Introduction to STATISTICS

Mortuza Ahmmed

obooko

Introduction to Statistics

By

Mortuza Ahmmed

Statistics

Statistics is the study of the collection, organization, analysis, and interpretation of data. It deals with all aspects of this, including the planning of data collection in terms of the design of surveys and experiments.

Applications of Statistics

Agriculture

What varieties of plant should we grow? What are the best combinations of fertilizers, pesticides and densities of planting? How does changing these factors affect the course of the growth process?

Business and economics

Which companies are likely to go out of business in the next year? What is the likely tourist flow next year? What causes companies to choose a particular method of accounting? How have living standards changed over the last six months?

Marketing Research

What makes advertisements irritating? Is an irritating ad a bad ad? Are telephone calls the best way to collect market data? What share of the television market does Sony have? Do higher prices signal higher quality?

Education

Does a course on classroom behavior for teachers purchased by the authority of IUBAT have an effect on the teacher's classroom performance? Do boys perform better than girls in the examinations? Is there evidence of sex bias in admissions to IUBAT?

Medicine

What are the important risk factors for bone cancer? What determines long term survival after open heart surgery? Would nationwide screening for breast cancer be effective? What projections can we make about the course of the AIDS epidemic? Is there a relationship between drinking alcohol and breast cancer in women?

Population

The entire set of individuals or objects of interest is called population.

Example

In IUBAT, there are 6000 students. All of them constitute a population.

Sample

A small but representative part of the population is called sample.

Example

In IUBAT, there are 6000 students. If we take 100 students randomly from them, these 100 students will constitute a sample.

Variable

A variable is a characteristic under study that assumes different values for different elements.

Example

Shirt size, height of students, age, colors, sex and so on.

Qualitative Variable

Qualitative variables take on values that are names or labels.

Example

Religion, colors, sex would be examples of qualitative or categorical variable.

Quantitative Variable

A variable that can be measured numerically is called quantitative variable.

Example

Shirt size, height of students, age would be examples of quantitative variable.

Discrete variable

A variable whose value is countable is called discrete variable.

Example

Number of mobile sets sold in a store last month, Number of patients admitted in a hospital last month would be examples of discrete variable.

Continuous variable

A variable which can't be counted and can assume any between two numbers is called continuous variable.

Example

Age, weight, height would be examples of continuous variable.

Independent variable

The variable that is use to describe the factor that is assumed to cause the problem is called independent variable.

Example

Smoking causes cancer - here smoking is the independent variable.

Dependent variable

The variable that is used to describe the problem under study is called dependent variable.

Example

Smoking causes cancer – here cancer is the dependent variable.

Scales of Measurement

Nominal scale

The variable under this measurement scale can be classified and counted but no ordering is possible.

Example

Sex, religion, marital status

Ordinal Scale

The variable under this scale can only be classified, counted and ordering is possible.

Example

Economic status, exam grade, academic result

Interval scale

Along with all the characteristics of nominal scale and ordinal scale it includes the different between the values which is constant.

Example

Temperature, calendar date

Ratio scale

This is the best measurement scale. It satisfies all the four properties of measurement: identity, magnitude, equal intervals and an absolute zero.

Example

Age, height, weight, length

Separate the following variables into discrete (D) and continuous(C)

Number of phone calls received in a day, Time taken to serve a customer, Weight of a customer, Volume of a 3c.c. bottle of medicine, Size of shoes produced by BATA

D	C	C	C	D

Identify whether each of the following constitutes a population (P) or sample(S)

Kilograms of wheat collected by all farmers in a village, Credit card debt of 50 families selected from a city, Ages of all members of a family, Number of parole violations by all 2147 parolees in a city, Amount spent on prescription drugs by 200 citizens in a city

P	S	P	P	S

Classify the following into nominal (N), ordinal (O), interval (I) and ratio(R)

Age of the pupils, Gender of the students, Health status (poor, average, well), Academic degree (primary, secondary, higher), Hair color, Weight, Disease status (diseased, non-diseased), Place of residence (urban-rural), Calendar time (3pm, 6pm. etc.), IQ test score.

R	N	O	O	N	R	N	N	I	I

Separate the following variables into quantitative (Qn) or qualitative (Ql)

Number of persons in a family, Color of cars, marital status of people, Length of frog's jump, Number of students in the class

Qn	Ql	Ql	Qn	Qn

PRESENTATION OF DATA

Frequency table

A table that shows the frequencies of each of the values of a variable under consideration is called frequency table.

