43905	46941	72300	11641	43548	30455	07686	31840	03261	89139
00504	48658	38051	59408	16508	82979	92002	63606	41078	86326
61274	57238	47267	35303	29066	02140	60867	39847	50968	96719
43753	21159	16239	50595	62509	61207	86816	29902	23395	72640
83503	51662	21636	68192	84294	38754	84755	34053	94582	29215
36807	71420	35804	44862	23577	79551	42003	58684	09271	68396
19110	55680	18792	41487	16614	83053	00812	16749	45347	88199
82615	86984	93290	87971	60022	35415	20852	02909	99476	45568
05621	26584	36493	63013	68181	57702	49510	75304	38724	15712
06936	37293	55875	71213	83025	46063	74665	12178	10741	58362
84981	60458	16194	92403	80951	80068	47076	23310	74899	87929
66354	88441	96191	04794	14714	64749	43097	83976	83281	72038
49602	94109	36460	62353	00721	66980	82554	90270	12312	56299
78430	72391	96973	70437	97803	78683	04670	70667	58912	21883
33331	51803	15934	75807	46561	80188	73440	29317	27971	16440
62843	84445	56652	91797	45284	25842	40938	73504	21631	81223
19528	15445	77764	33446	41204	70067	46926	70680	66664	75486
16737	01887	50934	43306	75190	86997	24857	79018	34273	25196
99389	06685	45945	62000	76228	60645	66318	46329	46544	95665
36160	38196	77705	28891	12106	56281	29579	66116	39626	06080
05505	45420	44016	79662	92069	27628	07440	32540	19848	27319
85962	19758	92795	00458	71289	05884	97407	23322	73243	98185
28763	04900	54460	22083	89279	43492	82470	40857	86568	49336
42222	40446	82240	79159	44168	38213	43797	26598	29983	67645
43626	40039	51492	36488	70280	24218	53872	04744	89336	35630
97761	43444	95895	24102	07006	71923	17132	32062	41425	66862
49275	44270	52512	03951	21651	53867	36136	70073	45542	22831
15797	75134	39856	73527	78417	36208	11283	76913	22499	68467
04497	24853	43879	07613	26400	17188	93285	66083	02196	10638



# Simple random sampling



A diagram of a spiral-bound notebook with a list of names on both sides of the page. The names are numbered from 1 to 48. The numbers 20, 27, 29, 32, and 40 are circled in red. The names corresponding to these circled numbers are Hélène H., Lucille L., Gilles D., Etienne M., and Alain M. respectively.

1	Albert D.	25	Monique Q.
2	Richard D.	26	Régine D.
3	Belle H.	27	Lucille L.
4	Raymond L.	28	Jérémy W.
5	Stéphane B.	29	Gilles D.
6	Albert T.	30	Renaud S.
7	Jean William V.	31	Pierre K.
8	André D.	32	Etienne M.
9	Jeremy W.	33	Marie M.
10	Anthony Q.	34	Gaétan Z.
11	James B.	35	Fidèle D.
12	Denis G.	36	Maria P.
13	Amanda L.	37	Anne Marie G.
14	Jennifer L.	38	Michel K.
15	Philippe K.	39	Gaston C.
16	Eve F.	40	Alain M.
17	Priscilla O.	41	Olivier P.
18	Robert D.	42	Geneviève M.
19	Brian F.	43	Berthe D.
20	Hélène H.	44	Jean Pierre P.
21	Isabelle R.	45	Jacques B.
22	Jean T.	46	François P.
23	Samanta D.	47	Dominique M.
24	Berthe L.	48	Antoine C.

# Systematic random sample

- Systematic sampling is a type of probability sampling method in which sample members from a larger population are selected according to a random starting point but with a fixed, periodic interval. This interval, called the sampling interval, is calculated by dividing the population size by the desired sample size
- We randomly select a first case and then proceed by selecting every  $k^{\text{th}}$  case thereafter.

Where  $k$  is determined by dividing the number of items in the population by the desired sample size

# Example

You want to select a sample of 10 children from the 80 children attending a school.

## Solution

- Divide the 80 (population size) children by 10 (sample size) = 8.
- so every 8th children is sampled.
- Select a number randomly between 1 and 8 first, and we then select every 8th children.
- Suppose we randomly select the number 5 from a random number table.
- Then, the systematic sample consists of children with ID numbers 5, 13, 21, 29, 37, 45, 53, 61, 69, 75, so; each subsequent number is determined by adding 8 to the last ID number.

# Stratified Random sample

- **Strata** are groups or classes inside a population that share a common characteristic.
- Stratified sample is used when the population is **diverse** and we wish the sample to represent the various strata of the population proportionately. A simple random sample is then selected from each strata.

# Procedure for selecting of stratified sampling

- The population is first divided into at least two distinct strata or groups.
- Then a random sample of a certain size is drawn from each stratum.
- The groups or strata are often sampled in proportion to their actual percentage of occurrence in the overall population.
- Combine results of all strata.

## Example (stratified Sample)

Select a sample of size 20 from a population consists of 40 females and 30 males.

$$\frac{40}{70} \times 20 = 11.42, \quad \text{take } n_{\text{female}} = 11$$

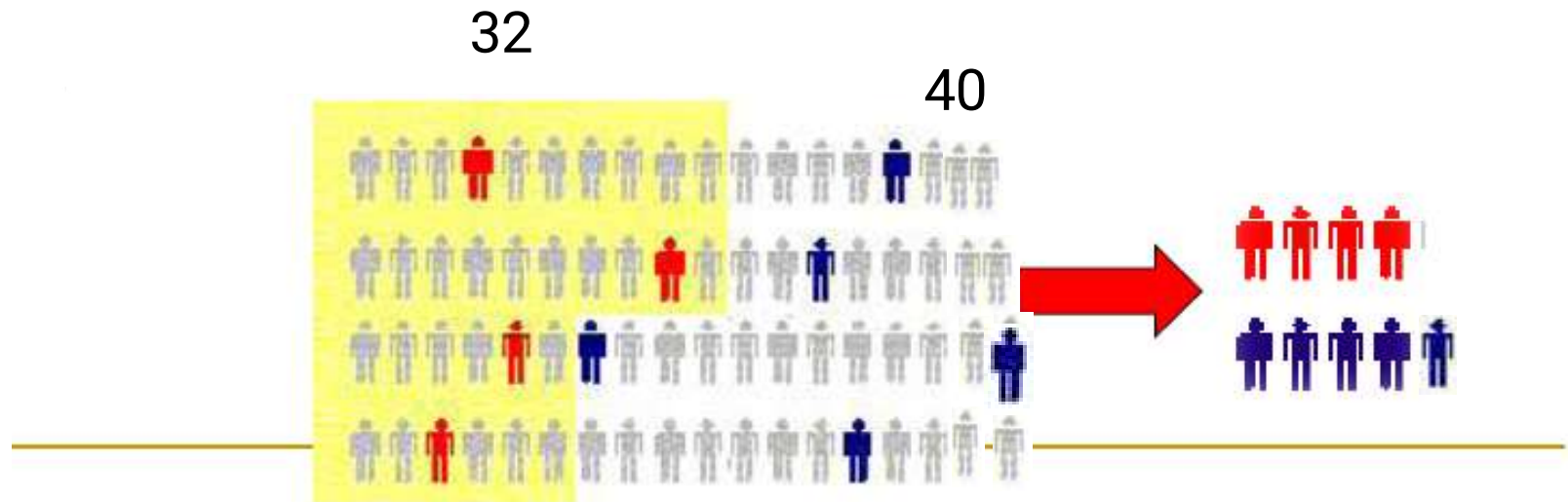
$$\frac{30}{70} \times 20 = 8.5, \quad \text{take } n_{\text{male}} = 9$$

The required sample consists of 11 female and 9 males. We select **a random sample** of size **11** from the group of females and **a random sample** of size **9** from that of males

# Stratified Samples

- Procedure: Divide the population into strata (mutually exclusive classes), such as men and women. Then randomly sample within strata.

**A stratified sample of size 9  
selected from a population of size 72**



# Cluster Sample

- It is a probability sampling technique that is commonly employed to study large populations that are geographically dispersed.
- It may be used when it is too expensive to draw a simple random or stratified random sample. The clusters are sampled randomly and all the members of cluster are sampled.
- It is economical and practical.

## Example:

In conducting a survey of school children in a large city, we could first randomly select 5 schools and then include all the children from each selected school. This technique is more economical than the random selection of persons throughout the city.



# Bias and Sampling error

**Bias** is any trend in the collection, analysis, interpretation, publication or review of data that can lead to conclusions that are systematically different from the truth. ( Las, 2001)

**Or** it is a systematic error in design or conduct of a study.

- inaccurate response (information bias).
- selection bias

**Sampling error** is the discrepancy between a sample statistic and its population parameter .

- Variability
- Sampling method
- Sample size

Defining and measuring sampling error is a large part of inferential statistics

# Factors affecting sample size

- 1. the population size:** The population size is important because the sample size must be sufficiently large that the results can be extrapolated to the population at large
- 2. The margin of error:** (also referred to as the confidence interval) measures the precision with which an estimate from a single sample approximates the population value. The margin of error is closely related to sample size. The less error you're willing to accept, the bigger the sample size needs to be.
- 3. Variability in the population:** The more heterogeneous your population, the bigger the required sample size. An initial estimate of this value is the standard deviation of one or more samples.

**4. The confidence level:** The confidence level is a measure of how certain the results are.

Confidence levels are also closely related to sample size. The greater the degree of confidence that the researcher wants to have in the results, the larger the sample size needs to be. A researcher that chooses a confidence level of 90% will need a smaller sample than a researcher who is required to be 99% confident that the population estimate lies within the margin of error

## 5. Other factors:

- Time and money constraints influence sample size.
- The lower your sampling error must be, the larger your sample must be.
- The more diverse your population is, the larger your sample must be.
- The more complex your analysis, the larger your sample must be.
- The stronger your expected relationships, the smaller your sample can be.

# Statistics and parameters

- **A parameter** is a measure that describes the population.
- **A statistic** is a measure that describes the sample.

The statistics obtained from a sample are used as estimates of the unknown parameters of the population.

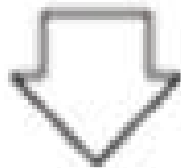
The value of the population parameter is constant but usually unknown.

The value of the statistic varies from sample to sample.

We want to know about these



*Population*



*Parameter*

$\mu$

*(Population mean)*

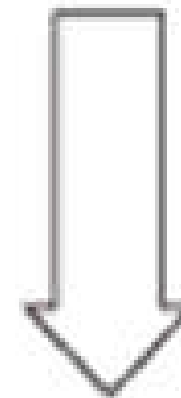
*Random selection*



We have these to work with



*Sample*



*Inference*

$\bar{X}$

*Statistic*

*(Sample mean)*

# Chapter 2

## Organizing and Displaying Data

# Types of numbers

**1. Nominals:** are used as names or identifiers of a person's status, category or attribute. **Gender, smoking status, marital status, telephone number, ID number, and student number** are all nominals

They don't represent an amount or quantity

e.g. marital status

0 = single

1 = married

2 = divorced

3 = widowed

It makes no sense to say divorced > married



**2. Ordinals:** that represent an ordered series of relationships, 1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup>, ... .

cardiovascular, cancer, cerebrovascular, accidents and injuries, chronic lung disease, pneumonia lung disease, Diabetes,...)) The ordinal number indicates the position in the ordered series but say nothing about the magnitude of the difference between any two successive entries. They may be applied for example to - **the ranked order of causes of death by type of disease**

- Educational level (1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup>, 4<sup>th</sup>, 5<sup>th</sup>)
- grade of a student at a university (excellent, very good, good, fair, fail)
- Degree of pain (severe, moderate, mild, none)

Mathematical operations cannot be performed on nominals and ordinals.

Statistics as mean, variance and standard deviation of nominals and ordinals are senseless

### 3. Numbers measured on an interval scale:

Its units can be added and subtracted but cannot be multiplied or divided.

We say the difference between  $100^{\circ}\text{C}$  and  $80^{\circ}\text{C}$  is  $20^{\circ}\text{C}$

But it's not correct to say that  $40^{\circ}\text{C}$  is twice as hot as  $20^{\circ}\text{C}$ .

Ratio of data is senseless since they have no absolute zero.

$0^{\circ}\text{C}$  does not represent that there is no heat but it is the freezing point of pure water.

Common statistics as mean, median and standard deviation can be computed

Day times and IQ "intelligent Quotient" are variables with interval scale.

## 4. Numbers measured on a ratio scale :

- Have the same properties as those measured on interval scale but have an absolute zero, so we can compare meaningfully with one another.

(say 50 kg is twice as 25 kg).

- It include all the usual measurement of **length, height, weight, area, volume, density, velocity, pressure, money and time.**

- Know what type of numbers you have to help you to select quickly the appropriate method of analysis.

# Variables







**A variable** : is a characteristic under study that assumes different values for different elements.  
e.g. sex, age, no. of children in a family , weight, pressure, intelligent quotient, ...

**Data set**: is a collection of observations on a variable or variables.

ID	Age	gender	weight	No. of decayed teeth
1	61	female	70	10
2	52	male	91	4

is an example of a data set.

# Primary Scales of Measurement

<b>Scale</b>				
<b>Nominal</b>	Numbers Assigned to Runners			 <b>Finish</b>
<b>Ordinal</b>	Rank Order of Winners	 <b>Third place</b>	 <b>Second place</b>	 <b>First place</b> <b>Finish</b>
<b>Interval</b>	Performance Rating on a 0 to 10 Scale	<b>8.2</b>	<b>9.1</b>	<b>9.6</b>
<b>Ratio</b>	Time to Finish, in	<b>15.2</b>	<b>14.1</b>	<b>13.4</b>

# Types of variables

## 1. Qualitative (Categorical) variables:

Are variables that yield observations on which individuals can be categorized according to some characteristic or quality

e.g. marital status, colour of eyes, degree of pain, gender, educational level , grade of a student, hair color, ethnic groups and other attributes of the population

## 2. Quantitative variables:

are variables that yield observations that can be measured numerically. They may be classified as

### a. Discrete variable:

must always be integers (0,1,2,...) e.g. **no. of children in a family, no. of times you visit a doctor and no. of missing teeth in a mouth of somebody, the number of bacteria which survive treatment with some antibiotic,..**

### b. Continuous variable

It can assume any numerical value over a certain interval or intervals. it may take on fractional values (e.g. 37.4, 138.9, and 112.1).

**Age, height, weight, time, pressure , IQ, stress score, cholesterol level** are referred to as continuous variables.

# Scales of Measure

- **Nominal** – qualitative classification of equal value: gender, race, color, city
- **Ordinal** - qualitative classification which can be rank ordered: socioeconomic status of families and grade of a student
- **Interval** - Numerical or quantitative data: can be rank ordered and sizes compared : temperature, day time and intelligent quotient .
- **Ratio** - Quantitative interval data along with ratio: time, age.



# Organizing and Displaying Data

**This section aims to:**

- Summarize and present data in different forms.
- Arrange and organize the raw data into an  $n$  array and construct the frequency distribution.
- Define, illustrate, and solve for the class limits, class boundaries and class marks.

Any survey or experiment yields a list of observations. These need to be organized and summarized in a logical fashion so that we may perceive the outcome clearly. **Tables, graphs** and **numerical** methods are popularly used to organize, summarize and describe data.

# Organizing and displaying Qualitative data

We can obtain frequencies of categorical data and summarize them in tables or graphs.

## 1. Frequency table:

Considerable information can be obtained from large masses of statistical data by grouping the data into classes and determining the number of observations that fall in each of the classes. Such an arrangement is called a **frequency distribution or frequency table**.

Frequency table may be the most convenient way of summarizing or displaying both **qualitative** and **quantitative** data.

# Frequency Distribution

**A frequency distribution** is a tabular summary of data showing the frequency (or number) of items in each of several non overlapping classes.

The objective is to provide insights about the data that cannot be quickly obtained by looking only at the original data

# Example

- The following data represents the status of 50 students at a university

F	F	SO	SE	F	F	SE	J	J	J
F	F	J	F	F	F	Se	SO	SE	J
J	F	SE	SO	SO	J	F	F	SE	SE
SO	SE	J	SO	SO	J	J	SO	F	SO
SE	SE	F	SE	J	SO	F	J	SO	SO

In this table, F, SO, J, and SE are the abbreviations for freshman, sophomore, junior, and senior, respectively. Prepare a frequency distribution.

# Frequency table for the status of students

status	Frequency f	Percentage frequency (rf)%
F	15	30
So	12	24
J	12	24
SE	11	22
Total	50	100

These figures, although presented in categories, do not allow for easy analysis. The reader must expend extra effort in order to compare (amounts spent or relate individual proportions to the total). For ease of analysis, these data can be presented pictorially.

# Bar-Chart

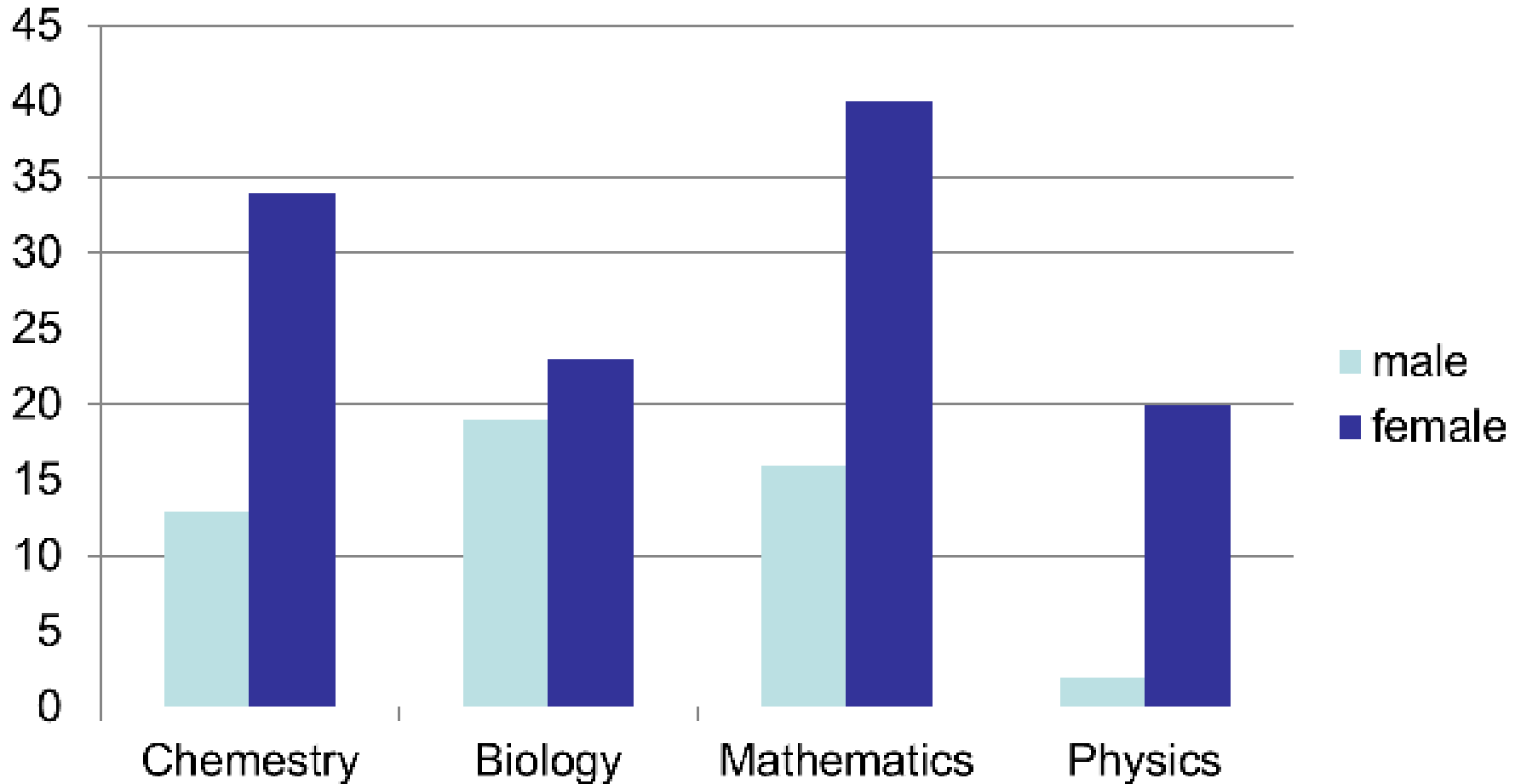
- It is used with categorical or numerical discrete data
- The height of the bar is the frequency or the percentage frequency
- Bars should be separated
- The vertical axis begins with **zero**.

# Bar graph of the status of the students





If we have a nominal categorical variable, divided into two categories, can show data with a grouped bar chart. It allows an easy comparison between groups.



The number of graduate students (College of science) at a university

# Pie chart

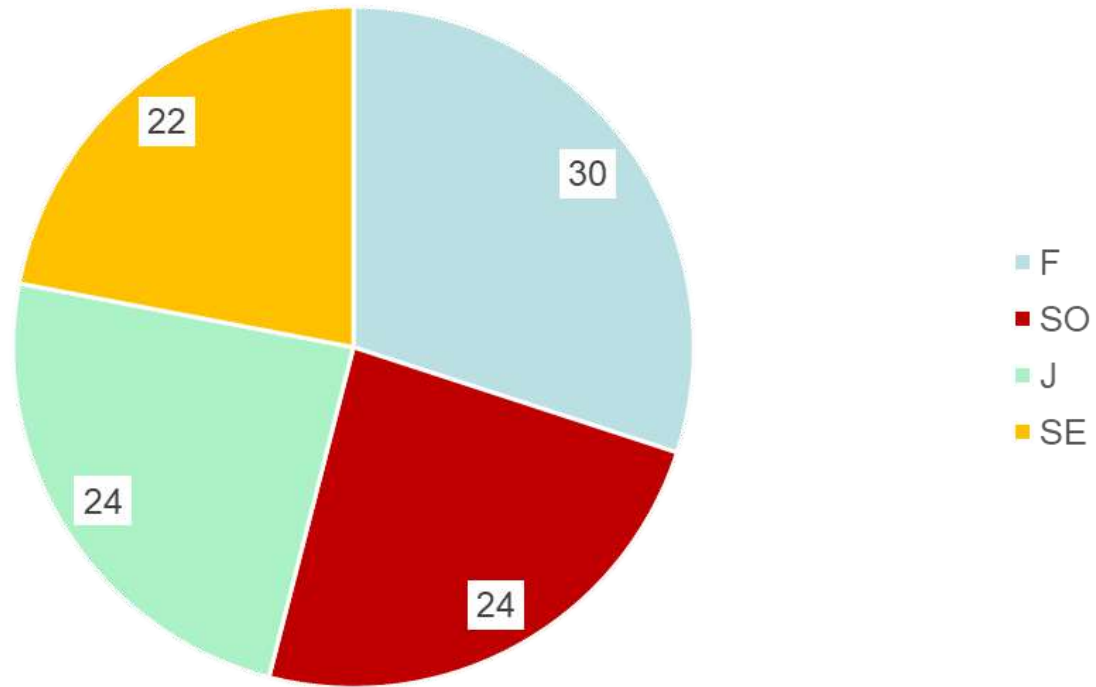
- The pie chart is a commonly used graphical device for presenting relative frequency distributions for qualitative data.
- It represents data in a circle, with “slices” corresponding to percentages of the whole.
- First draw a circle; then use the relative frequencies to subdivide the circle into sectors that correspond to the relative frequency for each class.
- Since there are 360 degrees in a circle, a class with a relative frequency of .30 would consume  $.30(360) = 108$  degrees of the circle.

# Frequency table for the status of students

stat us	Freque ncy f	Percentage frequency (rf)%	Angle of the slice
F	15	30	108 <sup>0</sup>
So	12	24	86.4 <sup>0</sup>
J	12	24	86.4 <sup>0</sup>
SE	11	22	79.2 <sup>0</sup>
Total	50	100	360 <sup>0</sup>

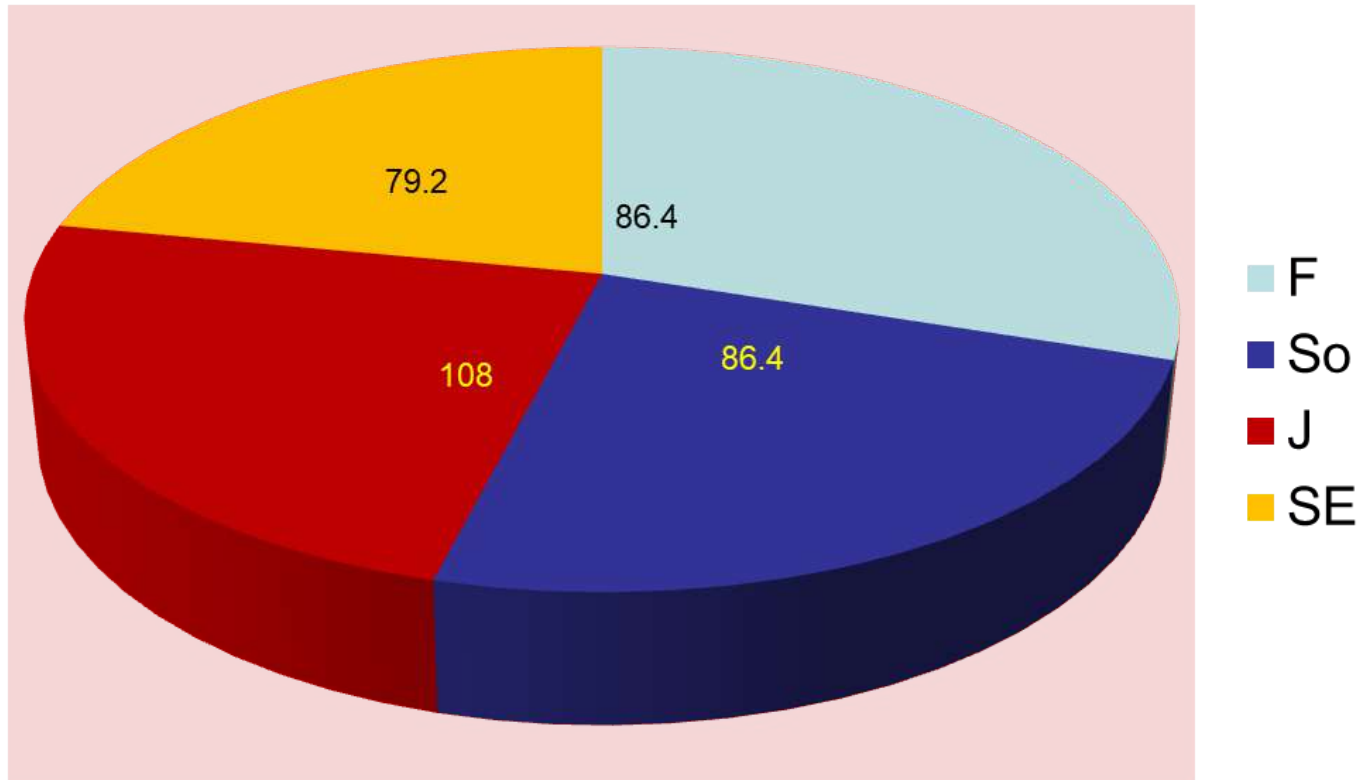
# Pie Chart of the status of the students

percentage frequency



# Pie Chart of the status of the students

**Status**



# Organizing and Graphing Quantitative Continuous Data

## Frequency Table (Distribution):

A grouped frequency distribution is obtained by constructing class intervals for the continuous data, and then listing the corresponding frequency (relative frequency, percentage frequency) of each interval.

table 2.1 represents data set for a sample of 100 individuals of the Honolulu Heart Study population of 7683 persons, 1969

Table 2.1

ID	Sex	E ducation level	W eight (K g)	Height (cm)	Age	Smoking status	Physical Activity	B lood Glucose	Serum Cholesterol	Syst. B lid Pressure	Salary \$	Ponderal Index
1	M	2	70	165	61	1	1	107	199	102	\$27,000	40.0361
2	M	1	60	162	52	0	2	145	267	138	\$18,750	41.3807
3	F	1	62	150	52	1	1	237	272	190	\$12,000	37.8990
4	F	2	66	165	51	1	1	91	166	122	\$13,200	40.8291
5	M	2	70	162	51	0	1	185	239	128	\$21,000	39.3081
6	M	4	59	165	53	0	2	106	189	112	\$13,500	42.3838
7	M	1	47	160	61	0	1	177	238	128	\$18,750	44.3357
8	F	3	66	170	48	1	1	120	223	116	\$9,750	42.0663
9	F	4	56	155	54	0	2	116	279	134	\$12,750	40.5137
10	F	2	62	167	48	0	1	105	190	104	\$13,500	42.1942
11	F	4	68	165	49	1	2	109	240	116	\$16,500	40.4248
12	M	1	65	166	48	0	1	186	209	152	\$12,000	41.2861
13	M	1	56	157	55	0	2	257	210	134	\$14,250	41.0365
14	F	2	80	161	49	0	1	218	171	132	\$16,800	37.3648
15	M	3	66	160	50	0	2	164	255	130	\$13,500	39.5918
16	M	4	91	170	52	0	2	158	232	118	\$15,000	37.7951
17	M	3	71	170	48	1	1	117	147	136	\$14,250	41.0547
18	M	5	66	152	59	0	2	130	268	108	\$27,510	37.6122
19	M	1	73	159	59	0	2	132	231	108	\$14,250	38.0443
20	F	4	59	161	52	0	1	138	199	128	\$11,550	41.3563
21	F	1	64	162	52	1	1	131	255	118	\$15,000	40.5000
22	M	3	55	167	52	1	1	88	199	134	\$12,750	43.9132
23	F	2	78	175	50	1	1	161	228	178	\$11,100	40.9581
24	F	2	59	160	54	0	1	145	240	134	\$9,000	41.0994
25	F	3	51	167	48	1	2	128	184	162	\$9,000	45.0325
26	M	3	83	171	55	0	1	231	192	162	\$12,600	39.2016
27	M	2	66	157	49	1	2	78	211	120	\$27,480	38.8495
28	M	4	61	165	51	0	1	113	201	98	\$14,250	41.9154
29	M	2	65	160	53	0	1	134	203	144	\$79,980	39.7938
30	M	3	75	172	49	0	1	104	243	118	\$14,250	40.7857
31	M	4	61	164	49	0	2	122	181	118	\$14,250	41.6614
32	M	1	73	157	53	1	2	442	382	138	\$45,000	37.5657
33	M	2	66	157	52	0	1	237	186	134	\$15,000	38.8495



Continued

ID	Sex	Education level	Weight (Kg)	Height (cm)	Age	Smoking status	Physical Activity	Blood Glucose	Serum Cholesterol	Syst. Blood Pressure	Salary \$	Ponderal Index
34	M	1	73	155	48	0	2	148	198	108	\$39,990	37.0872
35	M	2	61	160	53	0	1	231	165	96	\$30,000	40.6453
36	F	3	68	162	50	0	2	161	219	142	\$11,250	39.6898
37	M	2	52	157	50	0	2	119	196	122	\$13,500	42.0628
38	M	5	73	162	50	0	1	185	239	146	\$15,000	38.7621
39	M	1	52	165	61	1	2	118	259	126	\$15,000	44.2062
40	F	1	56	162	53	1	1	98	162	176	\$9,000	42.3434
41	F	3	67	170	48	1	2	218	178	104	\$11,550	41.8560
42	M	1	61	160	47	0	1	147	246	112	\$16,500	40.6453
43	M	3	52	166	62	1	2	176	176	140	\$14,250	44.4741
44	M	2	61	172	56	1	2	106	157	102	\$14,250	43.6937
45	M	3	62	164	55	1	2	109	179	142	\$13,500	41.4362
46	F	2	56	155	57	1	2	138	231	146	\$12,750	40.5137
47	F	1	55	157	50	0	2	84	183	92	\$16,500	41.2837
48	M	3	66	165	48	1	2	137	213	112	\$14,100	40.8291
49	M	1	59	159	51	0	2	139	230	152	\$16,500	40.8426
50	M	3	53	152	53	1	2	97	134	116	\$23,730	40.4655
51	M	5	71	173	52	0	2	169	181	118	\$15,000	41.7792
52	M	2	57	152	49	0	1	160	234	128	\$15,000	39.4959
53	M	2	73	165	50	1	1	123	161	116	\$26,250	39.4799
54	M	3	75	170	49	0	2	130	289	134	\$13,500	40.3115
55	M	3	80	171	50	1	2	198	186	108	\$15,000	39.6856
56	M	4	49	157	53	0	1	215	298	134	\$13,500	42.9043
57	M	4	65	162	52	0	1	177	211	124	\$15,750	40.2912
58	F	2	82	170	56	0	2	100	189	124	\$13,500	39.1301
59	M	3	55	155	52	0	2	91	164	114	\$14,250	40.7578
60	M	3	61	165	58	0	1	141	219	154	\$15,000	41.9154
61	M	2	50	155	45	1	2	139	287	114	\$9,750	42.0735
62	M	5	58	160	56	0	1	176	179	114	\$21,750	41.3343
63	M	1	55	166	50	1	2	218	216	98	\$26,250	43.6503
64	M	5	59	161	47	0	2	146	224	128	\$21,000	41.3563
65	M	2	68	165	53	1	1	128	212	130	\$14,550	40.4248
66	M	2	60	170	53	1	2	127	230	122	\$30,000	43.4242

ID	Sex	Education level	Weight (K.g)	Height (cm)	Age	Smoking status	Physical Activity	Blood Glucose	Serum Cholesterol	Syst. Blood Pressure	Salary \$	Ponderal Index
67	M	1	77	160	47	1	1	76	231	112	\$21,240	37.6088
68	M	5	60	155	52	0	1	126	185	106	\$21,480	39.5927
69	M	3	70	164	54	0	1	184	180	128	\$25,000	39.7934
70	M	2	70	165	46	0	1	58	205	128	\$20,250	40.0361
71	M	2	77	160	58	1	1	95	219	116	\$34,980	37.6088
72	F	5	86	160	53	0	2	144	286	154	\$18,000	36.2483
73	F	2	67	152	49	1	2	124	261	126	\$10,500	37.4242
74	F	3	77	165	53	1	1	167	221	140	\$19,500	38.7841
75	F	3	75	169	57	0	2	150	194	122	\$11,550	40.0743
76	F	2	70	165	52	0	2	156	248	154	\$11,550	40.0361
77	F	2	70	165	49	1	1	193	216	140	\$11,400	40.0361
78	F	1	71	157	53	0	1	194	195	120	\$10,500	37.9152
79	F	1	55	162	49	0	2	73	217	140	\$14,550	42.5985
80	F	2	59	165	53	1	2	98	186	114	\$18,000	42.3838
81	F	3	64	159	50	0	2	127	218	122	\$10,950	39.7500
82	F	1	66	160	54	0	1	153	173	94	\$14,250	39.5918
83	F	4	59	165	60	0	2	161	221	122	\$11,250	42.3838
84	F	3	68	165	57	0	1	194	206	172	\$10,950	40.4248
85	M	5	58	160	52	0	1	87	215	100	\$17,100	41.3343
86	M	1	57	154	65	1	1	188	176	150	\$15,750	40.0156
87	M	2	60	160	65	0	2	149	240	154	\$14,100	40.8698
88	M	2	53	162	62	0	1	215	234	170	\$28,740	43.1277
89	M	2	61	159	62	1	2	163	190	140	\$27,480	40.3912
90	F	1	66	154	62	0	1	111	204	144	\$9,750	38.1071
91	F	1	61	152	67	0	2	198	256	156	\$11,250	38.6130
92	F	2	52	152	66	0	2	265	296	132	\$10,950	40.7233
93	F	1	59	155	62	0	2	143	223	140	\$10,950	39.8151
94	F	1	63	155	62	1	1	136	225	150	\$10,050	38.9540
95	F	2	61	165	63	0	2	298	217	130	\$10,500	41.9154
96	M	2	68	155	67	0	2	173	251	118	\$15,000	37.9748
97	M	1	58	170	62	0	1	148	187	162	\$19,500	43.9177
98	M	3	68	160	55	0	1	110	290	128	\$15,000	66 39.1998
99	F	5	60	159	50	0	2	188	238	130	\$10,950	40.6144
100	M	2	61	160	54	1	1	208	218	208	\$27,480	40.6453

# How to construct a frequency table?

1. Determine the **range** from the difference between the smallest and largest value in the set of observations i.e.

$$\text{Range} = \text{Max. Value} - \text{Min. Value}.$$

2. Divide the range into a number of equal segments called **class intervals**.

## Note:

- The number of intervals in general should range from **5 to 15**.
- With **too many** class intervals, the data are **not summarized** enough for a clear visualization of how they are distributed
- With **too few**, the data are **over summarized** and some of the details of the distribution may be lost
- To find the number of classes,  $r$ , we can use the formula  $r = \sqrt{n}$ , where  $n$  is the number of the observations in the data set or sample size.

# Class interval

The length (width) of the class interval is determined by

$$\text{Class length} = \frac{\text{range}}{\text{number of classes}}$$

This value could be increased or decreased for convenience and clear representation.

This is illustrated by the following example.

## Example:

The following data represents Systolic Blood Pressure in mmHg for 37 smokers taken from table 2.1

102	122	116	116	136	118	134
178	162	120	138	126	176	104
140	102	142	146	112	116	116
108	114	98	130	122	112	116
126	140	140	114	150	140	150
208	190					

Prepare a frequency distribution.

# Frequency Table

- Range =  $208 - 98 = 110$
- Number of classes  $r = \sqrt{37} = 6.$
- Class width (length) =  $\frac{110}{6} = 18.33333$
- For easeiness, take  $L=20$
- Determine the starting point for the first class  
For easeiness, you can start with 90

# Frequency table of systolic Blood pressure for non smokers

Class Interval (Systolic Blood Pressure)	f (Frequency)	( <i>rf</i> ) % (Percentage Frequency)
90 – less than 110	5	14
110 – less than 130	15	41
130 – less than 150	10	27
150 – less than 170	3	8
170 – less than 190	2	5
190 – less than 210	2	5
<b>Total</b>	<b>37</b>	<b>100</b>



# Graphing Representation of Continuous Data

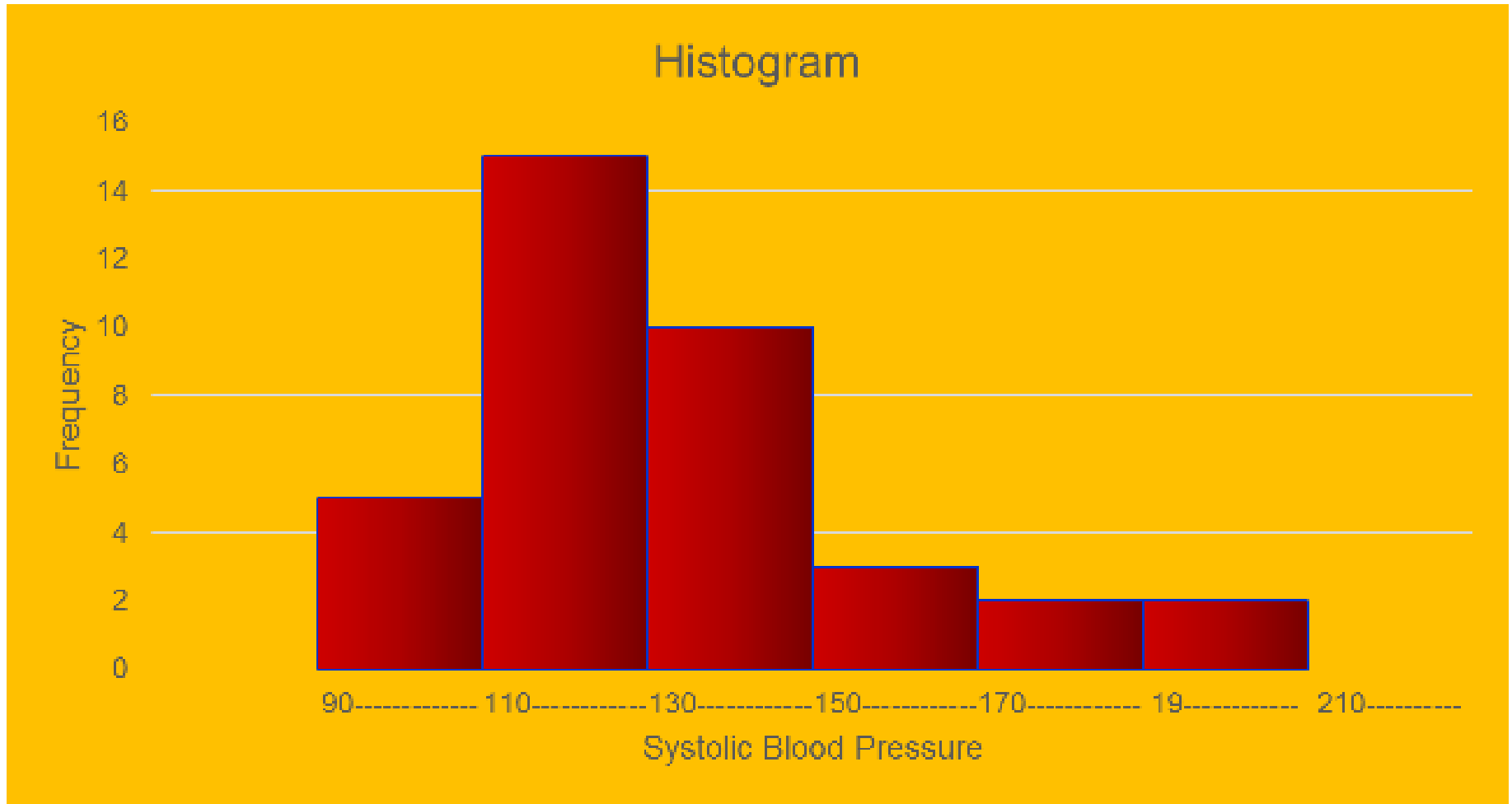
The information provided by a frequency distribution in tabular form is easier to grasp if presented graphically by any of

1. Histogram.
2. Frequency Polygon
3. Frequency Curve.
4. Ogive.

# Histogram

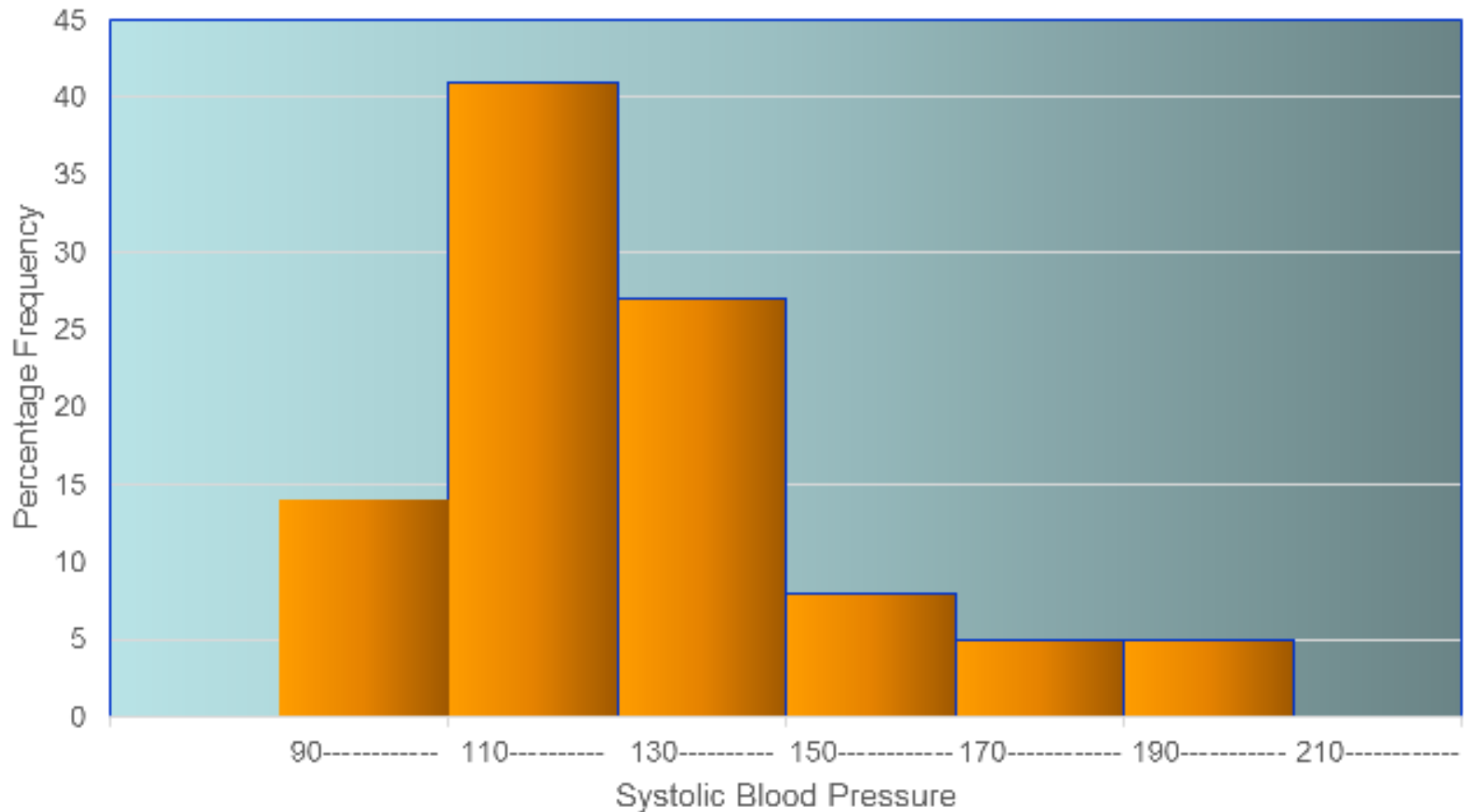
Is a graphical representation of tabulated frequencies, shown as adjacent rectangles, erected over discrete intervals, with an area equal to the frequency of the observations in the interval.

# Frequency Histogram



# Percentage Frequency Histogram

Percentage Histogram for Systolic Blood Pressure of Smokers



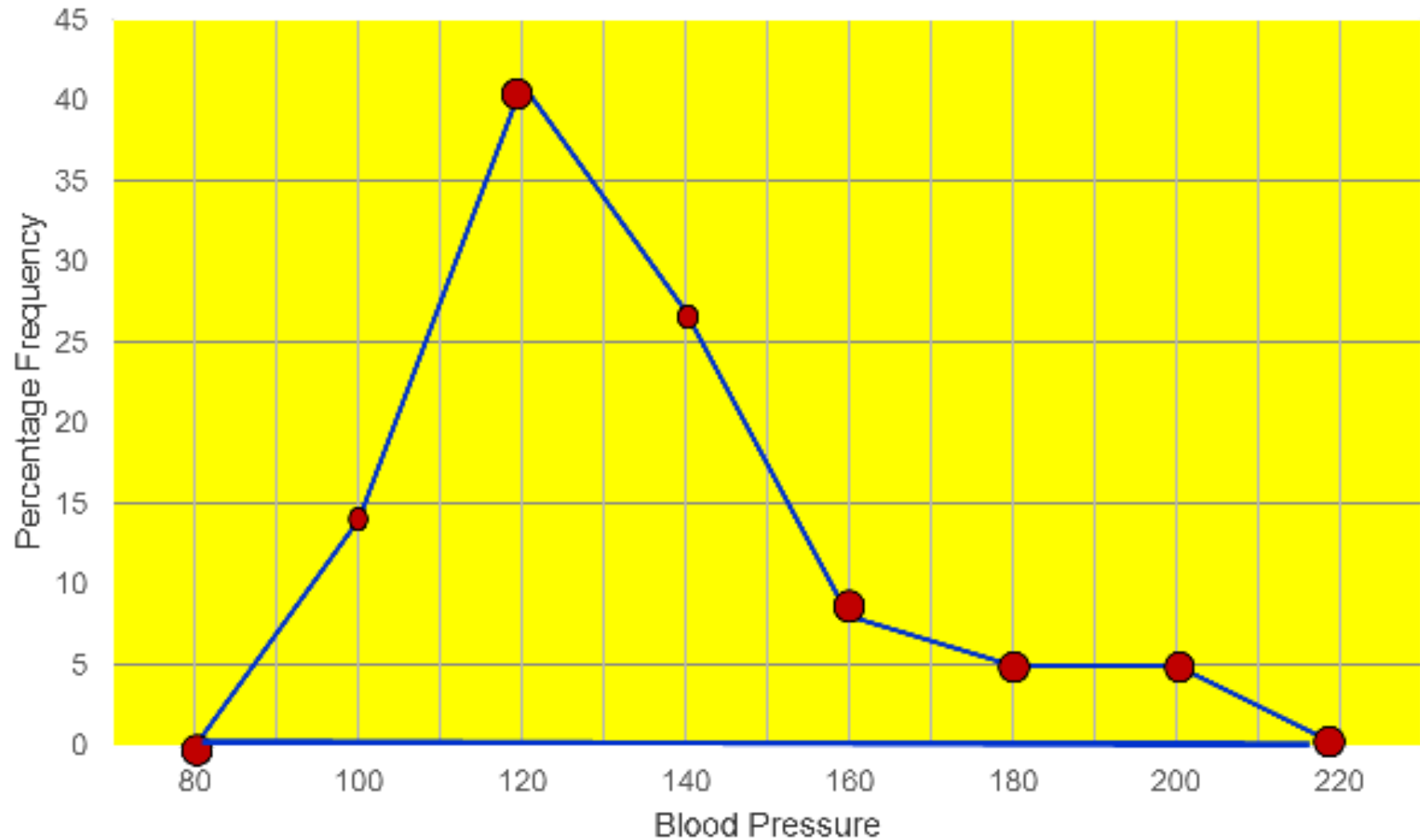
# Frequency Polygon

- It is constructed by making a dot over the ***class midpoint*** at the height of the class frequency. The coordinates of these dots are (***class midpoint, class frequency*** or ***percentage frequency***). These points are then connected with straight lines.
- Frequency polygons should be used to graph ***only quantitative data***.

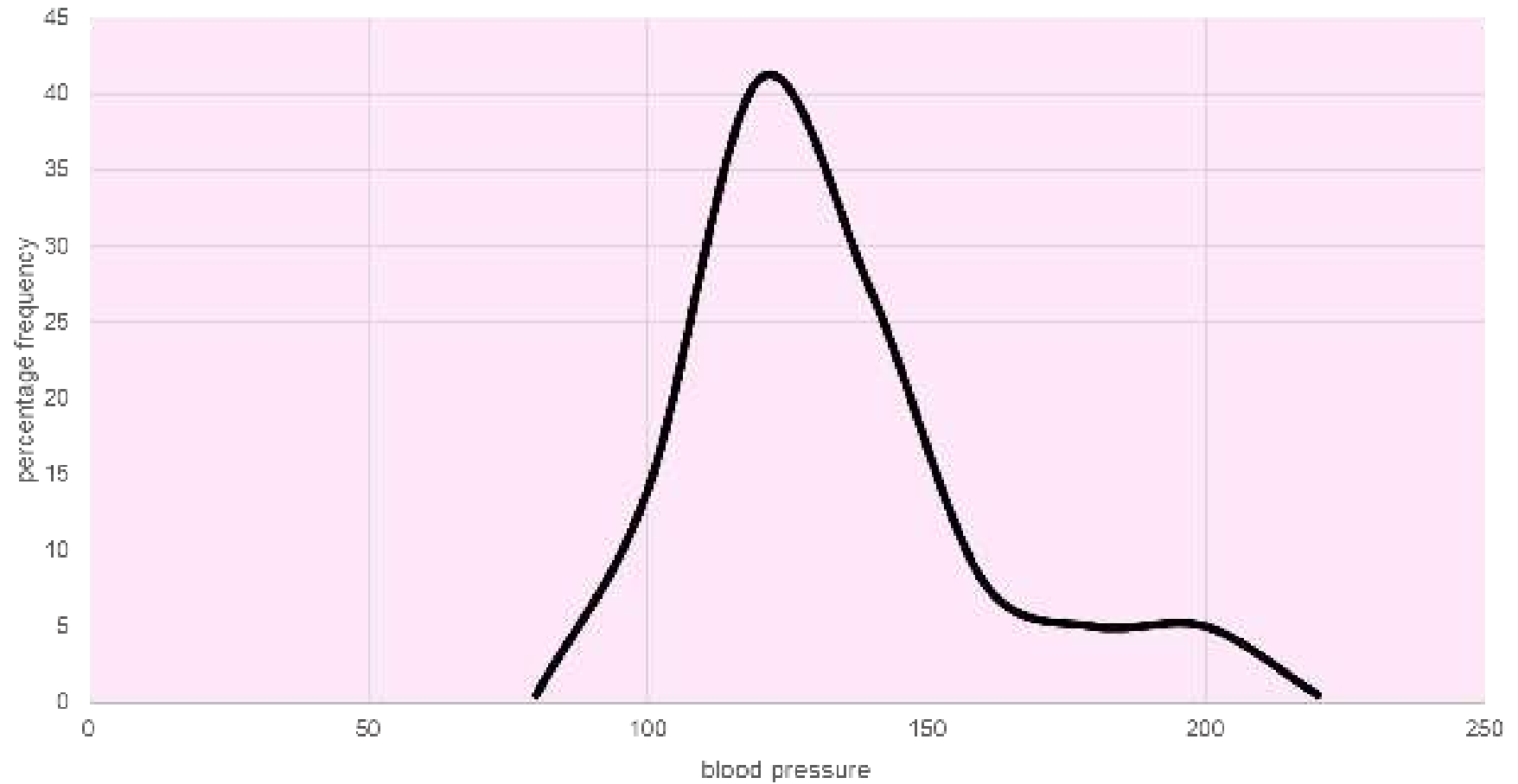
Class Interval	Frequency	midpoint	Percentage frequency
90 – less than 110	5	100	14
110- less than 130	15	120	41
130-less than 150	10	140	27
150- less than 170	3	160	8
170- less than 190	2	180	5
190- less than 210	2	200	5
<b>Total</b>	<b>37</b>		<b>100</b>

# Frequency Polygon

Frequency polygon for blood pressure of smokers



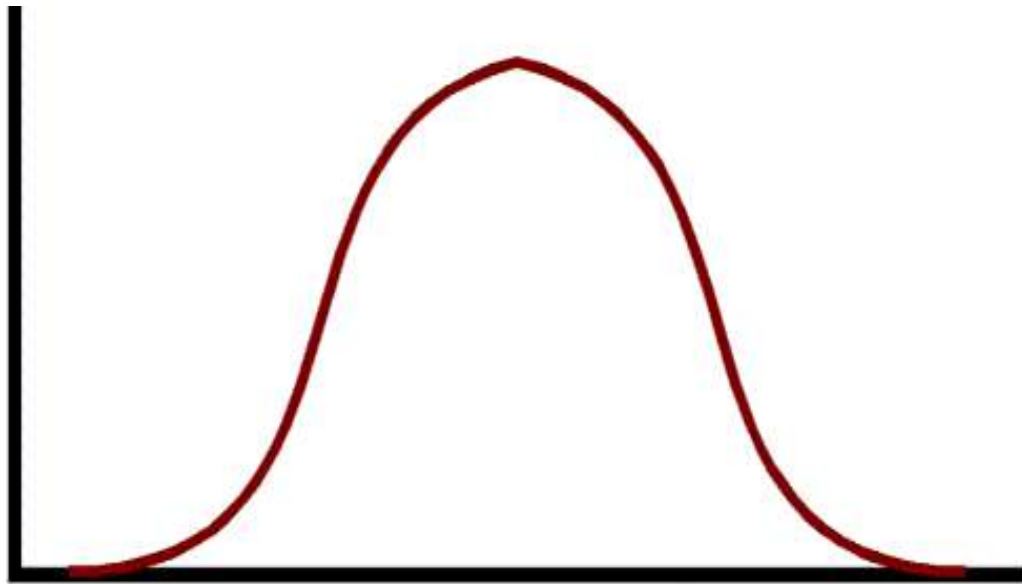
# Frequency curve





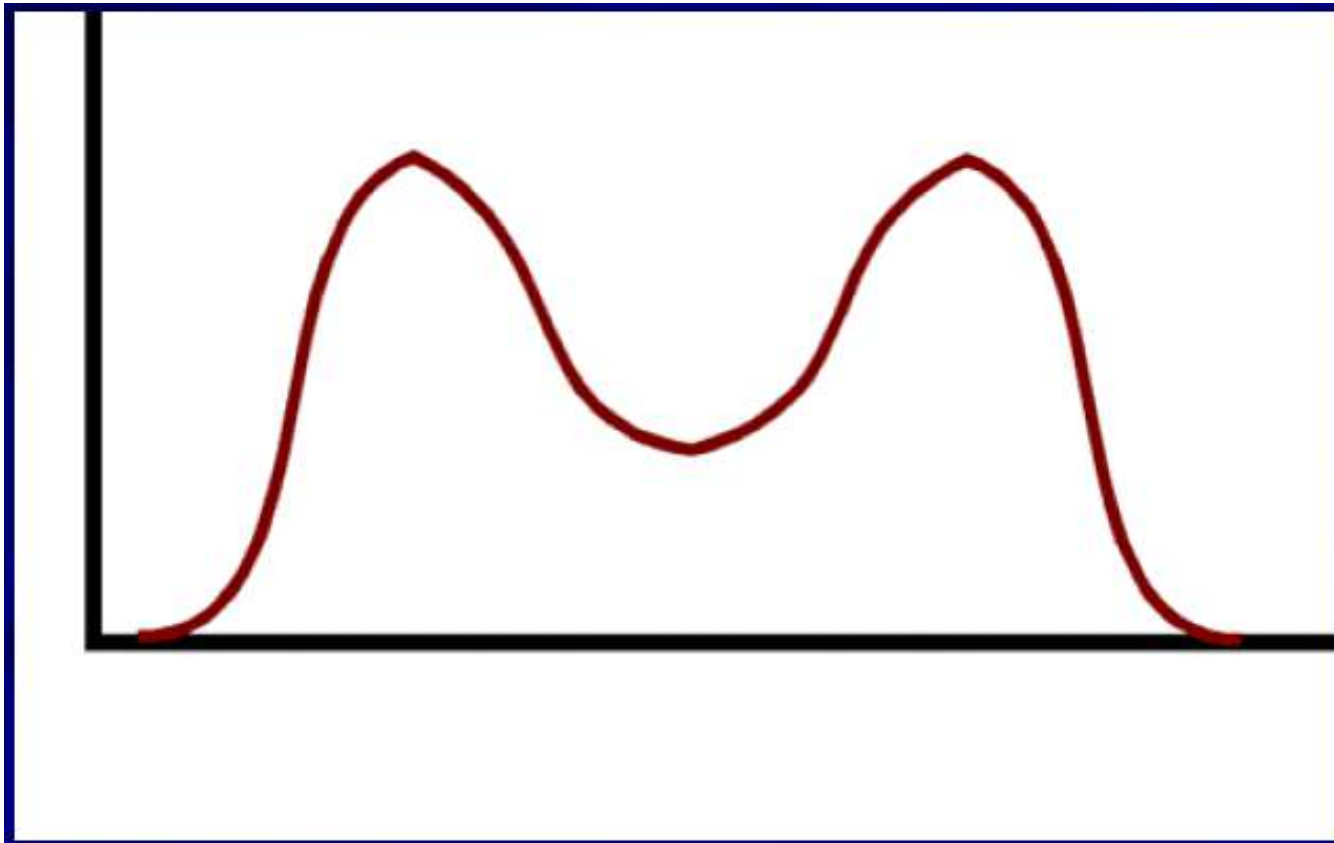
# Shapes of the Frequency Curve

1. (Bell – shaped) symmetrical curve



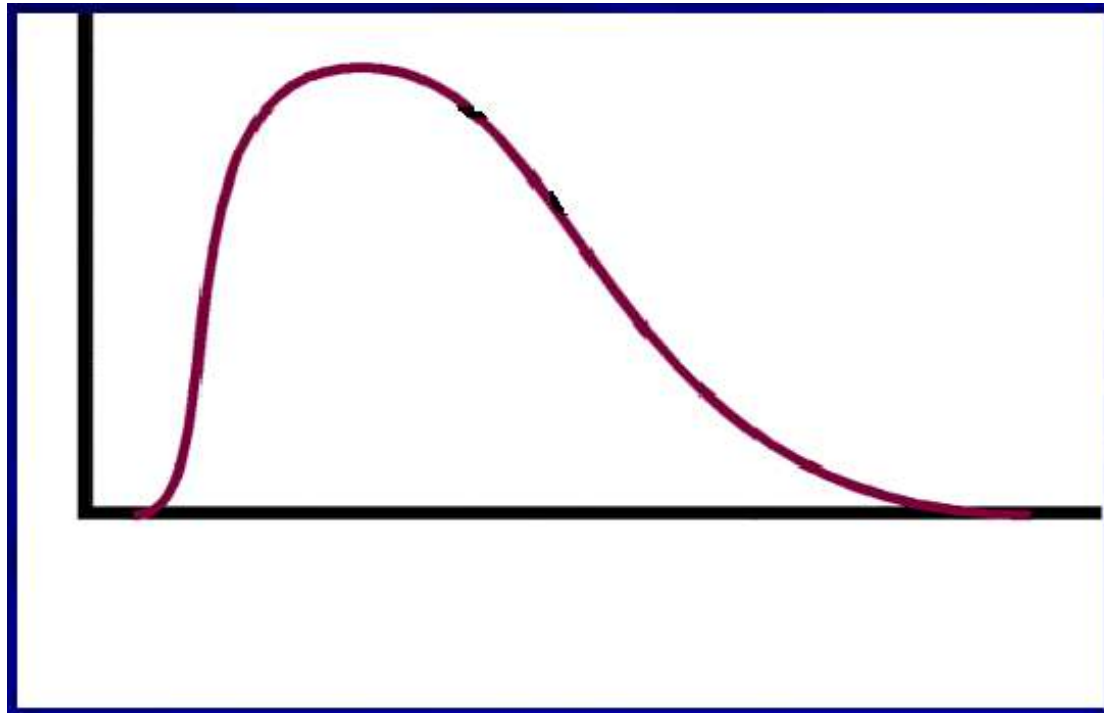
# Shapes of Frequency Distribution

## 2. Symmetric Bimodal Distribution (with two peaks)



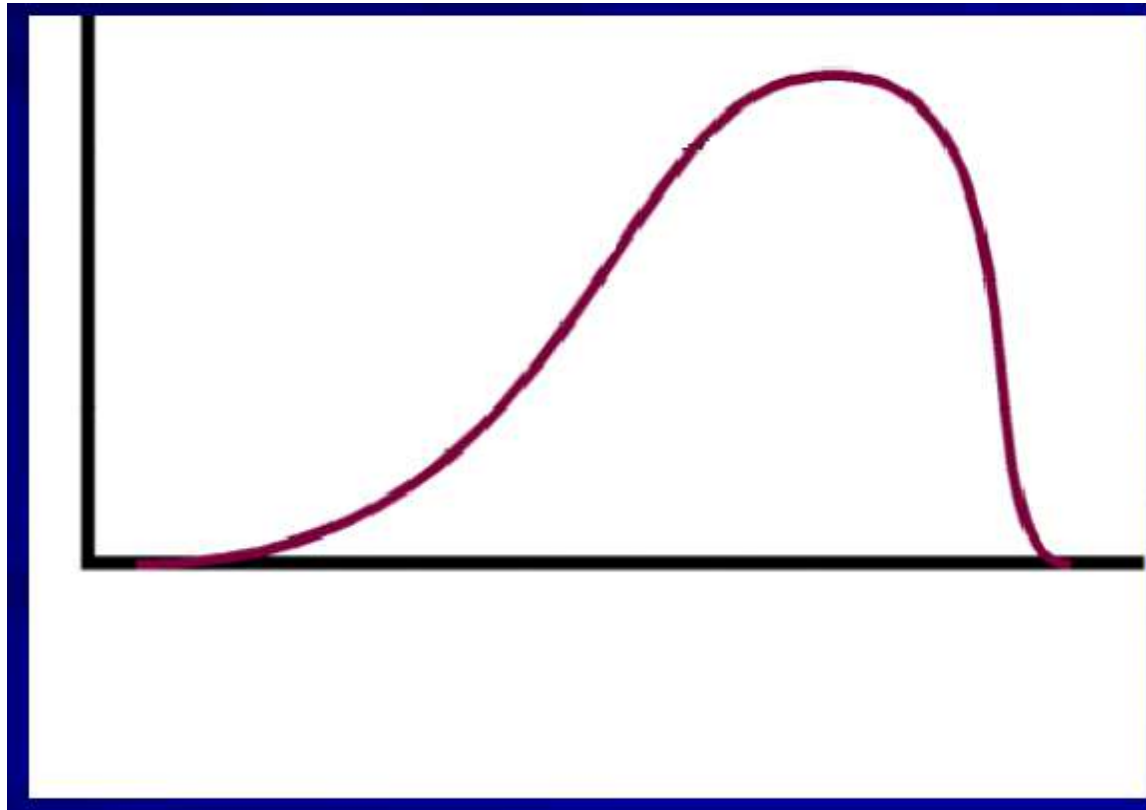
# Shapes of the Distribution

## 3. Right skewed distribution



# Shapes of the Frequency Distribution

## 4. Left skewed distribution



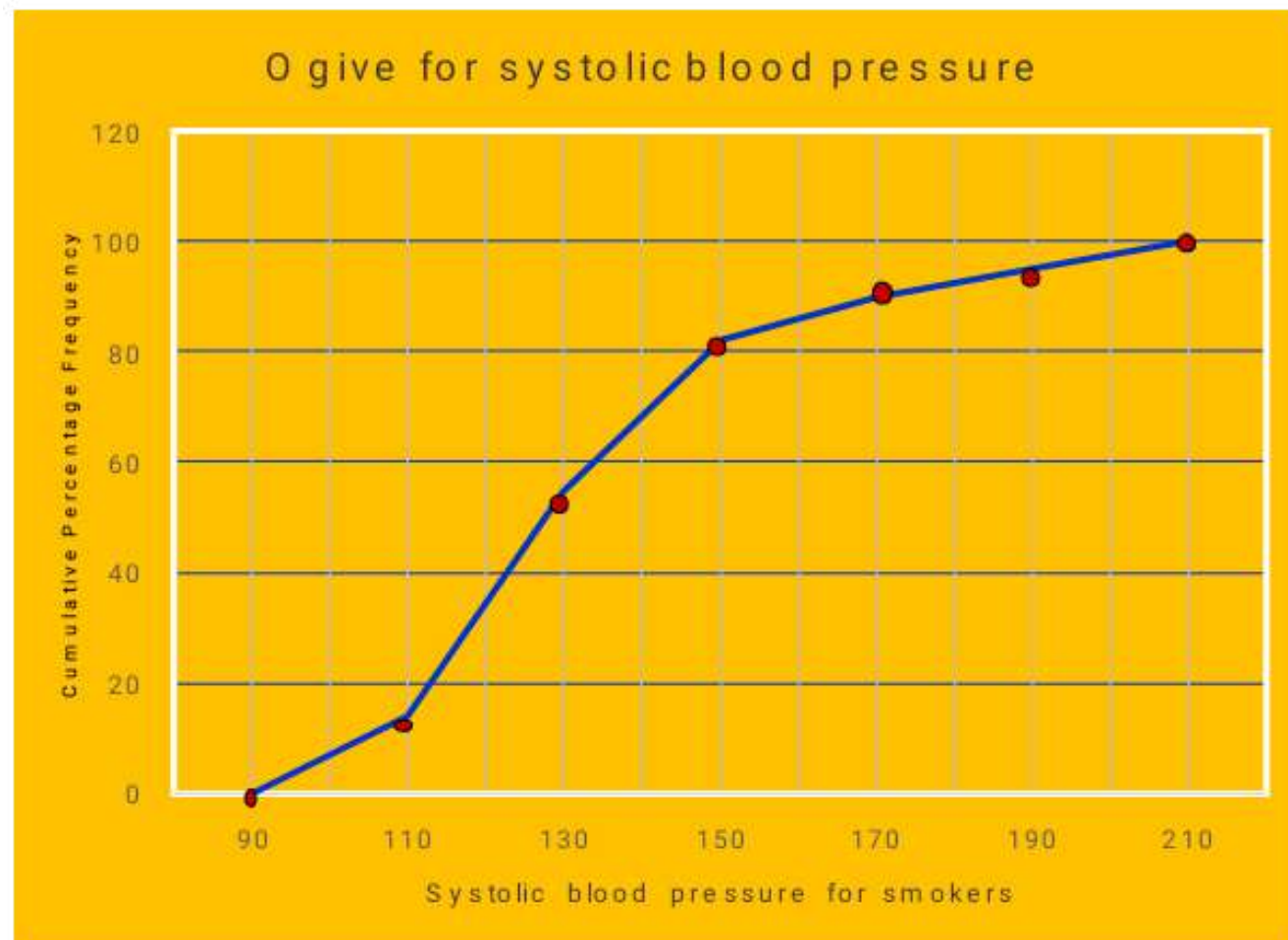
# Ogive

- The Ogive is a graph of a cumulative distribution, which explains data values on the horizontal plane axis and either the cumulative relative frequencies, the cumulative frequencies or cumulative per cent frequencies on the vertical axis.
- It is constructed by making a dot over the **upper class limit** at the height of the cumulative percentage frequency. The coordinates of these dots are (**class upper limit, cumulative percentage frequency**).

# Cumulative Frequency curve (Ogive)

Class interval	Percentage frequency (rf)%	Cumulative percentage frequency (crf)%
90-less than 110	14	14
110-less than 130	41	55
130-less than 150	27	82
150-less than 170	8	90
170-less than 190	5	95
190-less than 210	5	100

# Ogive



# Ogive

## Ogive can be useful in

- Comparing two sets of data, as, for example, blood pressure of smokers and non smokers of individuals.
- Finding the measures of positions as median, percentiles and quartiles.
- finding the percentage of observations in a certain interval.



# Example

Using the Ogive of blood pressure, find

a. the percentage of patients whose  
blood pressure is

i. Less than 140 mmHg

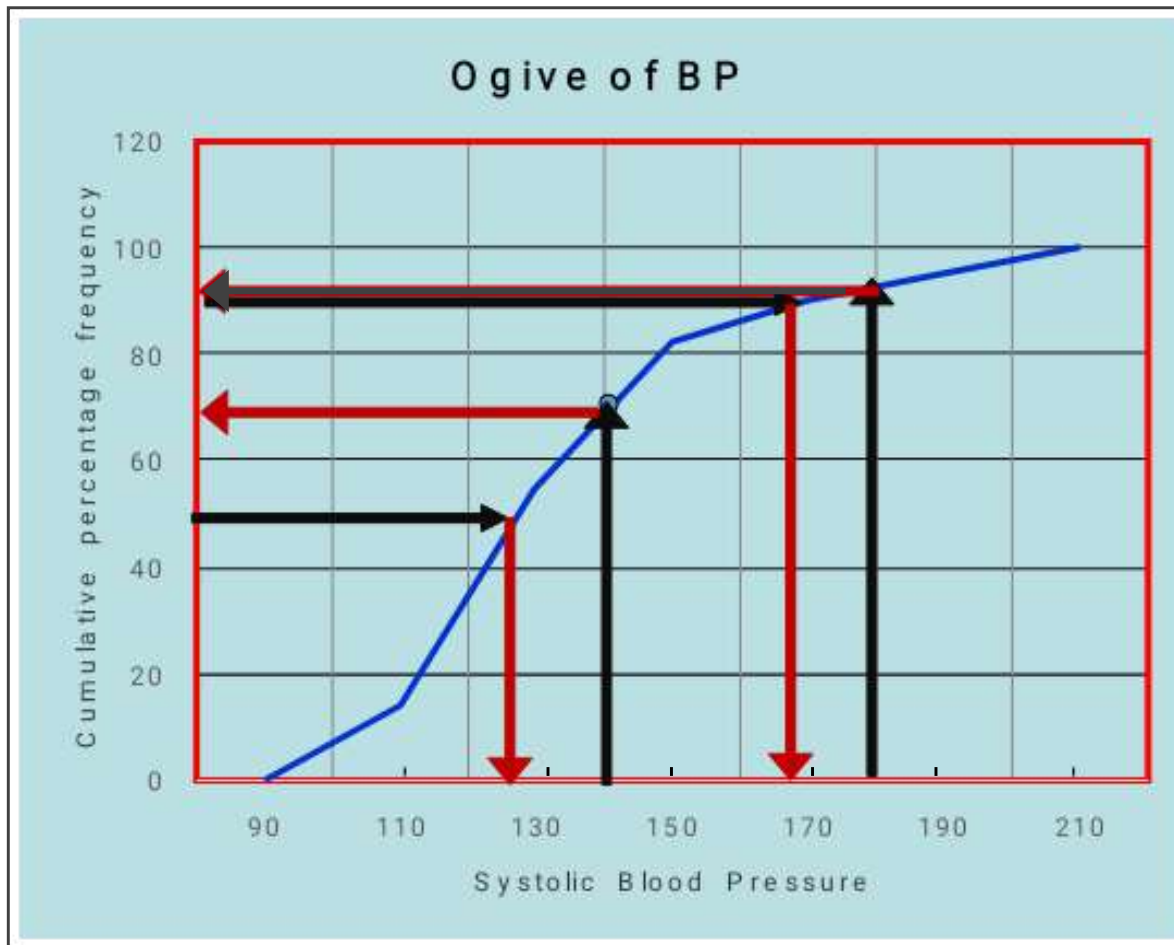
ii. Between 140 and 180 mmHg.

iii. More than 180 mmHg

b. the median

c. 90<sup>th</sup> percentile

# Example continued



# Answer:

(a)

i. Blood Pressure (BP) of about 70% of individuals is less than 140 mmHg.

ii. (BP) of about 93% of individuals is less than 180 mmHg.

So  $(93-70)\% = 23\%$  of individuals have BP between 140 and 180 mmHg.

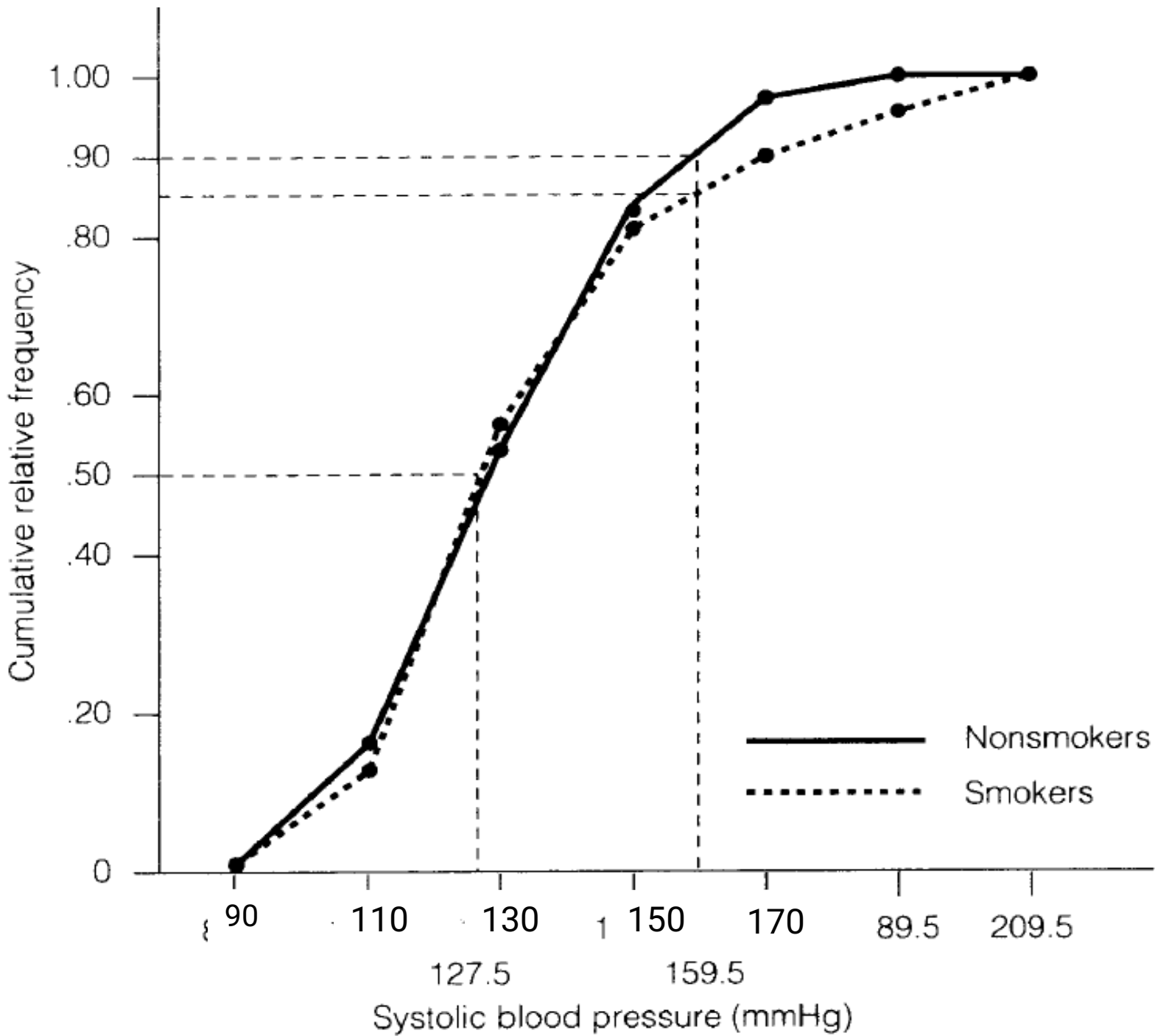
iii. About  $(100-93)\% = 7\%$  of individuals have BP higher than 180 mmHg.

(b) Median = 128 mmHg approximately. This means that about 50% of individuals have BP less than 128 mmHg

(c)  $P_{90} = 170$  mmHg.

# Comparison of systolic blood pressure between smokers and nonsmokers

Class interval pressure((Systolic blood	Nonsmokers (rf)%	Smokers (rf)%	Nonsmokers (crf)%	smokers (crf)%
90 – less than 110	16	14	16	14
110 – less than 130	38	41	54	55
130 – less than 150	29	27	83	82
150 – less than 170	14	8	97	90
170 – less than 190	3	5	100	95
190 – less than 210	0	5	100	100



of 37 smokers and 63 nonsmokers

## Comparison of systolic blood pressure between smokers and nonsmokers

By rapid comparison you can see that 97% of the nonsmokers in the sample have a systolic blood pressure below 169.5 and that 90% of the smokers have a blood pressure below the same level.

An alternate way of looking at this is to note that 3% of the nonsmokers and 10% of the smokers have a systolic blood pressure above 169.5 mmHg.

## Exercise:

The following data represents serum cholesterol level of 49 individuals of physical activity 1 taken from table 2.1

Serum cholesterol level in milligram percent						
199	272	166	239	238	223	190
209	171	147	199	255	199	228
240	192	201	203	243	186	165
239	162	246	234	161	298	211
219	179	212	231	185	180	205
219	221	216	195	173	206	215
176	234	204	225	187	290	218

# Exercise continued

- Prepare a frequency table
- Construct a histogram
- Construct a frequency curve and describe the distribution.
- Construct an Ogive
- Use the ogive to find
  1. The percentage of individuals whose cholesterol level is between 175 and 195 milligram percent.
  2. The percentage of individuals whose cholesterol level is more than 195 milligram percent.
  3. 75<sup>th</sup> percentile.
  4. Median



# Chapter 3

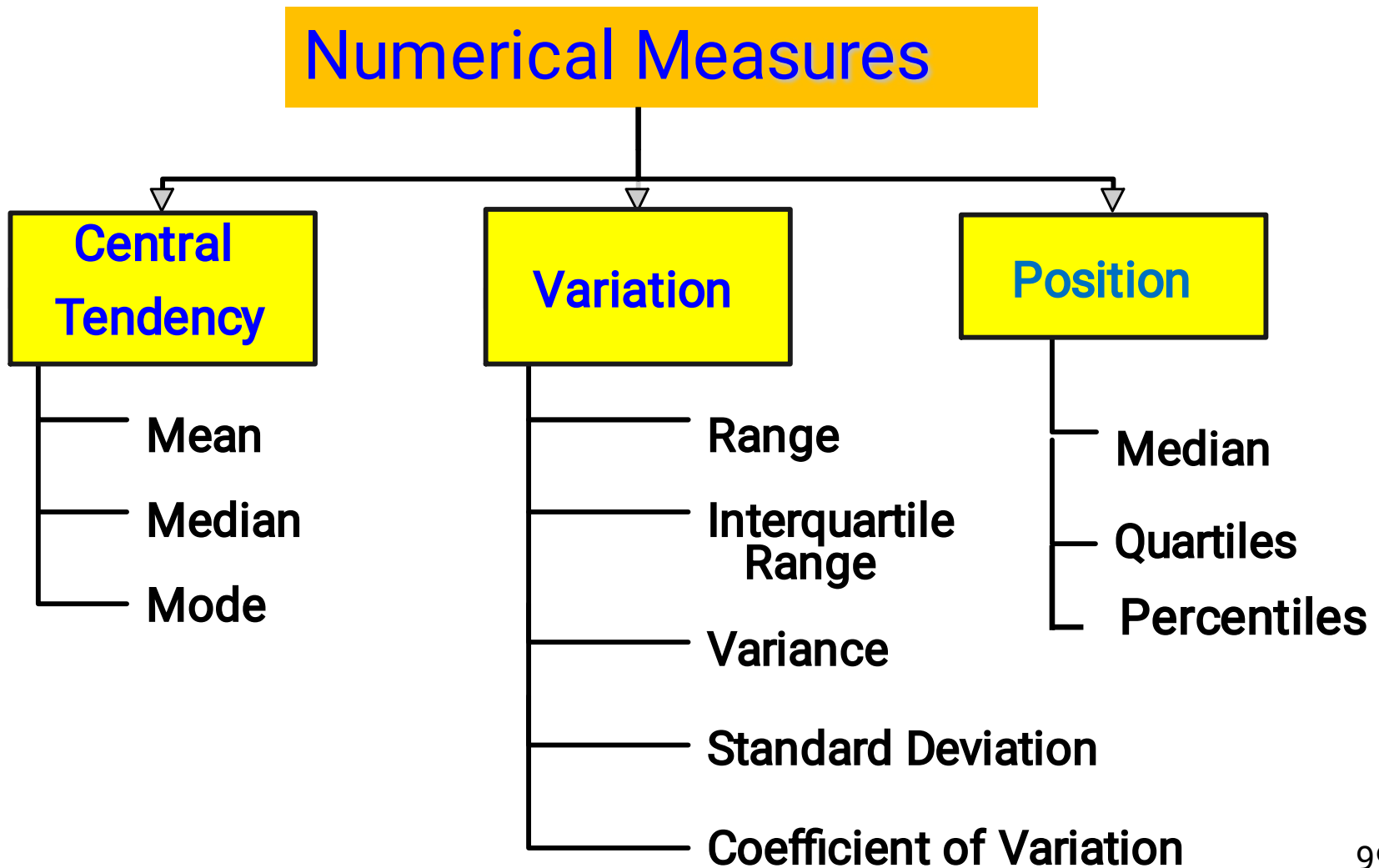
## Summarizing Data

# Summarizing Data

## This chapter aims to

- State and illustrate the definition of the measures both for grouped and raw data (ungrouped);
- Compute and distinguish between the uses of measures of central tendency.
- Compute some uses of measures of variation.
- Compare sets of data by computing and comparing their coefficients of variation.
- Be able to compute the mean and the standard deviation for grouped and ungrouped data.
- Understand the distinction between the population mean and the sample mean.

# Descriptive Statistics Measures



# Measures of Central tendency

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics.

The three most common values are the **mean**, the **median**, and the **mode**.

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others.

# Measures for ungrouped data

## Arithmetic Mean:

If  $x_1, x_2, \dots, x_n$  are the values of  $n$  observations, the sample arithmetic mean is

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

The sample mean is **a statistic**.

The most commonly used measure of central tendency is the **mean**.

# Example

- The following observations represent systolic blood pressure in mmHg of a sample of 12 non smokers selected at random from table 2.1

114 128 106 128 128 154 122 154 120 140  
122 172. The sample mean is

$$\bar{x} = \frac{114 + 128 + 106 + 128 + 128 + 154 + 122 + 154 + 120 + 140 + 122 + 172}{12}$$

$$= 132.33 \text{ mmHg.}$$

# The population arithmetic Mean

$$\mu = \frac{\sum_{i=1}^N x_i}{N} = \frac{\text{sum of all values in the population}}{\text{total number of observations}}$$

The population mean is a parameter.

# Weighted Mean

The weighted mean of a set of numbers  $x_1, x_2, \dots, x_n$  with weights  $w_1, w_2, \dots, w_n$  is given by the formula

$$\bar{x} = \frac{w_1x_1 + w_2x_2 + \dots + w_nx_n}{w_1 + w_2 + \dots + w_n}$$



## Example

- Al Quds Hospital at Gaza pays its hourly employees \$16.5, \$19, or \$25 per day. There are 26 employees, 14 of which are paid at the \$16.5 rate, 10 at the \$19, rate and 2 at the \$25 rate. What is the mean hourly rate paid the 26 employees. The weighted mean is

$$\bar{x}_w = \frac{14(16.5) + 10(19) + 2(25)}{26} = \$18.1154$$

## Exercise

Iman gets quiz grades of 80, 64, and 73. She gets a 72 on her final exam. Find the weighted mean score if the quizzes each count for 10% and the final exam counts for 70% of the final grade.

Answer:

$$\bar{x} = 10/100 * 80 + 10/100 * 64 + 10/100 * 73 + 30/100 * 72 = 72.1$$

**A combined mean** of more than one group is simply a weighted mean, where the weights are the size of each group

## Combined mean is the weighted mean

One property of the mean is that if we know the means and sample sizes of two (or more ) data sets, we can calculate the **combined mean** of the .

The combined mean of two data sets is given by the formula

Combined mean  $\bar{x} = \frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2}$

**Example:** The mean score of a 20 female students on statistics test is 77 and the mean score of a 15 male students on the same test is 71. Find the combined mean score

$$\bar{x} = \frac{20(77)+15(71)}{20+15} \simeq 74.43$$

# Outliers (Extreme Values)

The values that are very small or very large relative to the majority of the values in a data set are called outliers.

Outliers will be discussed later.

# Properties of the Arithmetic Mean

- Every set of data with interval or ratio scale has a mean.
- It is easily used in statistical analysis.
- All values in the data set are included in computing the mean.
- A set of data has a unique mean.
- The arithmetic mean is the only measure of central tendency where the sum of the deviations of each value from the mean is zero, that is

$$\sum_i(x_i - \mu)=0 \quad \text{and} \quad \sum_i(x_i - \bar{x})=0$$

# Disadvantages of the mean

- It is affected by extreme values.

$$(4.2 + 4.3 + 4.7 + 4.8 + 5.0 + 5.1 + 9.0) / 7 = 5.3$$

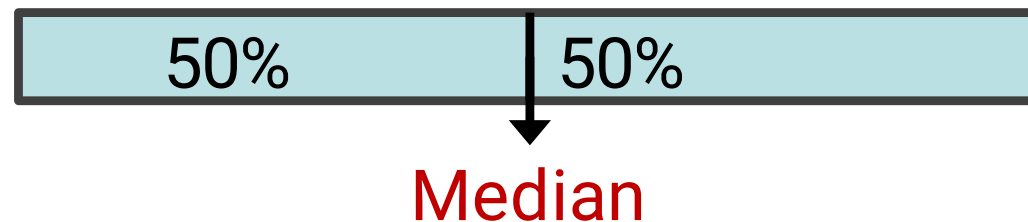
$$(4.2 + 4.3 + 4.7 + 4.8 + 5.0 + 5.1) / 6 = 4.7$$

It would be more representative to calculate the mean without including such an extreme value .

- The mean cannot be used to describe qualitative data.

# Median

- The **median** of a variable is the numerical value that lies in the middle of the data when arranged in ascending order. That is, half the data is below the median and half the data is above the median.



## Find the median in an ungrouped data set

1. Arrange the data in ascending order.
2. Determine the number of observation  $n$ .
3. Determine the observation in the middle of the data set.

If the number of observations is **odd**, then the median is the data value that is exactly in the middle of the data set. That is, it is the observation that lies in the  **$(n + 1)/2$  position**.

$$\text{i.e. median} = \begin{cases} \frac{x_{n+1}}{2}, & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}), & \text{If } n \text{ is even} \end{cases}$$



# Example

The following data represent blood glucose in milligrams percent of **15** individuals selected randomly from table 2.1

147 161 91 231 91 138 442 184  
123 109 136 165 146 132 116.

Ranked observations

91 91 109 116 123 132 136 **138**  
146 147 161 165 184 231 441.

Median =  $x_8$  = 138 mg percent

# Example

Find the median of

22, 19, 27, 32, 38, 25, 32, 26

Ranked data: 19, 22, 25, 26, 27, 32, 32, 38

$$\begin{aligned}\text{Median} &= \frac{x_4 + x_5}{2} \\ &= \frac{26 + 27}{2} = 26.5\end{aligned}$$

# Advantages of the median

- Easy And Simple
- Extreme values in data set do not affect the median as strongly as they do the mean. So it is better than the mean for skewed distribution.
- Median can be easily represented graphically.
- The median can be used with 3 types of variables namely quantitative continuous, quantitative discrete and qualitative ordinal.
- Median can be correctly calculated for open end distribution where mean cannot obtain accurate result.

# Example

Consider 7 physicians who practice in Gaza Strip are sampled and asked how much an office visit costs. Suppose we get the answers:

40, 40, 45, 50, 50, 55, and 200 NIS.

The mean charge for the sample of 7 doctors is approximately 68.57 NIS.

While the median is 50. This value is easily seen to be more representative of the values than was the sample mean, 46.67, which was affected by the extreme value of 200.

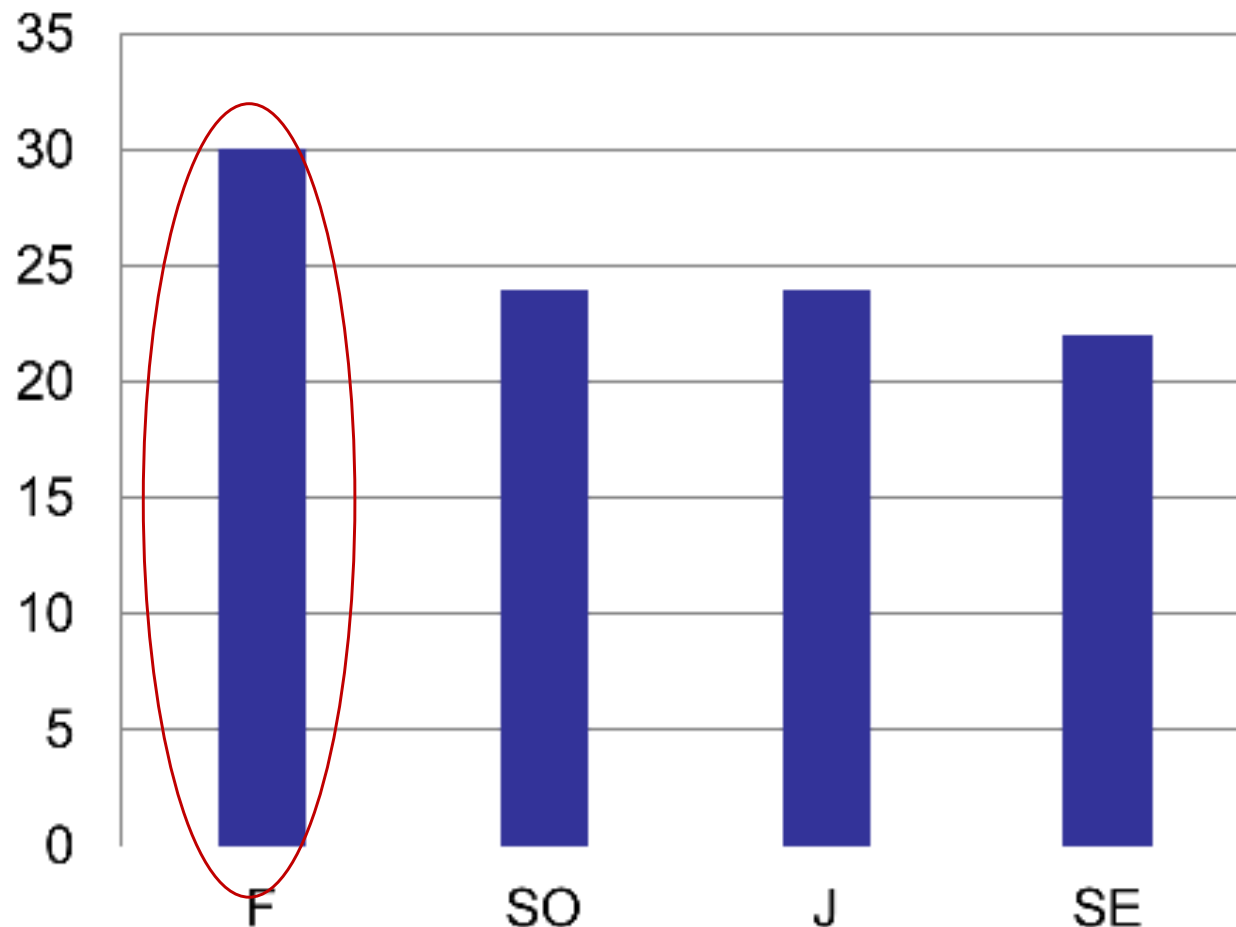
## Disadvantages of the median:

1. It does not take into account the precise value of each observation and hence does not use all information available in the data

2. Unlike mean, median is not amenable to further mathematical calculation and hence is not used in many statistical tests

## Mode:

The mode is the observation that occurs most frequently. i.e., is repeated most often in the data set. e.g. Freshman a is the mode of the observations in students status example.



**Multimodal distribution:** A data set may have several modes. In this case it is called multimodal distribution.

**Example:** The data set

20,22, 24, 24 ,24, 28, 29,29, 29,30

has two modes: 24 and 29.

This distribution is called **bimodal** distribution.

## **Advantages of the mode :**

- Like the median, the mode is **not** affected by extreme values.

For a given sample of size  $n=16$ :

33 35 36 37 38 38 38 **39 39 39 39** 40 40 41 120,

the mode is **39**.

- Easily determined for categorical data

- It can be used to describe both qualitative and quantitative data.

One of the **disadvantages** of the mod is that cannot be clearly defined in case of multi-model series.

# Selecting an Appropriate Measure of Central Tendency

There are two general criteria for choosing between the measures of central tendency

## 1. Scale of measurement

- **Nominal** scale data, you can only use the **Mode**
- **Ordinal** scale data, you can only use **Median** or **Mode**; Median is more informative
- **Interval** or **ratio** scale data, you can use **any one of the three**.

## 2. Extreme values

- Mean is more informative, if you don't have extreme values.
- If you have extreme values, you use the median in place of mean.

# Measures of Central Tendency for Grouped Data



# Mean of Grouped Data

1. Determine the mid-point  $m_i$  of each interval

$$m_i = \frac{\text{lower limit} + \text{upper limit}}{2}$$

2. Find the product  $m_i f_i$ , where  $f_i$  is the corresponding frequency.
3. Use the formula

$$\bar{x} = \frac{\sum m_i f_i}{\sum f_i}$$

## Example (Finding the Mean for Grouped Data)

Consider the table that represents Systolic blood pressure of 37 smokers in page

Class Interval	Frequency ( $f_i$ )	Midpoint ( $m_i$ )	$m_i f_i$
90 – less than 110	5	100	500
110- less than 130	15	120	1800
130-less than 150	10	140	1400
150- less than 170	3	160	480
170- less than 190	2	180	360
190- less than 210	2	200	400
<b>Total</b>	<b>37</b>		<b>4940</b>

$$\bar{x} = \frac{4940}{37} = 133.5 \text{ mmHg}$$

# Median for Grouped Data

In a grouped distribution, the following steps are followed:

1. Find the ascending cumulative frequency ( $cf$ ) value for each class in the table.
2. Find  $cf$  value that the first exceeds  $\frac{n}{2}$ , which identifies the median class  $M$ . i.e. the median class is that with least  $cf$  such that  $cf \geq \frac{n}{2}$ .

The median then is computed using the formula

$$\text{Median} = L + \left( \frac{n}{2} - cf_{M-1} \right) \frac{h}{f}, \text{ where}$$

$L$  is the lower limit of the median class

$cf_{M-1}$  is the cumulative frequency of the pre median class.

$h$  is the median class width.

$f$  is the frequency of the median class.

## Example (Find the median)

Consider the grouped data of systolic blood pressure

Class interval	Frequency ( $f$ )	Cumulative frequency ( $cf$ )
90-less than 110	5	5
110-less than 130	15	20
130-less than 150	10	30
150-less than 170	3	33
170-less than 190	2	35
190-less than 210	2	37

## Example cont.

$$\frac{n}{2} = \frac{37}{2} = 18.5$$

The median class is 110 -130

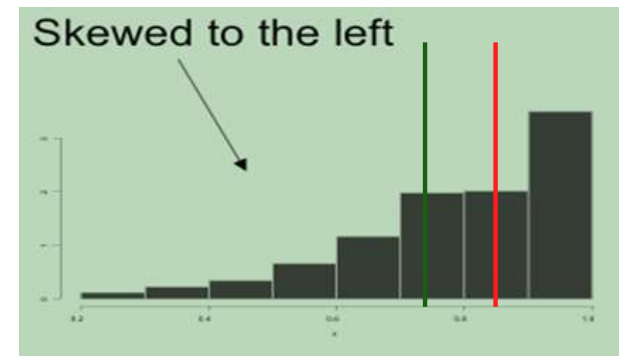
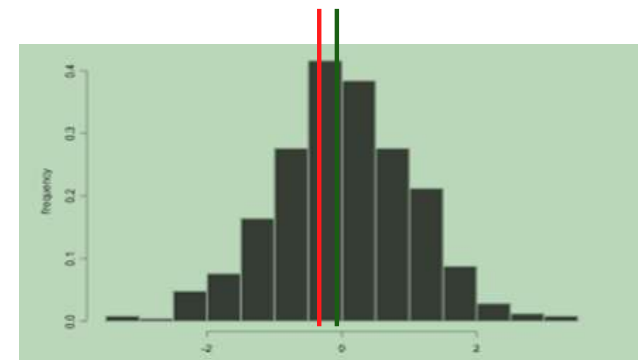
$$L = 110, cf_{M-1} = 5, h = 20, \text{ and } f = 15$$

$$\text{Median} = 110 + (18.5 - 5) \frac{20}{15} = 128 \text{ mmHg}$$

- We can find the mode class which is the class with the largest frequency
- The **mode class is 110 - 130**

# Effect of Asymmetry

- Symmetric Distributions
  - **Mean**  $\approx$  **Median** (approx. equal)
- Skewed to the Left
  - **Mean**  $<$  **Median**
  - Mean pulled down by small values
- Skewed to the Right
  - **Mean**  $>$  **Median**
  - Mean pulled up by large values

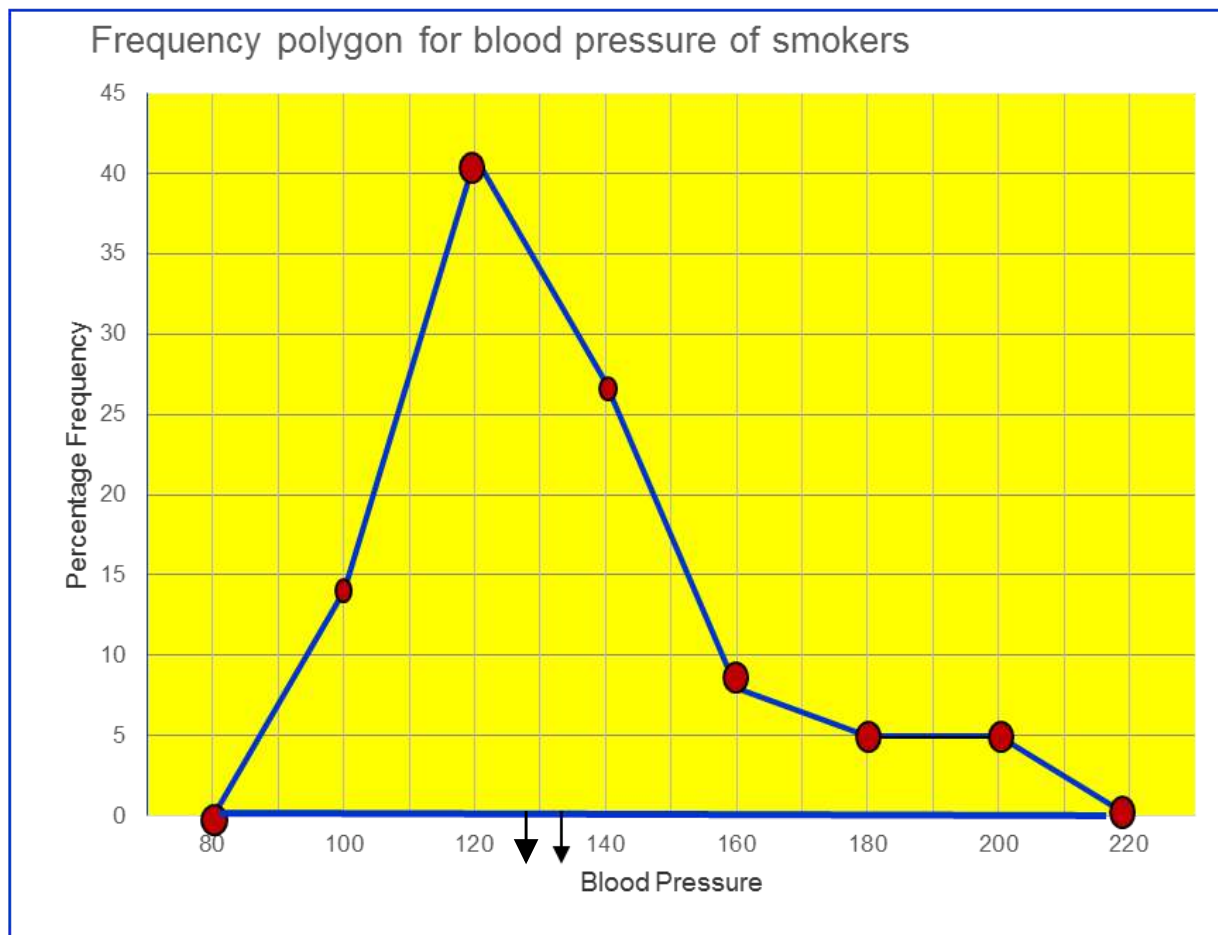


# Shape of the distribution

In the previous example,

**Mean = 133.5** and the **median = 128** mmHg i.e.

**mean > median** and we can see that the frequency distribution is skewed to the right



# Measures of Position (Location)

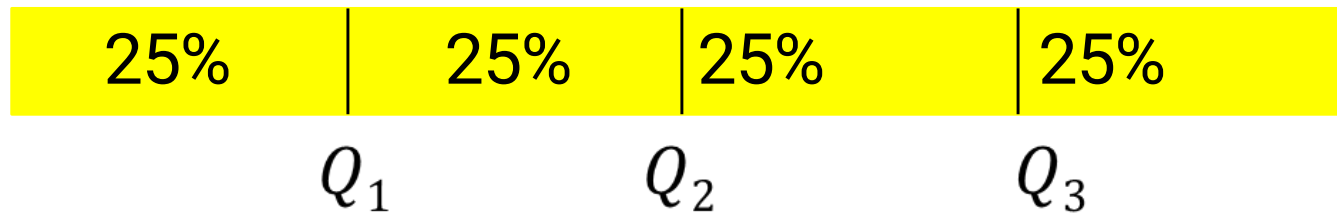
- A measure of position determines the position of a single value in relation to other values in a sample or a population data set.
- There are many measures of position; however, we discuss here quartiles and percentiles.

## Quartiles:

Are three summary measures that divide a ranked data set into 4 equal parts .

- The 2<sup>nd</sup> quartile ( $Q_2$ ) is the median of a data set.
- The 1<sup>st</sup> quartile ( $Q_1$ ) is the value of the middle term among the observations that are less than the median.
- The 3<sup>rd</sup> quartile ( $Q_3$ ) is the value of the middle term among the observations that are greater than the median.





- Approximately 25% of the values in a ranked data set are less than  $Q_1$  and about 75% are greater than  $Q_1$ .
- Approximately 50% of the values in a ranked data set are less than  $Q_2$  and about 50% are greater than  $Q_2$ .
- Approximately 75% of the values in a ranked data set are less than  $Q_3$  and about 25% are greater than  $Q_3$ .

## Interquartile Range (*IQR*)

$$IQR = Q_3 - Q_1$$

## Example (Quartiles)

The following data represent blood glucose in milligram per decilitre (mg%) of 15 individuals selected randomly from table 2.1

147 161 91 231 91 138 442 184 123 109 136 165  
146 132 116.

Ranked data:

91 91 109 116 123 132 136 138 146 147 161  
165 184 231 442.

$$Q_2 = \text{median} = x_{\frac{15+1}{2}} = x_8 = 138 \text{ mg\%}$$

$$Q_1 = 116 \text{ mg\%} \quad \text{and} \quad Q_3 = 165 \text{ mg\%}$$

## Example (Quartiles)

The following data represent blood glucose in milligram per decilitre (mg%) of 15 individuals selected randomly from table 2.1

147 161 91 231 91 138 442 184 123 109 136  
165 146 132 116.