Example

Consumers were asked to rate the taste of a new diet drink as being poor (P), good (G), excellent (E). The following data were obtained:

G	P	G	E	G	G	E	P	G	G
E	G	E	P	E	E	G	P	G	G
P	G	G	E	E					

(i) Construct a frequency table
(ii) Add a relative frequency table to the table

Rating of Drink	Tally marks	Frequency	Relative Frequency
P	IIII	05	05 / 25 = 0.20
G	IIII IIII II	12	12 / 25 = 0.48
E	IIII III	08	08 / 25 = 0.32
Total		25	1.00

Bar diagram

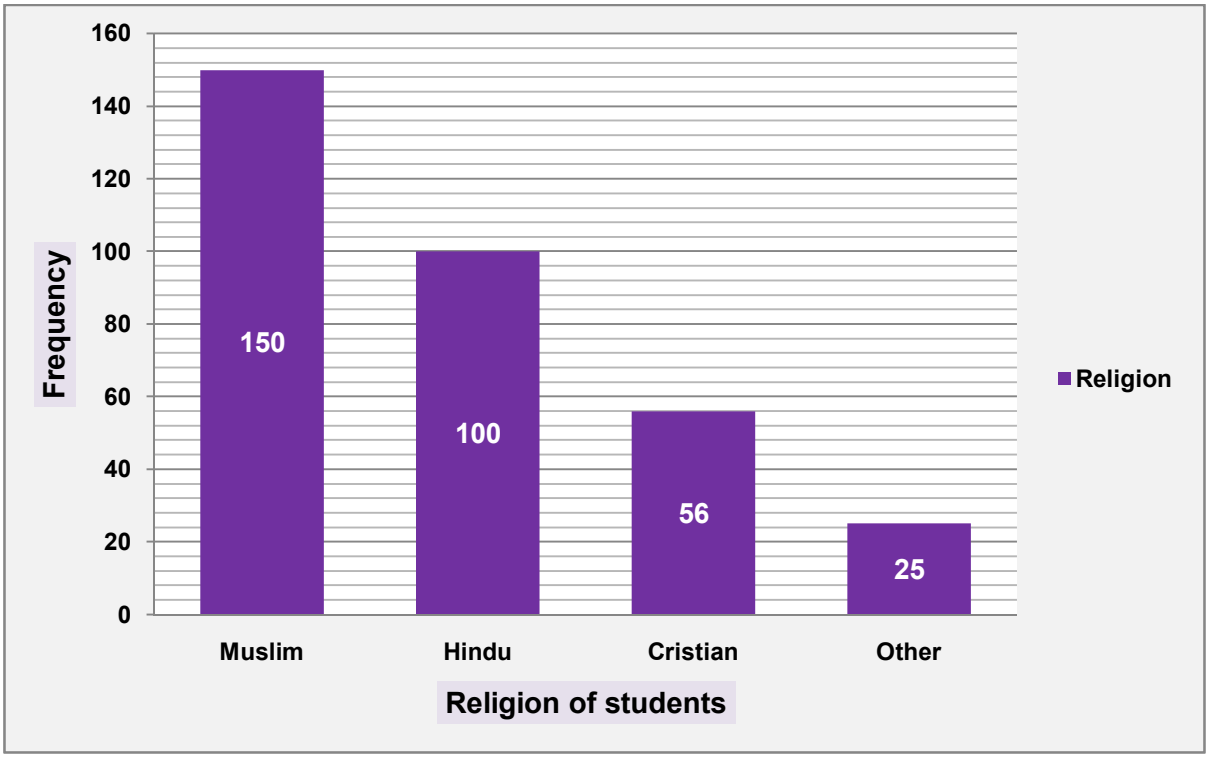
A graph in which the classes are represented on the horizontal axis and the class frequencies on the vertical axis is called bar diagram. Bar diagram is only used for the qualitative variable.