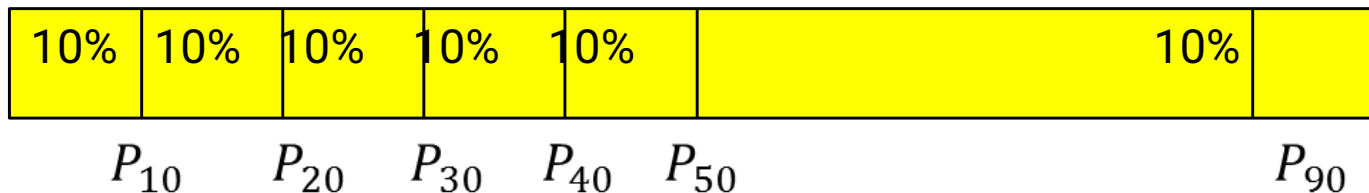
Ranked data:

91 91 109 116 123 132 136 138 146 147 161  
165 184 231 442.

- $Q_2 = \text{median} = x_{\frac{15+1}{2}} = x_8 = 138 \text{ mg}\%$
- $Q_1 = 116 \text{ mg}\%$  and  $Q_3 = 165 \text{ mg}\%$

# Percentiles

Percentiles are summary measures that divide a ranked data set (in ascending order) into 100 equal parts. Each data set has 99 percentiles  $P_1, P_2, \dots$ , and  $P_{99}$ .



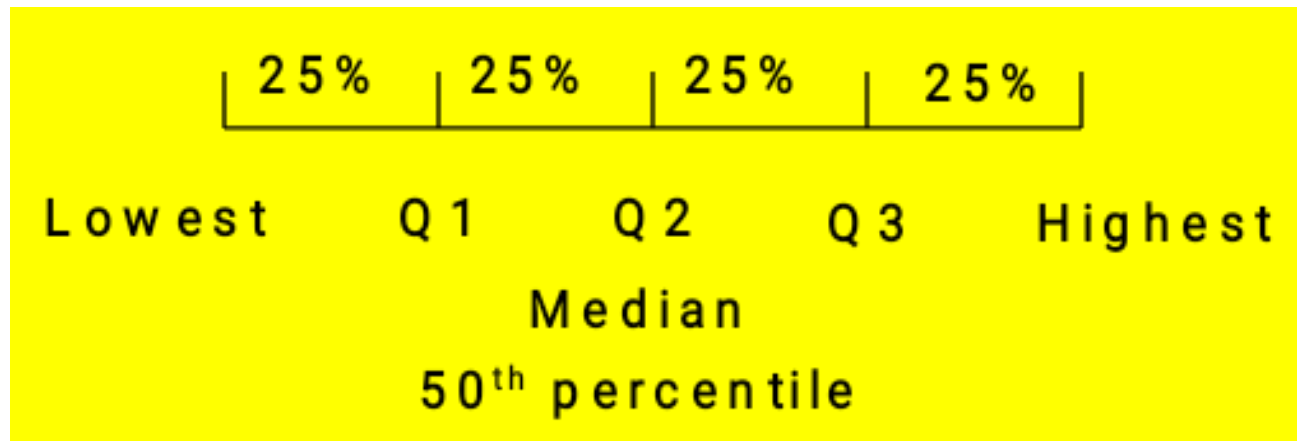
The  $k^{\text{th}}$  percentile ( $P_k$ ) is the value of the observation that is preceded by  $k\%$  and followed by  $(1 - k)\%$  of the observations in a ranked data set. The approximate value of ( $P_k$ ) is

$$(P_k) = x_{\frac{kn}{100}}, \text{ where } n \text{ is the sample size}$$

Note that  $P_{25} = Q_1$ ,  $P_{75} = Q_3$ , and  $P_{50} = Q_2 = \text{median}$

# Percentiles and Quartiles

- All the quartiles are percentiles. For example, the 3<sup>rd</sup> quartile is the 75<sup>th</sup> percentile  $P_{75} = Q_3$ , the 1<sup>st</sup> quartile is the 25<sup>th</sup> percentile  $P_{25} = Q_1$  and  $P_{50} = Q_2 = \text{median}$ .



## Example (Percentiles)

To find  $P_{20}$  of blood glucose in the previous example

$$\frac{kn}{100} = \frac{20 \times 15}{100} = 3, \text{ then } P_{20} = x_3 = 109 \text{ mg\%}.$$

This indicates that the blood glucose of about 20% of Individuals is less than 109mg%.

To find  $P_{90}$  ,

$$\frac{kn}{100} = \frac{90 \times 15}{100} = 13.5, \text{ then}$$

$$P_{90} = \frac{x_{13} + x_{14}}{2} = \frac{184 + 231}{2} = 207.5 \text{ mg\%}$$

This indicates that the blood glucose of about 10% of individuals in the sample is more than 208 mg%

# Measures of Position for Grouped Data

# Quartiles

## To find $Q_1$ for grouped data

$Q_1$  class is that with least  $cf$  such that  $cf \geq \frac{n}{4}$

$$Q_1 = L + \left(\frac{n}{4} - cf\right)\frac{h}{f}, \text{ where}$$

$L$  is the lower limit of  $Q_1$  class

$cf$  is the cumulative frequency of the pre  $Q_1$  class.

$h$  is the width of  $Q_1$  class

## To find $Q_3$ for grouped data

$Q_3$  class is that with least  $cf$  such that  $cf \geq \frac{3n}{4}$

$$Q_3 = L + \left(\frac{3n}{4} - cf\right)\frac{h}{f}, \text{ where}$$

$L$  is the lower limit of  $Q_3$  class

$cf$  is the cumulative frequency of the pre  $Q_3$  class.

$h$  is the width of  $Q_1$  class



# Percentiles

To find  $k^{th}$  Percentile ( $P_k$ ) for a grouped data,

We first find the  $P_k$  class, which is the class with least

$cf$  such that  $cf \geq \frac{kn}{100}$ . Then  $P_k$  is given by

$$P_k = L + \left( \frac{kn}{100} - cf_{p-1} \right) \frac{h}{f}, \text{ where}$$

$L$  is the lower limit of the  $P_k$  class

$cf_{p-1}$  is the cumulative frequency of the class preceding the  $P_k$  class

$h$  is the width of the  $P_k$  class

$f$  is the frequency of the  $P_k$  class

# Example (find Quartiles and Percentiles)

Consider the grouped data of systolic blood pressure

Find the **three quartiles**,  $P_{20}$ , and  $P_{90}$

Class interval	Frequency ( $f$ )	Cumulative frequency ( $cf$ )
90-less than 110	5	5
110-less than 130	15	20
130-less than 150	10	30
150-less than 170	3	33
170-less than 190	2	35
190-less than 210	2	37

## Example (continued)

$$Q_2 = \text{median} = 128 \text{ mmHg}$$

To find  $Q_1$ , we find  $Q_1$  class

$$\frac{n}{4} = 9.25, \quad \text{so, } Q_1 \text{ class is } 110-130$$

$$Q_1 = 110 + (9.25 - 5) \frac{20}{15} = 115.67 \text{ mmHg}$$

To find  $Q_3$ , we find  $Q_3$  class

$$\frac{3n}{4} = 27.75, \quad \text{so, } Q_3 \text{ class is } 130-150$$

$$Q_3 = 130 + (27.75 - 20) \frac{20}{10} = 145.5$$

## Example (continued)

To find  $P_{20}$

$\frac{20 \times 37}{100} = 7.4$ ,  $P_{20}$  class is 110 – 130

$$P_{20} = 110 + (7.4 - 5) \frac{20}{15} = 113.2 \text{ mmHg.}$$

To find  $P_{90}$

$\frac{90 \times 37}{100} = 33.3$ ,  $P_{90}$  class is 170 – 190

$$P_{90} = 170 + (33.3 - 33) \frac{20}{2} = 173 \text{ mmHg.}$$

# Measures of Variation (Dispersion)

Measures of variation measure its “spread”. When the variation is small, this means that the values are close together (but not the same).

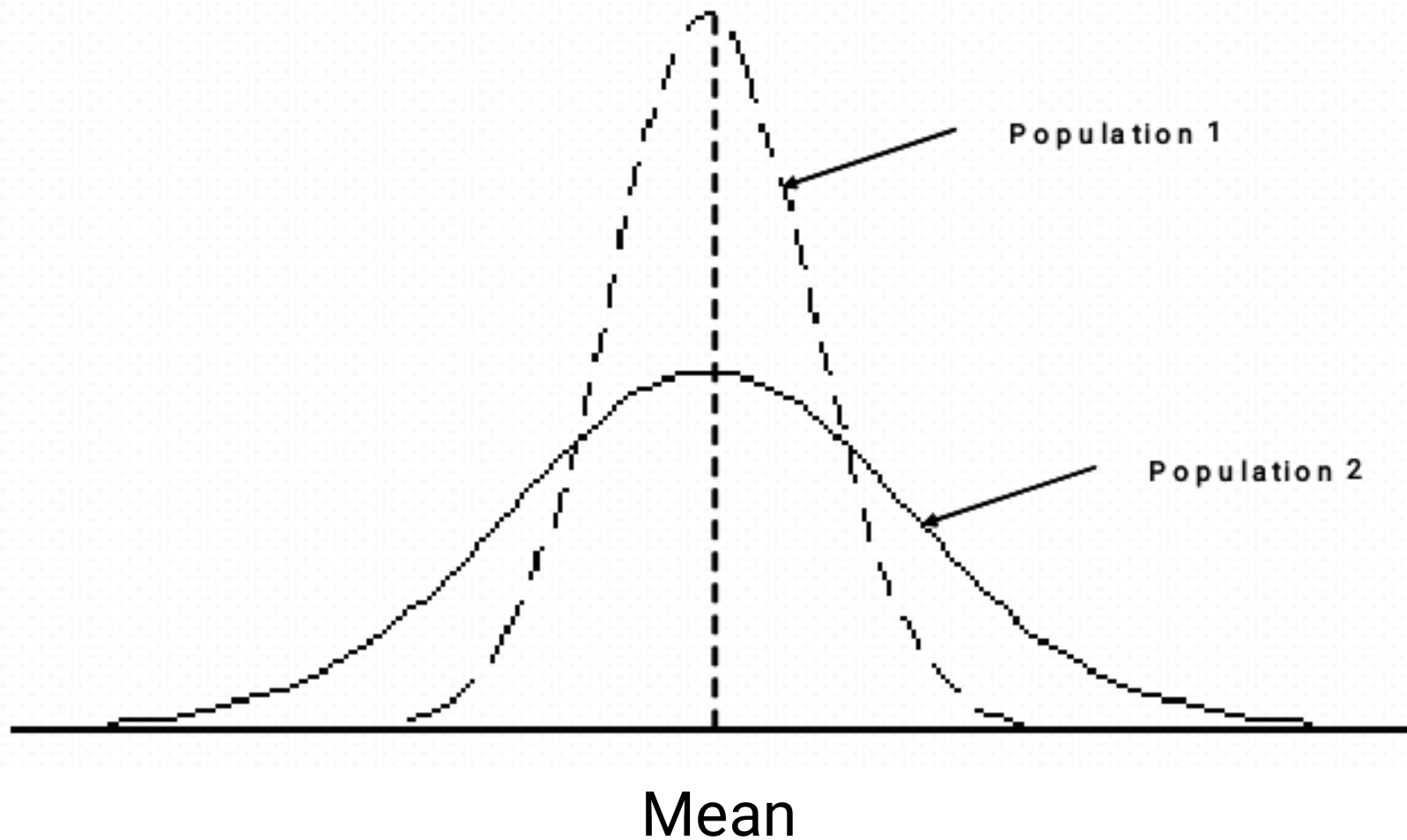
To understand Measures of Variation, consider the following example

### Example

Think of the difference between an exam with a mean mark of 65 in which scores ranged from (62 to 66) and an exam with an average score of 65 in which scores ranged from (30 to 90).

	Scores of two groups	
	Group (A)	Group (B)
	62	30
	65	45
	65	65
	66	75
	66	85
	66	90
<b>mean</b>	<b>65</b>	<b>65</b>

Two frequency distributions with equal means but different amounts of variation.



# Range

- The range is defined as the difference in value between the highest (maximum) and lowest (minimum) observation:

$$\text{Range} = x_{max} - x_{min}$$

- The range can be computed quickly, but it is not very useful since it **considers only the extremes** and does not take into consideration all of the observations.



# Variance

- The Variance is a measure which uses the mean as a point of reference.
- The Variance is less when all value are close to the mean while it is more when the values are spread out from the mean.

# Population variance

- The **population variance** of the observations  $x$  is defined by the formula

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

$x_i$  = the item or observation

$\mu$  = population mean

$N$  = total number of observations in the population

That is the population variance is the arithmetic mean of the sum of the squared deviations about the population mean.

# The standard deviation of a population

The standard deviation is the most commonly used in measures of variability. The standard deviation of the population is given by

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

Since most populations are large, the computation of  $\sigma^2$  and  $\sigma$  are rarely performed. In practice, the population variance (or standard deviation) is usually estimated by taking a sample from the population and using  $S^2$  and  $S$  as a estimate of  $\sigma^2$  and  $\sigma$  respectively.

# The sample variance

- The sample variance of the sample of the observations is defined the formula

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

OR

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

where,

$s^2$  = sample variance

$\bar{x}$  = sample mean

$n$  = number of observations in the sample

# Example

- The following data represent the weight in kg of 6 children.

14 , 22 , 16 , 17 , 20, 19

Find the standard deviation

$$\bar{x} = 18$$

$$S^2 = \frac{42}{5} = 8.4 \text{ kg}^2 \text{ or}$$

$$S^2 = \frac{1986 - \frac{108^2}{6}}{5} = 8.4 \text{ kg}^2$$

$$S = \sqrt{8.4} = 2.898 \text{ kg}$$

	$x$	$x - \bar{x}$	$(x - \bar{x})^2$	$x^2$
	14	-4	16	196
	22	4	16	484
	16	-2	4	256
	17	-1	1	289
	20	2	4	400
	19	1	1	361
total	108	0	42	1986

## Note:

- Standard deviation is never negative
- The unit of standard deviation is the same as that of the raw data, so it is used to compare data with the same units.
- The smaller is the standard deviation, the more homogeneous is the distribution.

Let the mean and the variance of a data set  $x_1, x_2, \dots, x_n$  be  $\bar{x}$  and  $S^2$ , respectively.

Consider the new data set obtained by the transformation

$$x' = ax + b,$$

Then the mean and variance of the new data set, respectively, are  $a\bar{x} + b$  and  $a^2S^2$ .

**Example:**

Assume that the mean score of students in a statistics class is 75 with standard deviation 10. The instructor used the transformation

$$x' = 0.78x + 20$$

What is the new mean. What is the new standard deviation

The new mean is  $a\bar{x} + b = 0.78 * 75 + 20 = 78.5$

The new variance is  $a^2S^2 = 60.84$

# Coefficient of Variation

One important application of the mean and the standard deviation is the coefficient of variation. It is defined as the ratio of the standard deviation to the value of the mean, expressed as a percentage.

$$\text{Coefficient of variation (cv)} = \frac{s}{|\bar{x}|} \times 100\%$$

Since both standard deviation and the mean are expressed in same units, therefore  $cv$  is unit less or dimensionless.

variation of even unrelated quantities (**with different units**). It also useful in comparing the variability among different variables that vary in magnitude of the values (elephant weight versus mouse weight) Therefore, it is possible to use it to compare the relative



# Interquartile range

- The interquartile range tells us about the spread of the middle half of the data.
- It is defined as the difference between the largest and smallest values in the middle 50% of a set of data. and it is defined by

$$IQR = Q_3 - Q_1$$

*IQR* is not affected by extreme values.

In our example of blood glucose of the ungrouped data, we have

$$Q_3 = 165 \quad \text{and} \quad Q_1 = 116 \text{ mg}\%, \quad \text{so}$$

$$IQR = 165 - 116 = 49 \text{ mg}\%$$

# Formula for calculating the standard deviation for grouped data

Sample standard deviation is given by

$$s = \sqrt{\frac{\sum_{i=1}^r m_i^2 f_i - \frac{\left(\sum_{i=1}^r m_i f_i\right)^2}{n}}{n-1}}$$

, where

$m_i$  is the midpoint of the  $i^{th}$  class.

$f_i$  is the frequency of the  $i^{th}$  class.

$r$  is the number of classes.

# Example (finding $S$ )

Consider the table that represents systolic blood pressure of 37 smokers

Class Interval	Frequency ( $f$ )	Midpoint ( $m_i$ )	$m_i f_i$	$m_i^2 f_i$
90 – less than 110	5	100	500	50,000
110- less than 130	15	120	1800	216,000
130-less than 150	10	140	1400	196,000
150- less than 170	3	160	480	76,800
170- less than 190	2	180	360	64,800
190- less than 210	2	200	400	80,000
<b>Total</b>	<b>37</b>		<b>4940</b>	<b>683,600</b>

$$S = \sqrt{\frac{\sum_{i=1}^r m_i^2 f_i - \frac{\left(\sum_{i=1}^r m_i f_i\right)^2}{n}}{n-1}} =$$

$$\sqrt{\frac{683,600 - \frac{(4940)^2}{37}}{36}} =$$

$$\sqrt{667.868} = 25.843 \text{ mmHg}$$

# Example (continued)

Find the coefficient of variation

$$cv = \frac{s}{\bar{x}} \% = \frac{25.843}{133.5} \% = 19.3\%$$

**Note** that there is some difference between results from computations ungrouped and grouped data. The size of the discrepancy depends on width of the class interval and on the number of observations within an interval. With short class intervals and large samples, the discrepancy is negligible.

# Outliers

Recall that

An **outlier** is a number that is so far above the data set or below most of the data set as to be considered abnormal and therefore of questionable accuracy.

## **Outliers may come from**

- data collection errors,
- data entry errors,
- or simply valid but unusual data values.

Regardless of the reason, it is important to identify the outliers in the data set and examine outliers carefully to determine if they are an error.

- An outlier is defined to be any data point that is **1.5 (IQR)** below the lower quartile or above the upper quartile.

## Example (Outliers)

91 91 109 116 123 132 136 138 146 147  
161 165 184 231 442

median = 138

lower quartile ( $Q_1$ ) = 116

upper quartile ( $Q_3$ ) = 165

$$IQR = 49$$

$$Q_1 - 1.5 (IQR) = 116 - 73.5 = 42.5$$

$$Q_3 + 1.5 (IQR) = 165 + 73.5 = 238.5$$

any number below  $Q_1 - 1.5 IQR = 42.5$  or any number  
above  $Q_3 + 1.5 (IQR) = 238.5$  is an outlier.

So, 442 is an outlier.

# Box- Plot

Is another tool, which uses quartiles of a set of measurements to describe the shape and the range of the distribution.

## To construct a box plot

1. Find the lower fence  $LF = Q_1 - 1.5(IQR)$
2. Find the upper fence  $UF = Q_3 + 1.5(IQR)$
3. Find the lower adjacent value  $LAV$  which is the smallest value in the data set  $\geq LF$ .
4. Find the upper adjacent value  $UAV$  which is the largest value in the data set  $\leq UF$ .

The five numbers ( $Q_1, Q_2, Q_3, LAV$ , and  $UAV$ ) can be used to create a **box-plot**.



## Example

To construct the box plot of the data in the previous example.

$$\text{Median} = 138, Q_1 = 116 \text{ and } Q_3 = 165$$

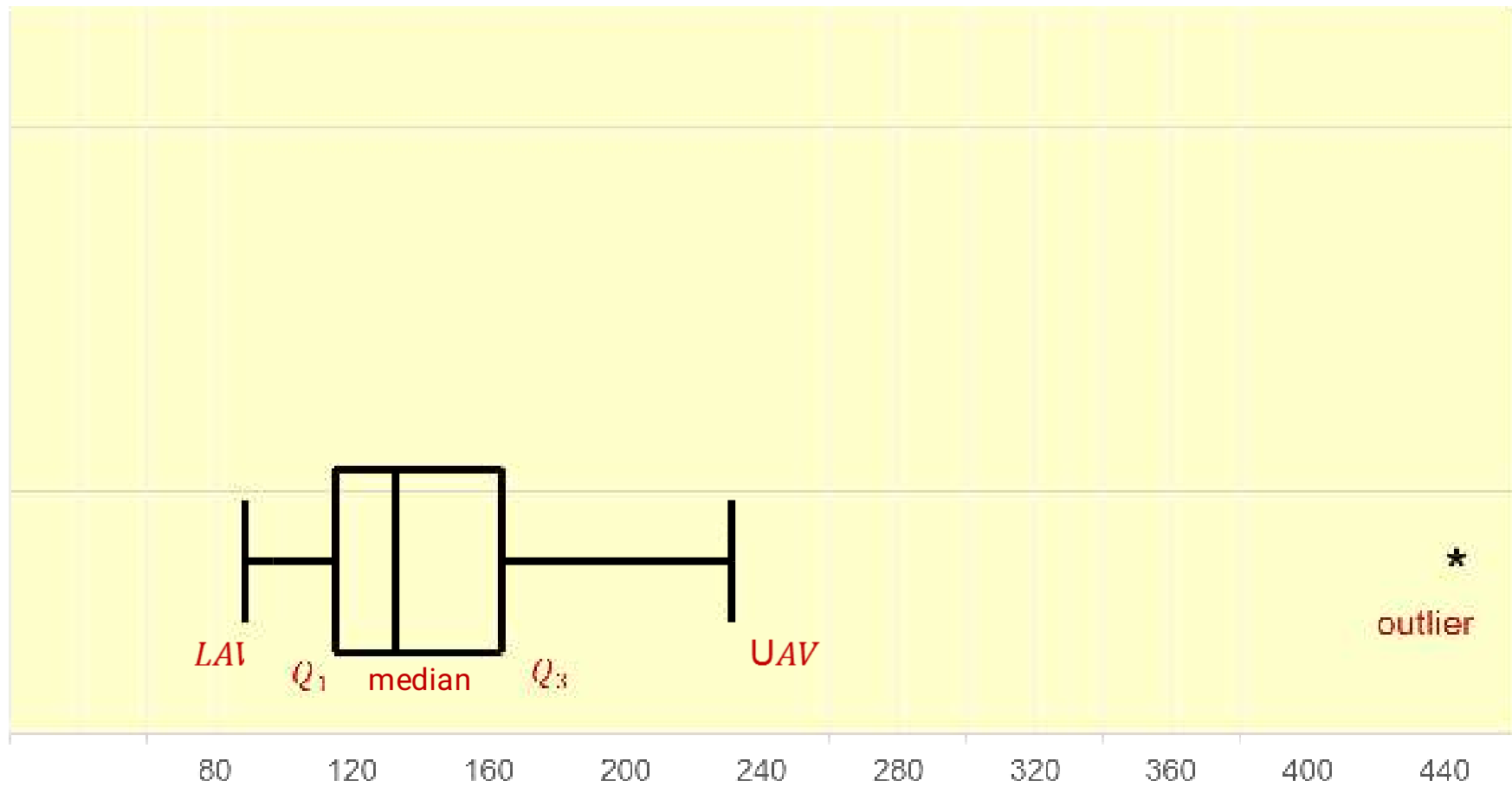
$$LF = Q_1 - 1.5(IQR) = 116 - 1.5(49) = 42.5$$

$$UF = Q_3 + 1.5(IQR) = 165 + 1.5(49) = 238.5$$

$$LAV = 91$$

$$UAV = 231$$

## Example(Boxplot of the Previous Data )



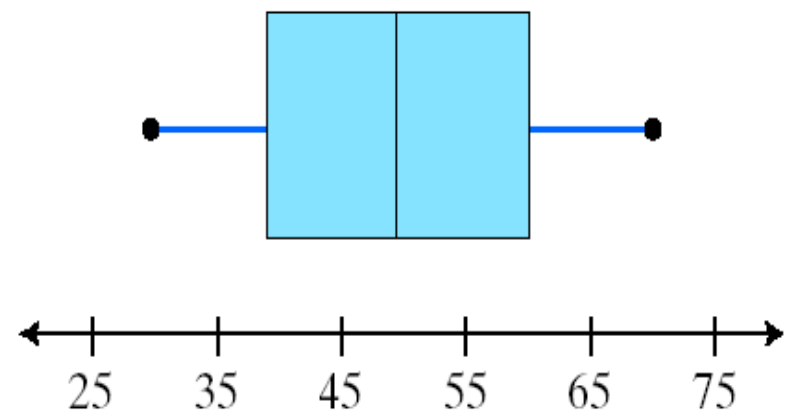
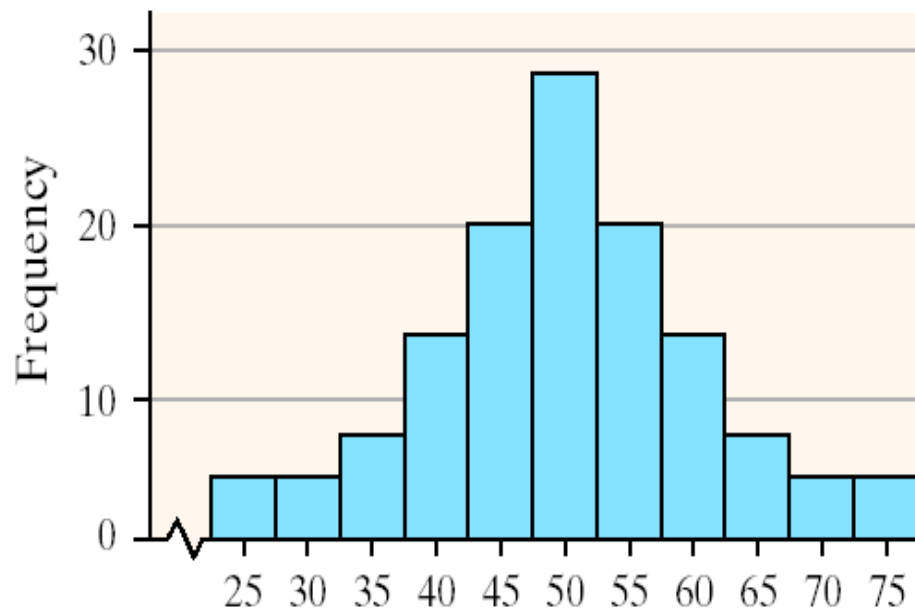
# Skewness of the distribution

positive skew: mean  $>$  median & high-score whisker is longer  
negative skew: mean  $<$  median & low-score whisker is longer

# Distribution Shape Based Upon Boxplot

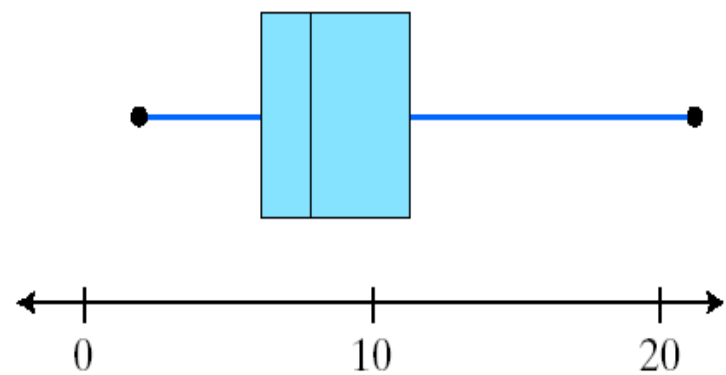
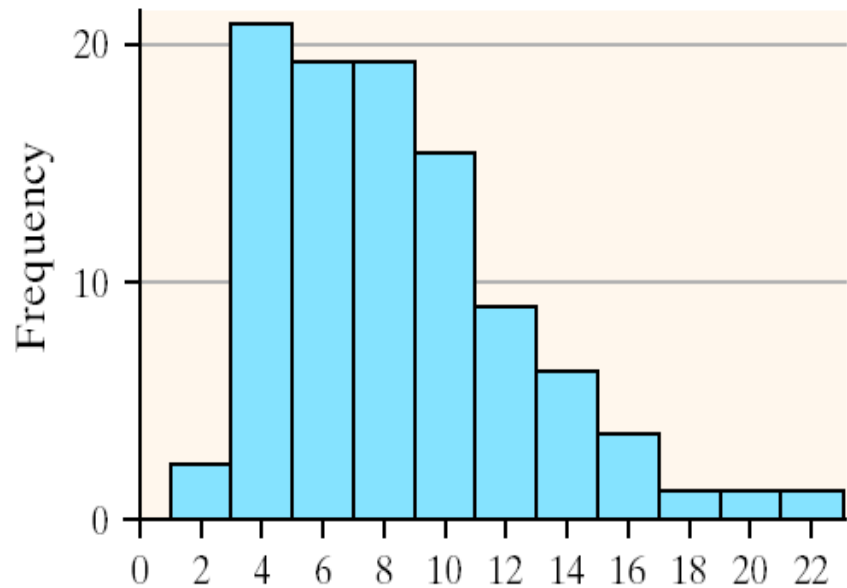
## 1. Symmetry

If the median is near the center of the box and each of the horizontal lines have approximately equal length, then the distribution is roughly symmetric



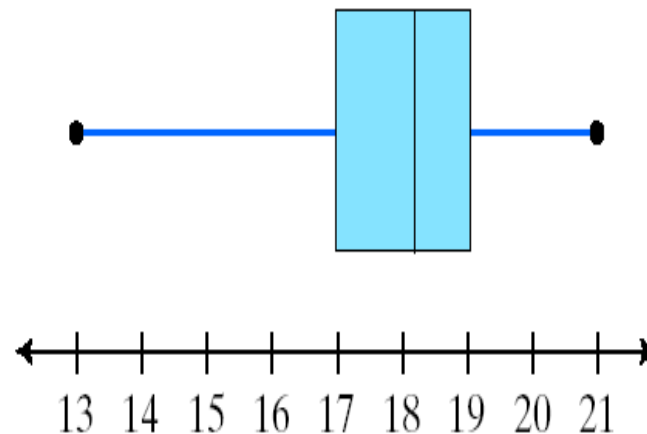
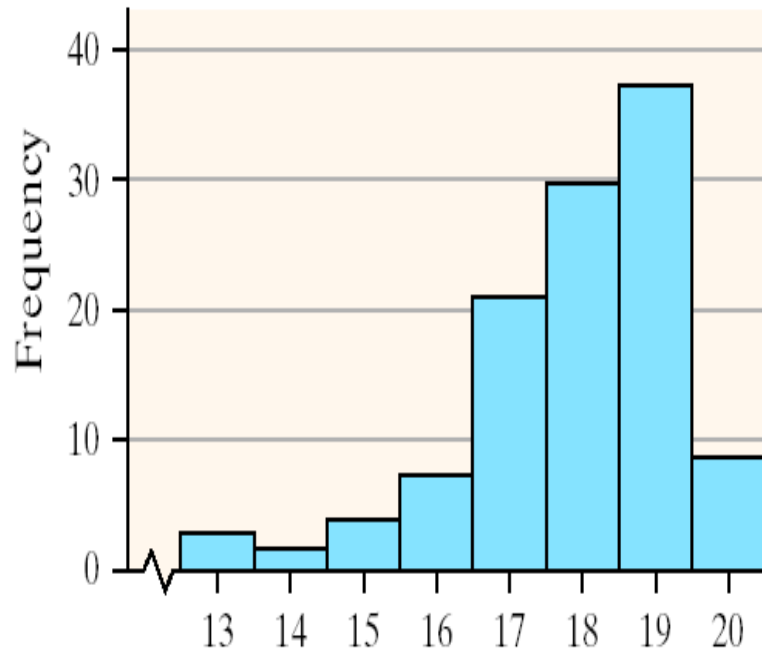
## 2. Right Skewed

If the median is left of the center of the box and/or the right line is substantially longer than the left line, the distribution is skewed to the right.



## Left skewed:

If the median is right of the center of the box and/or the left line is substantially longer than the right line, the distribution is skewed to the left.



## Exercise 1

Consider the following observations

63 58 28 65 55 83 57 61 61.

Find the mean, median, mode, variance, standard deviation, coefficient of variation, interquartile range, 90<sup>th</sup> percentile.

Construct a boxplot and find the outliers.

## Exercise 2

The following data represent the weight lost by 15 members of a club at the end of two months after joining the club.

5 10 8 7 25 12 5 14 11 10 21 9 8 11 18.

Construct a boxplot and find the outliers (if any).

## Exercise 3:

The following data represents serum cholesterol level of 49 individuals of physical activity 1 taken from table 2.1

Serum cholesterol level in milligram percent						
199	272	166	239	238	223	190
209	171	147	199	255	199	228
240	192	201	203	243	186	165
239	162	246	234	161	298	211
219	179	212	231	185	180	205
219	221	216	195	173	206	215
176	234	204	225	187	290	218

1. Prepare a frequency table



2. Find the mean, median and mode class.  
and describe the distribution
3. Find the three quartiles.
4. Find 90<sup>th</sup> percentile.
5. Find 10<sup>th</sup> percentile.
6. Construct a boxplot and find the outliers if any.

# Chapter 4

# Probability

Before we can move from descriptive statistics to inferential statistics, we need to have some understanding of probability.

# Counting Rules

## Rule 1. Fundamental Counting Principle or Multiplication Counting Rule

If an experiment consists of  $k$  steps and if the 1<sup>st</sup> step can result in  $m_1$  outcomes, the 2<sup>nd</sup> can result in  $m_2$  outcomes and the  $k^{\text{th}}$  step in  $m_k$  outcomes, then the total number of outcomes is  $m_1 \cdot m_2 \cdot \dots \cdot m_k$ .

### Examples:

1. Tossing a coin 3 times, the total number of outcomes is  $2 \cdot 2 \cdot 2 = 8$

2. Rolling 4 dice, the total number of outcomes is  $6 \cdot 6 \cdot 6 \cdot 6 = 6^4$

3. In a class of 20 the no. of ways for selecting president, vice-president, secretary, and treasurer is

$$20 \cdot 19 \cdot 18 \cdot 17 = 116280$$

4. If you had three different diet (D) choices by amount of protein (low, medium, high) and three different choices by amount of fat (low, medium, high),

there would be  $(n_1)(n_2) = (3)(3) = 9$  different *diets* :

$D_1$ : protein (low), fat (low)

$D_4$ : protein (low), fat (medium)

$D_2$ : protein (medium), fat (low)

$D_5$ : protein (medium), fat (medium)

$D_3$ : protein (high), fat (low)

$D_6$ : protein (high), fat (medium)

$D_7$ : protein (low), fat (high)

$D_8$ : proteins (medium), fat (high)

$D_9$ : protein (high), fat (high)

# Examples

1. How many samples of size 5 can be selected from a population of size 12?

**Answer:**

$$\text{No. of samples} = C(12, 5) = \frac{12!}{5! \cdot 7!} = 792$$

2. An English department at a university has 16 faculty members. Two members will be selected at random to represent the department. In how many ways can they be selected?

**Answer:**

$$\text{No. of ways} = C(16, 2) = \frac{16!}{2! \cdot 14!} = 120$$

3. Suppose that three patients with snakebites are brought to a physician. To his regret, he discovers that he has only two doses of antivenin. The three patients are a pregnant woman (w), a young child (c), and an elderly man (m). Before deciding which two to treat, he examines his choices:

$$C(3,2) = \frac{3!}{2!(3-2)!} = \frac{3 \cdot 2 \cdot 1}{2 \cdot 1} = 3$$

The three choices are **wc**, **wm**, **cm**.

Note that cw, mw, and mc are the same as the first three because order does not matter.

# Examples

4. A team comprising 6 people is to be selected randomly from 7 females and 8 males. Find the no. of possible selections if
- No conditions imposed
  - Half of the team is females.
  - The team must have at most two males

Answer: No. of selections is

a.  $C(15,6)=455$

b. There are  $n_1 = C(7,3) = 35$  ways to select 3 females of 7, there are  $n_2 = C(8,3) = 56$  ways to select 3 males of 8, so the total no. of selections =  $C(7,3).C(8,3)=35 \times 56=1960$

c.  $C(8,0).C(7,6)+C(8,1).C(7,5)+C(8,2).C(7,4) = 1155$



## Rule 3: Permutations

There are basically two types of permutations:

1. No repetition
2. Repetition is allowed

A permutation with no repetition  $p(n, r)$  is the no. of ways to select  $r$  different objects from  $n$  objects is defined by

$$p(n, r) = \frac{n!}{(n - r)!}$$

## Examples

1. If we wish to identify vials of a medication by using three different symbols, x, y, and z, how many different ways can the vials be identified?

The answer is

$$P(3,3)=3! =3.2.1=6$$

The six different identifications are

xyz, xzy, yxz, yzx, zxy, and zyx.

2. In a group of 10 people, a \$20, \$10, and \$5 prize will be given. In how many ways can the prizes be distributed

The answer is

$$P(10,3) = \frac{10!}{7!} = 10 \times 9 \times 8 = 720 \text{ ways}$$

# Permutations with Repetition

If you have  $n$  things to choose from, and you choose  $r$  of them, then the permutations are:

$$\underbrace{n \cdot n \cdot n \dots n}_{r\text{-times}} = n^r$$

**Example:** If there are 10 numbers to choose from  $\{0,1,..9\}$  and you choose 3 of them with repetition. Then, we have

$$10 \cdot 10 \cdot 10 = 10^3 = 1000 \text{ permutations.}$$

## Examples on combinations and permutations

1. In a group of 10 people, three \$5 prizes will be given. In how many ways can the prizes be distributed

**Answer:  $C(10,3) = \frac{10!}{3! \cdot 7!} = 120$**

2. In a group of 10 people, a \$20, \$10, and \$5 prize will be given. In how many ways can the prizes be distributed

**The answer is**

$$P(10,3) = \frac{10!}{7!} = 10 \times 9 \times 8 = 720 \text{ ways}$$

3. A local school board with 8 people needs to form a committee with 3 people. In how many ways can this committee be formed.

$$\text{Answer: } C(8,3) = \frac{8!}{3! \cdot (8-3)!} = 56$$

4. A local school board with 8 people needs to form a committee with 3 people, with 3 different responsibilities. In how many ways can this committee be formed

$$\text{Answer : } P(8,3) = \frac{8!}{(8-3)!} = 336.$$

5. If there are three effective ways of treating a cancer

## A random experiment:

- an experiment which can be repeated any number of times under the identical conditions
- The set of all possible outcomes is known in advance
- The outcome of a particular case is not known in advance

### Examples:

- Tossing a coin. Record outcome as “head” or “tail”
- Rolling a die. Record outcome as the number of spots facing up
- Performing a surgical operation. Record the outcome as “failure” or “success”
- Corona virus disease .Record outcome as “recovered”, “confirmed case” or “dead”
- Record outcome as from 7-20 gm/dl approximately  
Measures of Hemoglobin concentration in blood.

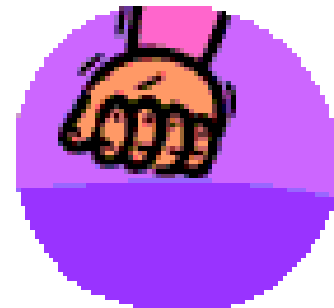
# Sample Space

The sample space **S** of statistical experiment is the set of all possible distinct outcomes of an experiment.

## Examples:

- consider a set of six balls numbered 1, 2, 3, 4, 5, and 6. If we put the six balls into a bag and without looking at the balls, we choose one ball from the bag, then, this is an experiment which has 6 possible outcomes i.e.

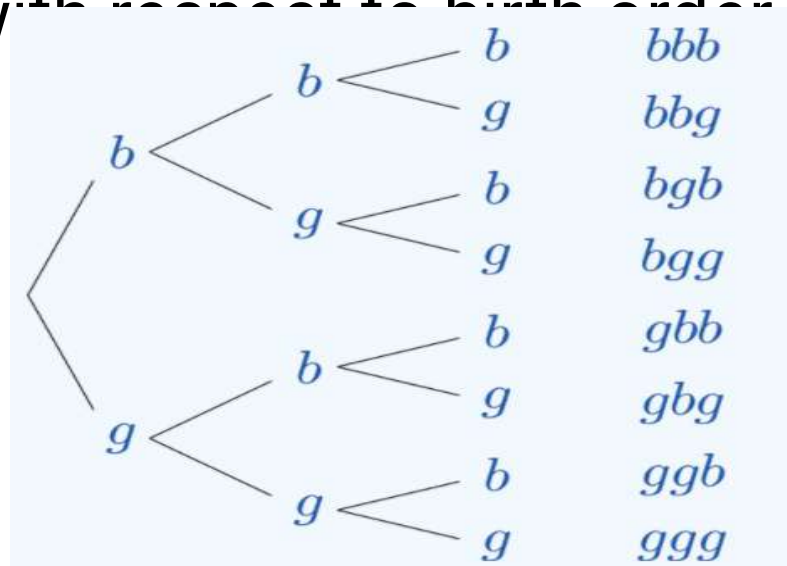
$$S = \{1, 2, 3, 4, 5, 6\}$$



- For a football game,  $S = \{\text{win, loss, tie}\}$ ,
- For a surgical operation,  $S = \{\text{success, failure}\}$
- For a coin toss,  $S = \{\text{head(H), tail(T)}\}$ ,
- For tossing 3 coins,  $S = \{\text{HHH, HHT, THH, HTH, TTH, THT, HTT, TTT}\}$
- For rolling two dice,  
 $S = \{(1,1), (1,2), \dots, (1,6), (2,1), (2,2), \dots, (2,6), \dots, (6,1), (6,2), \dots, (6,6)\}$
- For the number of heads in 4 tosses of a coin,  
 $S = \{0, 1, 2, 3, 4\}$
- For the number of females in a family of 3 children  
 $S = \{0, 1, 2, 3\}$
- For tossing a coin until a head appear and record the number of tosses  
 $S = \{1, 2, 3, 4, 5, \dots\}$



- A device that can be helpful in identifying all possible outcomes of a random experiment is what is called a **tree diagram**. It is described in the following example.
- Construct a sample space that describes all three - child families according to the genders of the children with respect to birth order.



$$S = \{bbb, bbg, bgb, bgg, gbb, gbg, ggb, ggg\}$$

## Event

An event is **a subset of the sample space** which all elements have some specified characteristic.

An event that includes one and only one of the (final) outcomes for an experiment is called a **simple event** and is denoted by  $E_i$ .

**Example**, when rolling a die,

$$S = \{1, 2, 3, 4, 5, 6\}$$

we might consider the following events

$A_1$  = the event of getting an even number =  $\{2, 4, 6\}$

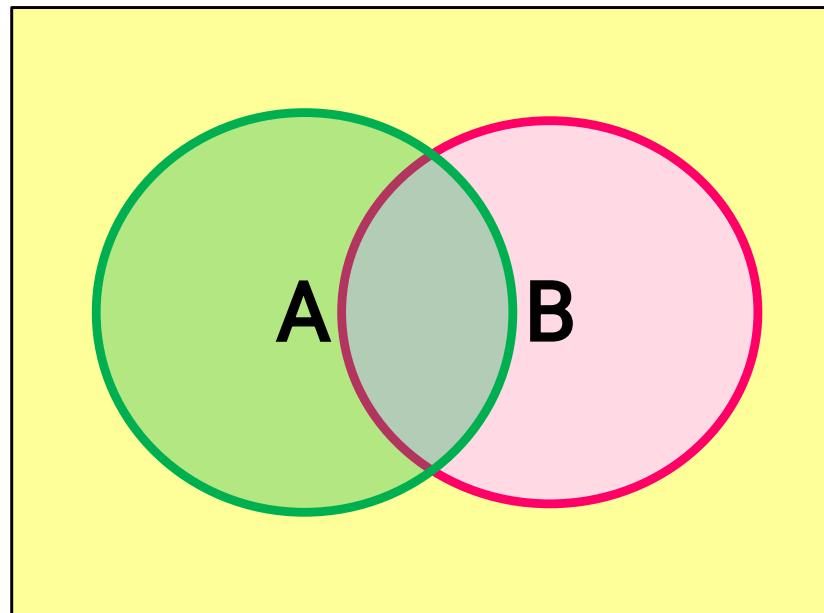
$A_2$  = the event of getting a number less than 4 =  $\{1, 2, 3\}$

$A_3$  = the event of getting an odd number =  $\{1, 3, 5\}$

$A_4$  = the event of getting an odd number greater than or equal 3 =  $\{3, 5\}$

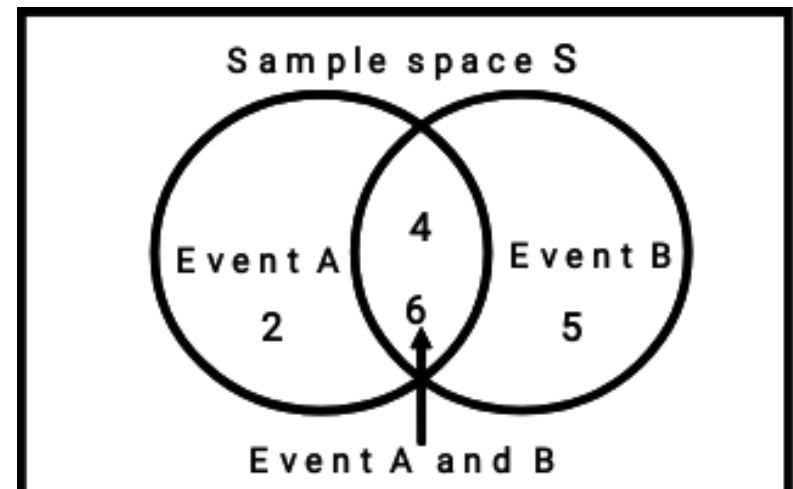
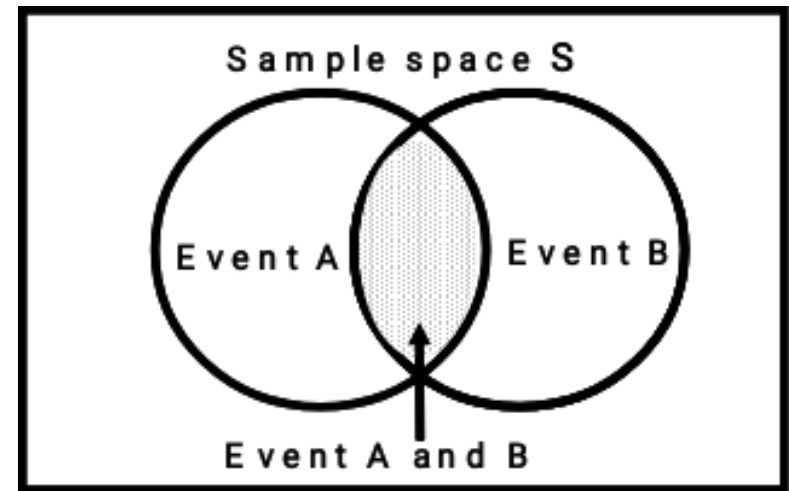
# The intersection of two events

- The intersection of two events A and B, denoted by  $(A \text{ and } B)$  or  $A \cap B$ , is the event containing all elements that are *common* to A and B.



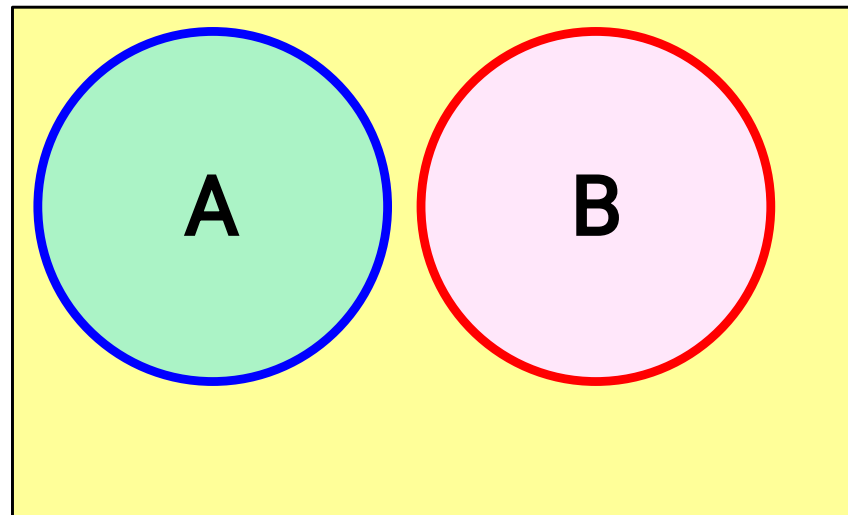
# Example

- Suppose that a die is tossed. Let A and B are events of the sample space  $S = \{1, 2, 3, 4, 5, 6\}$  where;
  - A = the event that an **even** number occurs =  $\{2, 4, 6\}$
  - B = the event that a number **greater than 3** =  $\{4, 5, 6\}$
- Then the event containing all **even** numbers (event A) that are **greater than 3** is  $\{4, 6\}$ , which is just the intersection of A and B or  $(A \cap B)$ .



# Mutually exclusive events

- Two events A and B are mutually exclusive if they cannot occur together (Simultaneously).
- Or if the event  $(A \cap B)$  contains no elements.



## Examples (Disjoint events)

- In rolling a die, consider the events A and B are subsets of the same sample space

$S = \{1, 2, 3, 4, 5, 6\}$  where;

A = the event that an **even** number occurs =  $\{2, 4, 6\}$

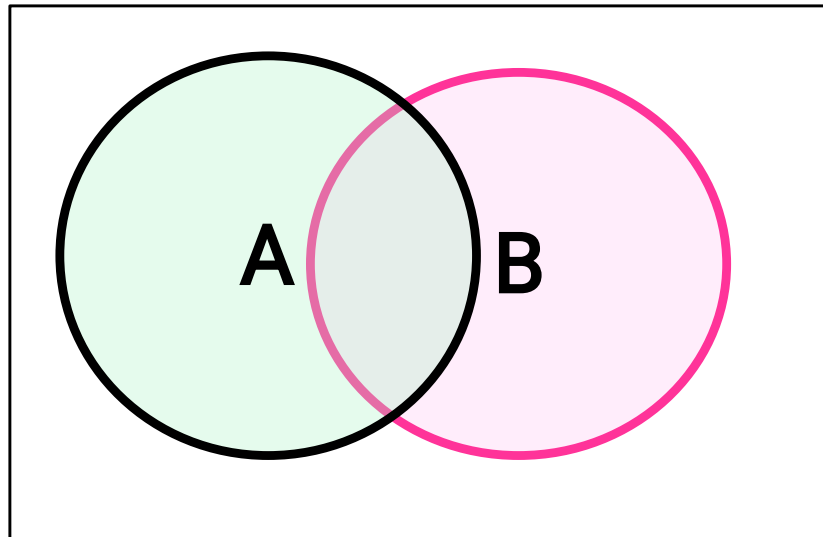
B = the event that an **odd** number occurs =  $\{1, 3, 5\}$

Therefore, the event (A and B) contain no elements, and subsequently the events A and B are **mutually exclusive** or **disjoint**.

- The event of getting grade A and the event of getting grade B in statistics quiz are disjoint.
- The event of having positivity for Covid-19 and that of having negativity

# The union of two events

The union of two events  $A$  and  $B$ , denoted by  $(A \text{ or } B)$  **OR**  $(A \cup B)$ , is the event containing all the elements that belong to  $A$  or to  $B$  or both.



# Example

- Let  $A$  be the event that a patient selected at random has a **hypertension**. Let  $B$  be the event that the patient selected has **high glucose** level. Then the event  $(A \text{ or } B) = (A \cup B)$  is the set of patients who **either** have a hypertension **or** have high glucose level, or who have **both**.



# Example

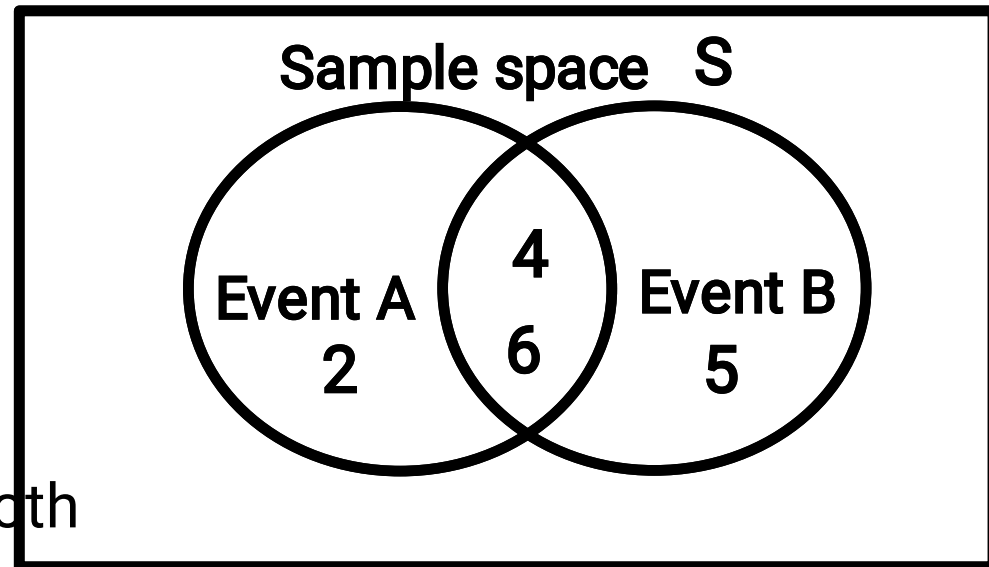
In tossing a die, consider the events A and B are subsets of the same sample space

$S = \{1, 2, 3, 4, 5, 6\}$  where;

A is the event that an even number occurs.  $A = \{2, 4, 6\}$

B is the event that a number greater than 3.  $B = \{4, 5, 6\}$

Then the event containing all the elements that belong to A or to B or both is  $A \cup B = \{2, 4, 5, 6\}$



# Probability of an event

- Probability of an event is a numerical measure of the likelihood that the event will occur
- Two Properties of Probability
  1. The probability of an event always lies in the range 0 to 1.  
 $0 \leq p(A) \leq 1$
  2. The sum of the probabilities of all simple events for an experiment, denoted by  $\Sigma P(E_i)$ , is always 1.  
$$P(E_1) + P(E_2) + \dots + P(E_n) = 1$$
- Two or more outcomes that have the same probability of occurrence are said to be equally likely outcomes

# Equally Likely Events

- Getting a 3 on the toss of a fair die and getting a 5 on the toss of a die are equally likely events.
- Getting an even number on the toss of a fair die and getting an odd number on the toss of a die are equally likely events.
- When throwing a matchbox, all the faces not equally likely.
- Getting a 3 on the toss of a loaded die and getting a 5 on the toss of a die are not equally likely events.

## Classical (Theoretical ) Probability Rule

$$P(E_i) = \frac{1}{\text{Total number of outcomes for the experiment}}$$

$$P(A) = \frac{\text{Number of outcomes favorable to } A}{\text{Total number of outcomes for the experiment}}$$

Note:

- It uses sample space to determine the numerical probability that an event will occur.
- We don't actually perform the experiment to determine the
- Theoretical probability.
- It assumes that all outcomes are equally likely to occur.

# Example

***Tow coins:*** when tossing two coins, four outcomes are possible:

		Second coin	
		H	T
First coin	H	HH	HT
	T	TH	TT

What is the probability of

- Flipping two heads?
- At least one head?
- No heads?
- One head and one tail?
- Not more than one tail?

## Answer:

- Flipping two heads

$$P(\{HH\})=1/4$$

- At least one head

$$P(\{HT,TH,HH\})=3/4$$

- No heads

$$P(\{TT\})=1/4$$

- One head and one tail

$$P(\{HT, TH\}) = 2/4 = 1/2$$

- Not more than one tail

$$P(\{HH,TH,HT\})=3/4$$

## Relative Frequency Concept of Probability

### Using Relative Frequency as an Approximation of Probability

If an experiment is repeated  $n$  times and an event  $A$  is observed  $f$  times where  $f$  is the frequency, then, according to the relative frequency concept of probability:

$$P(A) = \frac{f}{n} = \frac{\text{Frequency of } A}{\text{Sample Size}}$$

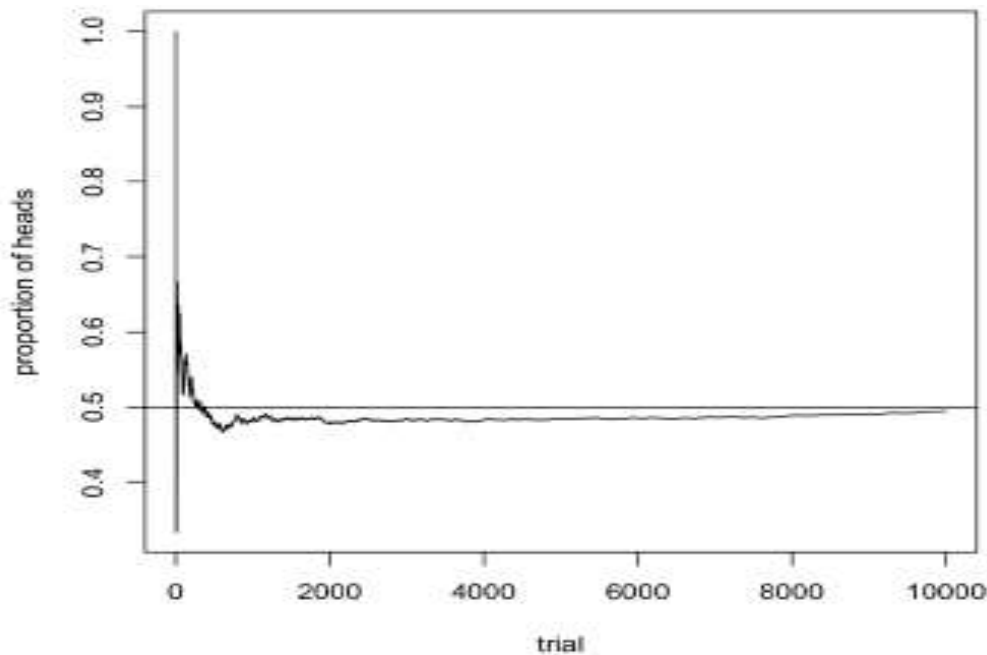
**Example:** The table represents the status of 200 people. If one is selected at random, what is the probability of positivity result and what is the probability of negativity result

Results for <i>Covid -19</i>	
Positive	59
Negative	151

$$p(\text{positive}) = \frac{59}{200} = 0.295$$
$$p(\text{negative}) = \frac{151}{200} = 0.705$$

- This is called the empirical probability
- as the number of trials increases the empirical probability gets closer to the theoretical (true) probability.

**Example:** Proportion of times a fair coin comes up as a “head”





**Subjective Probability** – uses a probability value based on an educated guess or estimate, employing opinions and inexact information.

Often, you cannot “repeat” the probability experiment.

**Example:** What is the probability you will pass this class?

**Example:** What is the probability that you will get a certain job when you apply?

# Notes

1. The term probability applies exclusively to a future event, never to a past event (even if its outcome is unknown).
2. Probability of event should be defined in the range of 0 to 1, never more and never less.
3. A probability of 1.0 means that the event will happen with certainty;  
A probability of 0 means that the event will not happen.  
If the probability is 0.5, the event should occur once in every two attempts on the average.

# Properties of probability for mutually exclusive events

1. The probability of an event is always between 0 and 1, it is never negative and never greater than 1.

$$0 \leq P(E_i) \leq 1$$

2. The sum of the probabilities of all mutually exclusive outcomes of an experiment is equal to 1.

$$P(E_1) + P(E_2) + \dots + P(E_n) = 1$$

**Example:** In a football game

$$P(\text{win}) + P(\text{loss}) + P(\text{tie}) = 1$$

In rolling a die

$$P(1) + P(2) + \dots + P(6) = 1$$

# Conditional Probability

- Conditional probability is denoted at  $P(A|B)$ . It is the probability that A occurs, **given that** B has occurred, and is given by the following ratio:

$$P(A | B) = \frac{P(A \text{ and } B)}{P(B)}$$

providing  $P(B)$  is not equal to zero

The vertical line in  $P(A|B)$  is read “given”

# Independent events

Two events are said to be independent if the occurrence of one event does not affect the occurrence of the other.

e.g. the events

- “Blood glucose of Ahmad is 120 mg percent” and “Blood glucose of Omar is 200 mg percent” are independent events.  
Also “ Ahmad is allergic to penicillin” and “ Omar is allergic to penicillin” are independent events
- The outcomes of repeated tosses of a coin illustrate independent events, for the outcome of one toss does not affect the outcome of any future toss.

# Example

- Suppose the table represents the no. of females and males in a statistics class of 45 students who are in favour of or against 2<sup>nd</sup> midterm exam

	Against (A)	In favour of (V)	
Male (M)	15	5	20
Female (F)	8	17	25
Total	23	22	45

If a student is selected at random, what is the probability that he is

1. Female.
2. Against exam.
3. Female and against exam.
4. Female given that he is against exam.

**Answer:**  $\frac{25}{45} = 0.556$

1.  $P(F) = \frac{23}{45} = 0.511$

2.  $P(A) = \frac{8}{45}$

3.  $P(F \cap A) = \frac{8}{45} = 0.178$

4.  $P(F|A) = \frac{8}{23} = 0.348$  **OR**  $P(F|A) = \frac{0.178}{0.511} = 0.348$

**Note**  $P(F|A) \neq P(F)$ . So  $F$  and  $A$  are dependent.

# Probability Rules

## Multiplication rule

If  $A$  and  $B$  are any two events, then

$$P(A \text{ and } B) = P(A \cap B) = P(A)P(B|A)$$

If  $A$  and  $B$  are **independent**, then  $P(B|A) = P(B)$  and

$$P(A \cap B) = P(A)P(B)$$

i.e., equal to the product of the probabilities of the two events.



# Example

In tossing two coins, what is the probability that a head will occur both on the first coin  $H_1$  and on the second coin  $H_2$

Answer:

$$P(H_1 \text{ and } H_2) =$$

$$P(H_1 \cap H_2) = P(H_1)P(H_2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

# Example

- If  $P(A) = 0.60$  ,  $P(B) = 0.50$  and  $P(A \cap B) = 0.10$ , determine whether  $A$  and  $B$  are independent.

**Answer:**

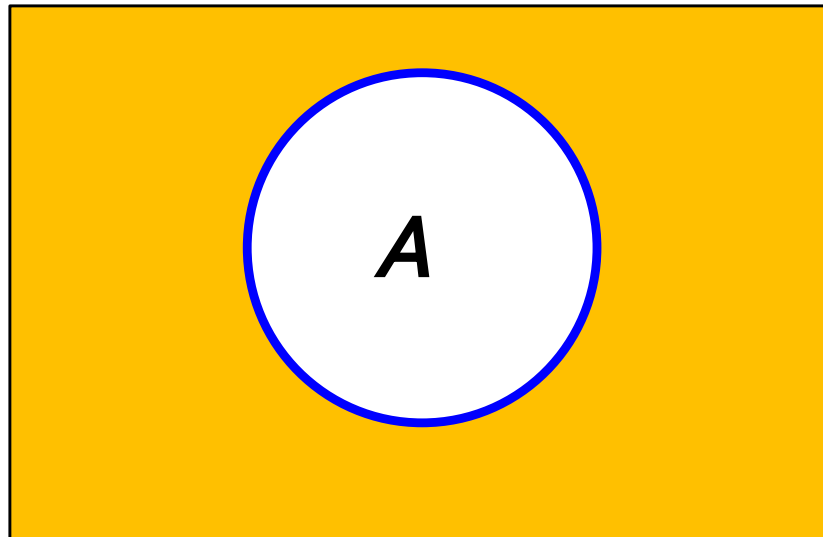
$$P(A) \cdot P(B) = 0.30 \neq P(A \cap B).$$

- So  $A$  and  $B$  are dependent

# Complementary events

The **complement** of an event  $A$ , denoted by  $\bar{A}$  is defined by  $A \cup \bar{A} = S$

And so  $P(\bar{A}) = 1 - P(A)$



# Example (independent events)

- The probability that a patient is allergic to penicillin is 0.20. Suppose the drug is administered to 3 patients, find the probability that
  1. All are allergic.
  2. None is allergic.
  3. At least one is allergic.
  4. Exactly 2 are allergic.

**Answer:**

Let  $A_1$ ,  $A_2$ , and  $A_3$  be the events that the 1<sup>st</sup>, the 2<sup>nd</sup> and the 3<sup>rd</sup> patients are allergic to penicillin, respectively

## Answer:

$$\begin{aligned} 1. P(A_1 \cap A_2 \cap A_3) &= P(A_1 A_2 A_3) \\ &= P(A_1) \cdot P(A_2) \cdot P(A_3) \\ &= (0.20) \cdot (0.20) \cdot (0.20) \\ &= 0.008 \end{aligned}$$

$$\begin{aligned} 2. P(\overline{A_1} \overline{A_2} \overline{A_3}) &= P(\overline{A_1}) P(\overline{A_2}) P(\overline{A_3}) \\ &= (0.80)(0.80)(0.80) = 0.512 \end{aligned}$$

$$3. P(\text{at least one is allergic}) = 1 - P(\text{none is allergic})$$

$$\begin{aligned} &= 1 - P(\overline{A_1} \overline{A_2} \overline{A_3}) \\ &= 1 - 0.512 = 0.488 \end{aligned}$$

# Example continued

4.  $P(\text{exactly 2 are allergic}) =$

$$P(A_1 A_2 \bar{A}_3) + P(A_1 \bar{A}_2 A_3) + P(\bar{A}_1 A_2 A_3)$$

$$= (0.20)(0.20)(0.80) + (0.20)(0.80)(0.20) + (0.80)(0.20)(0.80)$$

$$= 0.096$$

## Example (dependent events)

At a hospital, there are 4 boys and 6 girls. If we choose two children without replacement, what is the probability that

1. both are boys
2. Both are girls
3. One is a boy

**Answer:**

Probability of the first child being boy =  $4/10$

1. the probability of both children are boys is

$$P(B_1B_2) = \left(\frac{4}{10}\right) \cdot \left(\frac{3}{9}\right) = \frac{2}{15} = 0.133$$

$$2. P(G_1G_2) = \left(\frac{6}{10}\right) \cdot \left(\frac{5}{9}\right) = \frac{1}{3} = 0.333.$$

$$3. P(B_1G_2) + P(G_1B_2) = \frac{4}{10} \cdot \frac{6}{9} + \frac{6}{10} \cdot \frac{4}{9} = 0.533$$

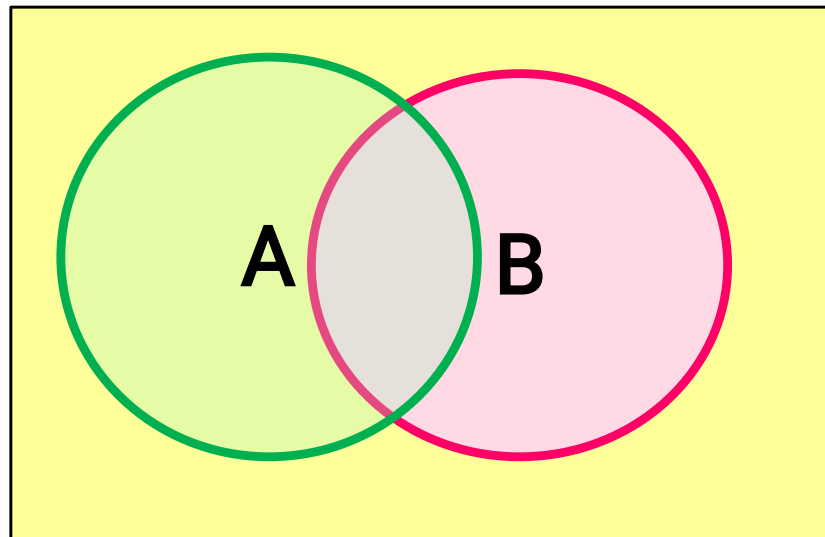


# Addition Rule

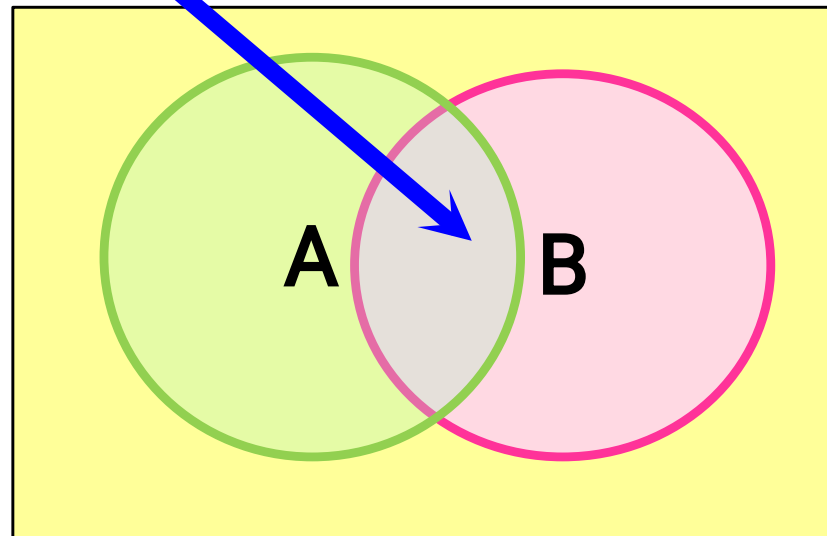
- The addition rule states that the probability that event  $A$  or event  $B$  (or both) will occur equals the sum of the probabilities of each individual event less the probability of both.

For any events  $A$  and  $B$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



- The reason for subtracting  $P(A \cap B)$  is that this portion would otherwise be included **twice**.



# Example

- In flipping two coins, you may wish to know the probability of having a head on the first coin  $H_1$ , or on the second  $H_2$ , or on both  $H_1$  and  $H_2$ .

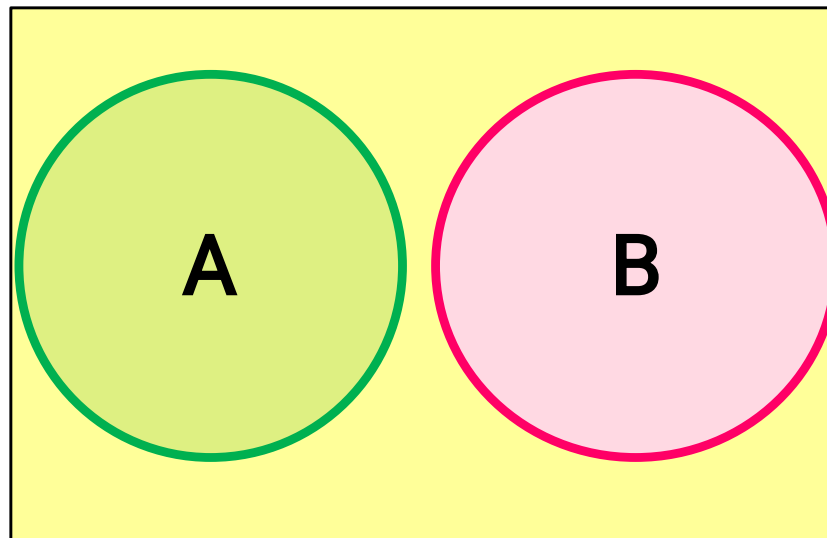
**Answer:** you use the addition rule

$$P(H_1 \cup H_2) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}$$

# Addition Rule For Mutually Exclusive events

For mutually exclusive events  $A$  and  $B$

$$P(A \cup B) = P(A) + P(B)$$



# Example

- In an experiment involving a toxic substance, the probability that a white mouse will be alive for 10 hours is  $\frac{7}{10}$  and the probability that a black mouse will be alive for 10 hours is  $\frac{9}{10}$ . Find the probability that, at the end of 10 hours,
  - a. Both mice will be alive.
  - b. Only the black mouse will be alive.
  - c. At least one will be a live.
  - d. Exactly one will be alive.

## Answer:

Let  $B$  be the event that the black mouse will be a live for 10 hours and

$W$  be the event that the white mouse will be a live for 10 hours. The events are independent

$$\text{a. } P(B \cap W) = P(B)P(W) = \frac{9}{10} \cdot \frac{7}{10} = 0.63$$

$$\text{b. } P(B \cap \bar{W}) = P(B)P(\bar{W}) = \frac{9}{10} \cdot \frac{3}{10} = 0.27$$

C.  $P$  (at least one will be a live)

$$= P(B \cup W) = P(B) + P(W) - P(B \cap W)$$

$$= \frac{9}{10} + \frac{7}{10} - 0.63 = 0.97$$

d.  $P$  ( exactly one will be a live)=

$$P(W \cap \bar{B}) + P(B \cap \bar{W}) = \frac{7}{10} \cdot \frac{1}{10} + \frac{9}{10} \cdot \frac{3}{10} = 0.34$$

# De Morgan's Law

1.  $\bar{A} \cap \bar{B} = \overline{A \cup B}$

2.  $\bar{A} \cup \bar{B} = \overline{A \cap B}$

So ,  $P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B)$

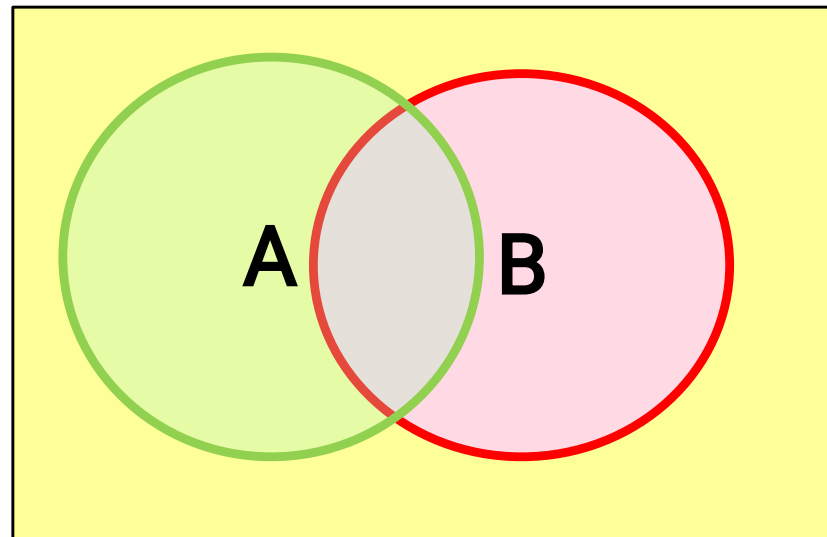
And

$$P(\bar{A} \cup \bar{B}) = P(\overline{A \cap B}) = 1 - P(A \cap B)$$



- $A = (A \cap B) \cup (A \cap \bar{B})$ , so

$$P(A) = P(A \cap B) + P(A \cap \bar{B})$$



# Example

- The probability that an adult is of physical activity 1 is 0.30, the probability that his blood glucose is more than 200 mg(%) is 0.13 and the probability of both is 0.08. if an adult is selected at random,
  - a. what is the probability that he is
    - i. **either** of physical activity1 **or** with blood glucose is more than 200 mg(%) .
    - ii. **Neither** of physical activity1 **nor** with blood glucose is more than 200 mg(%) .
    - iii. Only with blood glucose is more than 200 mg(%) .
  - b. If one is selected at random and it is found that his glucose is more than 200 mg%, what is the probability that he is of physical activity 1.

**Answer:**

Let  $H$  be the event that the adult is of physical activity 1.

Let  $B$  be the event that the blood glucose is more than 200 mg(%). So,

$$P(H) = 0.30 \quad P(B) = 0.13 \quad \text{and} \quad P(H \cap B) = 0.08$$

$$P(H) \cdot P(B) = 0.039 \neq P(H \cap B),$$

So, the events  $H$  and  $B$  are dependent

## Example continued

$$\begin{aligned} \text{a. } P(H \cup B) &= P(H) + P(B) - P(H \cap B) \\ &= 0.30 + 0.13 - 0.08 = 0.35 \end{aligned}$$

$$\text{b. } P(\bar{H} \cap \bar{B}) = 1 - P(H \cup B) = 1 - 0.35$$

$$\text{c. } P(B \cap \bar{H}) = P(B) - P(B \cap H) = 0.13 - 0.08$$

d.  $P(H \text{ given } B) =$

$$P(H | B) = \frac{P(H \cap B)}{P(B)} = \frac{0.08}{0.13} = 0.615$$

***Exercise:***

In a cafeteria, 80% of the customers order chips and 60% order buns. If 20% of those ordering buns do not want chips, find the probability that two customers chosen at random,

- i. both order chips but not buns.
- ii. exactly one of them orders a bun only.

***Exercise:***

In a cafeteria, 80% of the customers order chips and 60% order buns. If 20% of those ordering buns do not want chips, find the probability that two customers chosen at random,

- i. both order chips but not buns.
- ii. exactly one of them orders a bun only.

# Exercise

A hotel owner has determined that 83% of the hotel's guests eat either dinner or breakfast in the hotel restaurant. If 30% of the guests eat dinner and 60% eat breakfast.

- a. What is the proportion of the guests eat:
  - i. both breakfast and dinner.
  - ii. Neither dinner nor breakfast.
  - iii. Dinner and not breakfast.
- b. Determine whether the events “eat dinner” and “eat breakfast” are independent.
- c. If a guest is selected at random and it is found that he has eaten breakfast, what is the probability that he will eat dinner.

# Chapter 5

## Probability Distribution



# Random variables

- A random variable is a **numerical** outcome of a random process or random event
- Examples:
  1. three tosses of a coin
    - $S = \{HHH, THH, HTH, HHT, HTT, THT, TTH, TTT\}$
    - Random variable  $X =$  number of observed tails
    - Possible values for  $X = \{0, 1, 2, 3\}$
  2. If 4 balls are drawn without replacement from a box contains 4 red and 6 blue
    - Random variable  $X =$  number of red balls
    - Possible values for  $X = \{0, 1, 2, 3, 4\}$
  3. If 5 balls are drawn with replacement from a box contains 4 red and 6 blue
    - Random variable  $X =$  number of red balls
    - Possible values for  $X = \{0, 1, 2, 3, 4, 5\}$

# Types of Random Variables

1. A discrete random variable
2. A continuous random variable

**A discrete random variable is** a quantitative random variable that can take on only a finite number of values or a countable number of values.

## **Examples:**

- The number of children per family
- The number of cavities a patient has in a year.
- The number of bacteria which survive when treatment with some antibiotic.
- The number of times a person had a cold in Gaza Strip.
- The number of accidents in a certain way during a week.

## Continuous Random Variable

A continuous random variable is a quantitative random variable that can take infinite number of values within an interval

### Example:

- The time taken by a student to complete an exam
- The amount of rainfall in during the month of January
- The starting salaries of all college graduates with a computer science
- Intelligent quotient of students in a class.
- Stress score of a student before statistics exam

# Probability distribution

- **A probability distribution** is the listing of all possible outcomes of an experiment and the corresponding probability.
- Depending on the variable, the probability distribution can be classified into:
  1. Discrete probability distribution
  2. Continuous probability distribution

# Discrete probability distribution

## A discrete probability distribution

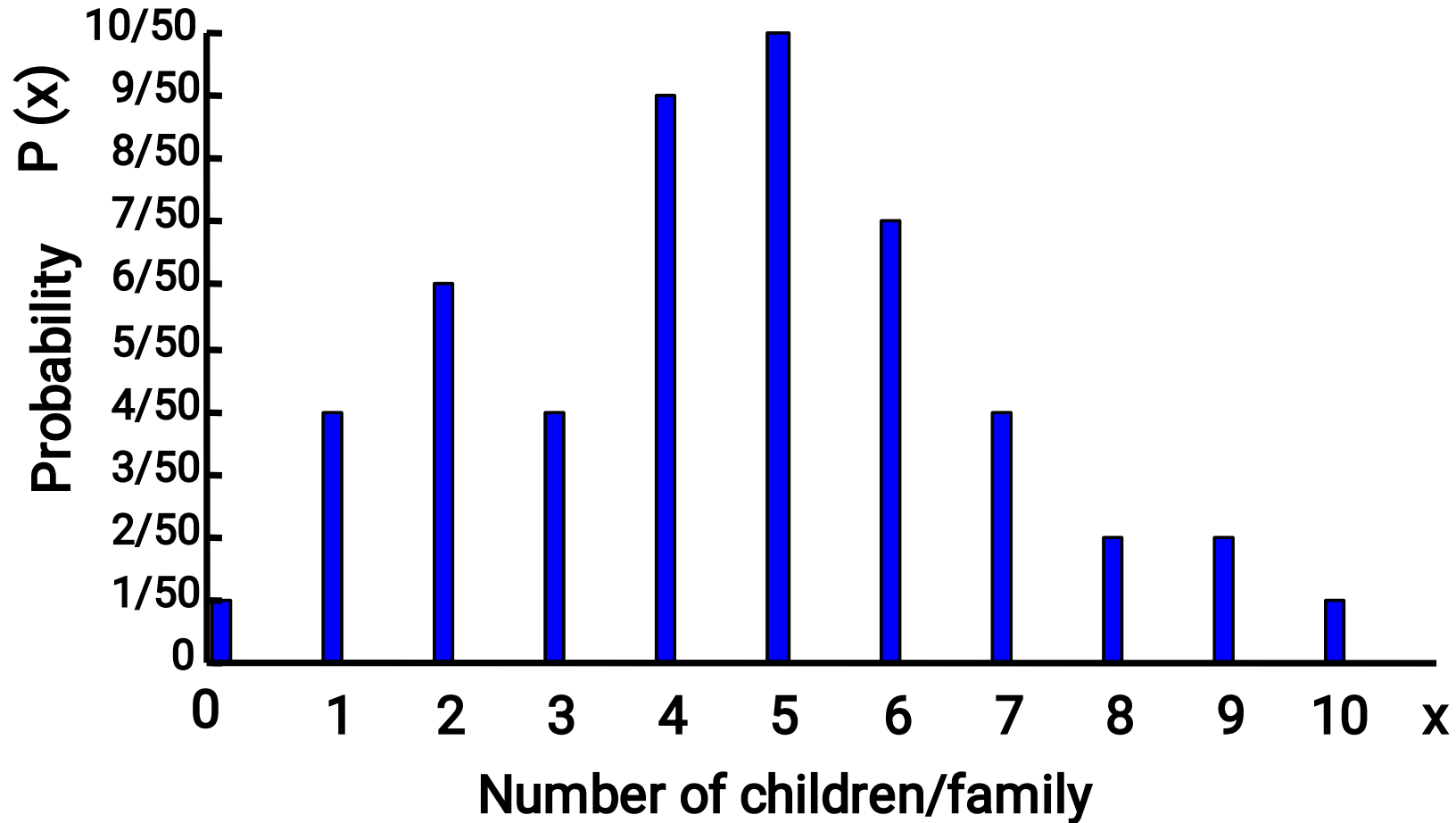
is a table, graph, formula, or other device used to specify all possible values of a discrete random variable along with their respective probabilities.

# Discrete probability distribution

$x$	Frequency of occurring of $x$	$P(X=x)$
0	1	1/50
1	4	4/50
2	6	6/50
3	4	4/50
4	9	9/50
5	10	10/50
6	7	7/50
7	4	4/50
8	2	2/50
9	2	2/50
10	1	1/50
	50	50/50

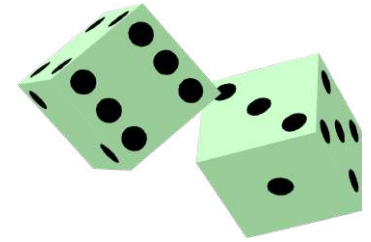
Probability distribution of number of children per family in a population of 50 families

# Discrete probability distribution



Bar chart Graphical representation of the Probability distribution of number of children per family for population of 50 families

# ((Two Dice Example



- Random variable  $X$  = the sum of two dice  
 $X$  takes on values from 2 to 12
- Use “equally-likely outcomes” rule to calculate the probability distribution:

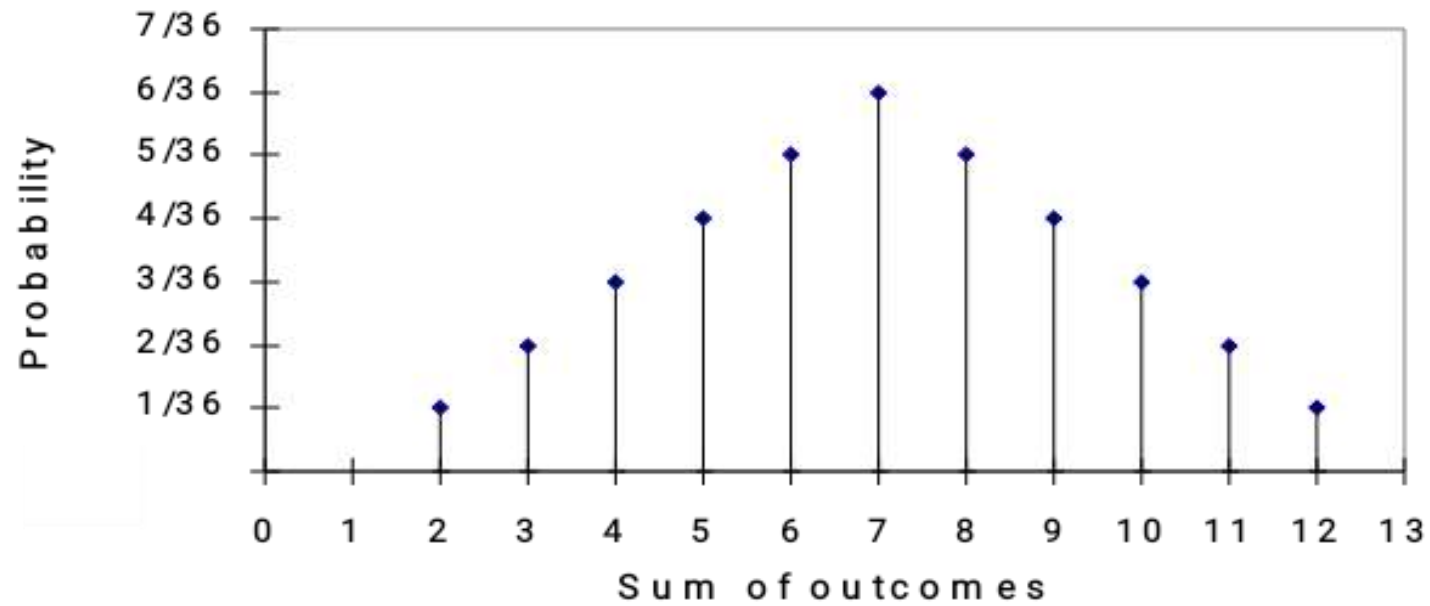
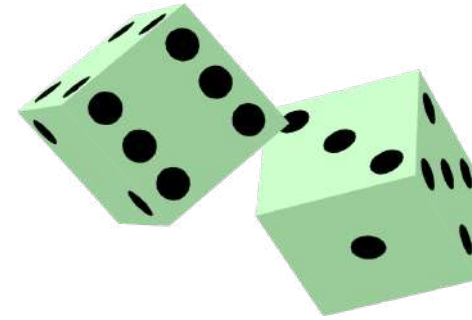
$X$	2	3	4	5	6	7	8	9	10	11	12
# of Outcomes	1	2	3	4	5	6	5	4	3	2	1
$P(X)$	$1/36$	$2/36$	$3/36$	$4/36$	$5/36$	$6/36$	$5/36$	$4/36$	$3/36$	$2/36$	$1/36$

Note that  $\sum P(X) = 1$

$$P(\text{sum} > 10) = P(\text{sum} = 11) + P(\text{sum} = 12) = 3/36$$



# Two Dice



# Examples of discrete probability distribution

Toss of Two Coins		Roll of a Die		Sex of Three-child Family	
E	$P(E)$	E	$P(E)$	E	$P(E)$
HH	1/4	1	1/6	3 boys	0.125
HT	1/4	2	1/6	2 boys, 1 girl	.375
TH	1/4	3	1/6	1 boy, 2 girls	.375
TT	1/4	4	1/6	3 girls	.125
	1.0	5	1/6		1.000
		6	1/6		
			1.0		

# Example

- A sample of two students is selected at random from a group of 4 males and 7 females. Let  $X$  be the number of males in this sample. List the probability distribution of  $X$ .
- Answer:  $X=0,1,2$

$$\begin{aligned}p(X=0) &= p(F_1F_2) \\ &= p(F_1)p(F_2 | F_1) \\ &= (7/11)(6/10) = 42/110.\end{aligned}$$

$$\begin{aligned}P(X=1) &= p(M_1F_2) + p(F_1M_2) \\ &= (4/11)(7/10) + (7/11)(4/10) \\ &= 56/110.\end{aligned}$$

$$\begin{aligned}P(X=2) &= p(M_1M_2) \\ &= (4/11)(3/10) = 12/110\end{aligned}$$

X	P(X)
0	42/110
1	56/110
2	12/110
	1

# Properties of a Discrete Distribution

**The main properties of a discrete probability distribution are:**

The probability of a particular outcome,  $P(X_i)$ , is between 0 and 1.00.

The sum of the probabilities of the various outcomes is 1.00.

That is,

$$P(X_1) + \dots + P(X_N) = 1$$

The outcomes are mutually exclusive. That is,

$$P(X_1 \text{ and } X_2) = 0 \text{ and}$$

$$P(X_1 \text{ or } X_2) = P(X_1) + P(X_2)$$

**Example:** Check whether the function given by

$$f(x) = \frac{x+2}{25}, x = 1,2,3,4,5,$$

can serve as the probability distribution of a discrete random variable.

- $f(x) > 0$  for  $x = 1,2,3,4,5$
- $f(1) + f(2) + f(3) + f(4) + f(5) = 1$

# Summary Measures

## 1. Mean (Expected value)

Mean of probability distribution

$$E(X) = \mu = \sum X P(X)$$

## 2. Variance

$$\sigma^2 = \sum (X - \mu)^2 P(X) = \sum X^2 P(X) - \mu^2$$

## 3. Standard deviation:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum X^2 P(X) - \mu^2}$$

# Discrete probability distribution

## Example

You toss 2 coins. You're interested in the number of tails. What are the expected value & standard deviation of this random variable, number of tails?

- Finding the mean:

$X$	$p(X)$	$X \cdot p(X)$
<b>0</b>	<b>0.25</b>	<b>0</b>
<b>1</b>	<b>0.50</b>	<b>0.50</b>
<b>2</b>	<b>0.25</b>	<b>0.50</b>
$\sum X \cdot p(X) =$		<b>1.0</b>

$$\mu = \sum X \cdot P(X) = 1.0$$

## Example:

You toss 3 coins. You're interested in the number of tails. What are the expected value & standard deviation of this random variable, number of tails?

### Finding the mean:

Let  $X$  be the number of tails

$$X = 0, 1, 2, 3$$

$$P(X = 0) = P(\text{HHH}) = 1/8$$

$$\begin{aligned} P(X = 1) &= P(\text{TTH}) + P(\text{THT}) + P(\text{HTT}) \\ &= 3/8 \end{aligned}$$

$$\begin{aligned} P(X = 2) &= P(\text{HTH}) + P(\text{HTT}) + P(\text{THT}) \\ &= 3/8 \end{aligned}$$

$$P(X = 3) = P(\text{TTT}) = 1/8$$

$$\mu = \sum X P(X) = 1.5.$$

$X$	$P(X)$	$X P(X)$
<b>0</b>	<b>1/8</b>	<b>0</b>
<b>1</b>	<b>3/8</b>	<b>3/8</b>
<b>2</b>	<b>3/8</b>	<b>6/8</b>
<b>3</b>	<b>1/8</b>	<b>3/8</b>
$\sum X P(X) =$		<b>12/8 = 1.5</b>



## Finding the standard deviation:

x	P(x)	XP(x)	X <sup>2</sup> P(x)
0	1/8	0	0
1	3/8	3/8	3/8
2	3/8	6/8	12/8
3	1/8	3/8	9/8
Total	1.00	12/8	24/8=3

$$\mu=1.5, \quad \sigma^2 = 3 - 1.5^2 = 0.75$$

$$\sigma = \sqrt{0.75} = 0.866$$

# Mean of Discrete random variables

- Example:  $X =$  sum of two dice

$$\begin{aligned}\mu &= \sum X_i P(X_i) \\ &= X_1 P(X_1) + X_2 P(X_2) \dots + X_{12} P(X_{12})\end{aligned}$$

X	2	3	4	5	6	7	8	9	10	11	12
P(X)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

$$\begin{aligned}\mu &= 2 \cdot \left(\frac{1}{36}\right) + 3 \cdot \left(\frac{2}{36}\right) + 4 \cdot (3/36) + \dots + 12 \cdot (1/36) \\ &= 252/36 = 7\end{aligned}$$

# Variance of Discrete Random Variable

- Example:  $X$  = sum of two dice

$X$	2	3	4	5	6	7	8	9	10	11	12
$P(X)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
$X^2P(X)$	$\frac{4}{36}$	$\frac{18}{36}$	$\frac{48}{36}$	$\frac{100}{36}$	$\frac{180}{36}$	$\frac{294}{36}$	$\frac{320}{36}$	$\frac{324}{36}$	$\frac{300}{36}$	$\frac{242}{36}$	$\frac{144}{36}$

$$\begin{aligned}\sigma^2 &= \frac{4}{36} + \frac{18}{36} + \frac{48}{36} + \frac{100}{36} + \frac{180}{36} + \frac{294}{36} + \frac{320}{36} + \frac{324}{36} + \frac{300}{36} + \frac{242}{36} + \frac{144}{36} - \mu^2 \\ &= \frac{1974}{36} - 49 = 54.833 - 49 = 5.833\end{aligned}$$

# Binomial Distribution $B(n, p)$

**A binomial experiment** has the following conditions:

1. There are  $n$  repeated identical trials.
2. Each trial has only two possible outcomes (success or failure, boy or girl, dead or a live, good or defective, yes or no).
3. The probability of a success,  $p$ , is constant from trial to trial
4. The outcome of each trial is independent of the outcomes of any other trial; that is, the outcome of one trial has no effect on the outcome of any other trial.

# Binomial distribution

- When the conditions of the binomial experiment are satisfied, our interest is in the **number of successes** occurring in the trials.

- Example

If a certain drug is known to cause a side effect 10% of the time and if **five** patients are given this drug, what is the probability that **four** or more experience the side effect?

## More Examples

- No. of Correct ion a 33 question exam.
- No. of H observed when tossing a fair coin 5 times.
- No. of defective items in a sample of 20 items from a large shipment

# The Binomial probability formula

- The probability of obtaining  $r$  successes in  $n$  trials with a probability  $P$  of success in each trial can be calculated using the formula;
- If  $X$  denotes the number of success, then

$$P(X = r) = C(n, r) P^r (1 - P)^{n-r}, \quad 0 \leq r \leq n$$

## Mean & Variance of the Binomial Distribution

The **mean** is found by:

$$\mu = nP$$

The **variance** is found by

$$\sigma^2 = nP(1 - P)$$

# Example

- In tossing a coin 3 times, let  $X$  be the no. of H observed. List the probability distribution of  $X$ .
- $X = 0, 1, 2, 3$ .  $p = 1/2$
- $p(X=0) = C(3,0) \cdot (1/2)^0 (1/2)^3 = 1/8$
- $p(X=1) = C(3,1) \cdot (1/2) (1/2)^2 = 3/8$
- $p(X=2) = C(3,2) \cdot (1/2)^2 (1/2)^1 = 3/8$
- $p(X=3) = C(3,3) \cdot (1/2)^3 (1/2)^0 = 1/8$

# Example

X	0	1	2	3	
P(X)	1/8	3/8	3/8	1/8	$\sum P(x)=1$

Find the probability of observing at least 2 H.

$$P(X \geq 2) = P(X = 2) + P(X = 3) = 3/8 + 1/8 = 1/2$$



# Exercise

- Tossing a fair coin 8 times. Let  $X$  be the number of H observed.
  - a. What is the probability of observing
    - i. Exactly 3 H.
    - ii. More than 2 H.
  - b. Find the mean and the standard deviation of the distribution.

# Example

It is known that approximately 10% of the population is hospitalized at least once during a year. If 10 persons in such a community are to be interviewed.

1. What is the probability that you will find
  - a. All have been hospitalized at least once during the year.
  - b. 50% have been hospitalized at least once during the year.
  - c. Exactly 2 have been hospitalized at least once during the year
  - d. At least 3 have been hospitalized at least once during the year .
2. Find the mean and the standard deviation of the distribution.

## Answer:

Let  $X$  be the number of persons have been hospitalized at least once during the year.

$X=0,1,2,3,4,5,6,7,8,9,10$ .

$$a. p(X=10) = C(10,10) \cdot (0.10)^{10} (0.90)^0 = 10^{-10}$$

$$b. p(X=5) = C(10,5) \cdot (0.10)^5 (0.90)^5 = 1.488(10^{-3})$$

$$c. p(X=2) = C(10,2) \cdot (0.10)^2 (0.90)^8 = 0.194.$$

$$d. p(\text{at least } 3) = p(X \text{ is greater than or equal } 3)$$

$$= 1 - p(X < 3)$$

$$= 1 - (p(X=0) + p(X=1) + p(X=2))$$

$$P(X=0) = C(10,0) \cdot (0.10)^0 (0.90)^{10} = 0.349$$

$$P(X=1) = C(10,1) \cdot (0.10)^1 (0.90)^9 = 0.387$$

$$p(\text{at least } 3) = p(X \geq 3) = 1 - p(X < 3)$$

$$= 1 - (0.349 + 0.387 + 0.194) = 0.07$$

- Mean =  $nP = 10(0.10) = 1$
- Variance  $\sigma^2 = nP(1 - P)$   
 $= 0.90$
- Standard deviation  $\sigma = \sqrt{0.90} = 0.949$

# Using the Binomial Probability Table

$n = 8, p = 0.7, \text{ find } P(r = 6):$

$n$	$r$	$p$																			
		.01	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
8	0	.923	.663	.430	.272	.168	.100	.058	.032	.017	.008	.004	.002	.001	.000	.000	.000	.000	.000	.000	.000
	1	.075	.279	.383	.385	.336	.267	.198	.137	.090	.055	.031	.016	.008	.003	.001	.000	.000	.000	.000	.000
	2	.003	.051	.149	.238	.294	.311	.296	.259	.209	.157	.109	.070	.041	.022	.010	.004	.001	.000	.000	.000
	3	.000	.005	.033	.084	.147	.208	.254	.279	.279	.257	.219	.172	.124	.081	.047	.023	.009	.003	.000	.000
	4	.000	.000	.005	.018	.046	.087	.136	.188	.232	.263	.273	.263	.232	.188	.136	.087	.046	.018	.005	.000
	5	.000	.000	.000	.003	.009	.023	.047	.081	.124	.172	.219	.257	.279	.279	.251	.208	.147	.084	.033	.005
	6	.000	.000	.000	.000	.001	.004	.010	.022	.041	.070	.109	.157	.209	.259	.296	.311	.294	.238	.149	.051
	7	.000	.000	.000	.000	.000	.000	.001	.003	.008	.016	.031	.055	.090	.137	.198	.267	.336	.385	.383	.279
8	.000	.000	.000	.000	.000	.000	.000	.000	.001	.002	.004	.008	.017	.032	.058	.100	.168	.272	.430	.663	
9	0	.914	.630	.387	.232	.134	.075	.040	.021	.010	.005	.002	.001	.000	.000	.000	.000	.000	.000	.000	
	1	.083	.299	.387	.368	.302	.225	.156	.100	.060	.034	.018	.008	.004	.001	.000	.000	.000	.000	.000	.000
	2	.003	.063	.172	.260	.302	.300	.267	.216	.161	.111	.070	.041	.021	.010	.004	.001	.000	.000	.000	.000
	3	.000	.008	.045	.107	.176	.234	.267	.272	.251	.212	.164	.116	.074	.042	.021	.009	.003	.001	.000	.000
	4	.000	.001	.007	.028	.066	.117	.172	.219	.251	.260	.246	.213	.167	.118	.074	.039	.017	.005	.001	.000
	5	.000	.000	.001	.005	.017	.039	.074	.118	.167	.213	.246	.260	.251	.219	.172	.117	.066	.028	.007	.001
	6	.000	.000	.000	.001	.003	.009	.021	.042	.074	.116	.164	.212	.251	.272	.267	.234	.176	.107	.045	.008
	7	.000	.000	.000	.000	.000	.001	.004	.010	.021	.041	.070	.111	.161	.216	.267	.300	.302	.260	.172	.063
	8	.000	.000	.000	.000	.000	.000	.000	.001	.004	.008	.018	.034	.060	.100	.156	.225	.302	.368	.387	.299
9	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.002	.005	.010	.021	.040	.075	.134	.232	.387	.630	
10	0	.904	.599	.349	.197	.107	.056	.028	.014	.006	.003	.001	.000	.000	.000	.000	.000	.000	.000	.000	
	1	.091	.315	.387	.347	.268	.188	.121	.072	.040	.021	.010	.004	.002	.000	.000	.000	.000	.000	.000	.000
	2	.004	.075	.194	.276	.302	.282	.233	.176	.121	.076	.044	.023	.011	.004	.001	.000	.000	.000	.000	.000
	3	.000	.010	.057	.130	.201	.250	.267	.252	.215	.166	.117	.075	.042	.021	.009	.003	.001	.000	.000	.000
	4	.000	.001	.011	.040	.088	.146	.200	.238	.251	.238	.205	.160	.111	.069	.037	.016	.006	.001	.000	.000
	5	.000	.000	.001	.008	.026	.058	.103	.154	.201	.234	.246	.234	.201	.154	.103	.058	.026	.008	.001	.000
	6	.000	.000	.000	.001	.006	.016	.037	.069	.111	.160	.205	.238	.251	.238	.200	.146	.088	.040	.011	.001
	7	.000	.000	.000	.000	.001	.003	.009	.021	.042	.075	.117	.166	.215	.252	.267	.250	.201	.130	.057	.010
	8	.000	.000	.000	.000	.000	.000	.001	.004	.011	.023	.044	.076	.121	.176	.233	.282	.302	.276	.194	.07
	9	.000	.000	.000	.000	.000	.000	.000	.000	.002	.004	.010	.021	.040	.072	.121	.188	.268	.347	.387	.315
10	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.003	.006	.014	.028	.056	.107	.197	.349	.599	

# Binomial Probability Distribution

- To construct a **binomial distribution**, let
  - $n$  be the number of trials
  - $r$  be the number of observed successes
  - $P$  be the probability of success on each trial
- The formula for the binomial probability distribution is:

$$p(r) = \left( \frac{n!}{r!(n-r)!} \right) \cdot p^r \cdot q^{n-r}$$

# Exercise

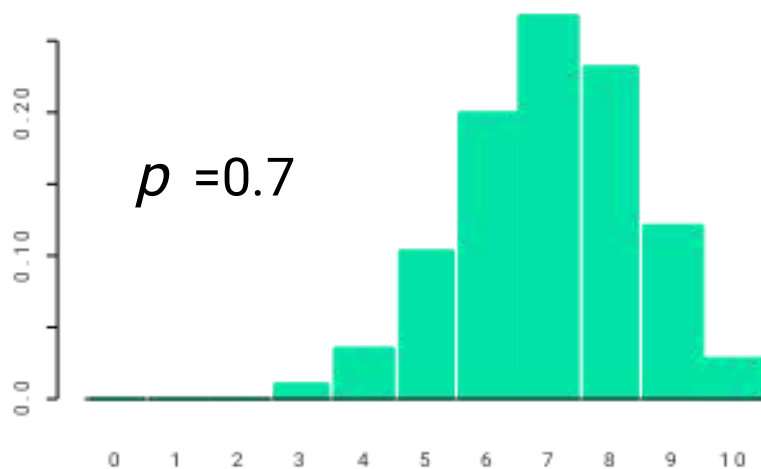
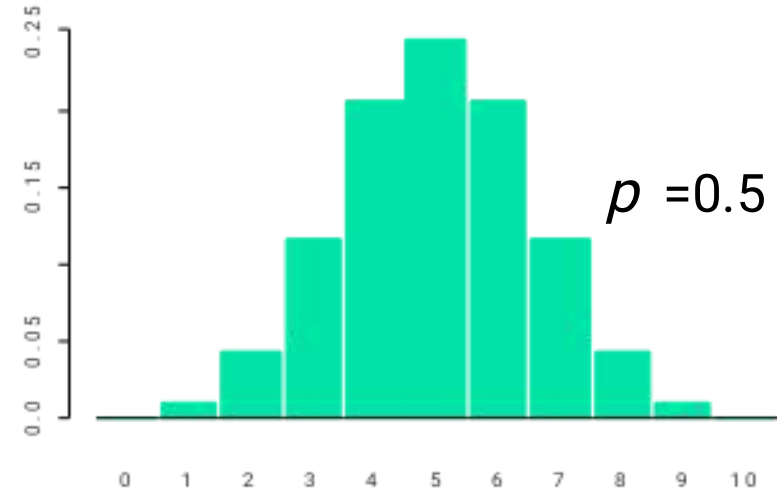
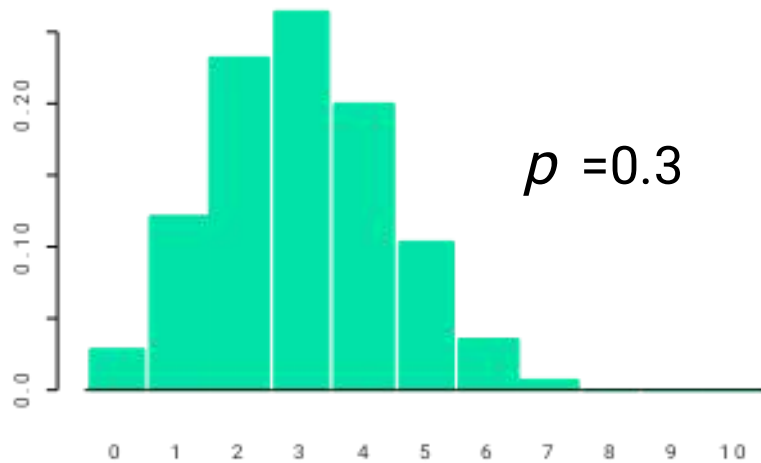
A biologist is studying a new hybrid tomato. It is known that the seeds of this hybrid tomato have probability 0.70 of germinating. The biologist plants 10 seeds.

a. What is the probability that

- i. exactly 8 seeds will germinate?
- ii. at least 8 seeds will germinate?
- iii. at least 2 seeds will germinate?

b. What is the expected number of seeds in this sample to germinate.

# Characteristics of Binomial Distribution



Notice that when  $p < 0.5$  the distribution is skewed right, and when  $p > 0.5$  the distribution is skewed left. When  $p = 0.5$ , the distribution is symmetric.