Example

Students of BBA department of IUBAT are classified as follows

Religion of students	Frequency
Muslim	150
Hindu	100
Christians	056
Others	025

Construct a simple bar diagram



Component bar Diagram

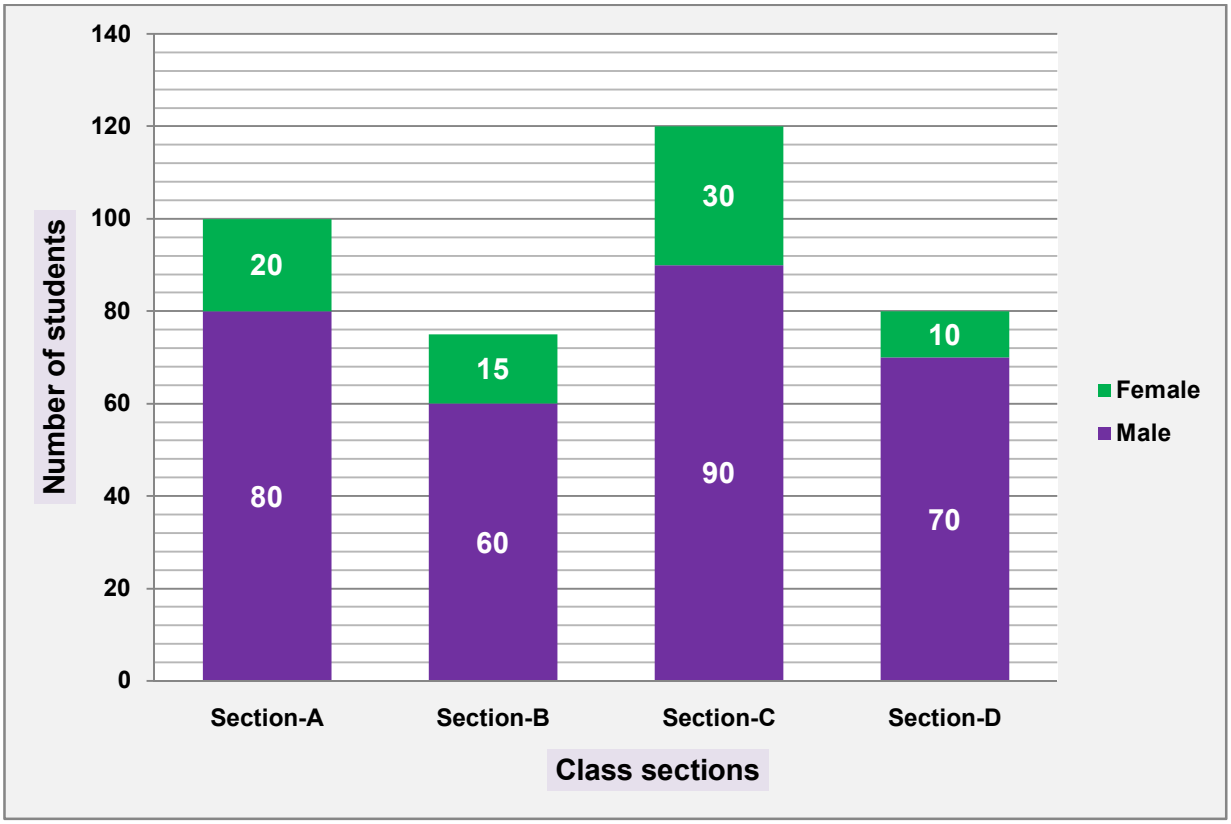
Here bar is sub-divided into as many parts as there are components. Each part of the bar represents component while the whole bar represents the total value.

Example

Students of the course STA 240 of summer semester are classified as follows

	Section A	Section B	Section C	Section D
Male	80	60	90	70
Female	20	15	30	10

Construct a component bar diagram



Multiple bar Diagram

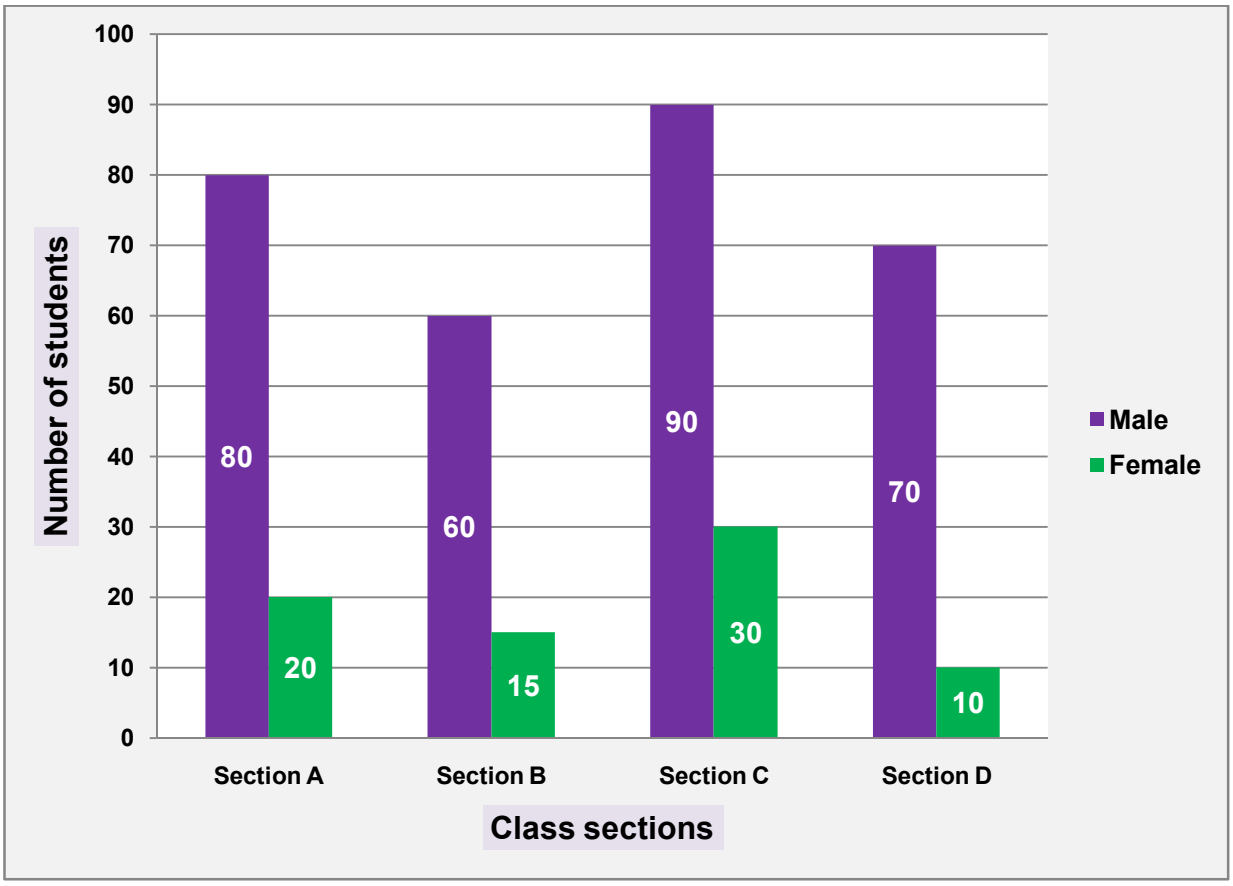
A multiple bar graph contains comparisons of two or more categories or bars.