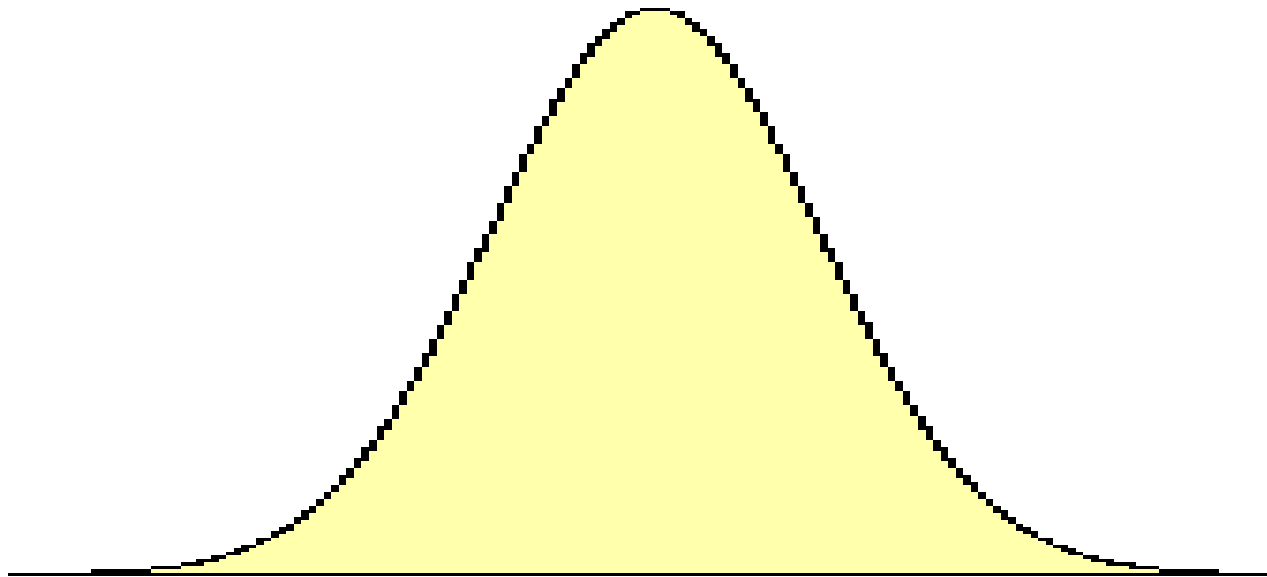


Common English expression and corresponding inequalities  
(consider a binomial experiment with  $n$  trials and  $r$  success)

Expression	Inequalities
Four or more successes	$r \geq 4$
At least four successes	That is, $r = 4, 5, 6, \dots, n$
No fewer than four successes	
Not less than four successes	
Four or fewer successes	$r \leq 4$
At most four successes	That is, $r = 0, 1, 2, 3, \text{ or } 4$
No more than four successes	
The number of successes does not exceed four	
More than four successes	$r > 4$
The number of successes exceeds four	That is, $r = 5, 6, 7, \dots, n$
Fewer than four successes	$r < 4$
The number of successes is not as large as four	That is, $r = 0, 1, 2, 3$

# Continuous probability distribution

- Continuous random variables have a **non-countable** number of values
- Can't list the entire probability distribution, so we use a **density curve** instead of a histogram

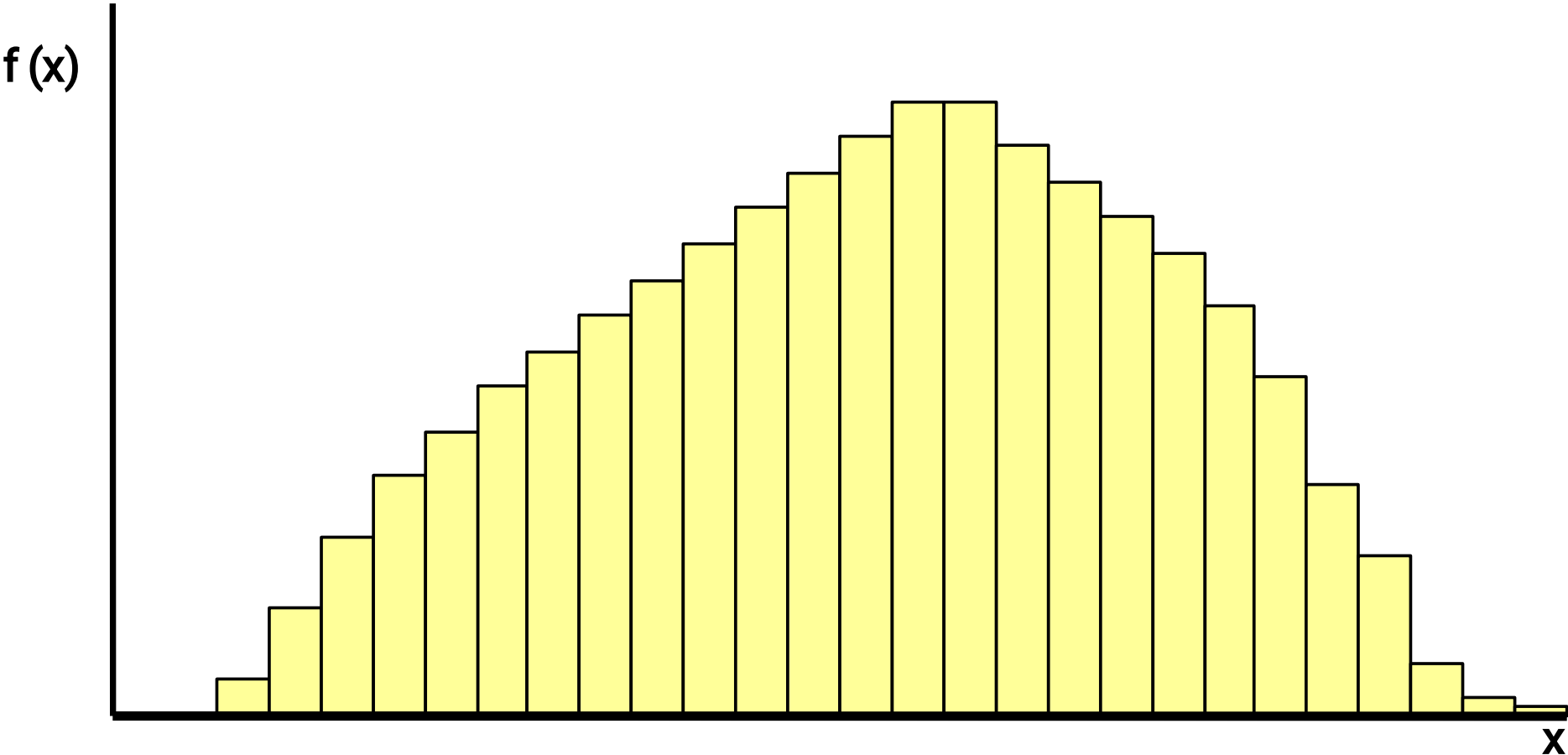


# Recall

A **continuous probability distribution** can assume an **infinite** number of values within a given range – for variables that take continuous values.

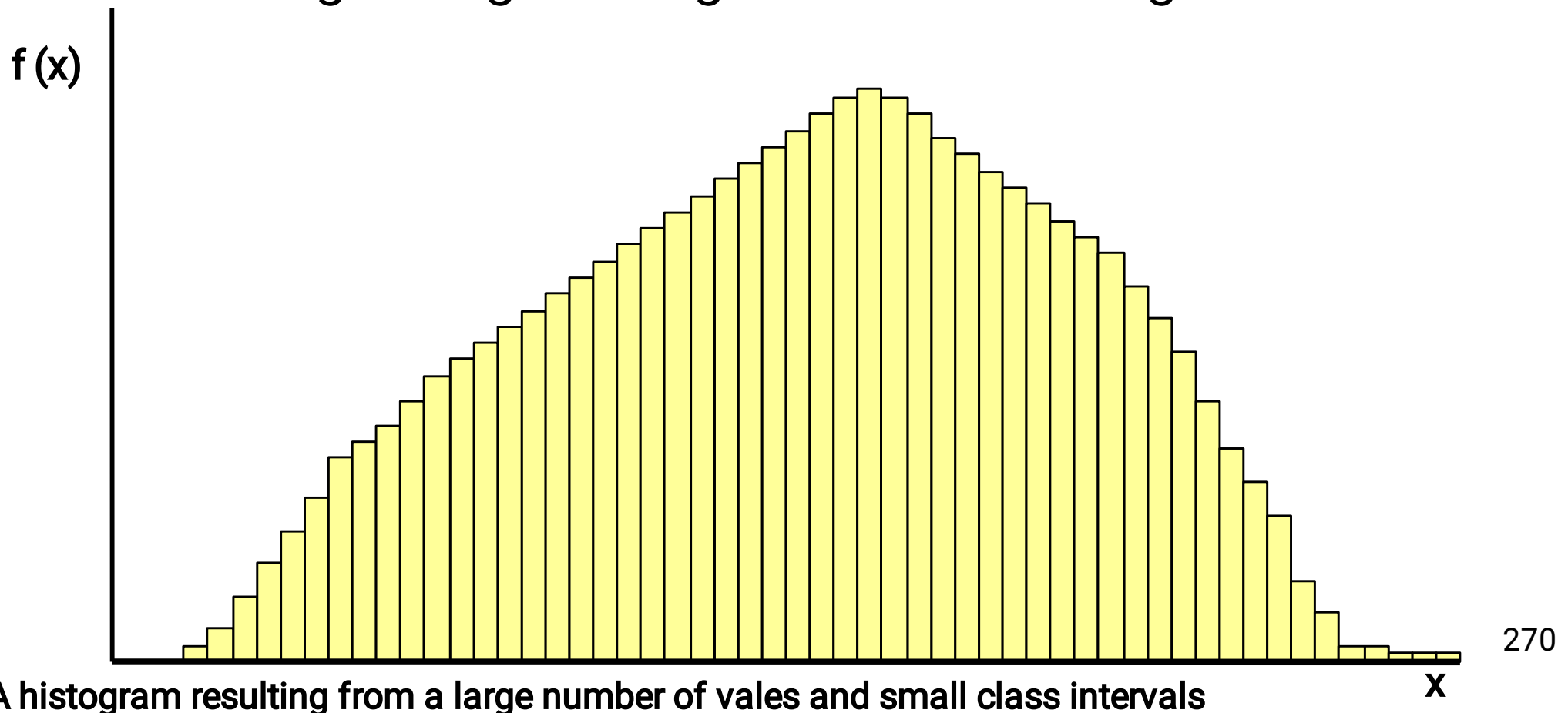
- The distance students travel to class.
- The time it takes an executive to drive to work.
- The length of an afternoon nap.
- The length of time of a particular phone call.
- The size of some kind of fruits.

# Probability distributions of continuous random variables



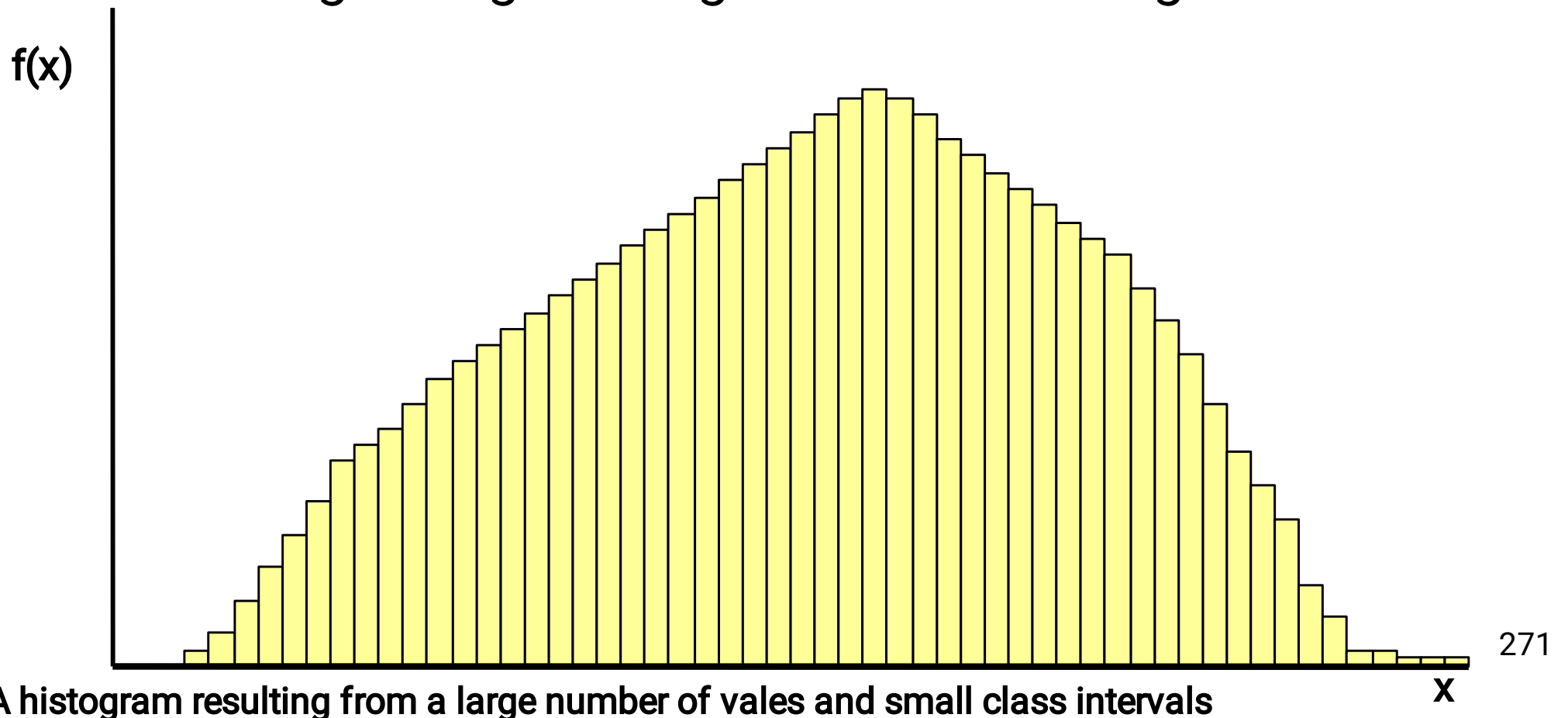
# Probability distributions of continuous random variables

- Imagine the situation where the number of observations is very large and the width of class intervals is made very small. The resulting histogram might look like this figure



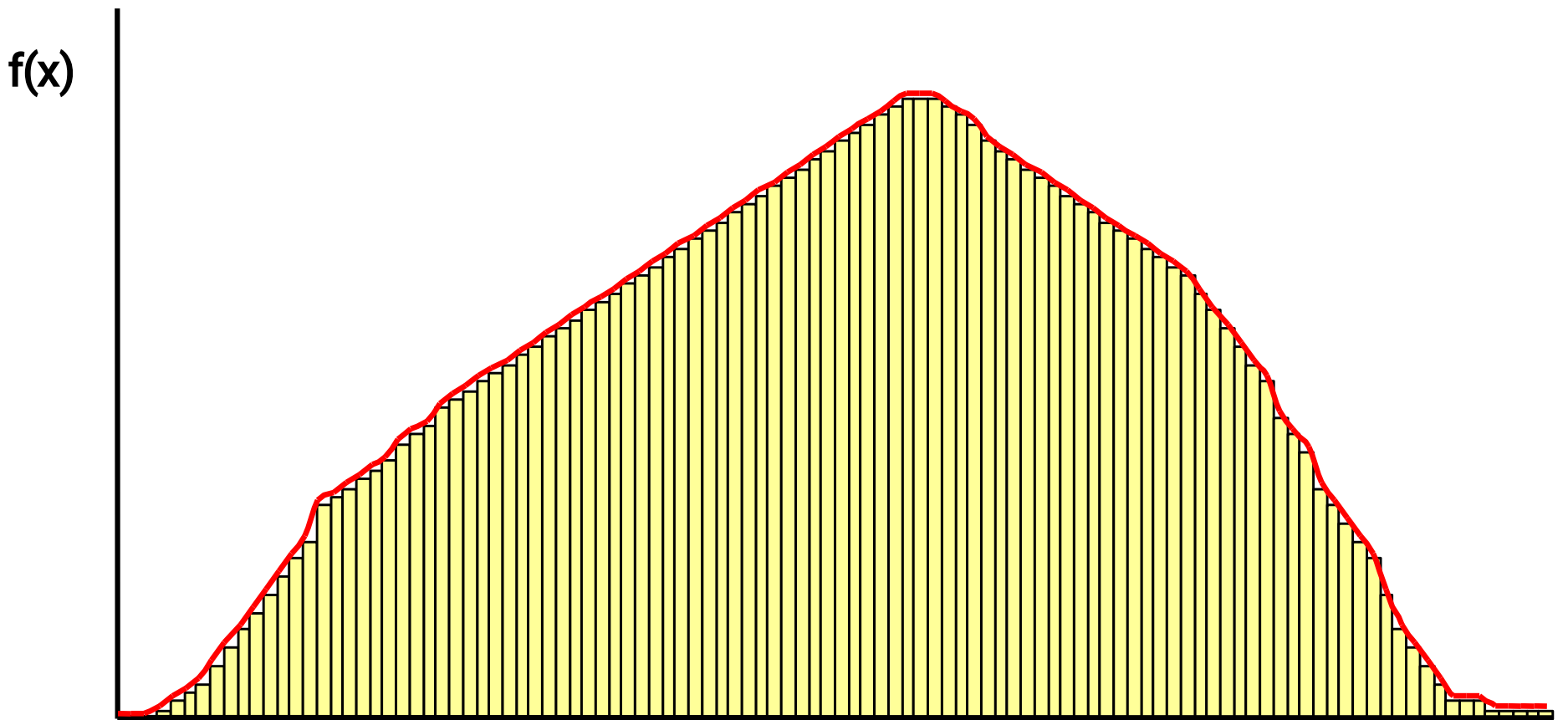
# Probability distributions of continuous random variables

- Imagine the situation where the number of observations is very large and the width of class intervals is made very small. The resulting histogram might look like this figure



# Probability distributions of continuous random variables

- As the  $n$  of observations used to construct the histogram approaches **infinity**, and the width of the class intervals goes to **zero**, we will arrive at a smooth curve superimposed on the histogram called a density curve.





# Normal probability distribution

Many continuous variables are approximately normally distributed

- Physical and mental properties of people as (height, weight, body temperature, blood pressure, IQ,... ).
- Size of apples, oranges,....
- Weight of a born baby.

# Normal Probability Distribution

$$N(\mu, \sigma^2)$$

## Normal Probability Density Function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where:

$\mu$  = mean

$\sigma$  = standard deviation

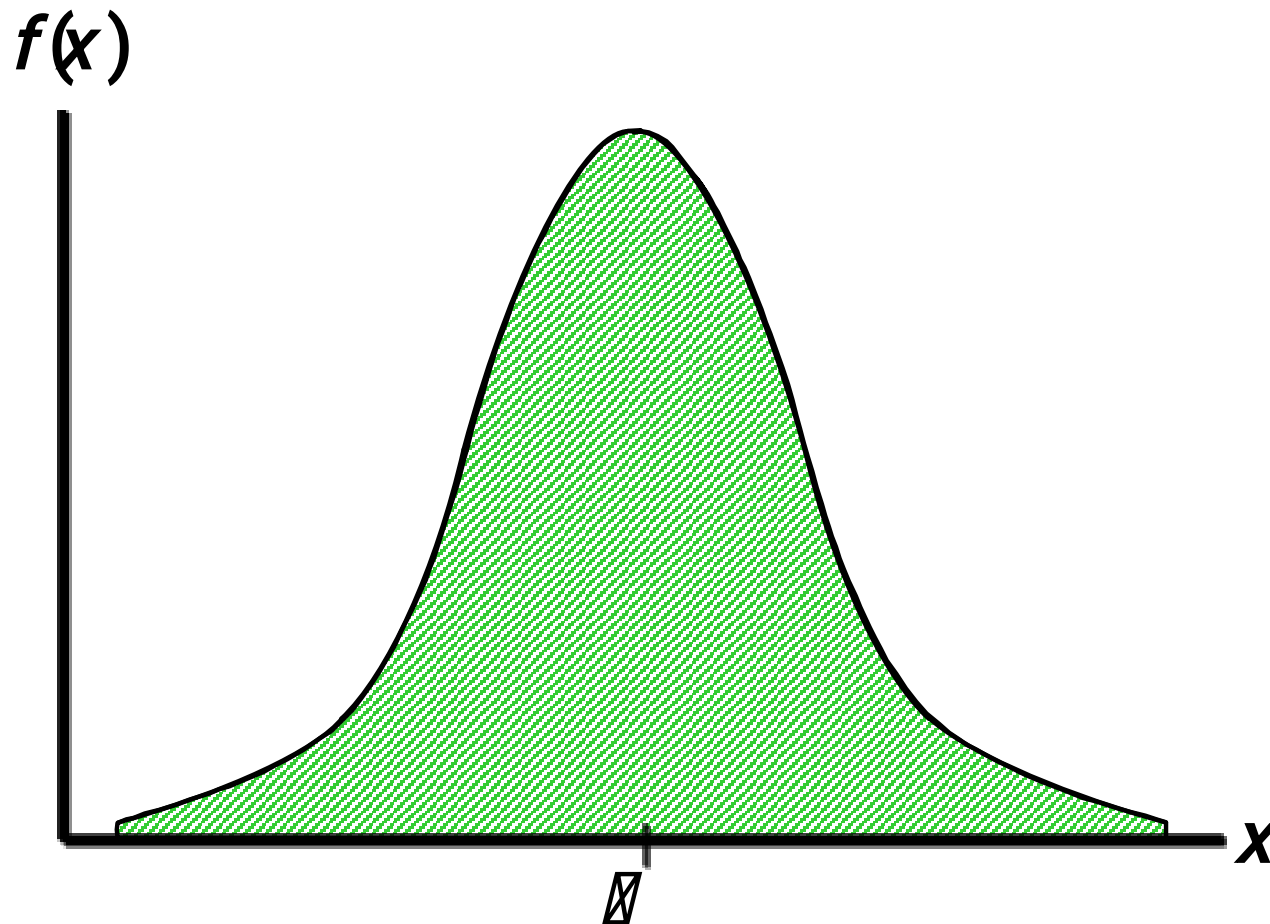
$\pi \cong 3.14159$

$e \cong 2.71828$

This formula generates the *density curve* which gives the shape of the normal distribution.

# Normal Probability Distribution

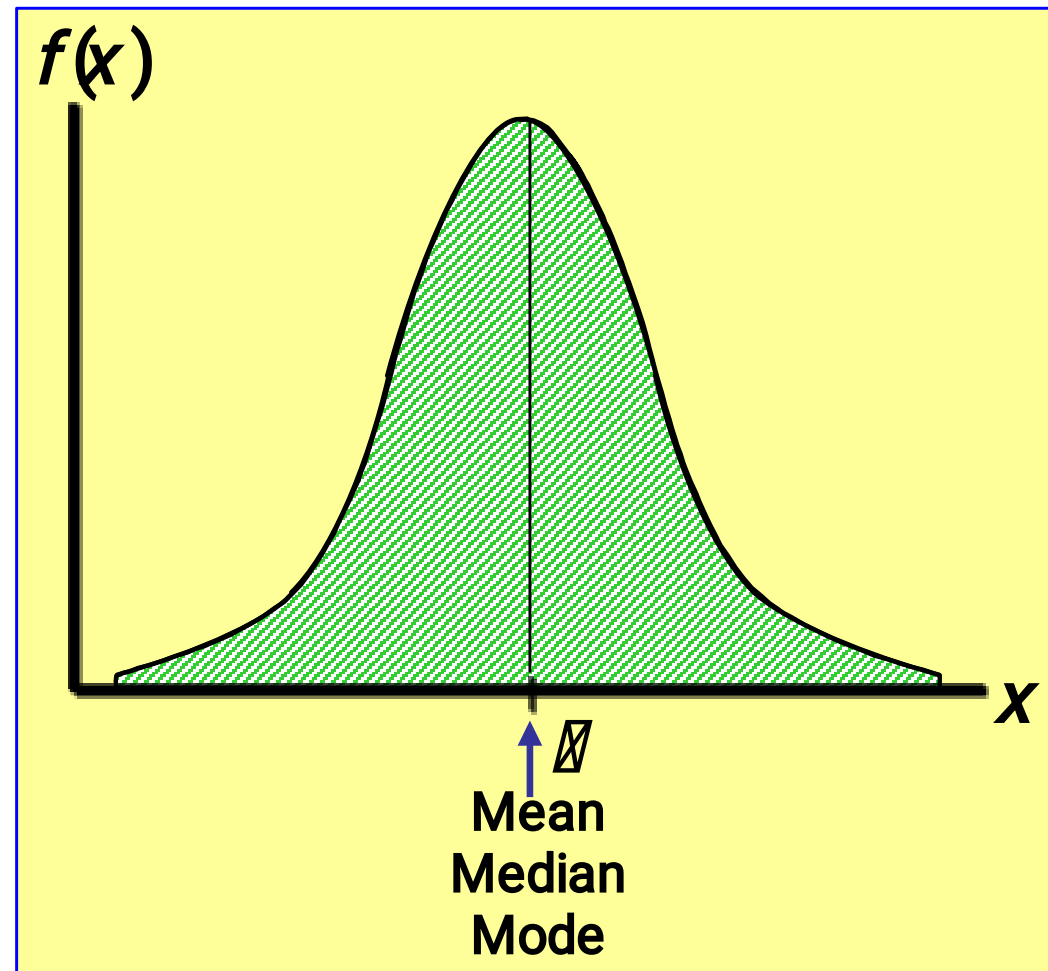
- Graph of the Normal Probability Density Function



# Properties of the Normal Distribution

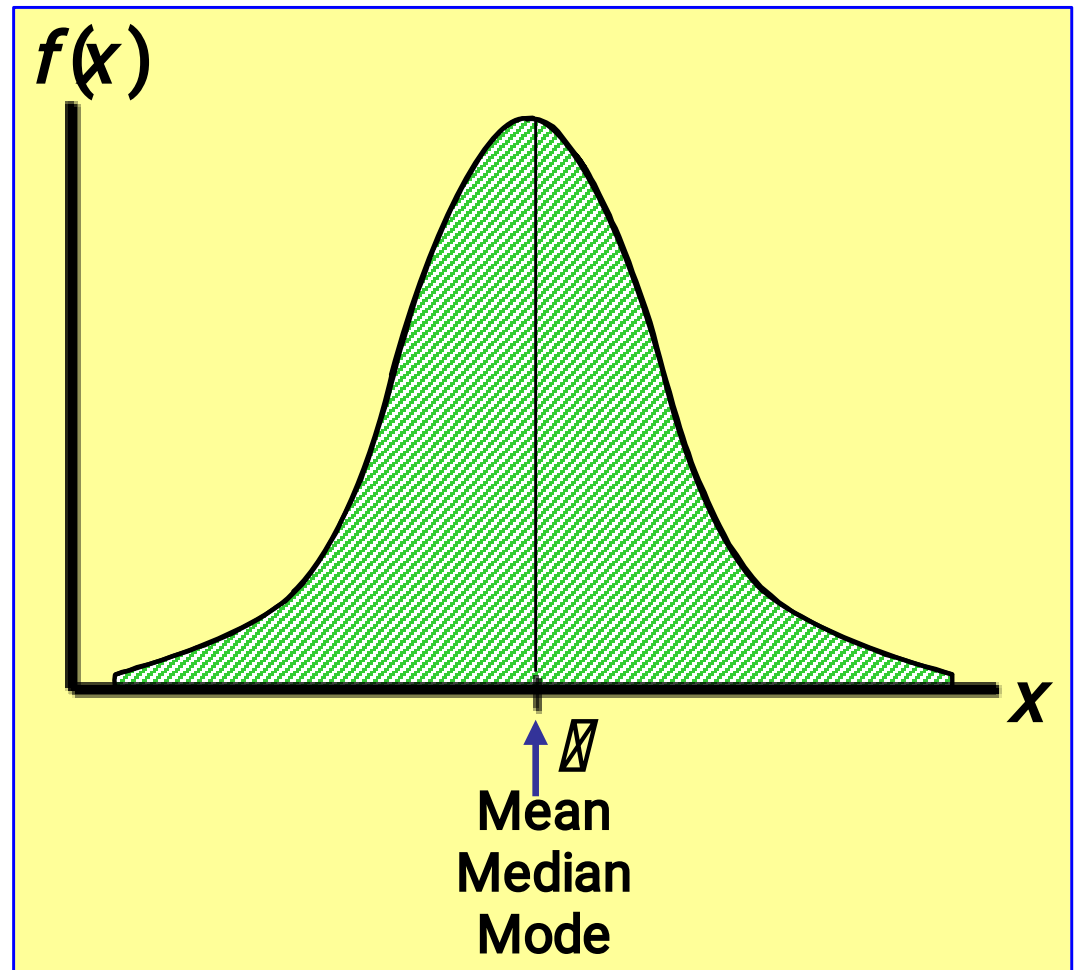
There are several characteristics that make the normal distribution very important for statisticians

- “Bell shaped” and Symmetrical about the mean  $\mu$ .
- Unimodal and its mode occurs at  $x = \mu$ .
- Mean, median and mode are equal
- Interquartile range equals **1.33  $\sigma$**
- Random variable has infinite range from  $-\infty$  to  $\infty$ .
- The total area under the curve and above the horizontal axis is equal to 1.



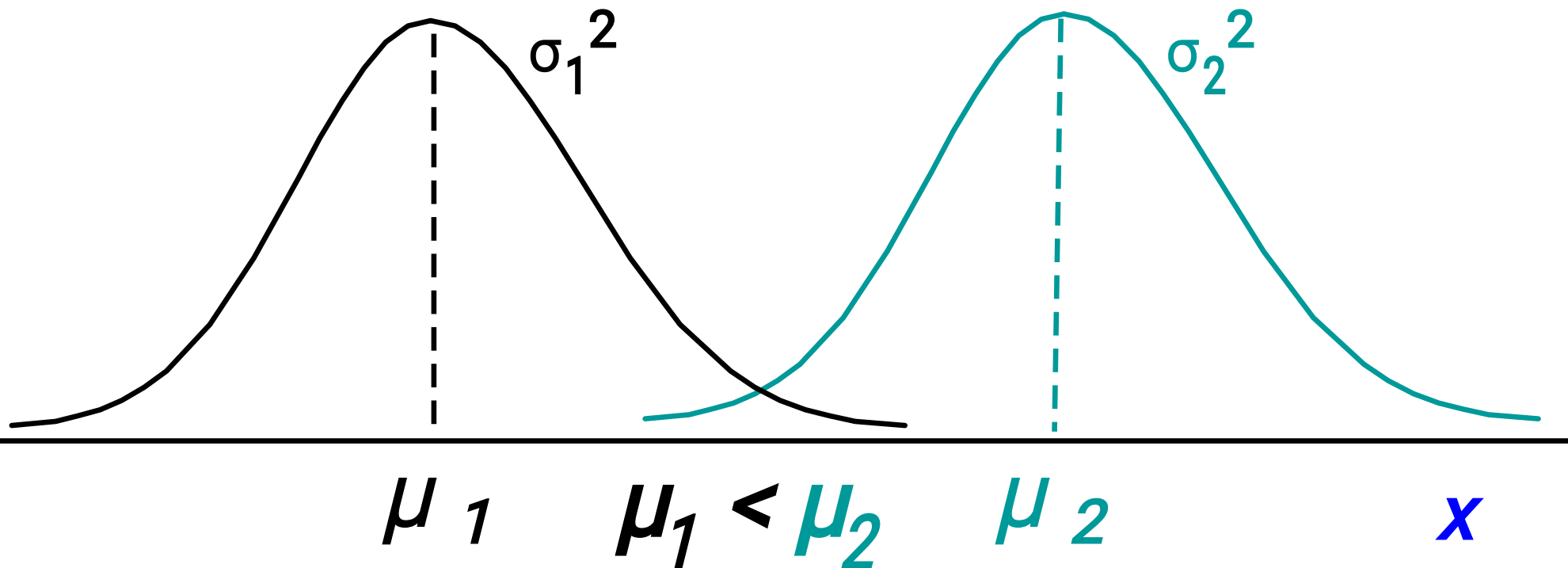
# Properties of the Normal Distribution

Most observations in the distribution are close to the mean, with gradually fewer observations further away



## Different Normal Distribution

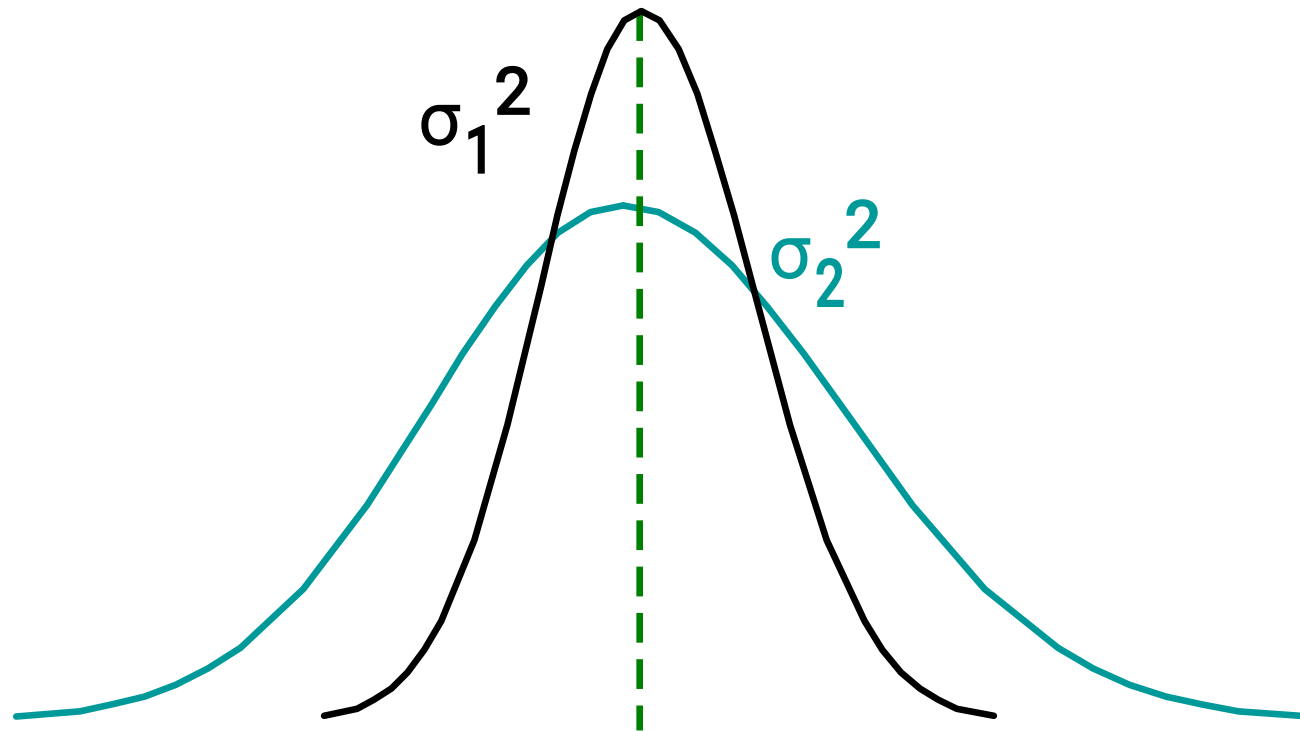
The normal distribution depends on the values of the parameters  $\mu$ , the population mean and  $\sigma$ , the population variance.



- a. Two normal curves, which have the same standard deviation but different means

# Properties of the Normal Distribution

- b. Two normal curves with the same mean but different standard deviations



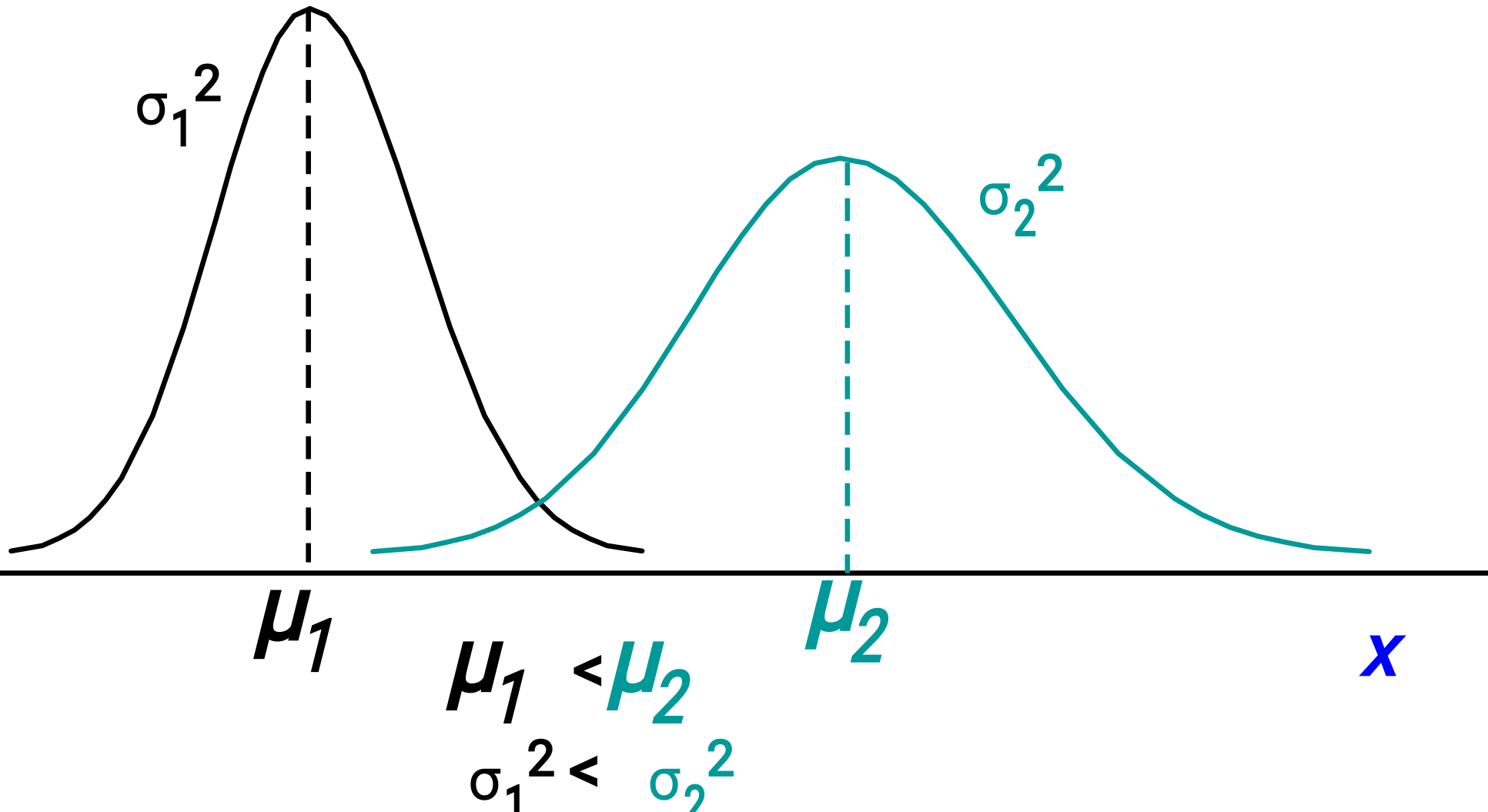
$$\mu_1 = \mu_2$$

$$\sigma_1^2 < \sigma_2^2$$

 $x$

# Properties of the Normal Distribution

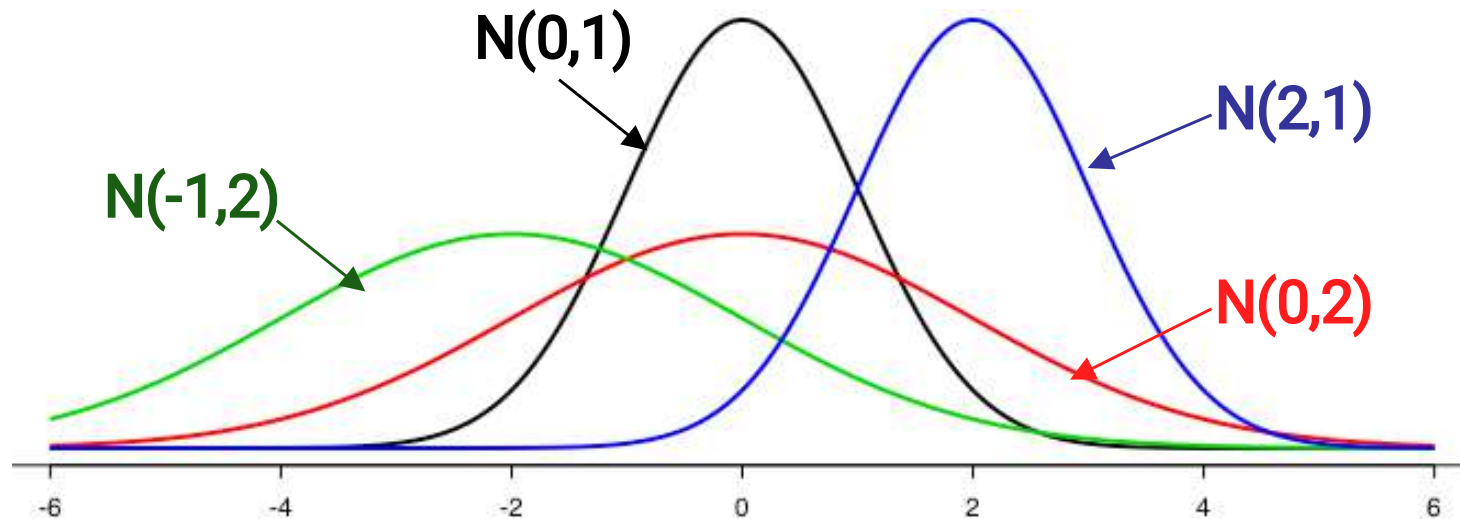
- c. Two normal curves that have different means and different standard deviations.





# Different Normal Distributions

- Each different value of  $\mu$  and  $\sigma^2$  gives a different Normal distribution, denoted  $N(\mu, \sigma^2)$



- We can adjust values of  $\mu$  and  $\sigma^2$  to provide the best approximation to observed data
- If  $\mu = 0$  and  $\sigma^2 = 1$ , we have the **Standard Normal** distribution

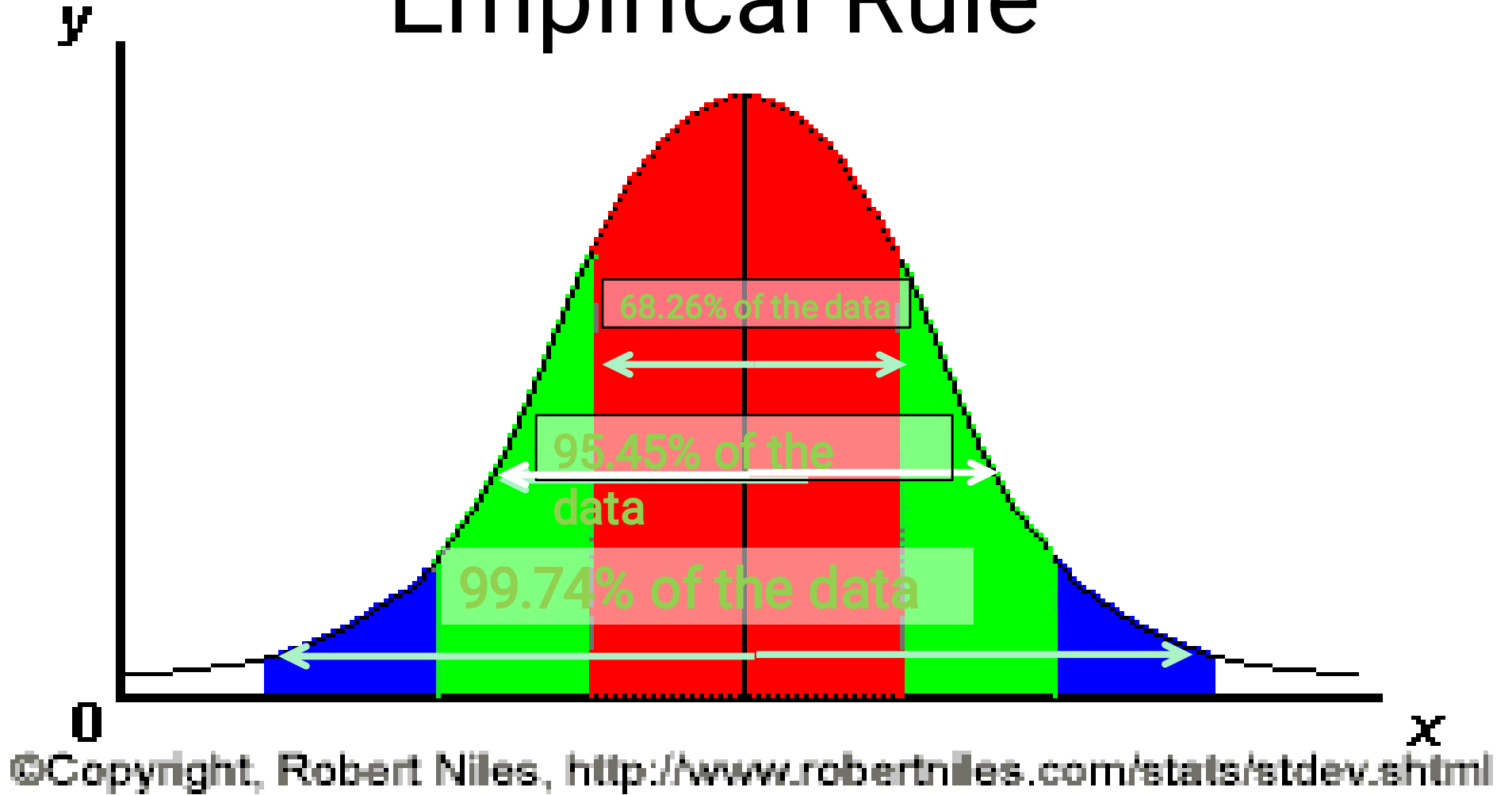
# Property of Normal Distributions

## Empirical Rule

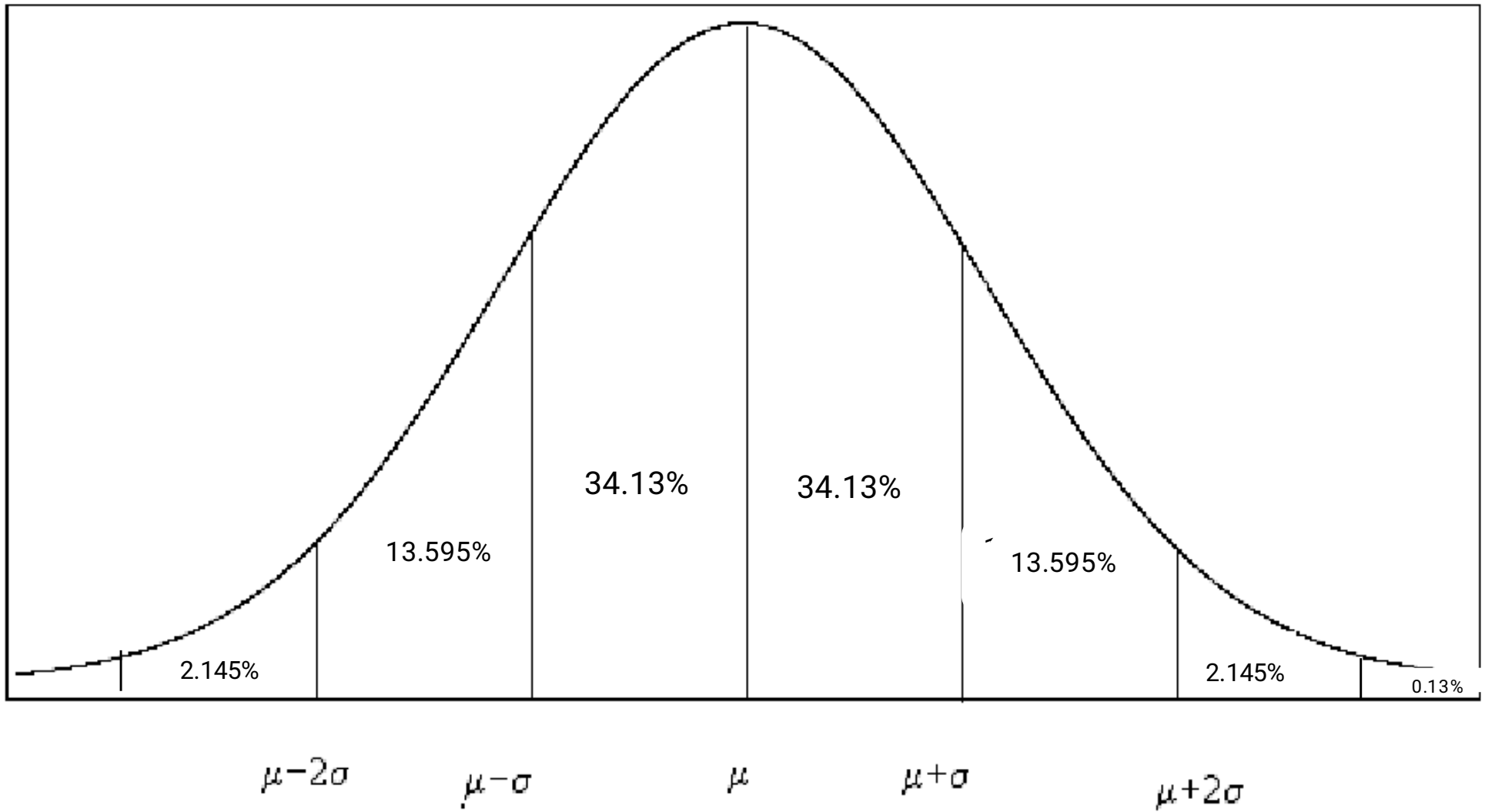
### About

- 68.26% of observations are between  $\mu - \sigma$  and  $\mu + \sigma$
- 95.45% of observations are between  $\mu - 2\sigma$  and  $\mu + 2\sigma$
- 99.74% of observations are between  $\mu - 3\sigma$  and  $\mu + 3\sigma$

# Empirical Rule



# Normal Distribution



$$P(X \geq \mu) = 0.50 \quad P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.6846 \quad P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.9545$$

# Application of the Empirical Rule

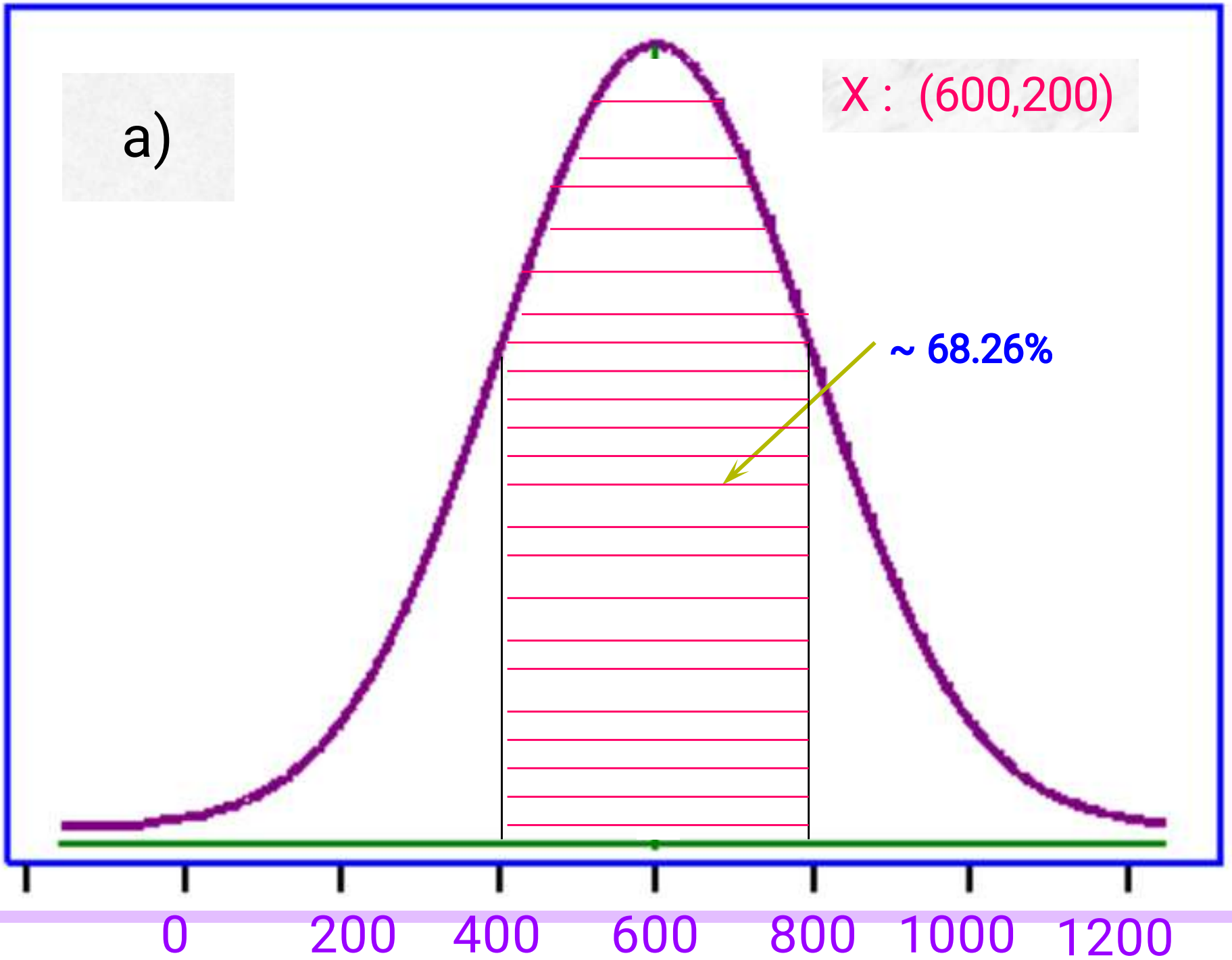
Women participating in a three-day experimental diet regime have been demonstrated to have normally distributed weight loss with mean 600 g and a standard deviation 200 g.

- a) What percentage of these women will have a weight loss between 400 and 800 g?
- b) What percentage of women will lose weight too quickly on the diet (where too much weight is defined as  $>1000\text{g}$ )?

a)

$X : (600, 200)$

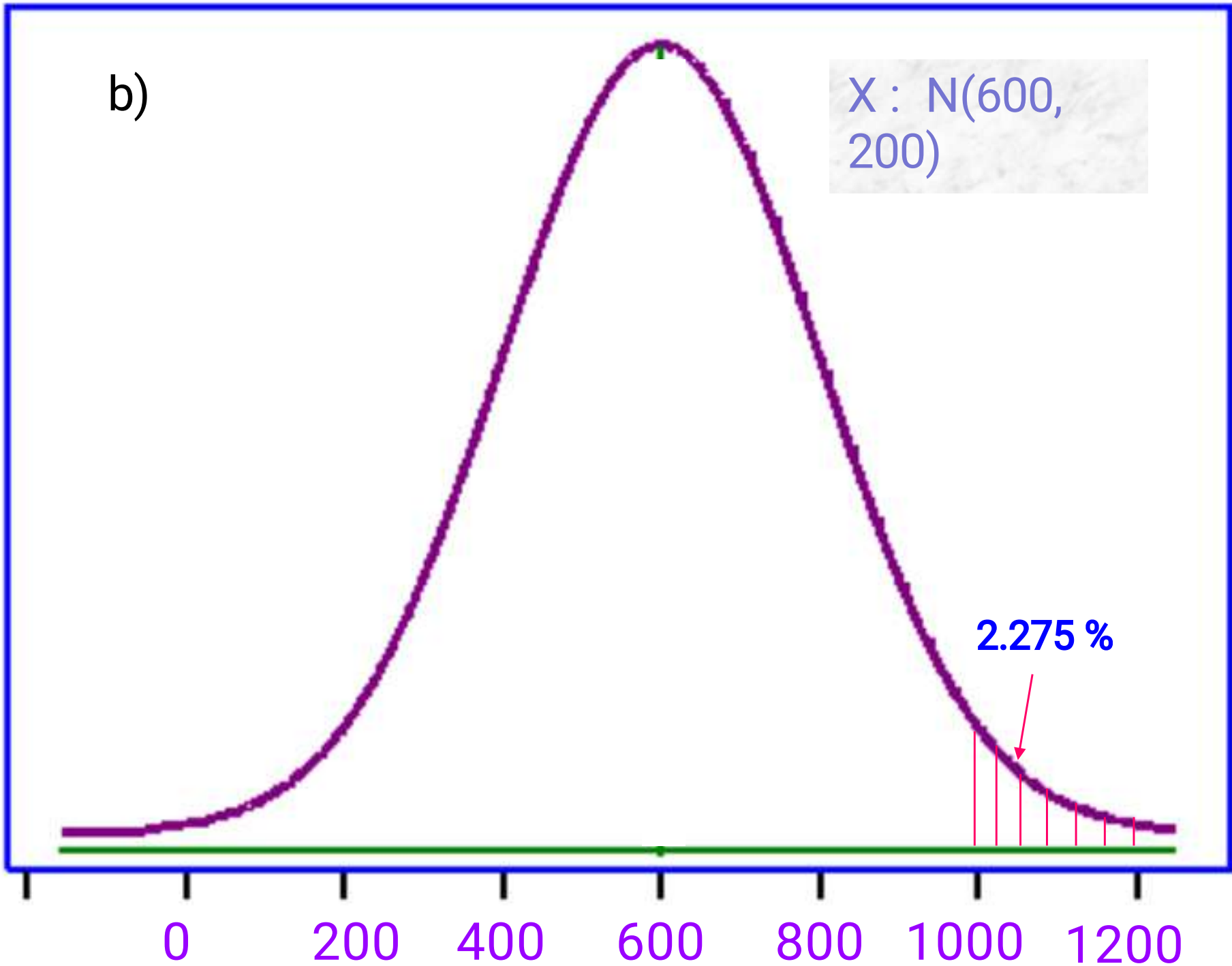
$\sim 68.26\%$



b)

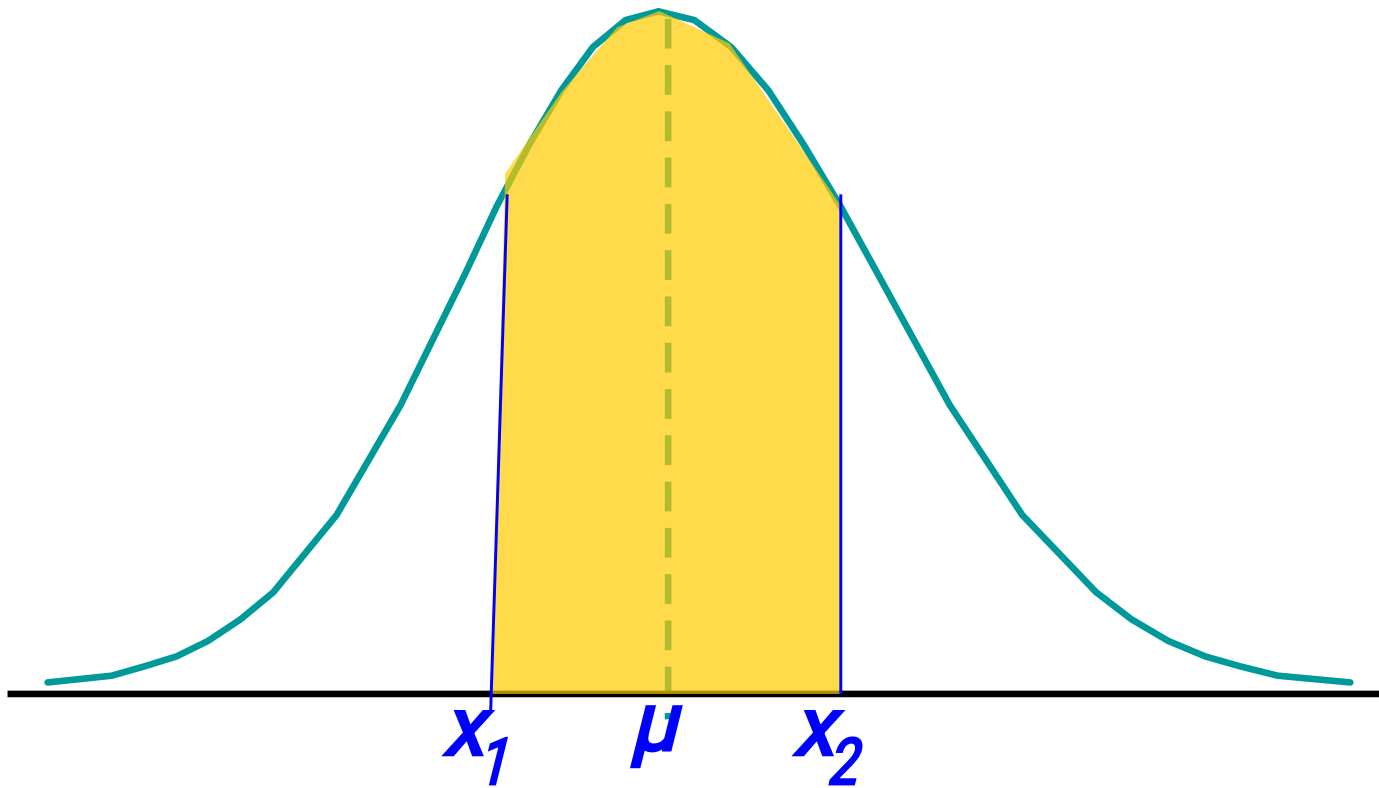
$X : N(600, 200)$

2.275 %



# Areas under the Normal Curve

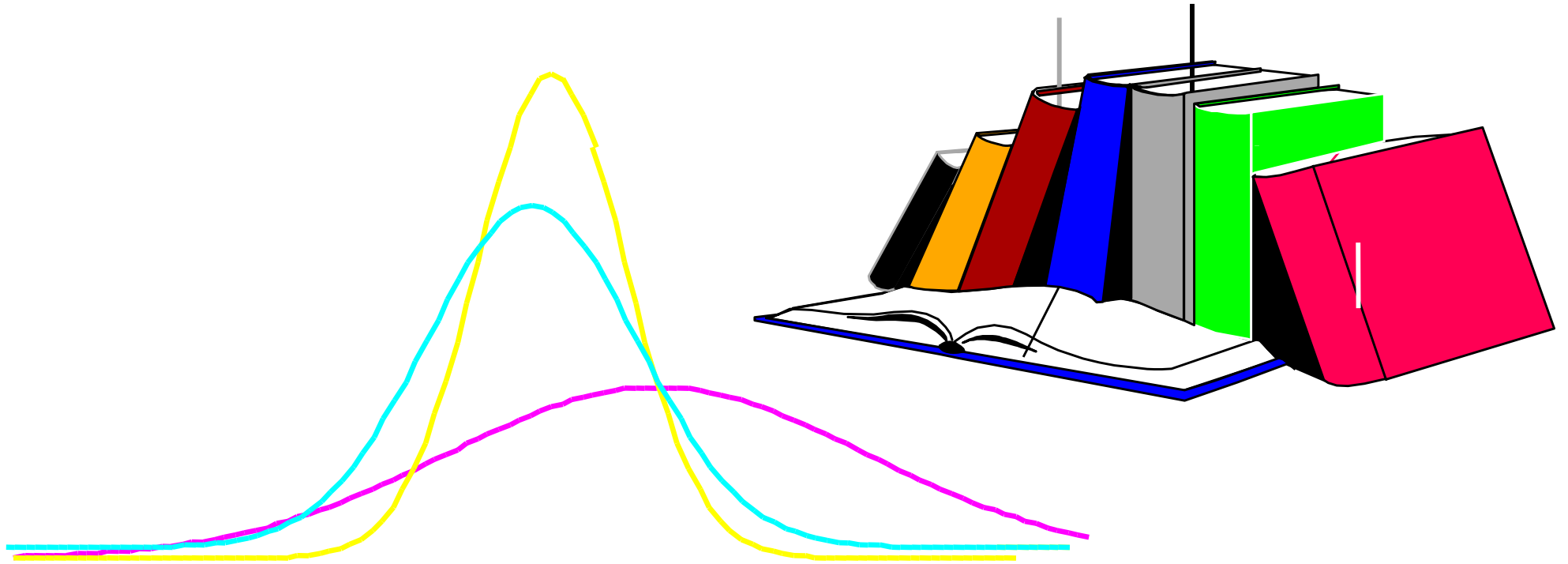
- The area under the curve bounded by the two ordinates  $X = X_1$  and  $X = X_2$  equals the **probability** that the **random variable  $X$**  assumes a value between  $X = X_1$  and  $X = X_2$ . Thus, for the normal curve in the Figure below, the  $P(X_1 < X < X_2)$  is represented by the area of the





# Which Table to Use?

Yet we must use tables if we hope to avoid the use of integral calculus



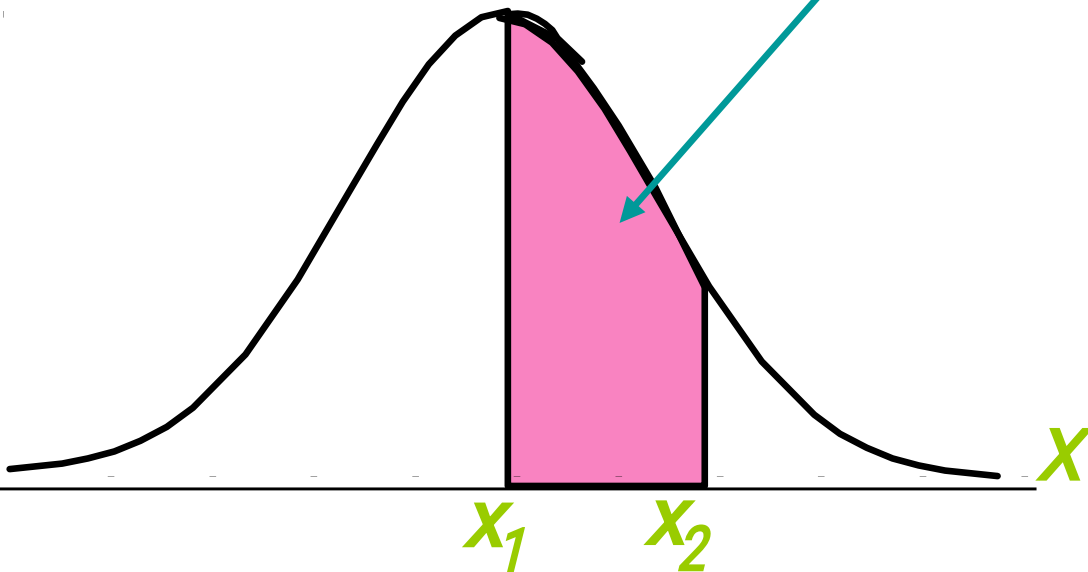
An infinite number of normal distributions means an infinite number of tables to look up!

It would be hopeless task to set up separate tables of normal curve areas for every conceivable value of  $\mu$  and  $\sigma$

# Finding Probabilities

Probability is the area under the curve!

$$P(x_1 < X < x_2) = ?$$



# The Standard Normal Distribution

**Fortunately**, all the observations of any normal random variable  $X$  could be transformed to a new set of observations of a normal random variable  $Z$  with **mean zero** and **variance 1**, by using the transformation

$$Z = \frac{X - \mu}{\sigma}$$

So, if  $X$  is normally distributed with mean  $\mu$  and standard deviation  $\sigma$ , then  $Z$  is normally distributed with mean 0 and standard deviation 1

# standardization

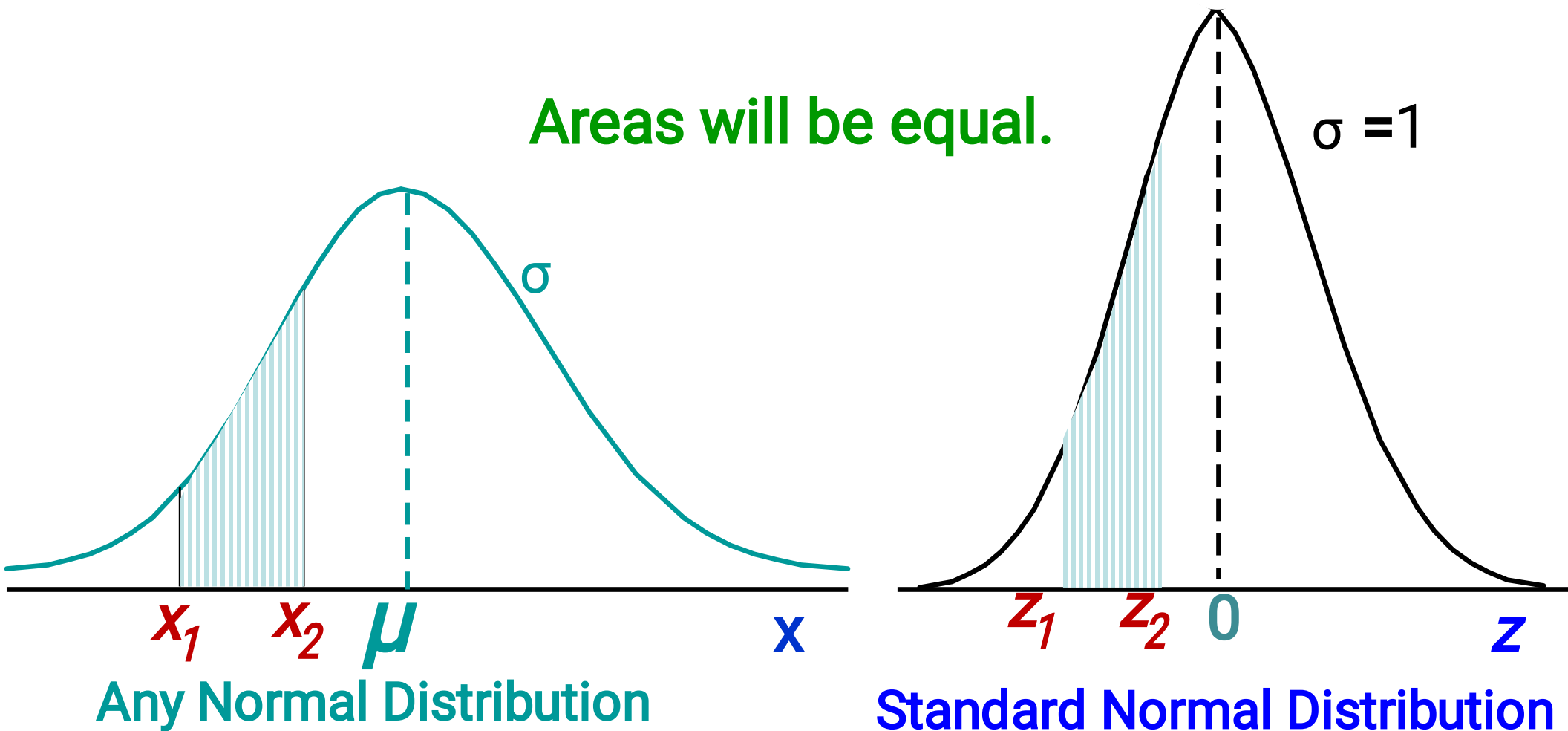
- Since we only have a standard normal table, then we need to *transform* our non-standard normal distribution into a standard one by the formula

$$Z = \frac{X - \mu}{\sigma}$$

- This process is called **standardization**

# The Standard Normal Distribution

- The new distribution is called **Standard Normal Distribution**, with mean equal to 0 and its standard deviation equal to 1



## ***z-scores:***

z-scores are "standard scores".

A z-score states the position of a raw score in relation to the mean of the distribution, using the standard deviation as the unit of measurement.

$$z = \frac{\text{raw score} - \text{mean}}{\text{standard deviation}}$$

for a population :

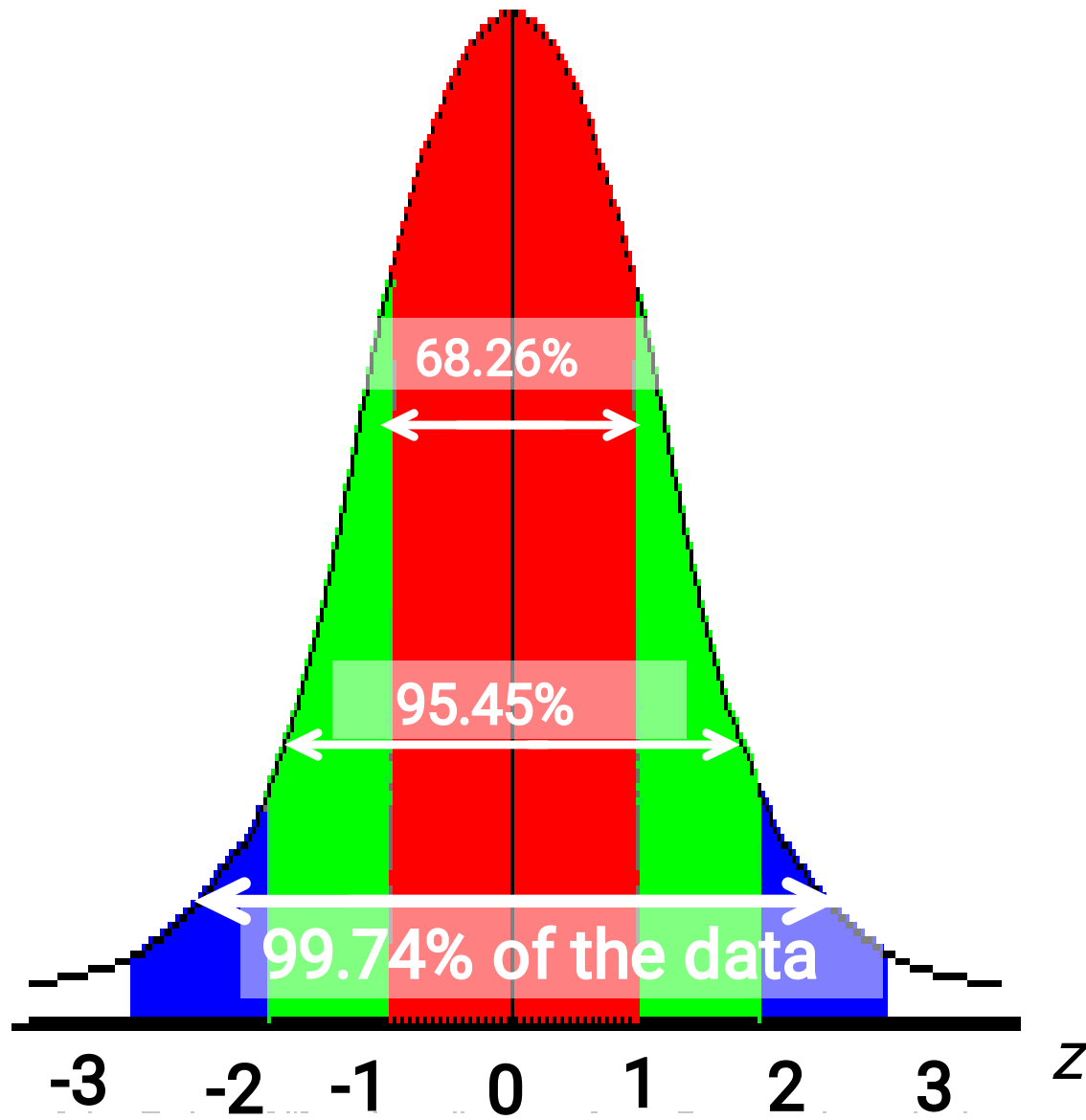
$$z = \frac{X - \mu}{\sigma}$$

for a sample :

$$z = \frac{X - \bar{X}}{s}$$

1. Find the *difference* between a score and the mean of the set of scores.
2. Divide this difference by the SD (in order to assess how big it really is).

U



## Why use z-scores?

1. z-scores make it easier to compare scores from distributions using different scales.

e.g. **two tests:**

Test A: Ahmad scores 78. Mean score = 70, SD = 8.

Test B: Ahmad scores 78. Mean score = 66, SD = 6.

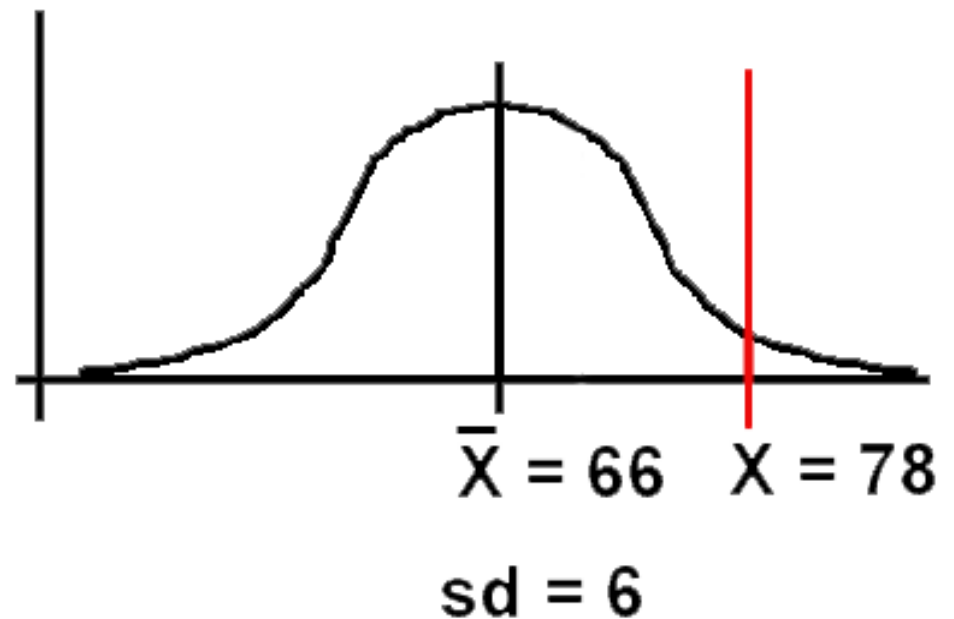
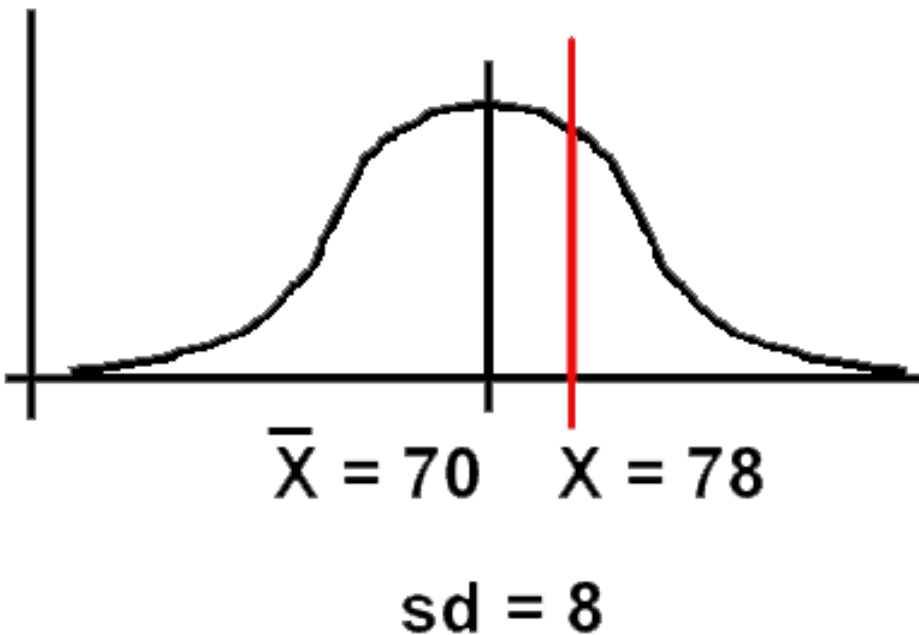
Did Ahmad do better or worse on the second test?



Test A: as a z-score,  $z = (78-70) / 8 = 1.00$

Test B: as a z-score,  $z = (78 - 66) / 6 = 2.00$

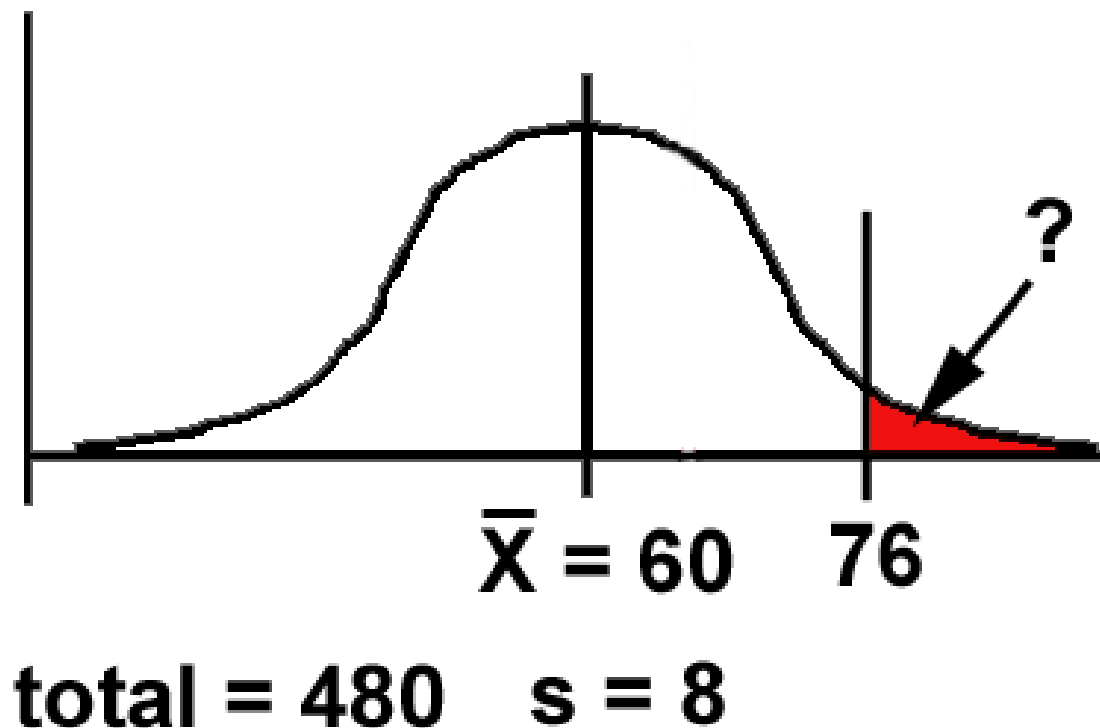
Conclusion: Ahmad did much better on Test B.

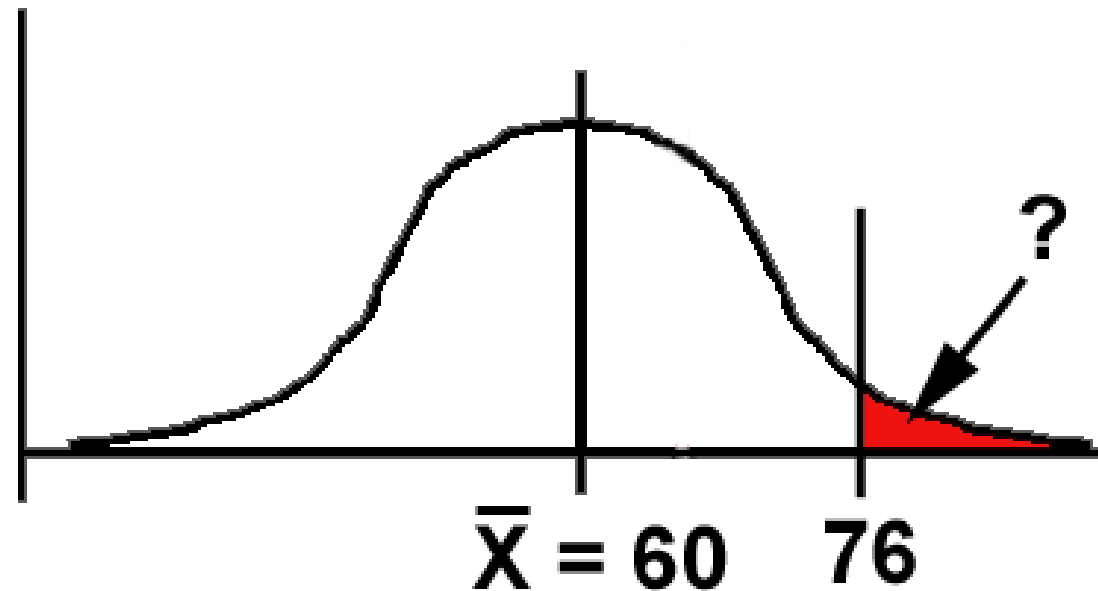


2. z-scores enable us to determine the relationship between one score and the rest of the scores, using just one table for all normal distributions.

e.g. If we have 480 scores, normally distributed with a mean of 60 and an SD of 8, how many would be 76 or above?

(a) Graph the problem:





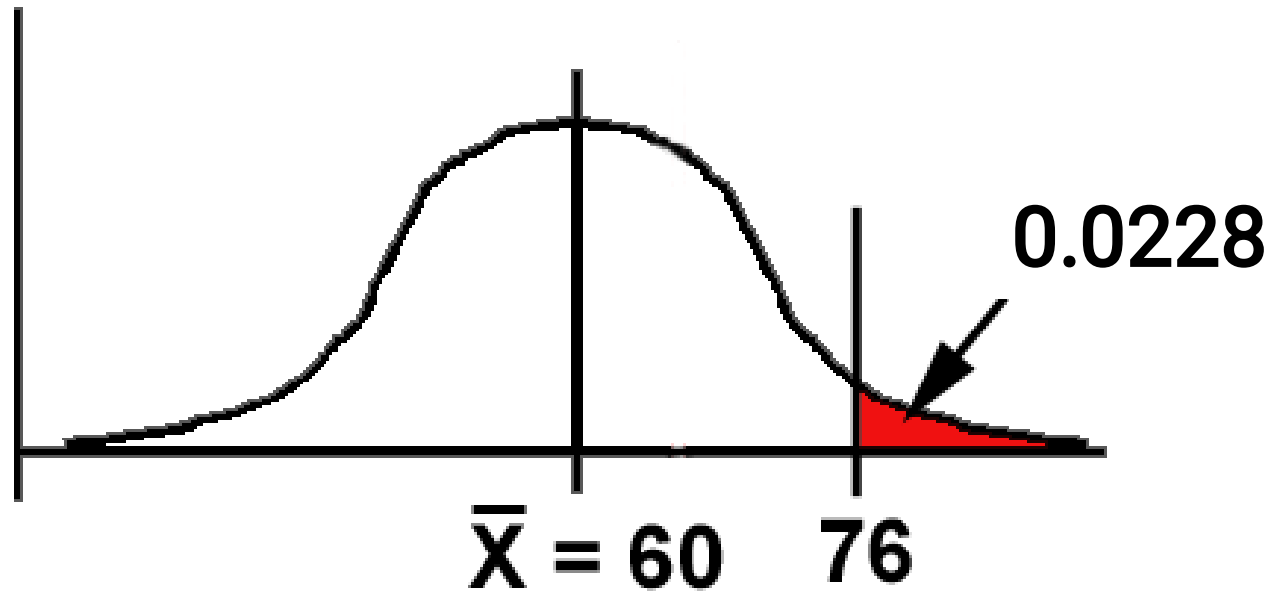
**total = 480    s = 8**

(b) Work out the z-score for 76:

$$z = (X - \bar{X}) / s = (76 - 60) / 8 = 16 / 8 = 2.00$$

(c) We need to know the size of the **area beyond z** (remember - the area under the normal curve corresponds directly to the proportion of scores).

**area beyond z=2 is 0.50-0.4772=0.0228**



**total = 480    s = 8**

**(d) So: as a proportion of 0.0228 of scores are likely to be 76 or more.**

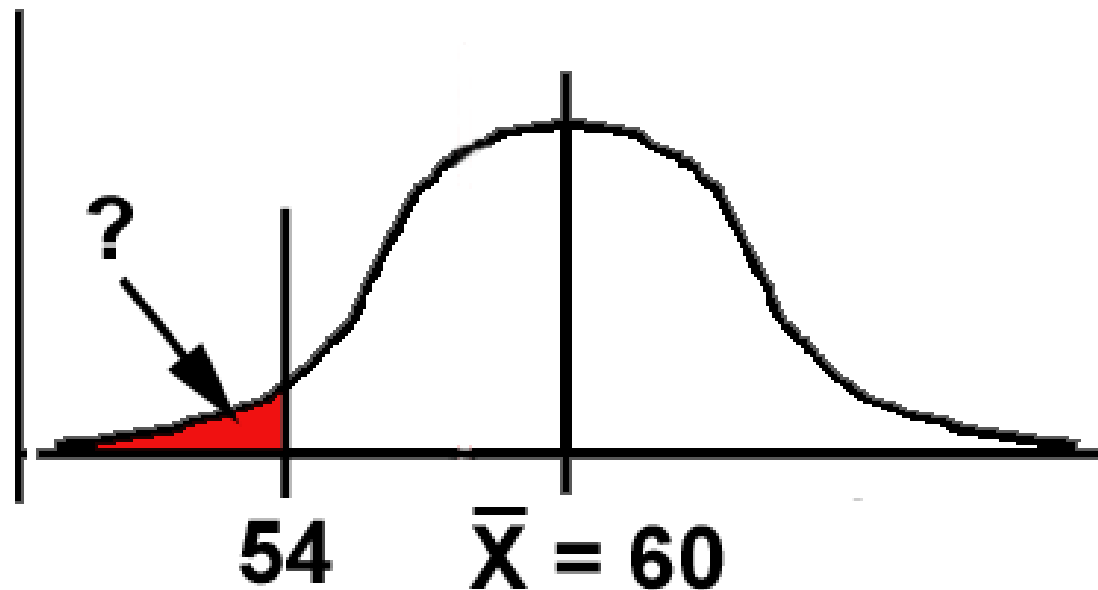
**As a percentage, = 2.28%**

**As a number,  $0.0228 * 480 = 10.94$  scores.**

How many scores would be 54 or less?

301

Graph the problem:



**total = 480    s = 8**

$$z = (X - \bar{X}) / s = (54 - 60) / 8 = -6 / 8 = -0.75$$

Use table by *ignoring* the sign of  $z$ : “area beyond  $z$ ” for  $0.75 = 0.2266$ . Thus 22.7% of scores (109 scores) are 54 or less.

# Standard Normal Distribution Tables

- By standardising any normally distributed random variable, we can use just the table namely,

*Areas Under the Normal Curve*

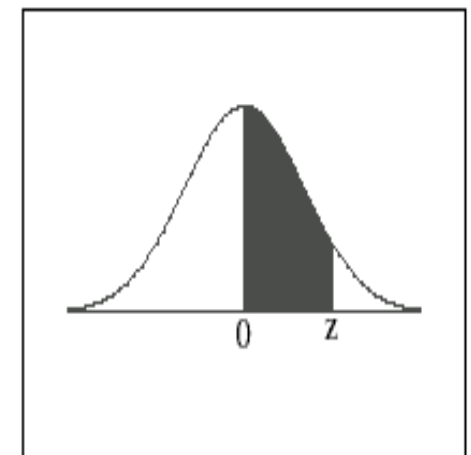
*Or*

*Areas of a Standard Normal Distribution*

Such tables are usually found in the Appendix of any statistics book.

# Z-Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998



# Use of the Normal Probability Table

Entries give the probability that a standard normally distributed random variable will assume a value between 0 and a given  $z$  value.

## Example:

Find the probability that  $z$  is between 0 and 1.74.

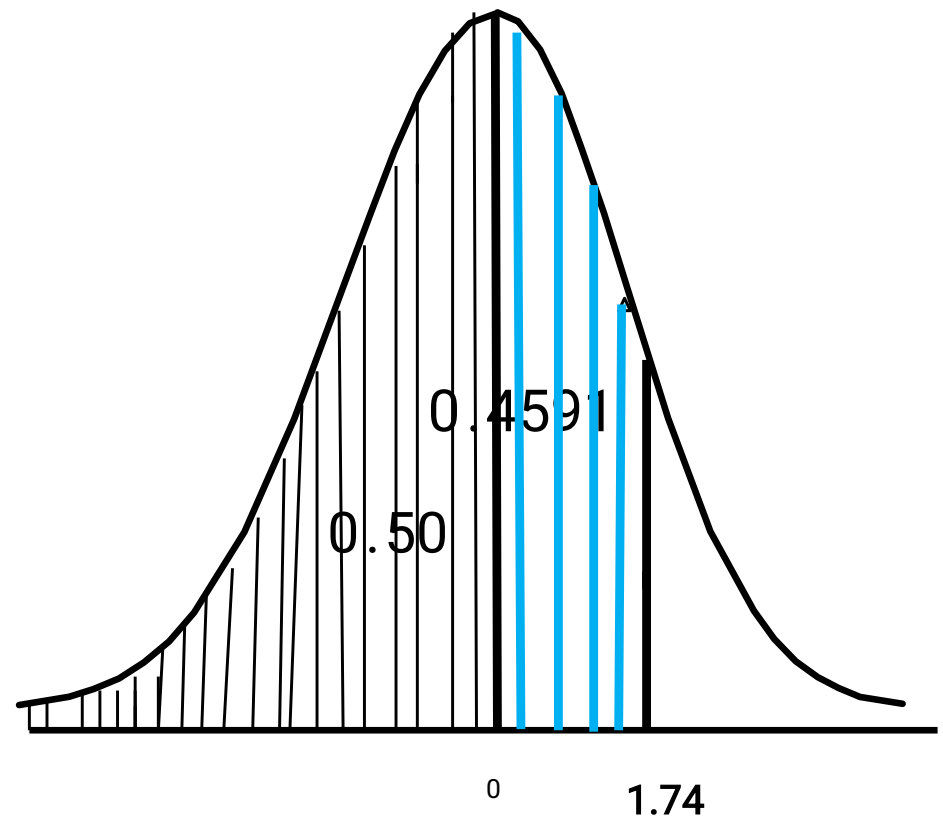
- Locate a value of  $z$  equal to 1.7 in the left column.
- Move across the row to the column under 0.04, where we read 0.9591.

Therefore,

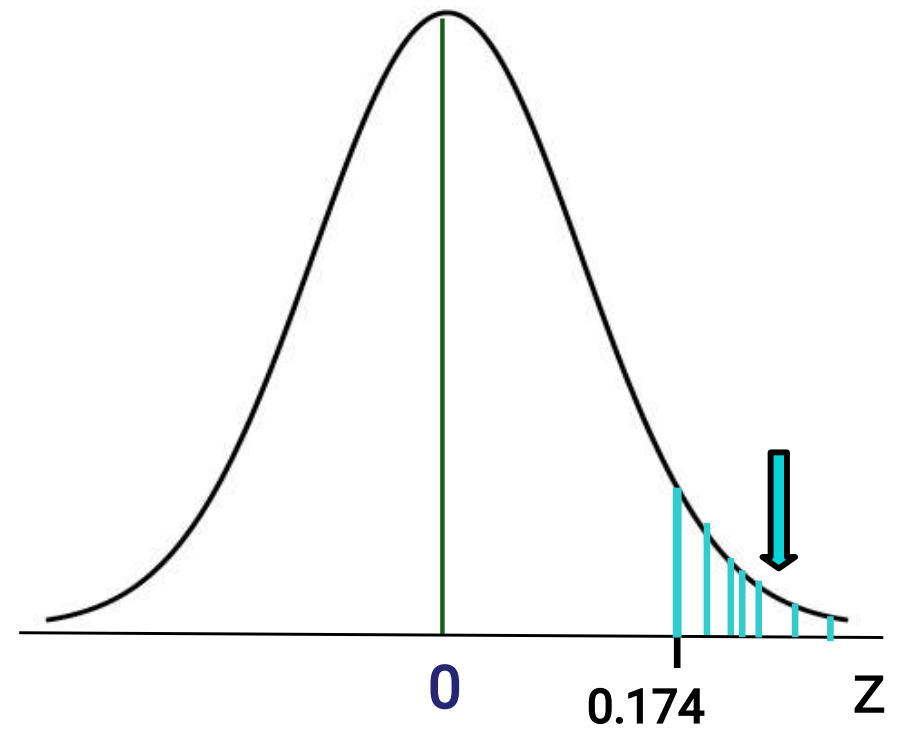
$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633



$$\begin{aligned} \text{ii. } P(z < 1.74) &= 0.50 + p(0 < z < 1.74) \\ &= 0.50 + 0.4591 = 0.9591 \end{aligned}$$



$$p(z > 1.74) = 0.50 - p(0 < z < 1.74) = 0.50 - 0.4591 = 0.0409$$



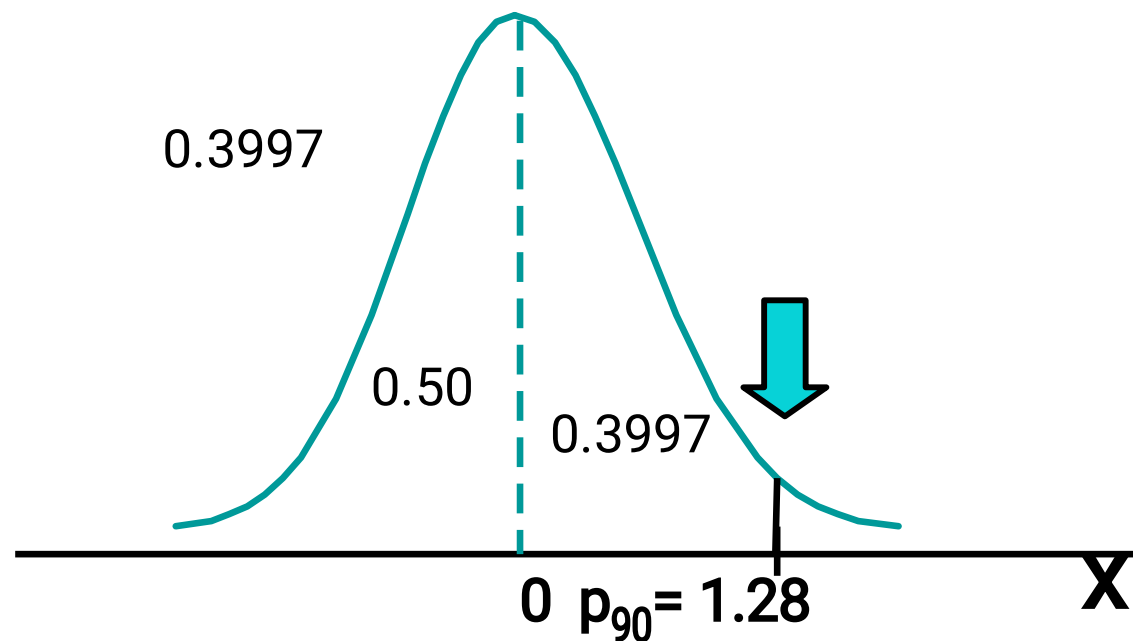
- $p(z < 1.28) = 0.5 + 0.3997 = 0.8997 \approx 0.90$
- $P(z > 1.28) = 0.5 - 0.3997 = 0.1003$
- $P(z < -1.28) = 0.1003 \approx 10\%$
- So the z value corresponding to the  $P_{90} = 1.28$ , and the z value corresponding to the  $P_{10} = -1.28$
- $p(z < 1.64) = 0.5 + 0.4495 = 0.9495 \approx 95$ .  
So the z value corresponding to the  $P_{95} = 1.64$

## Exercise

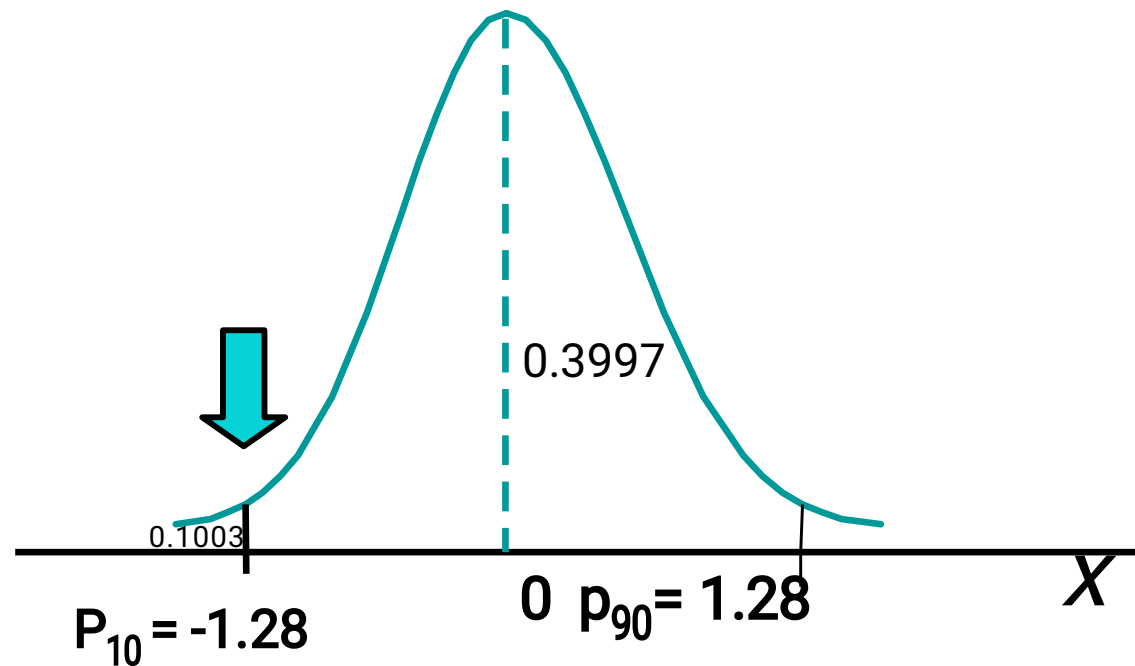
Find the z value corresponding to  $Q_1$  and  $Q_3$

To find percentiles

$P_{90}$  is the value that is preceded by 90% and followed by 10% of the ranked data



$P_{10}$  is the value that is preceded by 10% and followed by 90% of the ranked data



# Computing Normal Probabilities

1. State the problem.
2. What is the appropriate probability statement?
3. Draw a picture and shade required area
4. Convert to a standard normal distribution
5. Find the probability in the standard normal table

**Example:** Suppose the number of a particular type of bacteria in samples of 1-ml of drinking water tend to be approximately normally distributed with  $\mu = 85$  and  $\sigma^2 = 81$ .

a. What is the probability that a given 1-ml sample will contain

i. more than 100 bacteria?

ii. Between 90 and 100

iii. Between 70 and 100

iv. Less than 90

v. Less than 70.

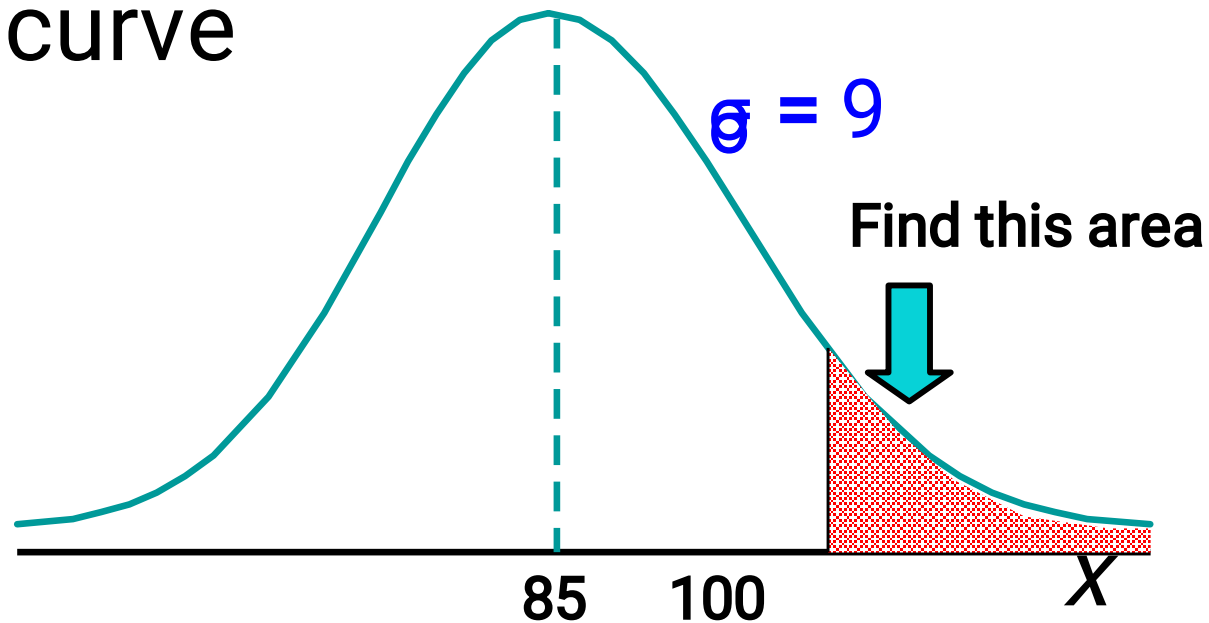
b. Is it possible that the number of this type of bacteria to be more than 125 in a sample of 1-ml of this water?

a.

i. More than 100

We are to find the  $P(x > 100)$

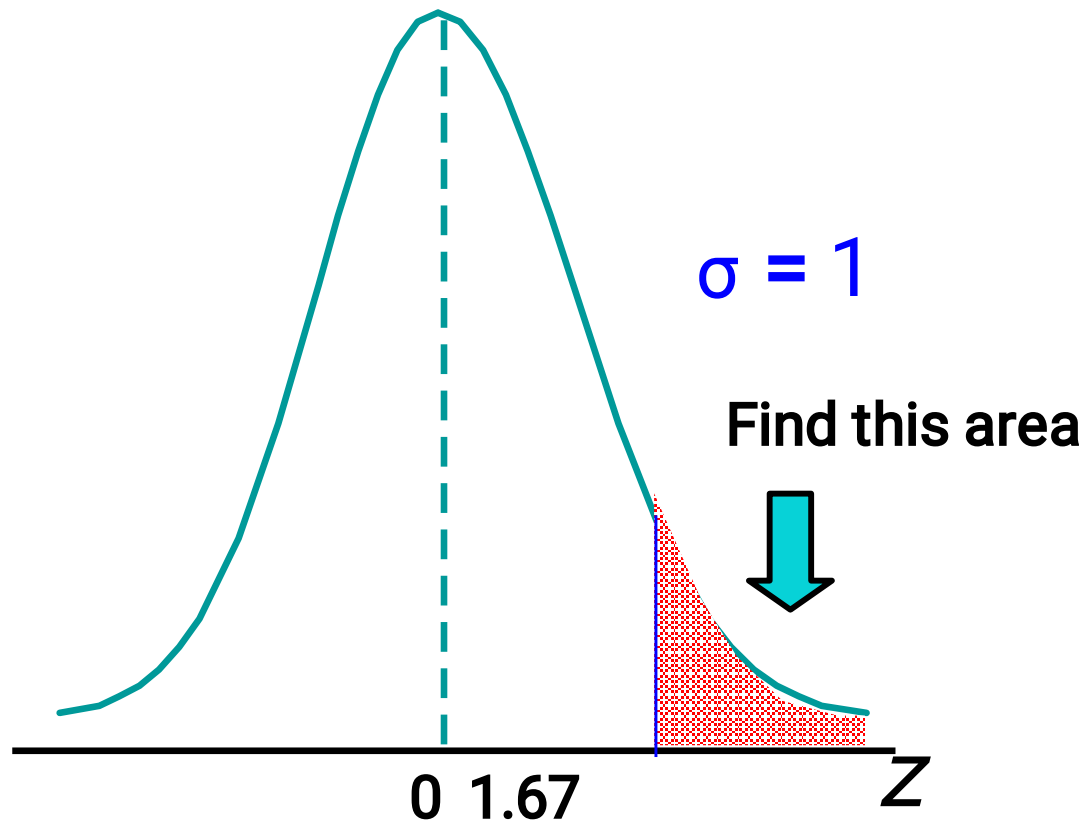
Sketch a curve



To find the  $P(x > 100)$ , we need to evaluate the area under the normal curve to the right of  $x = 100$ .