Example

Students of the course STA 240 of summer semester are classified as follows

	Section A	Section B	Section C	Section D
Male	80	60	90	70
Female	20	15	30	10

Construct a multiple bar diagram



Pie Chart:

A pie chart displays data as a percentage of the whole. Each pie section should have a label and percentage.

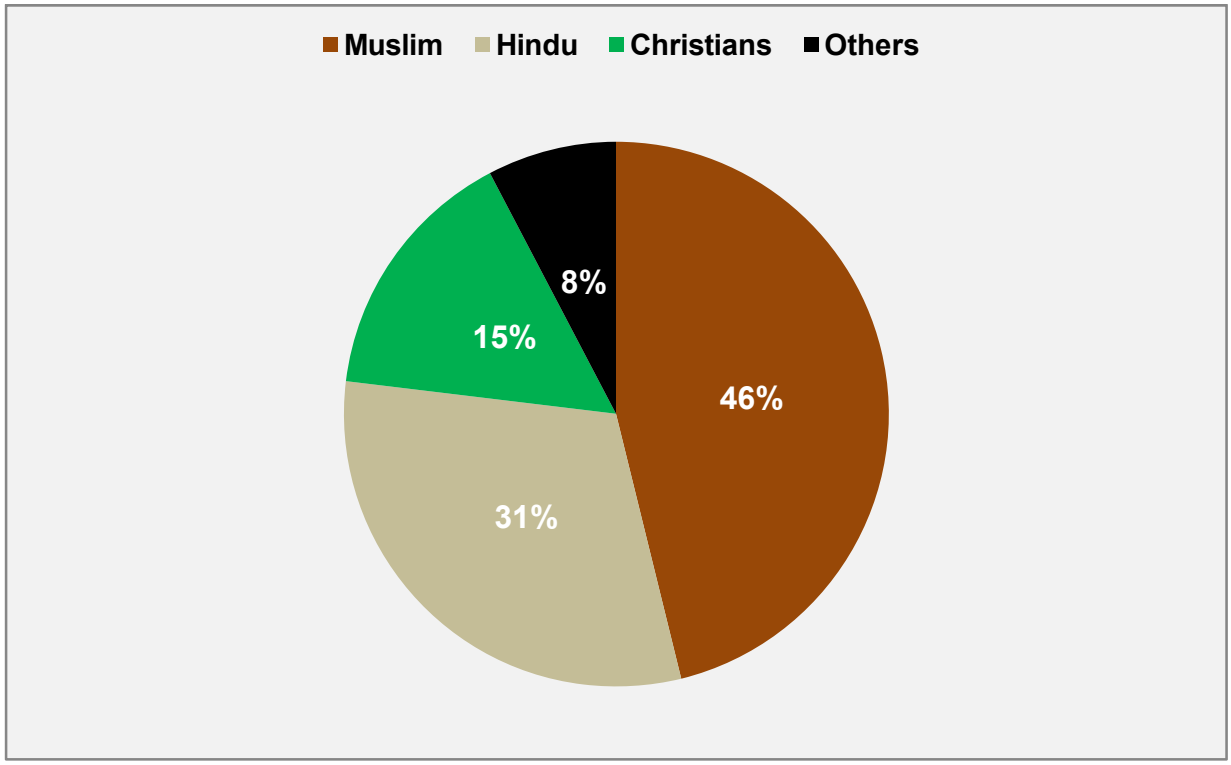
Example

Students of BBA department of IUBAT are classified as follows

Religion of students	Frequency
Muslim	150
Hindu	100
Christians	050
Others	025

Construct a pie chart

Religion	Frequency	Percentage	Angle
Muslim	150	150 / 325 x 100 = 46	46% of 360° = 166°
Hindu	100	100 / 325 x 100 = 31	31% of 360° = 111°
Christians	050	050 / 325 x 100 = 15	15% of 360° = 054°
Others	025	025 / 325 x 100 = 08	08% of 360° = 029°
Total	325	100	360°