To find the  $P(x > 100)$ , we need to evaluate the area under the normal curve to the right of  $z = 1.67$

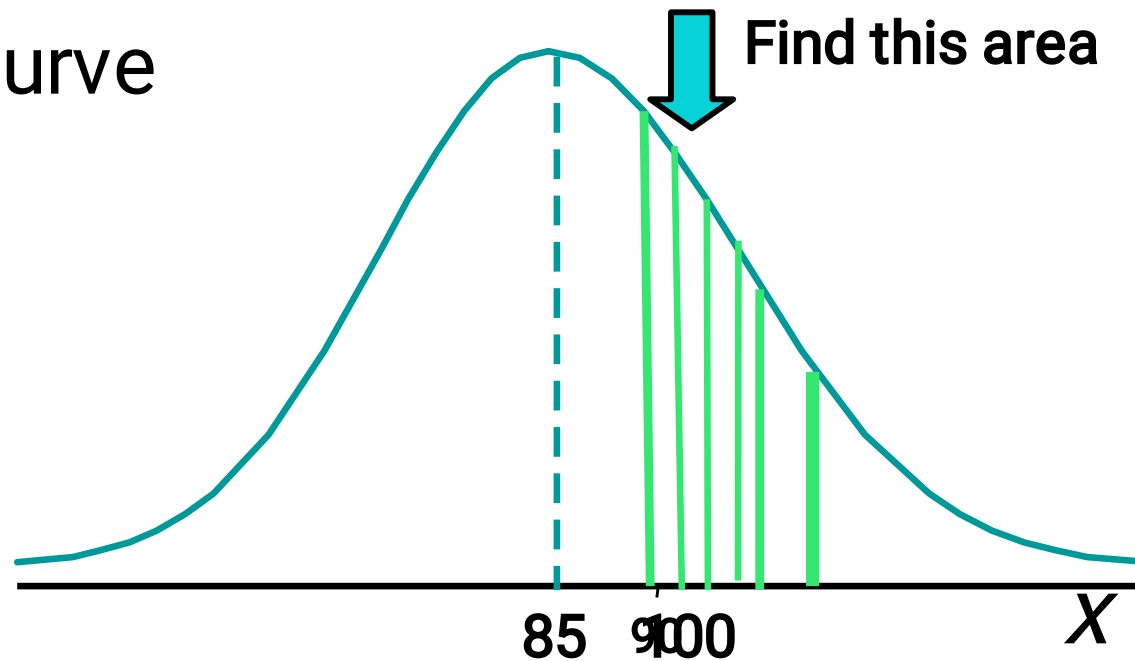


$$\begin{aligned} P(X > 100) &= p\left(z > \frac{100 - 85}{9}\right) = P(z > 1.670) \\ &= 0.50 - 0.4525 = 0.0475 \end{aligned}$$

## ii. Between 90 and 100

We are to find the  $P(90 < x < 100)$

Sketch a curve



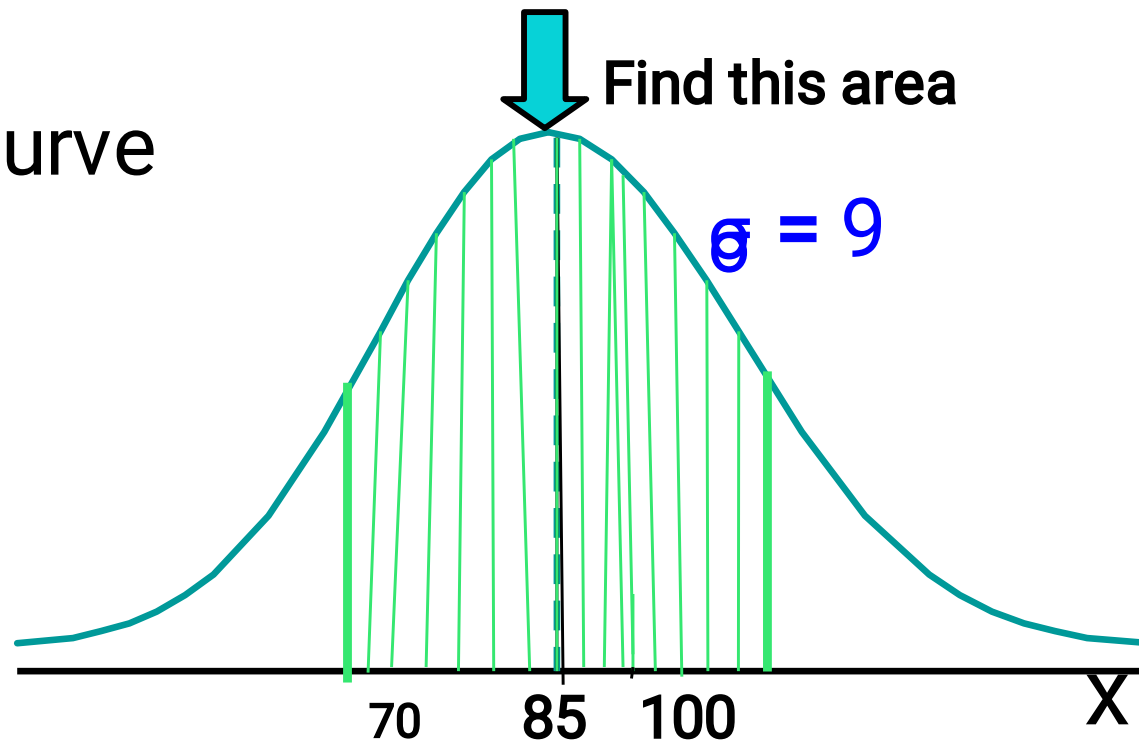
To find the  $P(90 < x < 100)$ , we need to evaluate the area under the normal curve from  $x = 90$  to  $x = 100$ .

$$\begin{aligned} &P(90 < X < 100) \\ &= P\left(\frac{90 - 85}{9} < z < \frac{100 - 85}{9}\right) \\ &= P(0.56 < z < 1.67) \\ &= P(0 < z < 1.67) - P(0 < z < 0.56) \\ &= 0.4525 - 0.2123 = 0.2402 \end{aligned}$$

### iii. Between 70 and 100

We are to find the  $P(70 < x < 100)$

Sketch a curve



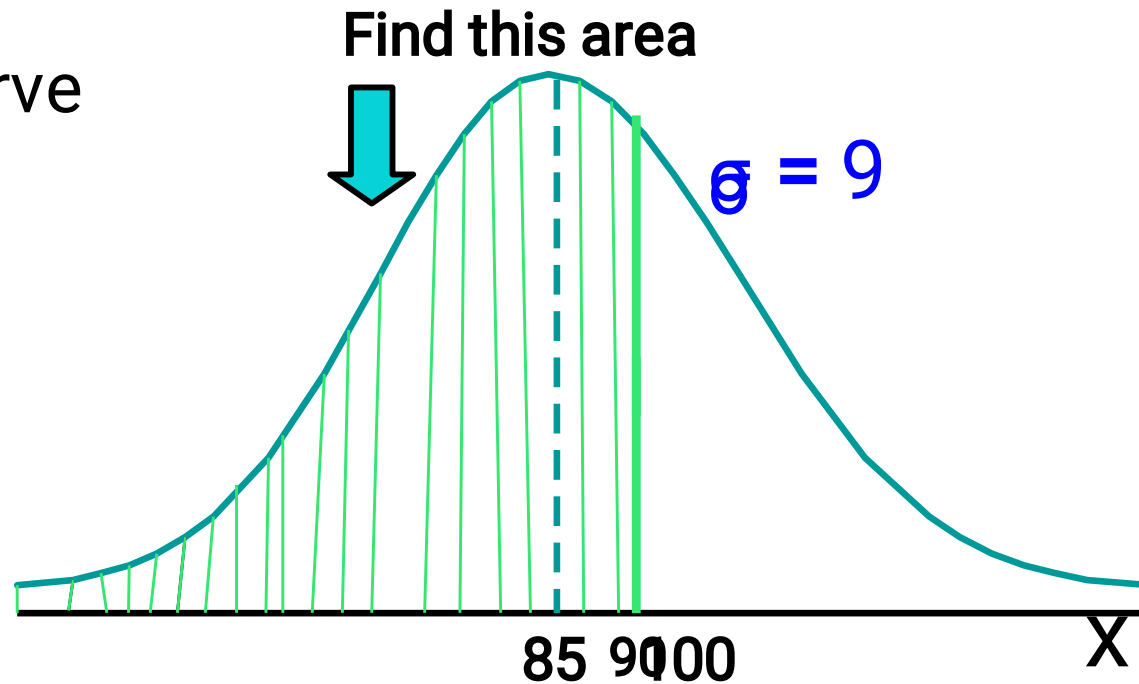
To find the  $P(70 < x < 100)$ , we need to evaluate the area under the normal curve from  $x = 70$  to  $x = 100$ .

$$\begin{aligned} &P(70 < X < 100) \\ &= P\left(\frac{70 - 85}{9} < z < \frac{100 - 85}{9}\right) \\ &= P(-1.67 < z < 1.67) \\ &= 2P(0 < z < 1.67) = 2(0.4525) \\ &= 0.9050 \end{aligned}$$

## iv. Less than 90

We are to find the  $P(x < 90)$

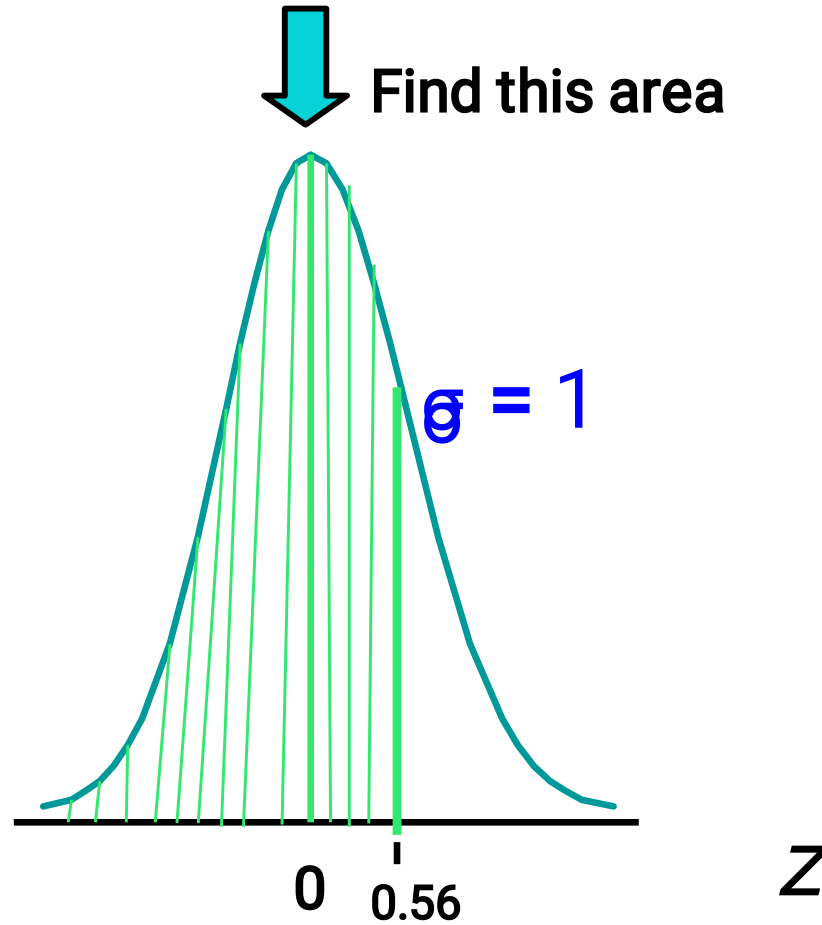
Sketch a curve



To find the  $P(x < 90)$ , we need to evaluate the area under the normal curve to the left of  $x = 90$ .

# Less than 90

Sketch a curve



To find the  $P(z < 90)$ , we need to evaluate the area under the normal curve to the left of  $z = 0.56$ .

Less than 90

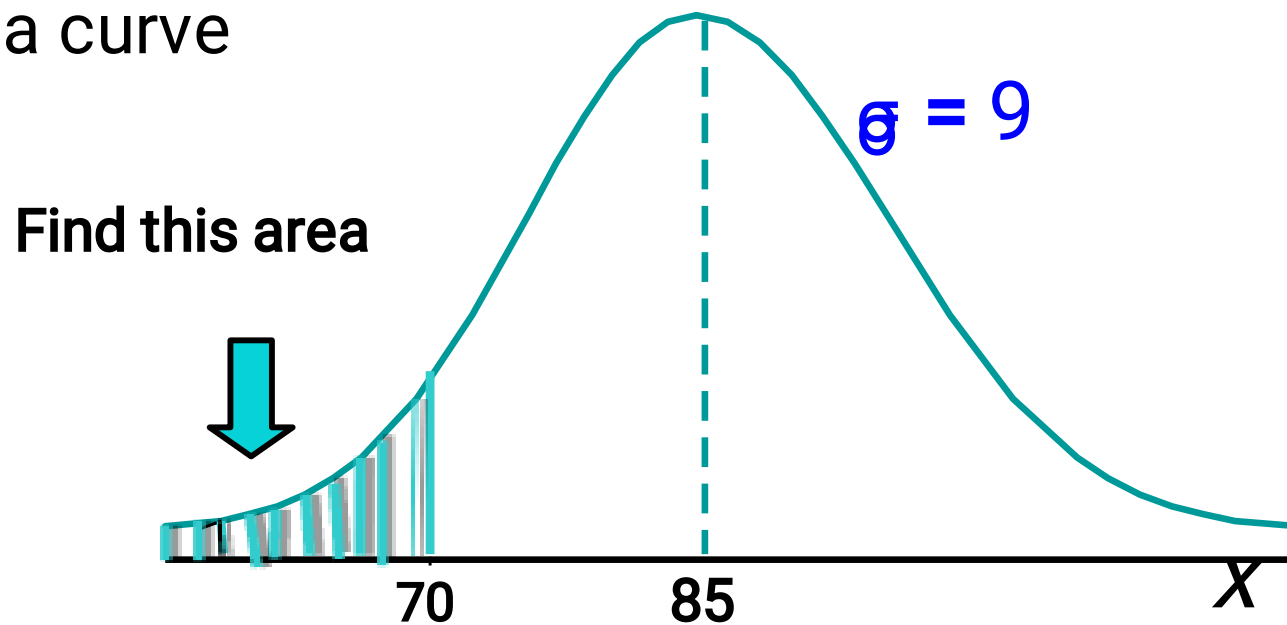
$$\begin{aligned} P(X < 90) &= P\left(z < \frac{90 - 85}{9}\right) \\ &= P(z < 0.56) = 0.5 + 0.2123 = 0.7123 \end{aligned}$$



## v. Less than 70

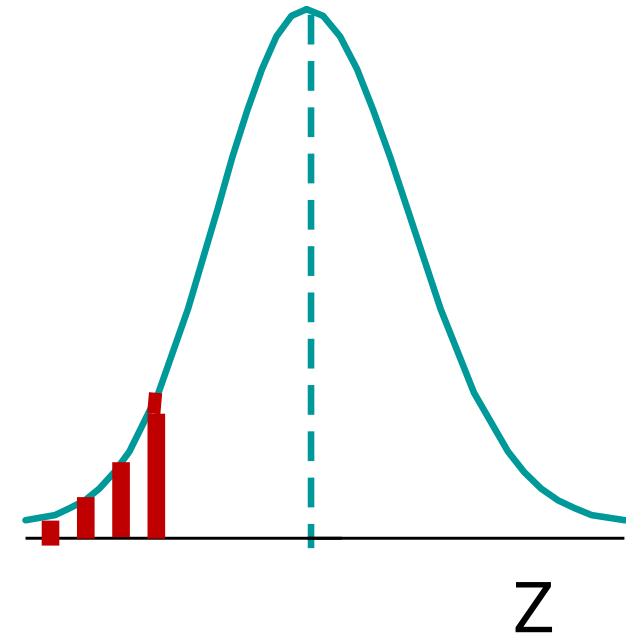
We are to find the  $P(x < 70)$

Sketch a curve



To find the  $P(x < 70)$ , we need to evaluate the area under the normal curve to the left of  $x = 70$ .

- $$\begin{aligned}
 P(X < 70) &= P\left(z < \frac{70-85}{9}\right) \\
 &= P(z < -1.67) \\
 &= 0.50 - P(0 < z < 1.67) \\
 &= 0.50 - 0.4525 \\
 &= 0.0475
 \end{aligned}$$



$$\text{b. } p(X > 125) - p\left(X > \frac{125 - 85}{9}\right) = p(X > 4.44) = 0$$

So it is impossible to have a number of bacteria more than 125.

### Finding $x$ value when probabilities (areas) are given

1. State the problem
2. Draw a picture
3. Use table to find the probability closest to the one you need
4. Read off the z-value
5. use the formula  $x = \mu + z\sigma$

# Example

For the previous example with  $\mu=85$  and standard deviation  $\sigma =9$ , **a.** find the value of  $x$  that

**i. 45% of the area below it**

**ii. 14% of the area above it.**

**b. Find  $P_{90}$ ,  $P_{10}$  for the normal distribution**

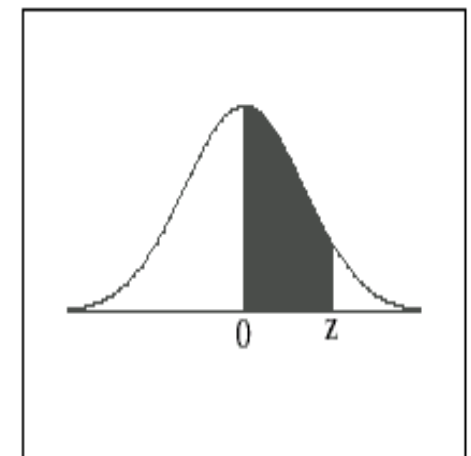
## Solution

In this problem we reverse the process and begin with a known area or probability, find the  $z$  value, and then determine  $x$  by the formula

$$x = z\sigma + \mu$$

# Z-Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998



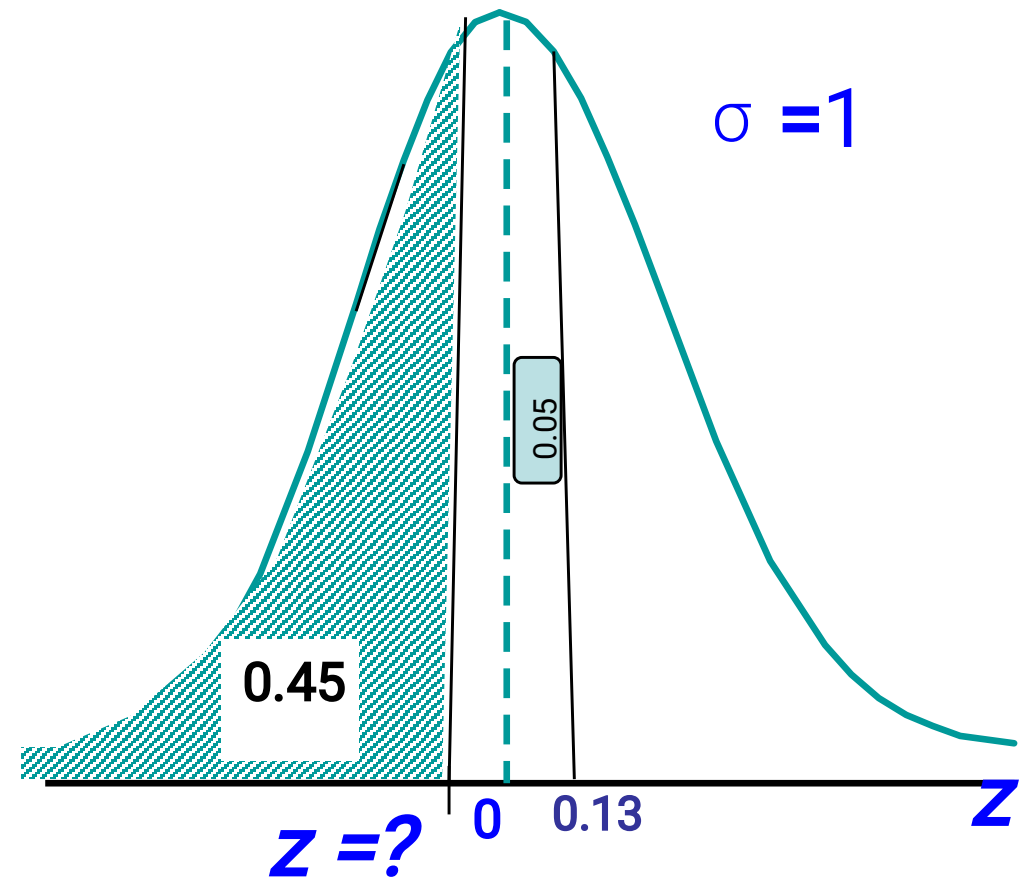
# Example cont.

a.

i. We require a  $z$  value that leaves an area of 0.45 to the left.

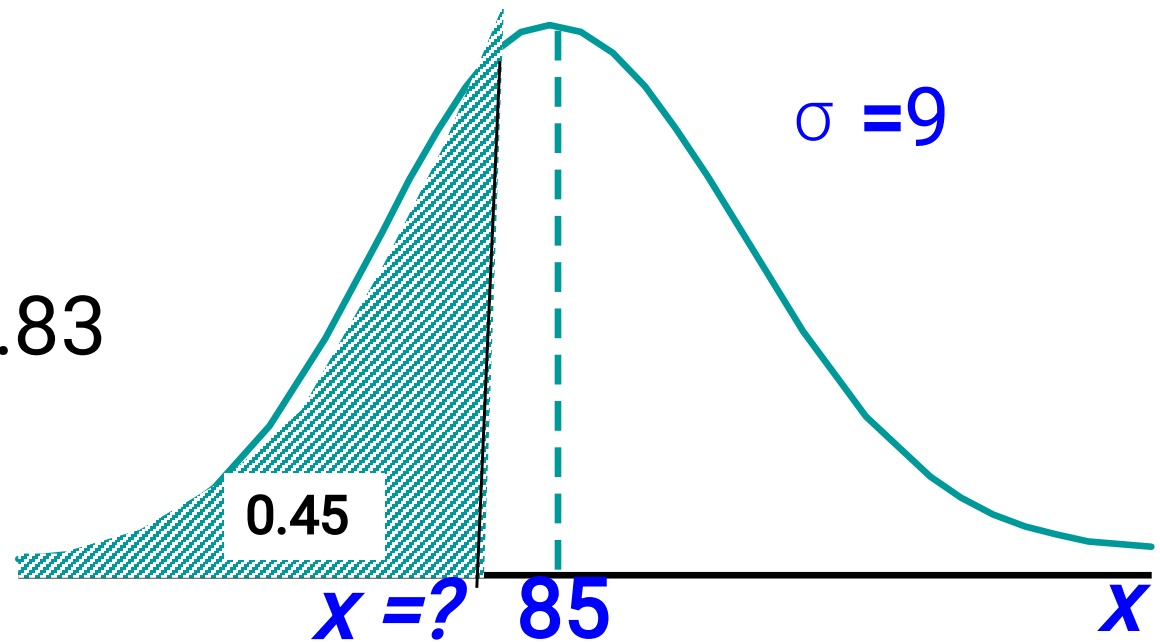
From the table we find  $P(0 < z < 0.13) = 0.0517$  so that the desired  $z$  value is  $-0.13$ .

$$x = z\sigma + \mu$$



# Example cont.

$$x = z\sigma + \mu$$
$$= (-0.13 \times 9) + 85 = 83.83$$



# Example cont.

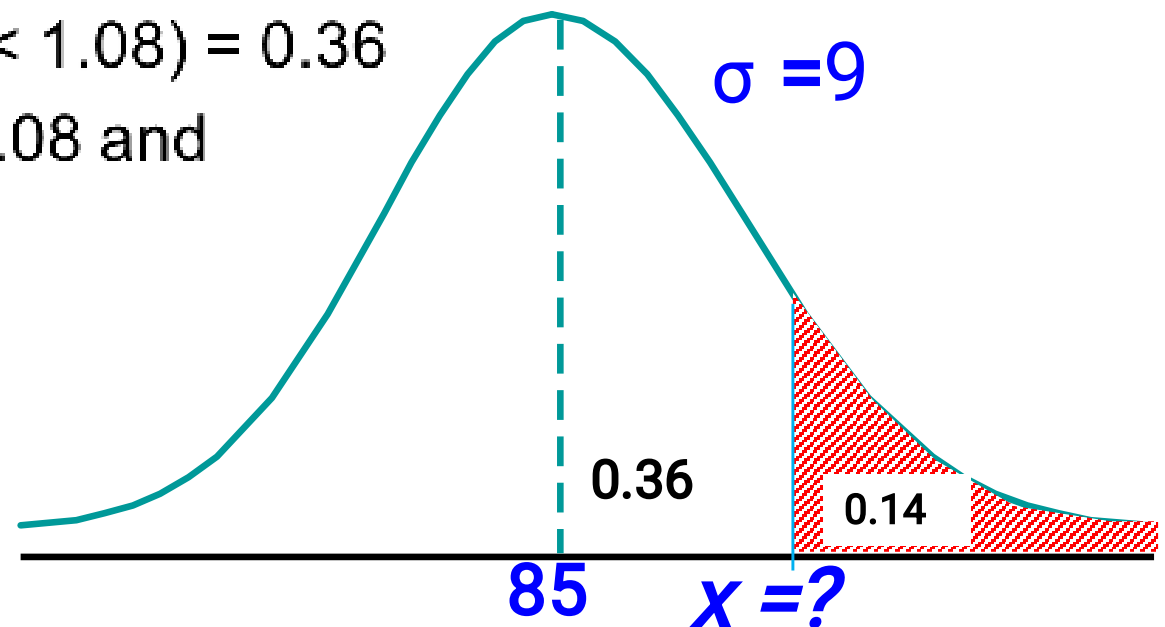
## ii. 14% of the area above it.

This time we require a z value that leaves 0.14 of the area to the right and hence an area of 0.36 between z and 0.

From the table we find  $P(0 < z < 1.08) = 0.36$   
so that the desired z value is 1.08 and

$$x = z\sigma + \mu$$

$$\begin{aligned} x &= (1.08)(9) + 85 \\ &= 94.72 \end{aligned}$$



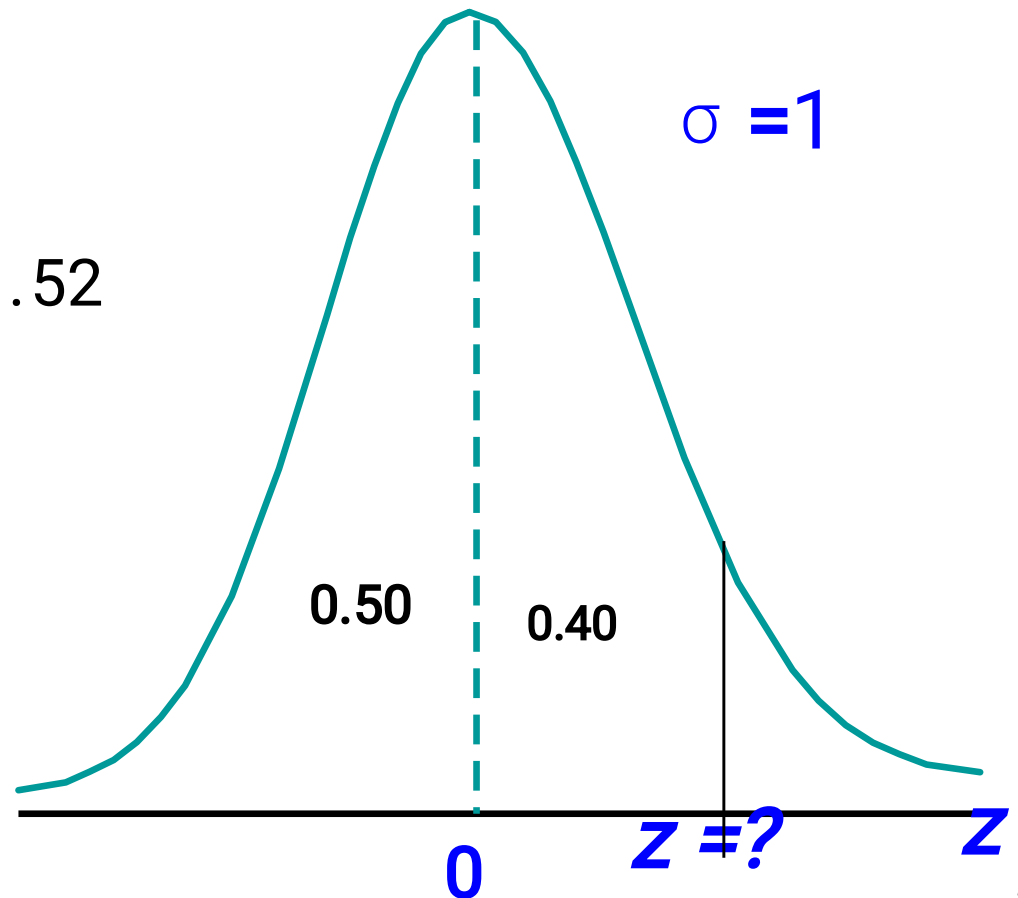


# Example cont.

b. We require a  $z$  value that leaves an area of 0.90 to the left and hence an area of 0.40 between  $z$  and 0.

From the table we find  $P(0 < z < 1.28) = 0.3997$ , so that the desired  $z$  value is **1.28**. So

$$\begin{aligned}x &= z\sigma + \mu \\ &= 1.28(9) + 85 = 96.52\end{aligned}$$

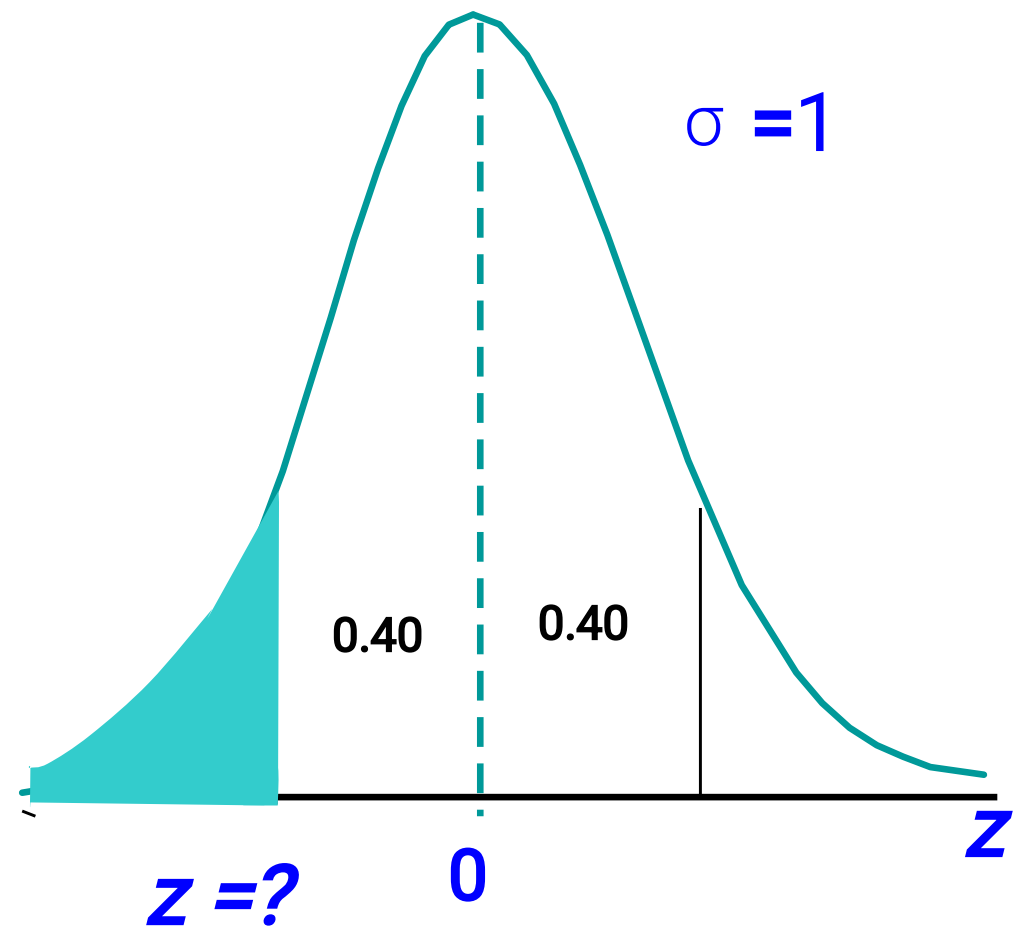


## Example cont.

b. We require a  $z$  value that leaves an area of 0.10 to the left and hence an area of 0.40 between  $z$  and 0.

From the table we find  $P(0 < z < 1.28) = 0.3997$ , so that the desired  $z$  value is  $-1.28$ . So

$$\begin{aligned} x &= z\sigma + \mu \\ &= -1.28(9) + 85 = 73.48 \end{aligned}$$



# Exercise

If the total cholesterol values for a certain target population are approximately normally distributed with a mean of 200 (mg/100 mL) and a standard deviation of 20 (mg/100 mL),

- a. What is the probability that a person picked at random from this population will have a cholesterol value
  - i. greater than 240 (mg/100 mL)?
  - ii. between 180 and 240
  - iii. Less than 180.
- b. what is the value of cholesterol above which 5% of this population have.
- c. Is it possible for a person in such a population to have cholesterol level more than 300.
- c. Find IQR for cholesterol level of tis population.
- d. If two persons are picked at random from this population, what is the probability that the have cholesterol level greater than 240

# Chapter 6

## Correlation and Regression

# Correlation and Regression

In analyzing data for branches of science, we find that it is frequently desirable to investigate the relationship between two or more variables. Correlation and regression analysis are two statistical techniques that are used to examine the **nature and strength** of the relationships between two variables.

e.g.

- Blood pressure and age
- Height and weight
- The concentration of an injected drug and heart rate
- The consumption level of some nutrient and weight gain.
- Income and food Expenditure.
- Stress score before an exam and test score.
- IQ of the mother and IQ of her children.
- mother's pregnancy weight and infant's birth weight

# Correlation

- Correlation analysis is concerned with measuring the strength and the nature of the relationship between variables.
- When we compute measures of correlation from a set of data, we are interested in the degree of the correlation between variables.

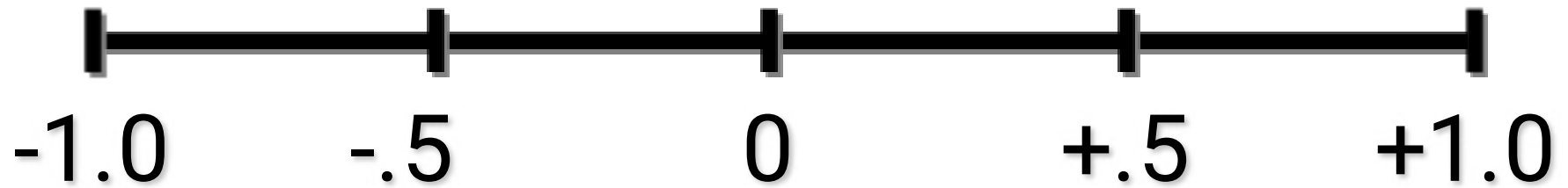
- **Correlation** : is concerned with measuring the strength and the nature of the relationship between variables.
- **Regression** : is used to predict or estimate the value of one variable corresponding to a given value of another variable.
- **Scatter Diagram**: is a chart that portrays the relationship between the two quantitative variables.
- One is called independent variable  $x$  and the second is called dependent variable  $y$

# Correlation coefficient

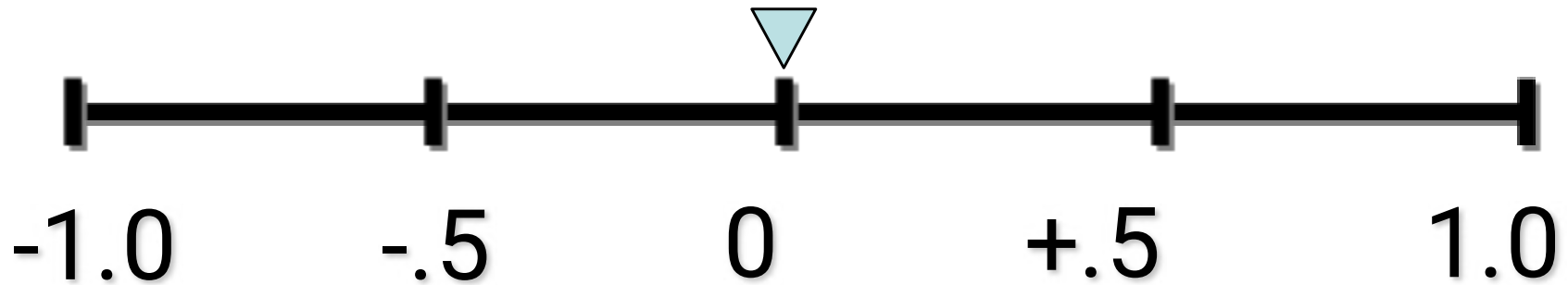
- Correlation coefficient of variables  $x$  and  $y$  shows how strongly the values of these variables are related to one another.
- It is denoted by  $r$ , where  $r \in [-1, 1]$
- If the correlation coefficient is positive, then both variables are simultaneously increasing (or simultaneously decreasing).  
The relation is positive (direct)
- If the correlation coefficient is negative, then one variable increases while the other decreases, and reciprocally. The relation is negative (inverse)



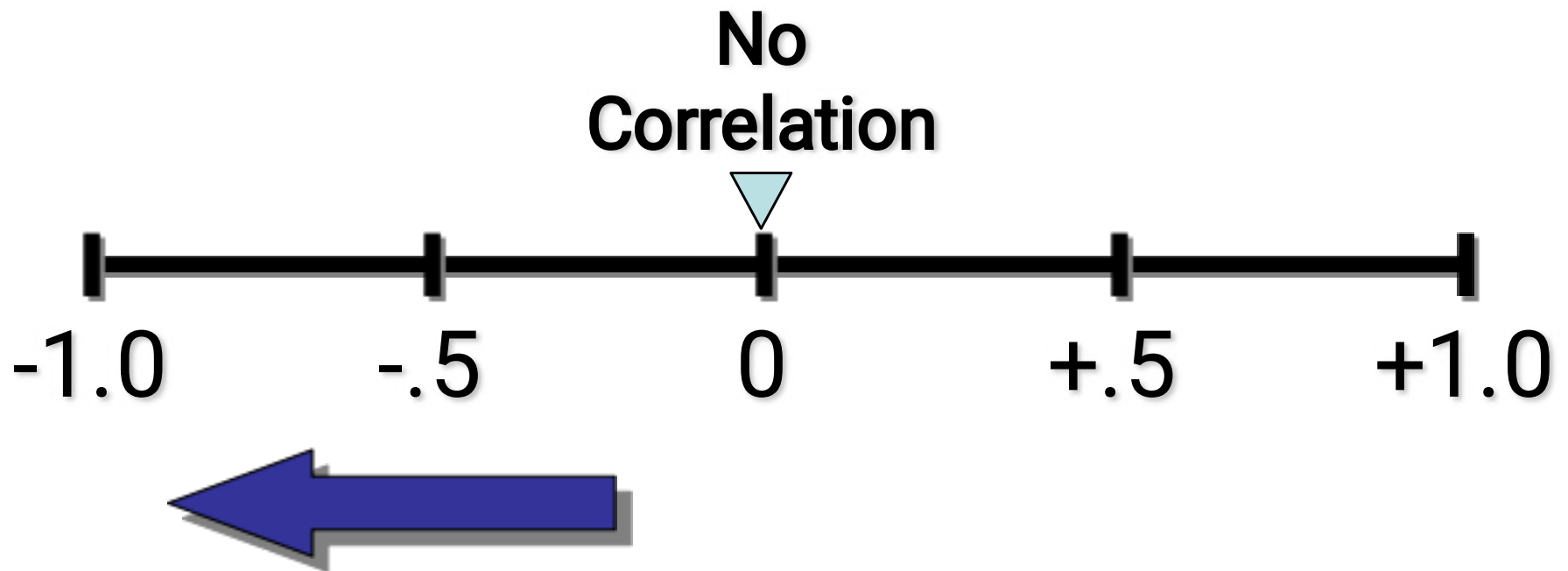
# Coefficient of Correlation Values



No correlation



# Coefficient of Correlation Values

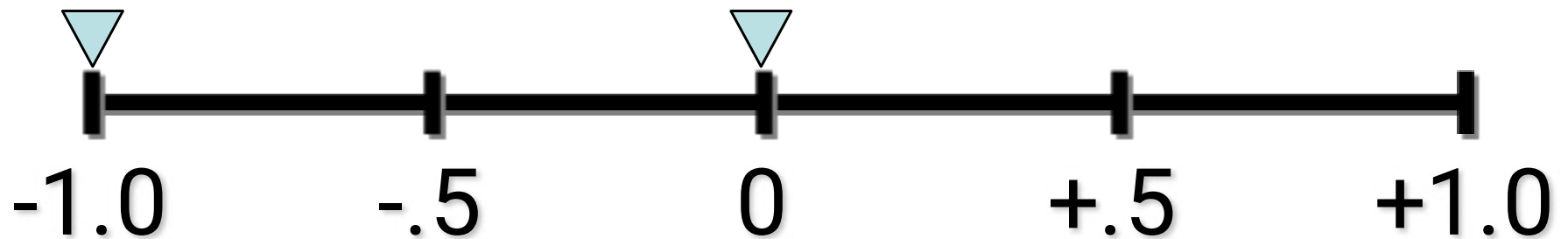


**Increasing degree of  
negative correlation**

# Coefficient of Correlation Values

Perfect  
Negative  
Correlation

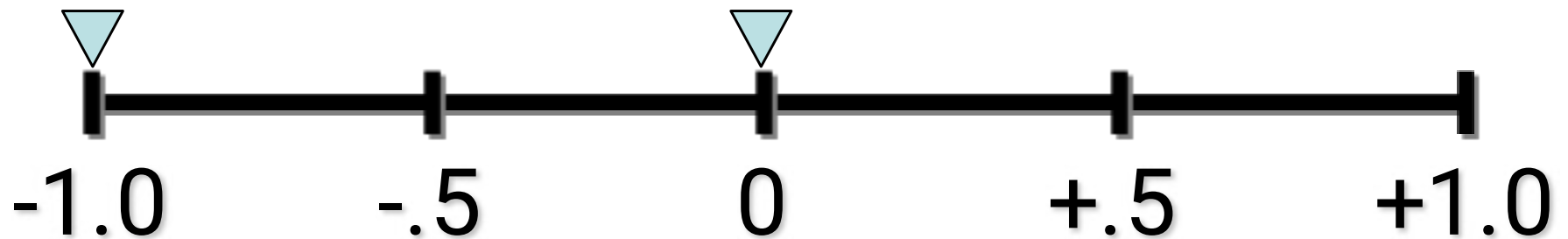
No Correlation



# Coefficient of Correlation Values

Perfect  
Negative  
Correlation

No Correlation



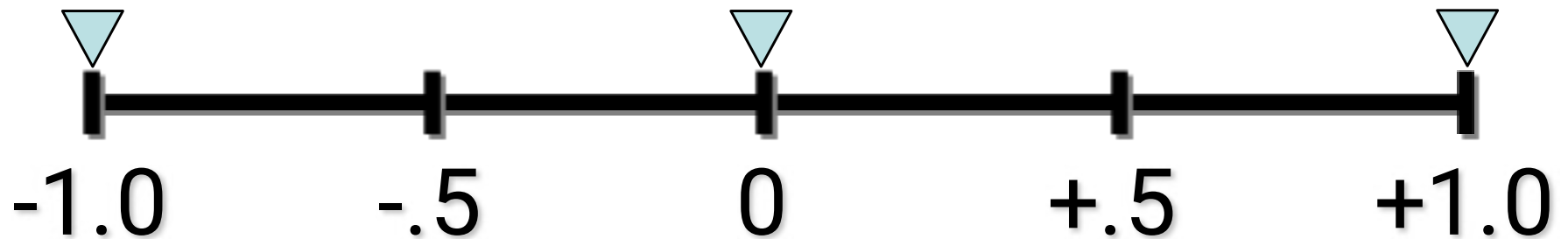
Increasing degree of  
positive correlation

# Coefficient of Correlation Values

**Perfect  
Negative  
Correlation**

**No Correlation**

**Perfect  
Positive  
Correlation**



# Characteristics of correlation coefficient

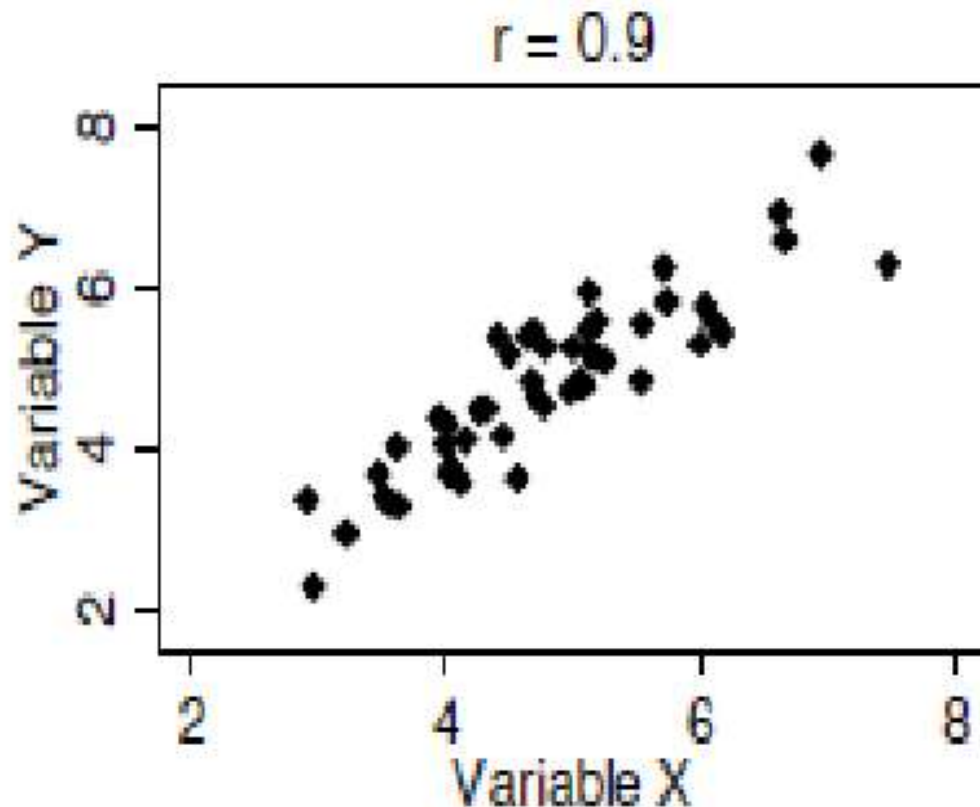
Although there is no fixed rule or interpretation of the strength of a correlation, we will say that the relation is

- Strong if  $r \in [0.8, 1]$  or  $r \in [-1, -0.8]$ ,
- Moderate if  $r \in (0.5, 0.8)$  or  $r \in (-0.8, -0.5)$ ,
- Weak if  $r \in [-0.5, 0.5]$ .

We will also add the words positive or negative to indicate the type of correlation.

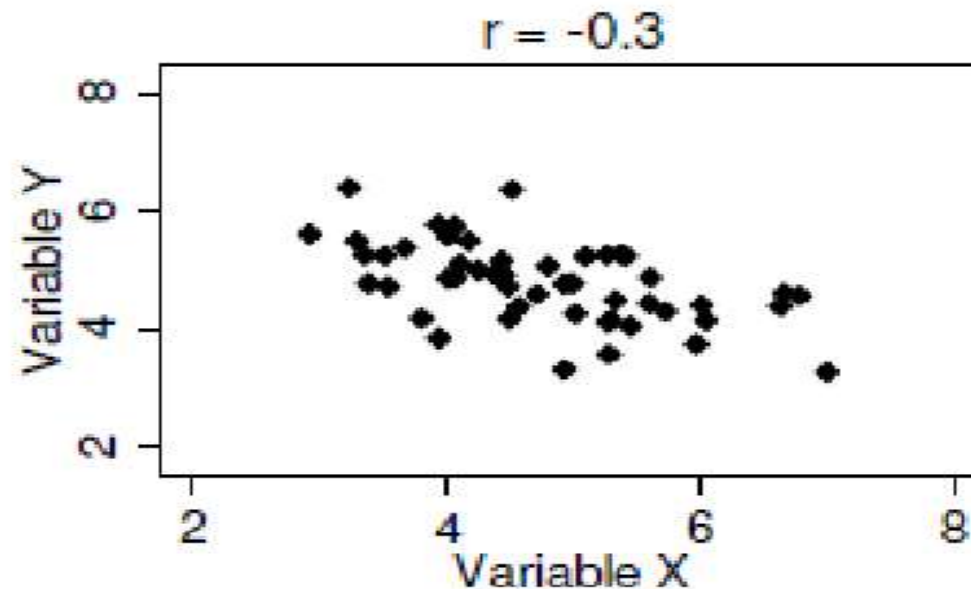
# Correlation coefficient

- Positive when large values of one variable are associated with large values of the other.