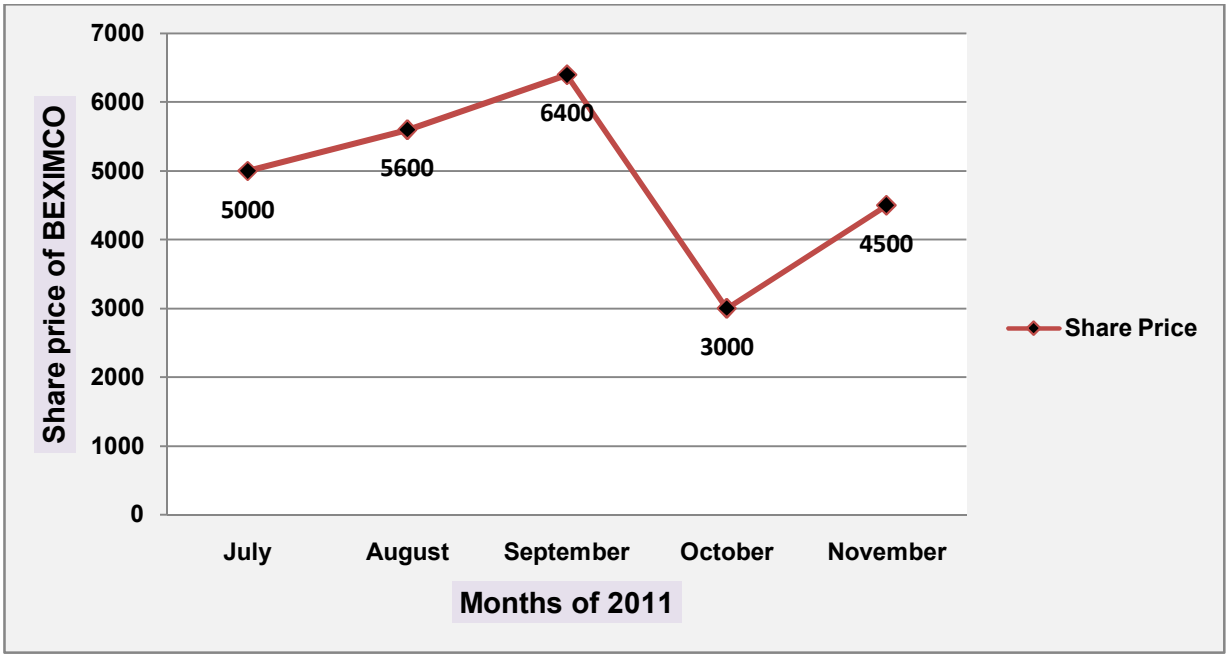
Line graph

A line chart or line graph is a type of graph, which displays information as a series of data points connected by straight line segments. It is created by connecting a series of points that represent individual measurements. A line chart is often used to visualize a trend in data over intervals of time.

Example

Construct a line chart for the following data provided by DSE

Months of 2011	Share price of BEXIMCO
July	5000
August	5600
September	6400
October	3000
November	4500



Histogram

A histogram displays continuous data in ordered columns. It is constructed by placing the class boundaries as the horizontal axis and the frequencies of the vertical axis.

Bar diagram vs. histogram

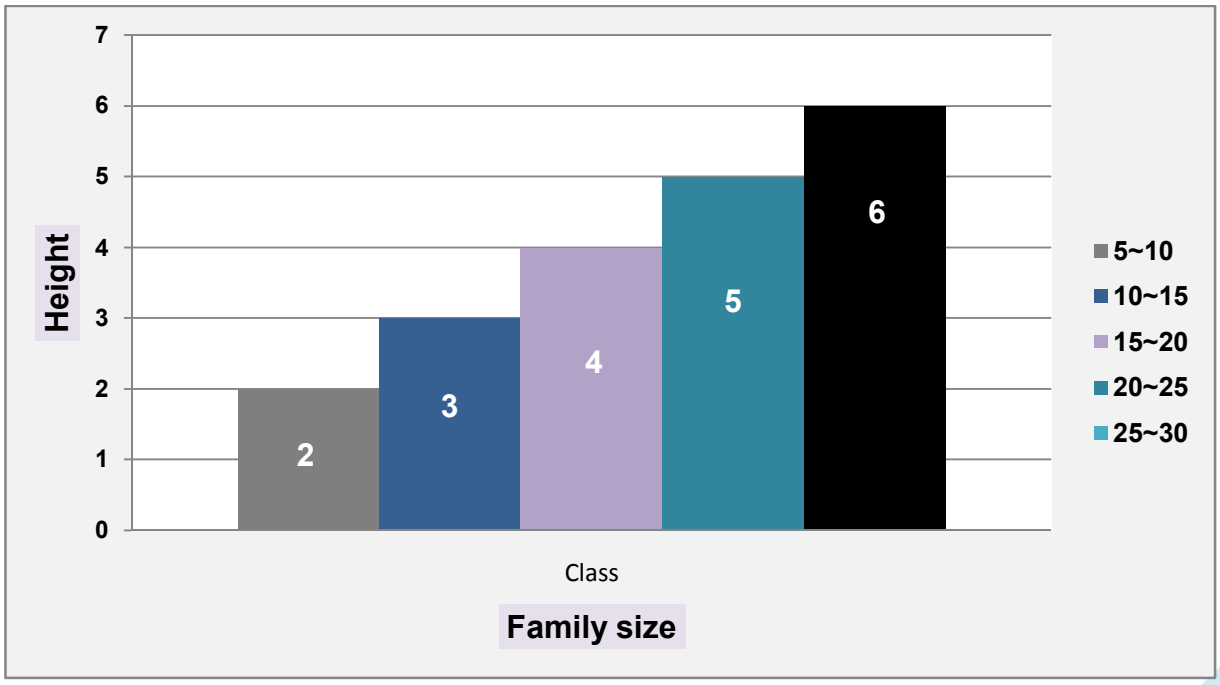
Histogram	Bar diagram
Area gives frequency	Height gives frequency
Bars are adjacent to each others	Bars are not adjacent to each others
Constructed for quantitative data	Constructed for qualitative data

Example

Construct a histogram for the following data provided by BBS

Family size	Number of families
05 –10	10
10 –15	15
15 – 20	20
20 – 25	25
25 – 30	30

Family size	Number of families	Height
05 –10	10	10 / 5 = 2
10 –15	15	15 / 5 = 3
15 – 20	20	20 / 5 = 4
20 – 25	25	25 / 5 = 5
25 – 30	30	30 / 5 = 6



Stem and leaf plot

It is a graphical technique of representing data that can be used to examine the shape of a frequency distribution as well as range of the value.

Example

Construct a stem and leaf plot for the following list of grades on a recent test

11,14,17,19,21,23,24,27,29,31,33,37,39,41,43,44,47,51,53,54,59,61,63,64,67

Stem	Leaf
1	1 4 7 9
2	1 3 4 7 9
3	1 3 7 9
4	1 3 4 7
5	1 3 4 9
6	1 3 4 7

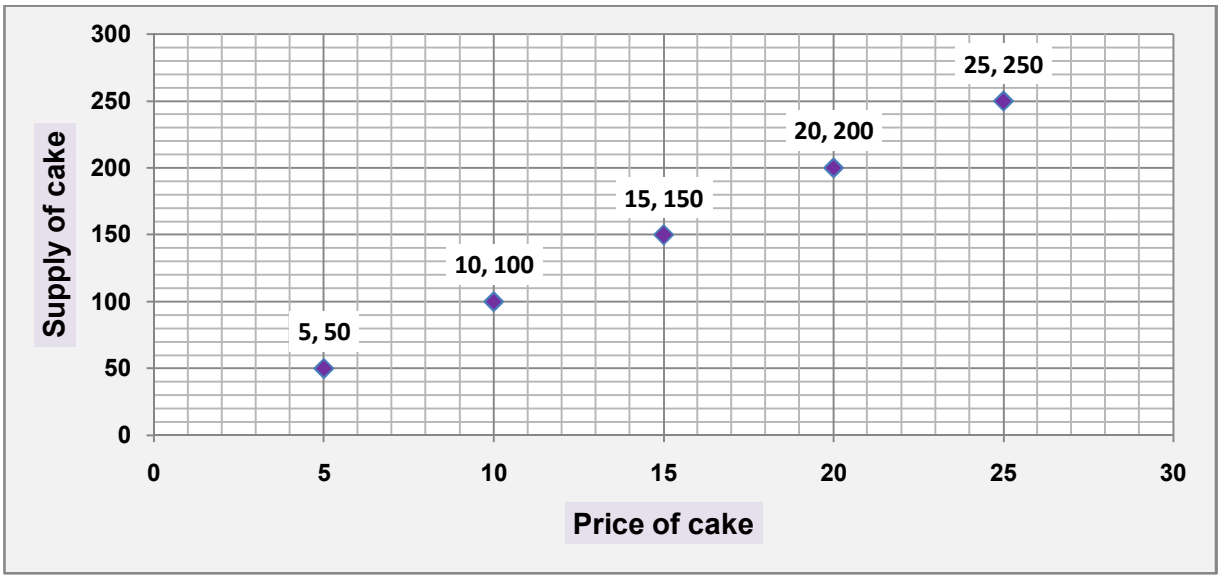
Scatter diagram

Scatter diagrams are used to represent and compare two sets of data. By looking at a scatter diagram, we can see whether there is any connection (correlation) between the two sets of data.

Example

Construct a scatter diagram for the following data provided by IUBAT cafeteria

Price of cake (Taka)	Supply of cake
5	50
10	100
15	150
20	200
25	250



Comparison among the graphs

Graph	Advantages	Disadvantages
Pie chart	Visually appealing Shows percent of total for each category	Hard to compare 2 data sets Use only with discrete data
Histogram	Visually strong Can compare to normal curve	More difficult to compare 2 data sets Use only with continuous data
Bar diagram	Visually strong Can compare 2 or 3 data sets easily	Use only with discrete data
Line graph	Can compare 2 or 3 data sets easily	Use only with continuous data
Scatter plot	Shows a trend in the data relationship Retains exact data values and sample size	Hard to see results in large data sets Use only with continuous data
Stem and Leaf Plot	Can handle extremely large data sets Concise representation of data	Not visually appealing Does not easily indicate measures of centrality for large data sets

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. They are also classed as summary statistics. The measures are:

Arithmetic mean (AM)	Geometric mean (GM)	Harmonic mean (HM)	Median	Mode

Arithmetic mean (AM)

The arithmetic mean (or average) is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data. It is equal to the sum of all the values in the data set divided by the number of values in the data set.