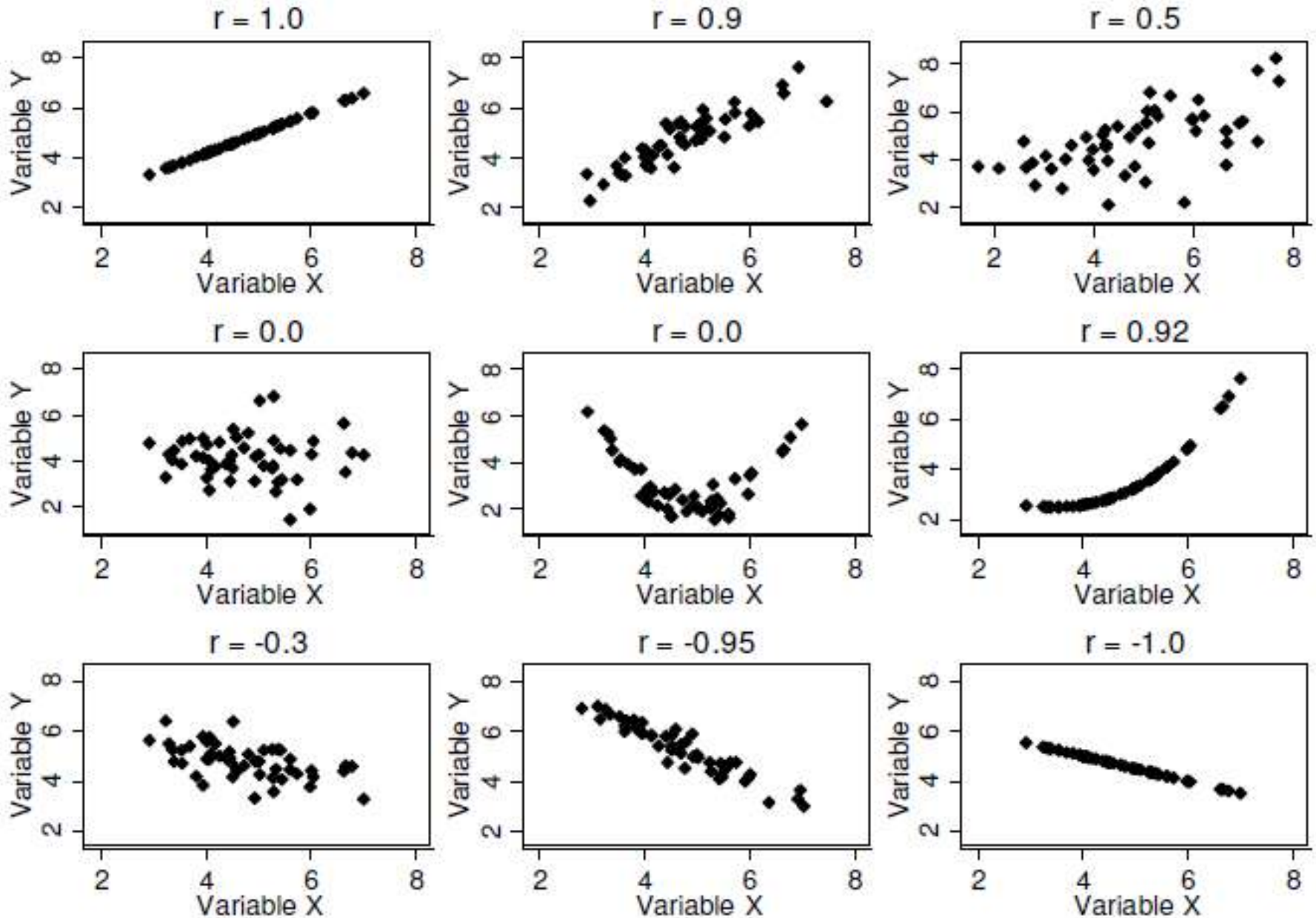
# Correlation coefficient

- Negative when large values of one variable are associated with small values of the other.





# Correlation coefficient



# Simple Correlation coefficient

- It is also called **Pearson's correlation coefficient**, it measures the nature and strength between two variables of the quantitative type.
- The simple correlation coefficient is obtained using the following formula:

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) \left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

- where  $n$  is the sample size,  $x$  is the independent variable and  $y$  is the dependent variable.

# Example

We find the correlation coefficient between PH and optical density of materials.

pH ( $x$ )	Optical density ( $y$ )	$xy$	$x^2$	$y^2$	
3	0.1	0.3	9	0.01	
4	0.2	0.8	16	0.04	
4.5	0.25	1.125	20.25	0.0625	
5	0.32	1.6	25	0.1024	
5.5	0.33	1.815	30.25	0.1089	
6	0.35	2.1	36	0.1225	
6.5	0.47	3.055	42.25	0.2209	
7	0.49	3.43	49	0.2401	
7.5	0.53	3.975	56.25	0.2809	
<b>Total</b>	<b>49</b>	<b>3.04</b>	<b>18.2</b>	<b>284</b>	<b>1.1882</b>

# Coefficient of Correlation

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) \cdot \left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

$$r = \frac{18.2 - \frac{(49)(3.04)}{9}}{\sqrt{\left(284 - \frac{(49)^2}{9}\right) \cdot \left(1.1882 - \frac{(3.04)^2}{9}\right)}} = 0.989$$

$$r = 0.989 \approx 0.99$$

# Strength of the relation:

Note that

$r = 0.99$ , so there is

a strong direct relation between the variables

- The correlation coefficient  $r$  measures the strength of the relationship between two variables.
- Just because two variables are related does not imply that there is a cause-and-effect relationship between them.

# Regression analysis

- The ultimate objectives when this method of analysis is employed usually is to **predict** or **estimate** the value of one variable corresponding to a given value of another variable.

# Spearman Rank Correlation Coefficient

- It is a non-parametric measure of correlation used in the case of ordinal or qualitative (ratio or relative) variables.
- This procedure makes use of the two sets of ranks that may be assigned to the sample values of  $x$  and  $y$ .
- Spearman Rank correlation coefficient could be computed in the following cases:
  - Both variables are quantitative.
  - Both variables are qualitative ordinal.
  - One variable is quantitative and the other is qualitative ordinal.

- **Procedure:**

- Rank the values of  $X$  from 1 to  $n$ , where  $n$  is the numbers of pairs of values of  $X$  and  $Y$  in the sample.
- Rank the values of  $Y$  from 1 to  $n$ .
- Compute the value of  $d_i$  for each pair of observations by subtracting the rank of  $Y$  from the rank of  $X$ .
- Square each  $d_i$  and compute  $\sum d_i^2$  which is the sum of the squared values.
- Apply the following formula
- $$r_s = 1 - \frac{6 \sum (d_i)^2}{n(n^2 - 1)}$$



- **Example:** In a study of the relationship between education level and health awareness, the following data was obtained. Find the relationship between them and comment.

No.	Education level ( x )	Health awareness ( y )
1	preparatory.	25
2	primary.	10
3	university.	8
4	secondary	10
5	secondary	15
6	illiterate	50
7	university.	60

No.	( x )	( y )	Rank ( x )	Rank ( y )	d	
1	Preparatory	25	5	3	2	4
2	Primary	10	6	5.5	0.5	0.25
3	University	8	1.5	7	-5.5	30.25
4	secondary	10	3.5	5.5	-2	4
5	secondary	15	3.5	4	-0.5	0.25
6	illiterate	50	7	2	5	25
7	university	60	1.5	1	0.5	0.25
Total						64

$$r_s = 1 - \frac{6 \times 64}{7(48)} = -0.1$$

There is an indirect weak correlation between education level and health awareness

# Simple linear regression

- In simple linear regression we are interested in two variables  $x$  and  $y$ .
- The variable  $x$  is usually referred to as the *independent variable*, since frequently it is controlled by the investigator; that is; values  $x$  of  $x$  may be selected by the investigator and, corresponding to each preselected value  $x$  of  $x$ , one -or more- value  $y$  of  $y$  is obtained.
- The other variable,  $y$ , accordingly, is called the *dependent variable*, and we speak of the regression of  $y$  on  $x$ .

# The regression equation

In simple linear regression the object of the researcher's interest is the **regression equation** that describes the true relationship between the dependent variable  $y$  and the independent variable  $x$ .

$x$

# Scatter diagram

- A first step that is usually useful in studying the relationship between two variables is to prepare a **scatter diagram** of the data.
- The points are plotted by assigning values of the independent variable  $x$  to the horizontal axis and values of the dependent variable  $y$  to the vertical axis.
- The pattern made by the points plotted on the scatter diagram usually suggests the basic nature and the strength of the relationship between two variables.

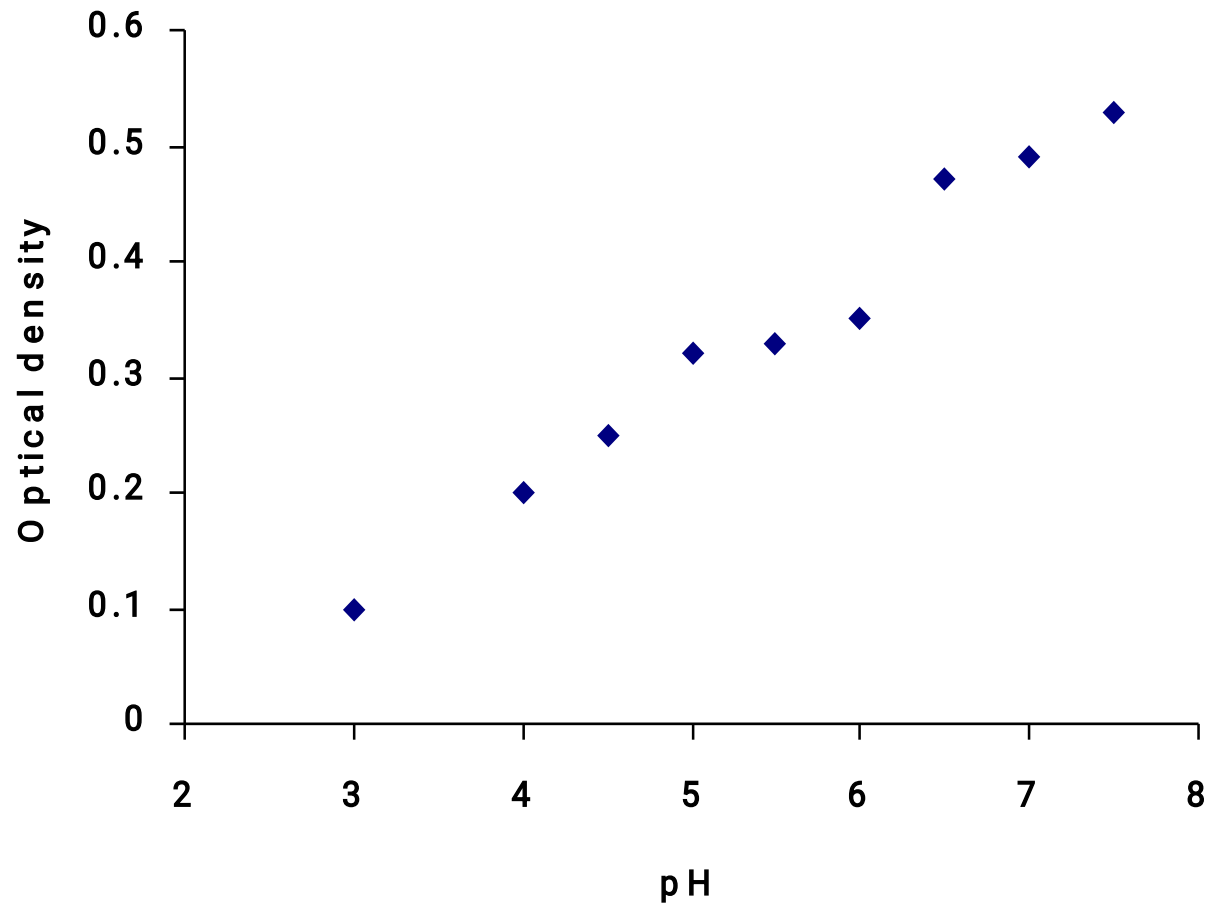
# Example

Relationship between pH and optical density

<b>pH</b>	<b>Optical density</b>
<b>3</b>	<b>0.1</b>
<b>4</b>	<b>0.2</b>
<b>4.5</b>	<b>0.25</b>
<b>5</b>	<b>0.32</b>
<b>5.5</b>	<b>0.33</b>
<b>6</b>	<b>0.35</b>
<b>6.5</b>	<b>0.47</b>
<b>7</b>	<b>0.49</b>
<b>7.5</b>	<b>0.53</b>

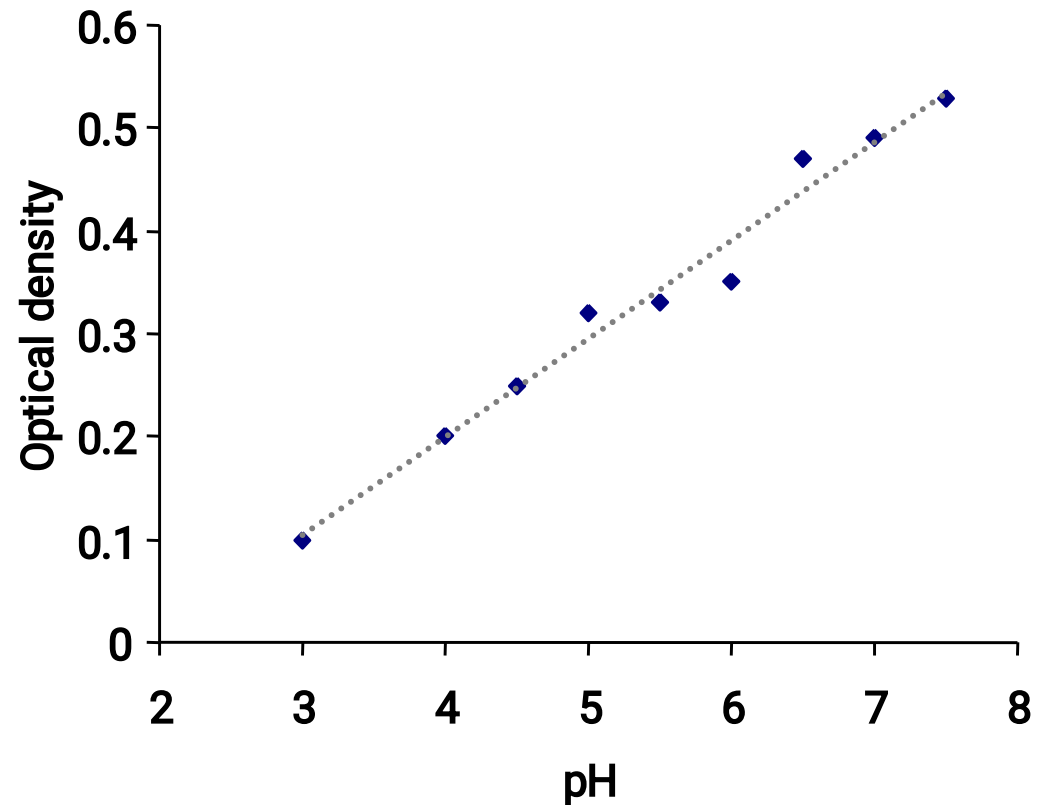
# Scatter Diagram

Relationship between pH and optical density



# Notes

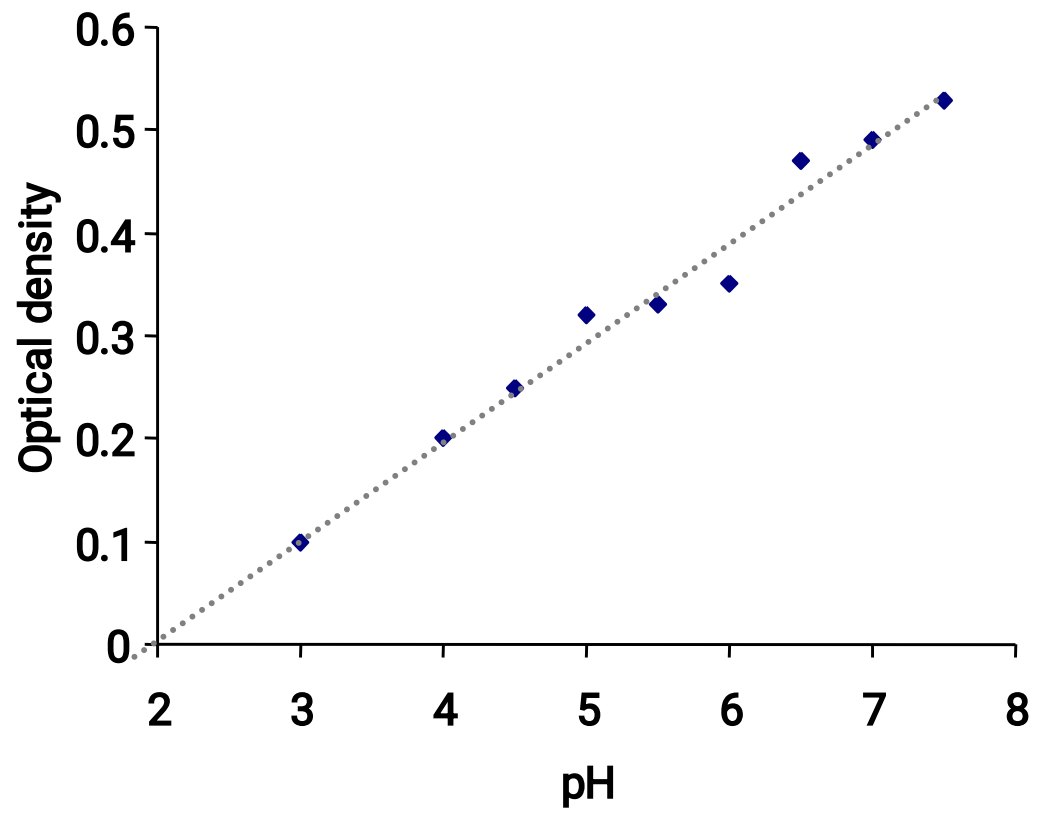
- The points in the figure seem to be scattered around an invisible straight line.
- The scatter diagram also shows that, in general, high pH also has high optical density reading.





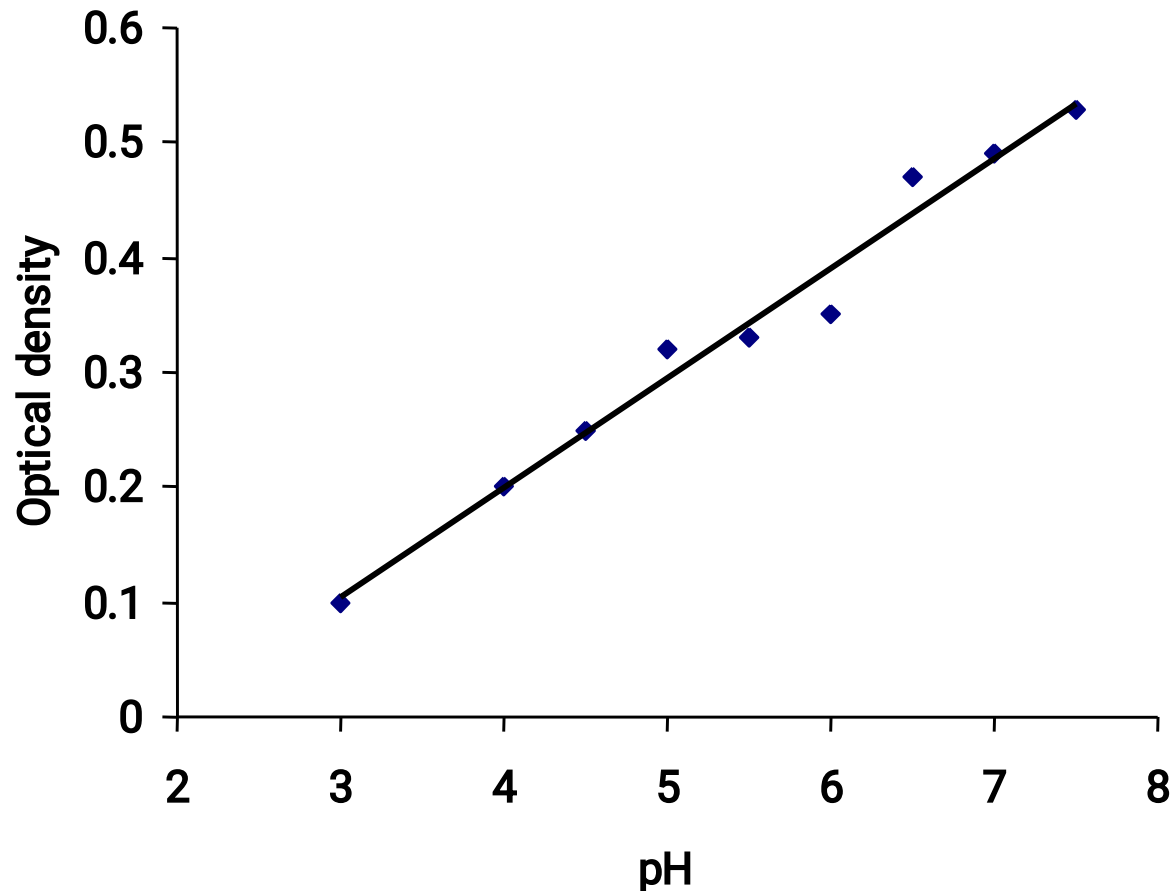
- These impressions suggest that the relationship between points in the two variables may be described by a straight line crossing the  $y$ -axis near the origin and making approximately a 45 degree angle with the  $x$ -axis.
- It looks as if it would be simple to draw, freehand, through the data points the line that describe the relationship between and .

$y$   $x$



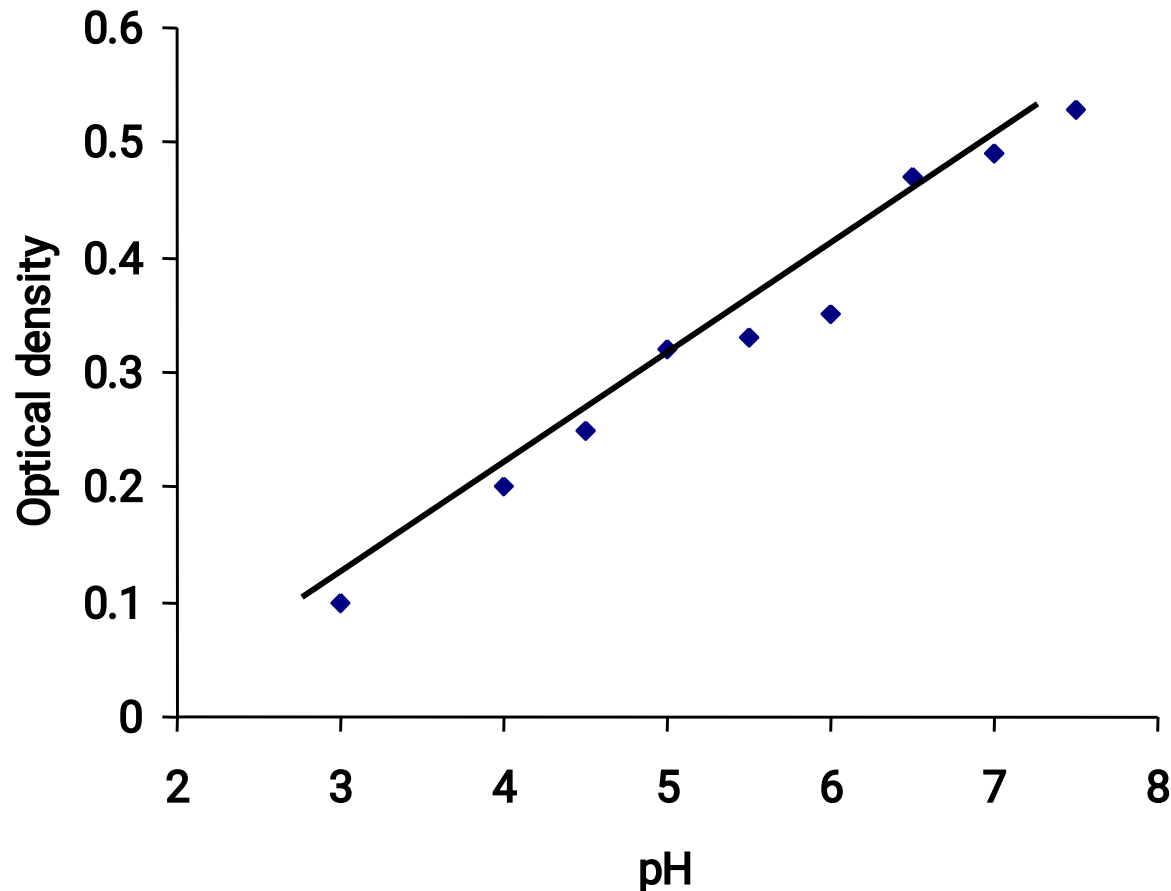
# Thinking Challenge

- For every person drawing such a line by eye, or freehand, we would expect a slightly different line.



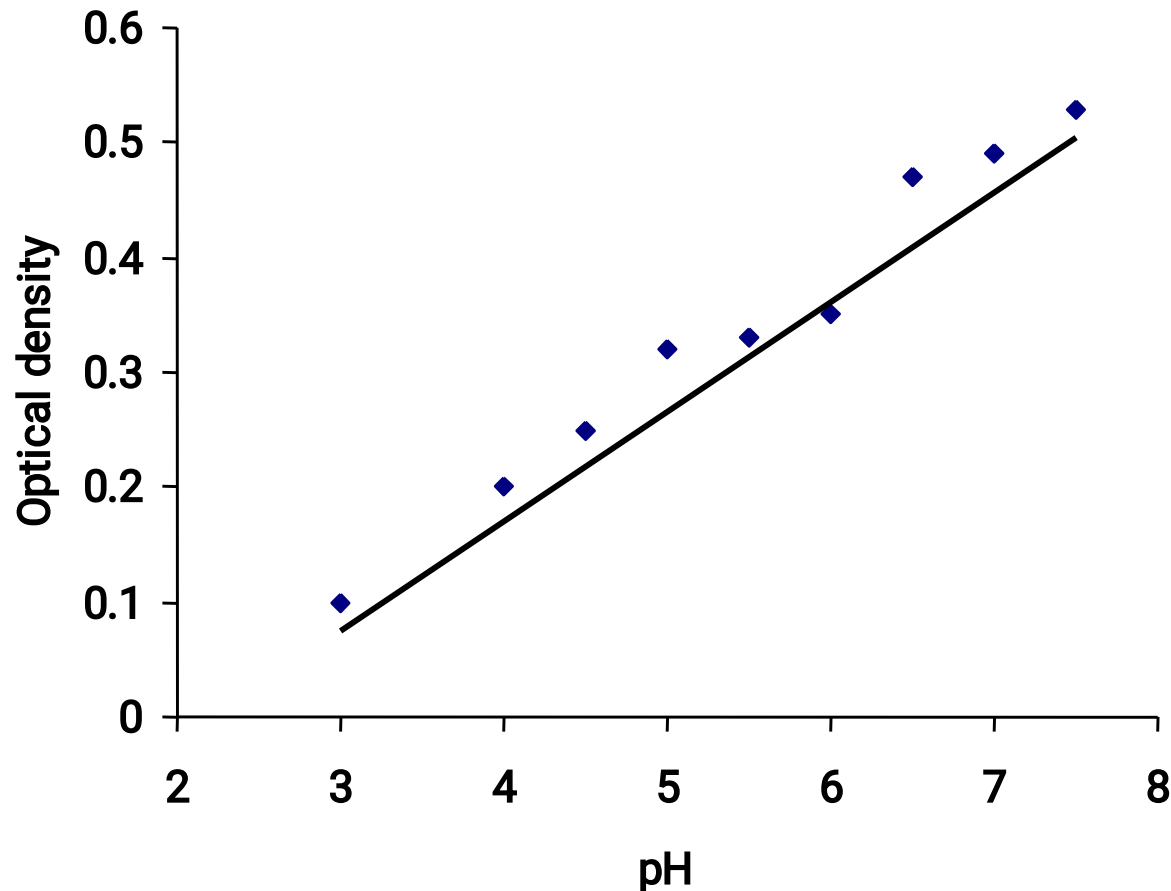
# Thinking Challenge

- For every person drawing such a line by eye, or freehand, we would expect a slightly different line.



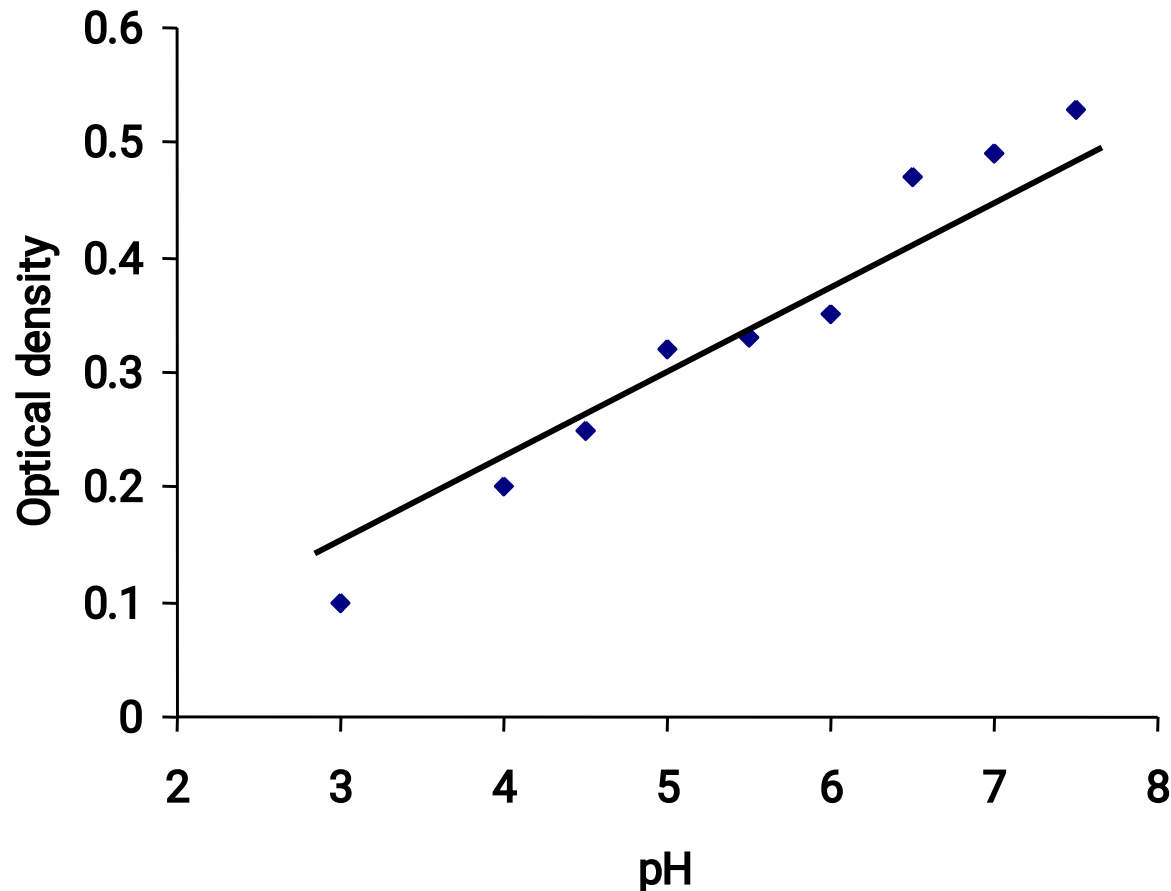
# Thinking Challenge

- For every person drawing such a line by eye, or freehand, we would expect a slightly different line.



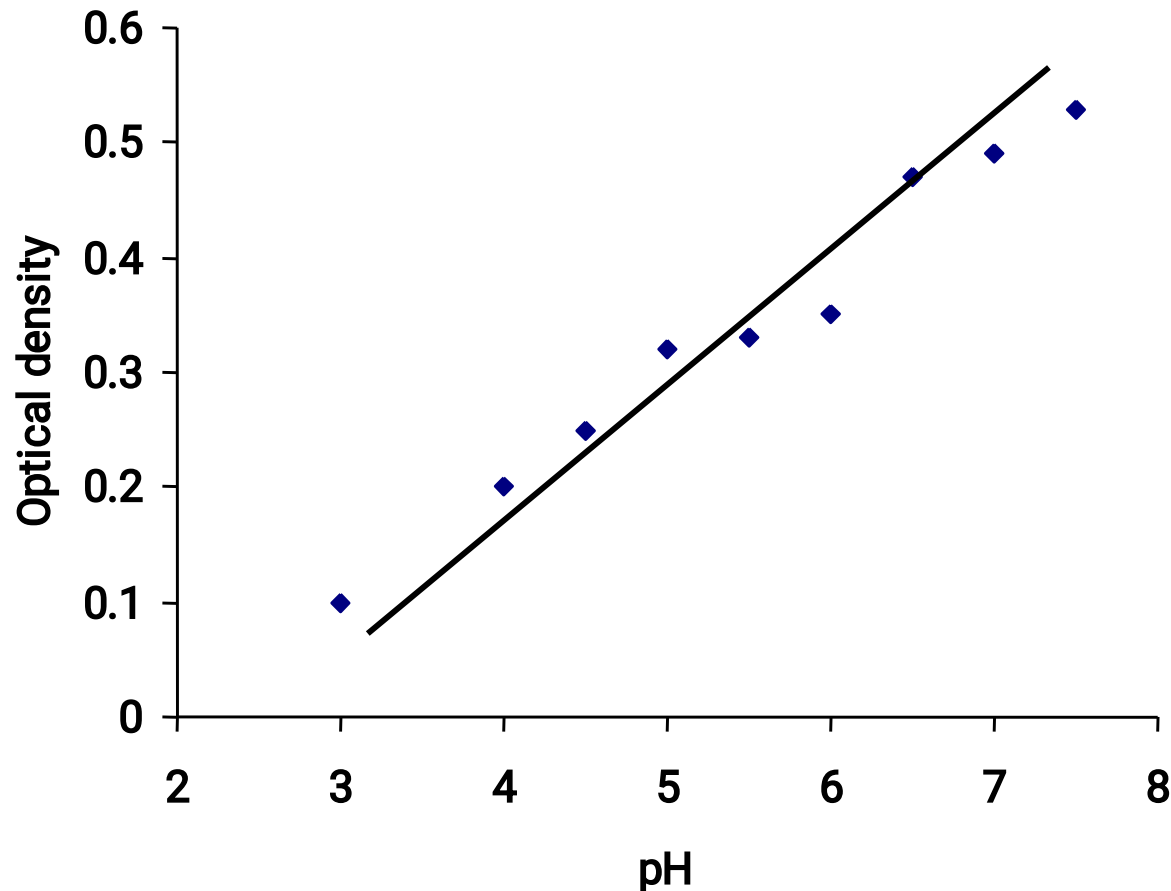
# Thinking Challenge

- For every person drawing such a line by eye, or freehand, we would expect a slightly different line.



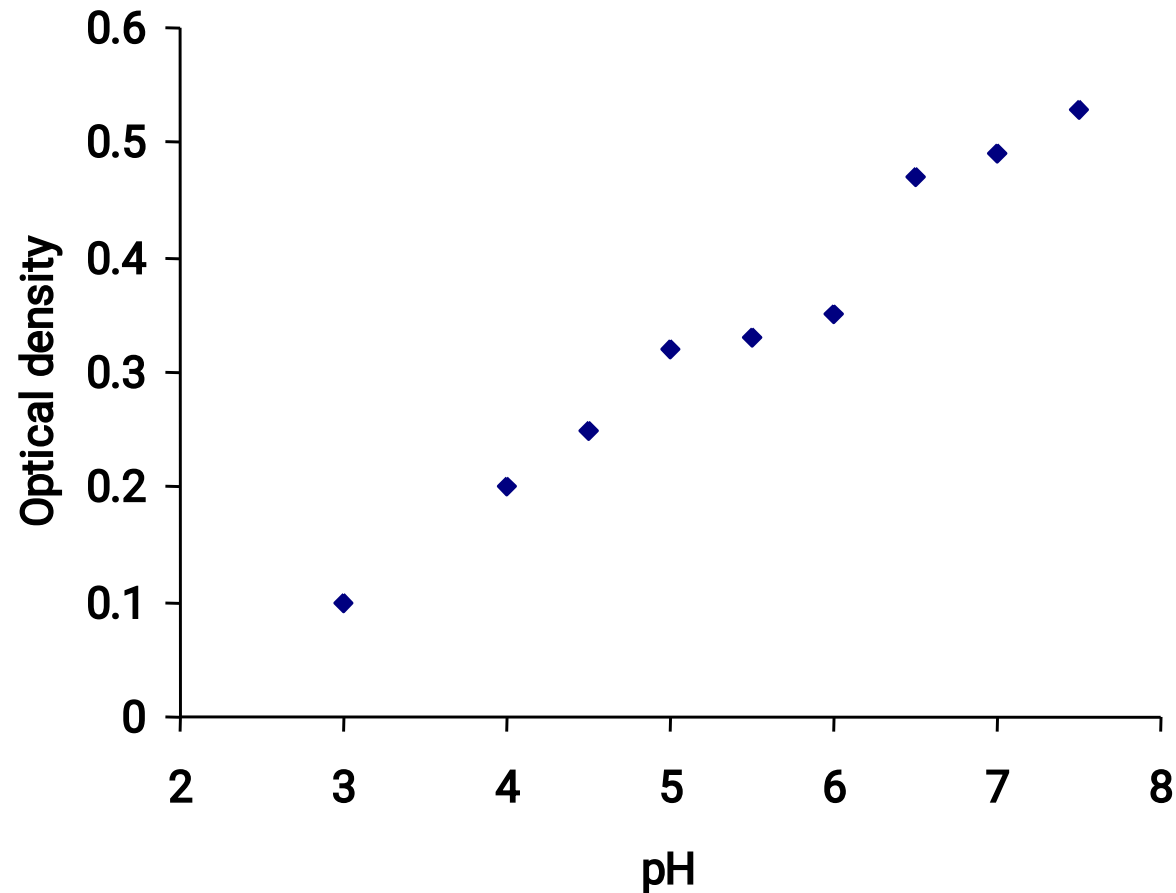
# Thinking Challenge

- For every person drawing such a line by eye, or freehand, we would expect a slightly different line.



# Thinking Challenge

Which line best describes relationship between the variables?



## Answer

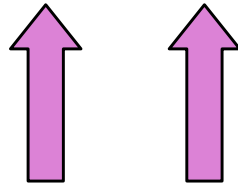
- We need to employ a method known as the **method of least squares** for obtaining the desired line, and the resulting line is called the **least-squares regression line**.
- The reason for calling the method by this name will be explained in the discussion that follow.



# Equation for straight line

- Now, recall from algebra that the general equation for straight line is given by

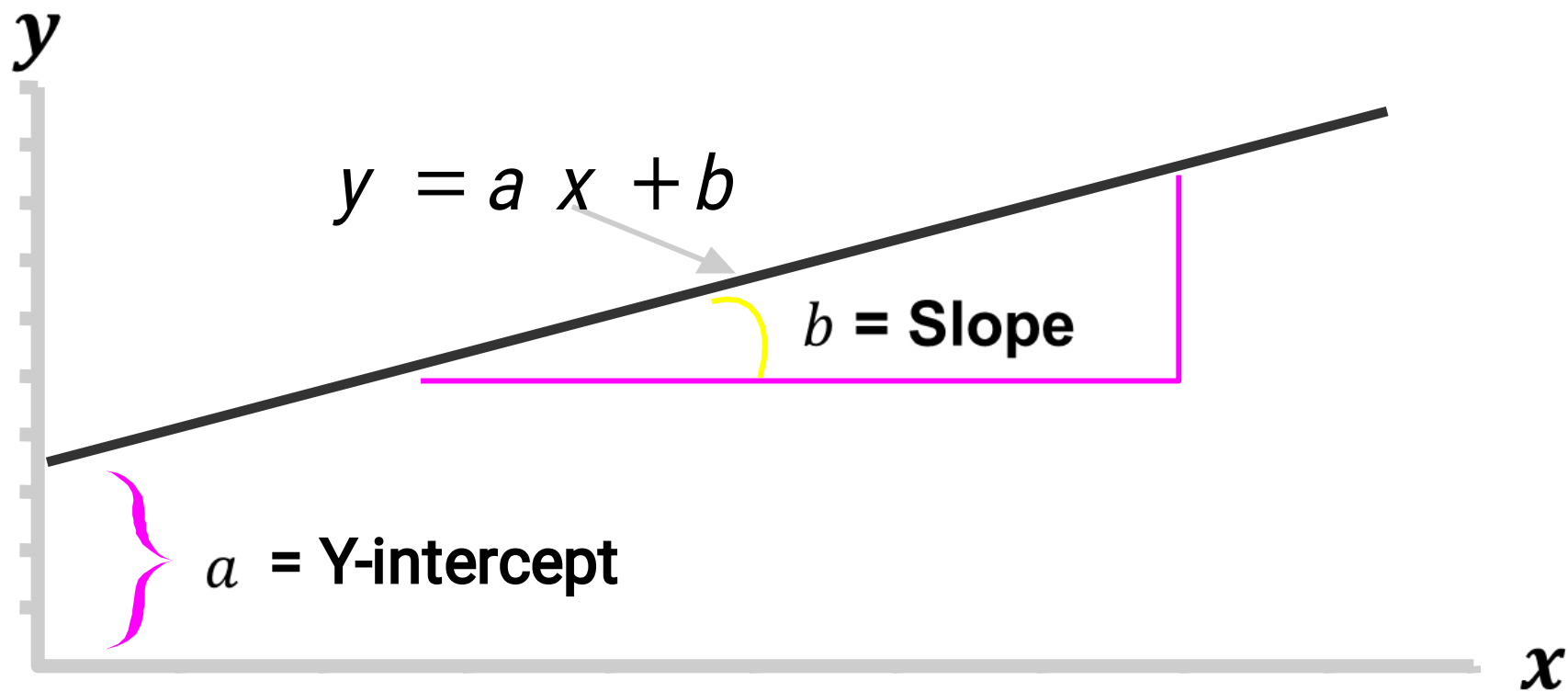
$$y = a + bx$$



**a = the y-intercept**

**b = the slope**

# Linear Equations



**b** is referred to as the slope of the line.

# Linear Equations

- To draw a line based on the equation, we need the numerical values of the constants  $a$  and  $b$ .
- Given these constants, we may substitute various values of  $x$  into the equation to obtain corresponding values of  $y$ .

$$\hat{y} = a + b x$$

- The resulting points may be plotted.

How to determine our “best” line ?

i.e. best regression coefficients  $a$  and  $b$  ?

There is a mathematical procedure that minimizes the Estimated error (residual)  $e = (y - \hat{y})$ .

It is known as the least- squares method since it minimizes

$$\sum e^2 = \sum (y - \hat{y})^2$$

This procedure uses equations that estimate  $a$  and  $b$  the equations are

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \quad \text{and}$$

$$a = \bar{y} - b\bar{x}$$

It can be shown that the sum of the residuals equal zero

i.e.  $\sum (y - \hat{y}) = 0$ .

# Example

	<b>pH ( <math>x</math> )</b>	<b>Optical density ( <math>y</math> )</b>	$x^2$	$y^2$	$xy$
	<b>3</b>	<b>0.1</b>	<b>9</b>	<b>0.01</b>	<b>0.3</b>
	<b>4</b>	<b>0.2</b>	<b>16</b>	<b>0.04</b>	<b>0.8</b>
	<b>4.5</b>	<b>0.25</b>	<b>20.25</b>	<b>0.0625</b>	<b>1.125</b>
	<b>5</b>	<b>0.32</b>	<b>25</b>	<b>0.1024</b>	<b>1.6</b>
	<b>5.5</b>	<b>0.33</b>	<b>30.25</b>	<b>0.1089</b>	<b>1.815</b>
	<b>6</b>	<b>0.35</b>	<b>36</b>	<b>0.1225</b>	<b>2.1</b>
	<b>6.5</b>	<b>0.47</b>	<b>42.25</b>	<b>0.2209</b>	<b>3.055</b>
	<b>7</b>	<b>0.49</b>	<b>49</b>	<b>0.240</b>	<b>3.43</b>
	<b>7.5</b>	<b>0.53</b>	<b>56.25</b>	<b>0.281</b>	<b>3.975</b>
<b>Total</b>	$\Sigma x = 49$	$\Sigma y = 3.04$	$\Sigma x^2 = 284$	$\Sigma y^2 = 1.1882$	$\Sigma xy = 18.2$
<b>Mean</b>	$\bar{X} = 5.444$	$\bar{Y} = 0.3378$			

## Finding the value of b

**b =**

$$\frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{18.2 - \frac{(49)(3.04)}{9}}{284 - \frac{(49)^2}{9}}$$
$$= 0.0957$$

# Finding the value of a

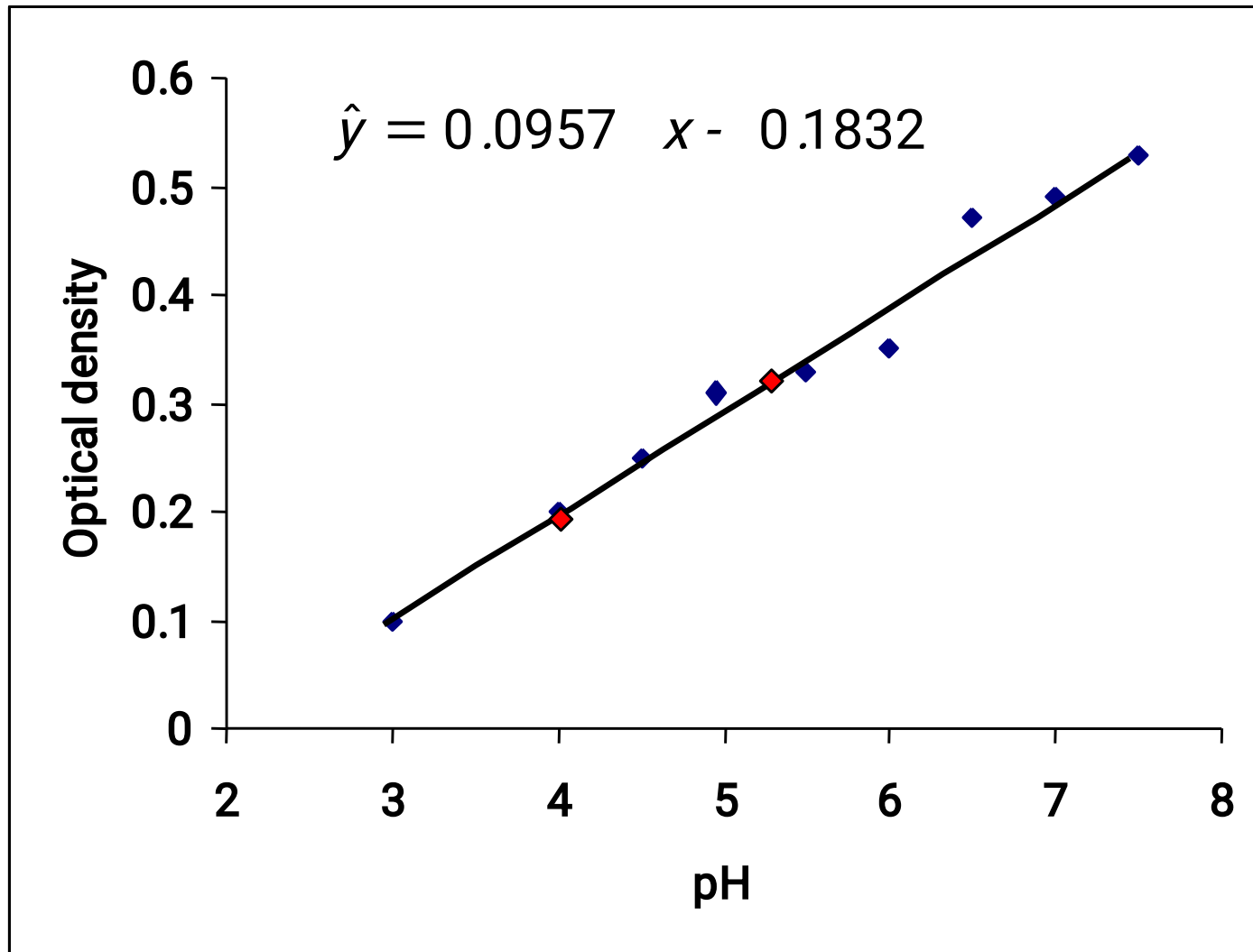
$$\begin{aligned} a &= \bar{y} - b\bar{x} \\ &= 0.3378 - (0.0957)(5.444) \\ &= -0.1832 \end{aligned}$$

The equation for the least squares line is

$$\begin{aligned} \hat{y} &= a + bx \\ &= -0.1832 + 0.0957x \end{aligned}$$

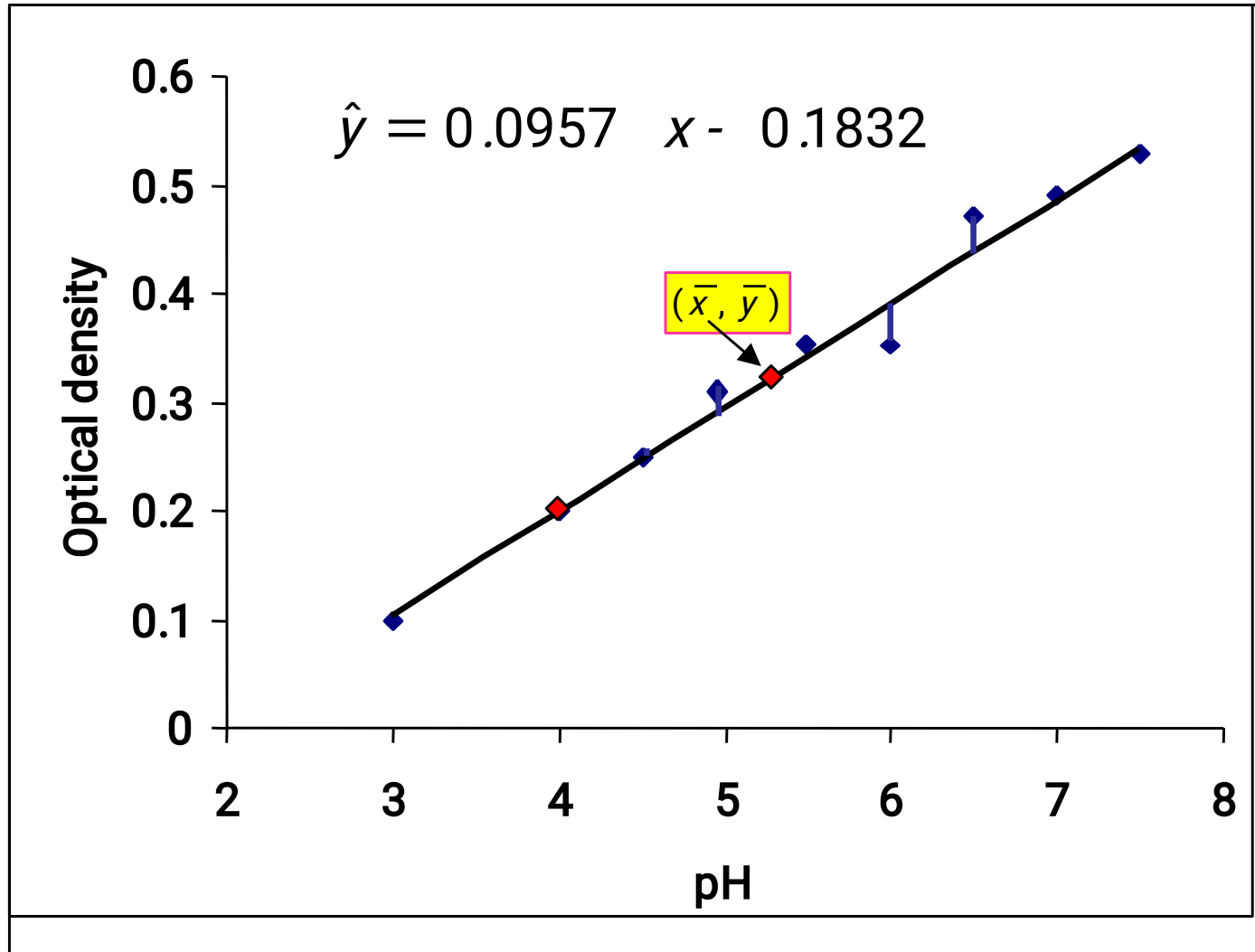
Note that we use the symbol  $\hat{y}$  because this value is computed from the equation and is not an observed or actual value of  $y$ .

# Sketching the Line Using the Points (5.444 , 0.3378) and (4 , 0.1996)





# Sketching the Line Using the Points (5.444 , 0.3378) and (4 , 0.1996)



# Using the Regression Equation

## Predicting $y$ for a given $x$

- Choose a value for  $x$  (within the range of  $x$  values).
- Substitute the selected  $x$  in the regression equation.
- Determine corresponding value of  $y$ .

- The regression equation:  $\hat{y} = 0.0957 x - 0.1832$

Substitute  $x = 6.8$ :

$$\begin{aligned}\hat{y} &= 0.0957 (6.8) - 0.1832 \\ &= 0.46756\end{aligned}$$

- According to the equation, a pH of 6.8 would have a 0.46756 optical density.
- Predict the optical density of a material with PH=6 and find the corresponding error.

The predicted value of  $y$  is

$$\hat{y} = 0.0957 (6) - 0.1832 = 0.39$$

- The corresponding error is the difference between the actual value and the predicted value

$$\begin{aligned}e &= y - \hat{y} = 0.35 - 0.39 \\ &= -0.04\end{aligned}$$

**Exercise:** The following are the age (in years) and systolic blood pressure of 20 apparently healthy adults.

- Find the correlation between age and blood pressure .
- Describe the relation between the two variables
- Find the regression equation.
- What is the predicted blood pressure for a man aging 25 years?
- Predict blood pressure of an adult aging 31 years and find the corresponding error

Age (x )	B.P ( y)	Age ( x)	B.P (y)
20	120	46	128
43	128	53	136
63	141	60	146
26	126	20	124
53	134	63	143
31	128	43	130
58	136	26	124
46	132	19	121
58	140	31	126
70	144	23	123