So, if we have n values in a data set and they have values x_1, x_2, ..., x_n, then the sample mean, usually denoted by $\bar{x}$ (pronounced x bar), is:

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

Example

Find the average of the values 5, 9, 12, 4, 5, 14, 19, 16, 3, 5, 7.

Average = (5 + 9 + 12 + 4 + 5 + 14 + 19 + 16 + 3 + 5 + 7) / 11 = 99 / 11 = 9

The mean weight of three dogs is 38 pounds. One of the dogs weighs 46 pounds. The other two dogs, Eddie and Tommy, have the same weight. Find Tommy's weight.

Let x = Eddie's weight = Sandy's weight

Now,

(x + x + 46) / 3 = 38

=> 2x + 46 = 114

=>2x = 68

=> x = 34

So, Tommy weighs 34 pounds.

On her first 5 math tests, Zany received scores 72, 86, 92, 63, and 77. What test score she must earn on her sixth test so that her average for all 6 tests will be 80?

Let x = Test score Zany must earn on her sixth test

Now,

(72 + 86 + 92 + 63 + 77 + x) / 6 = 80

=> 390 + x=480

=> x=90

So, Zany must get 90 on her sixth test.

Affect of extreme values on AM

Let us consider the wages of staff at a factory below

Staff	1	2	3	4	5	6	7	8	9	10
Salary($)	15	18	16	14	15	15	12	17	90	95

The mean salary is $30.7. However, this mean value might not be the best way to reflect the typical salary of a worker, as most workers have salaries in the $12 to $18 ranges. The mean is being skewed by the two large salaries. Therefore, we would like to have a better measure of central tendency.

Median would be a better measure of central tendency in this situation.

Calculation of AM for grouped data

Number of alcoholic beverages consumed by IUBAT students last weekend

x	f	fx
0	05	00
1	10	10
2	05	10
3	10	30
4	05	20
10	02	20
Total	N = 37	90

AM = $\sum fx$ / N = 90 / 37 = 2.43

Median

The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data.

Example

In order to calculate the median, suppose we have the data below

65	55	89	56	35	14	56	55	87	45	92

We first need to rearrange that data into order of magnitude (smallest first)

14	35	45	55	55	56	56	65	87	89	92

Our median mark is the middle mark - in this case 56.

This works fine when we have an odd number of scores but what happens when we have an even number of scores? What if we had only 10 scores? Well, we simply have to take the middle two scores and average the result.

So, if we look at the example below

65	55	89	56	35	14	56	55	87	45

We again rearrange that data into order of magnitude (smallest first)

14	35	45	55	55	56	56	65	87	89	92

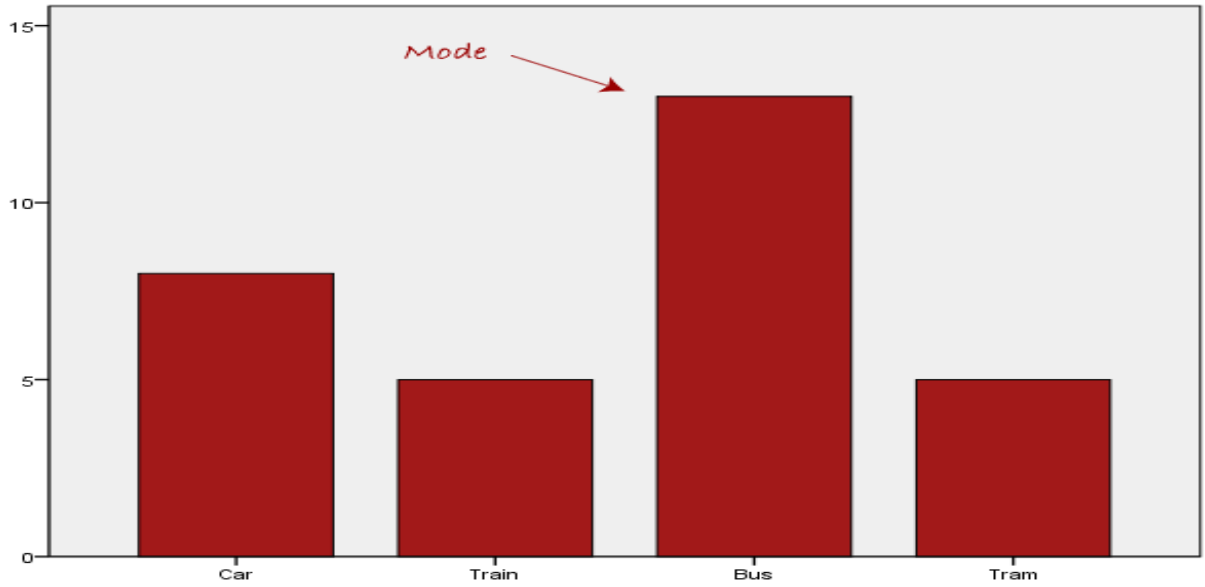
Now we have to take the 5th and 6th score in our data set and average them [(56+56)/2] to get a median of 55.5.

Mode

The mode is the most frequent score in our data set. It represents the highest bar in a bar diagram or histogram.

Example

For the values 8, 9, 10, 10, 11, 11, 11, 12, 13, the mode is 11 as 11 occur most of the time.



Summary of when to use the mean, median and mode

Use the mean to describe the middle of a set of data that *does not* have an outlier. Use the median describes the middle of a set of data that *does* have an outlier. Use the mode when the data is non-numeric or when asked to choose the most popular item.

Type of Variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

Measures of central tendency when we add or multiply each value by same amount

	Data	Mean	Mode	Median
Original Data Set	6, 7, 8, 10, 12, 14, 14, 15, 16, 20	12.2	14	13
Add 3 to each data value	9, 10, 11, 13, 15, 17, 17, 18, 19, 23	15.2	17	16
Multiply 2 times each data value	12, 14, 16, 20, 24, 28, 28, 30, 32, 40	24.4	28	26

When added, since all values are shifted the same amount, the measures of central tendency all shifted by the same amount. If you add 3 to each data value, you will add 3 to the mean, mode and median.

When multiplied, since all values are affected by the same multiplicative values, the measures of central tendency will feel the same affect. If you multiply each data value by 2, you will multiply the mean, mode and median by 2.

Calculation of mean, median and mode for series data

For a series 1, 2, 3 ….n, mean = median = mode = (n + 1) / 2

Geometric mean (GM)

The geometric mean of n numbers is obtained by multiplying them all together and then taking the nth root, that is,

$$GM = \sqrt[n]{\quad\quad}$$

Example

The GM of two numbers 2 and 8 is $\sqrt{\quad}$ = $\sqrt{\quad}$ = 4.

It is useful when we expect that changes in the data occur in a relative fashion. For zero & negative values, geometric mean is not applicable.

Harmonic mean

Harmonic mean for a set of values is defined as the reciprocal of the arithmetic mean of the reciprocals of those values, that is,

$$HM = \frac{\quad\quad}{\frac{\quad}{\quad} \quad \frac{\quad}{\quad} \quad\quad \frac{\quad}{\quad}}$$

Example

The HM of 1, 2, and 4 is

$$\frac{3}{\frac{1}{1}+\frac{1}{2}+\frac{1}{4}} = \frac{1}{\frac{1}{3}\left(\frac{1}{1}+\frac{1}{2}+\frac{1}{4}\right)} = \frac{12}{7}.$$

HM is better when numbers are defined in relation to some unit, like averaging speed.

Nishi has four 10 km segments to her car trip. She drives her car 100 km/hr for the 1st 10 km, 110 km/hr for the 2nd 10 km, 90 km/hr for the 3rd 10 km, 120 km/hr for the 4th 10 km. What is her average speed?

$$HM = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{4}{\frac{1}{100} + \frac{1}{110} + \frac{1}{90} + \frac{1}{120}} = 103.8$$

So, her average speed is 103.8 km/hr.

AM X HM = (GM)²

For any 2 numbers a and b,

AM = (a + b) / 2

GM = $\sqrt{ab}$

HM = 2 / (1 / a + 1 / b) = 2ab / (a + b)

AM X HM = (a + b) / 2 X 2ab / (a + b)

= ab

= (GM) ²

Example

For any two numbers, AM = 10 and GM = 8. Find out the numbers.

$\sqrt{ab}$ = 08

=> ab = 64

(a + b) / 2 = 10

=> a + b = 20 . . . . .(1)

$(a - b)^2 = (a + b)^2 - 4ab$

$= (20)^2 - 4 \times 64$

= 144

=> a - b = 12 . . . . .(2)

Solving (1) and (2)

(a, b) = (16, 4)

Example

For any two numbers, GM = 4√3 and HM = 6. Find out AM and the numbers.

AM = (GM) ² / HM

= (4√3) ² / 6

= 8

√ab = 4√3

=>ab = 48

(a + b) / 2 = 8

=> a + b = 16 . . . .(1)

$(a - b)^2 = (a + b)^2 - 4ab$

$= (16)^2 - 4 \times 48$

= 64

=> a - b = 08 . . . . .(2)

Solving (1) & (2)

(a, b) = (12, 4)

AM ≥ GM ≥ HM

For any two numbers a and b

AM = (a + b) / 2

GM = $\sqrt{ab}$

HM = 2 / (1 / a + 1 / b)

= 2ab / (a + b)

So, AM ≥ GM ≥ HM

$(\sqrt{a} - \sqrt{b})^2 \geq 0$

=> $a + b - 2\sqrt{ab} \geq 0$

=> $a + b \geq 2\sqrt{ab}$

=> $(a + b) / 2 \geq \sqrt{ab}$

=> AM ≥ GM

Multiplying both sides by $2\sqrt{ab} / (a + b)$

$\sqrt{ab} \geq 2ab / (a + b)$

=> GM ≥ HM

Criteria for good measures of central tendency

- Clearly defined
- Readily comprehensible
- Based on all observations
- Easily calculated
- Less affected by extreme values
- Capable of further algebraic treatment

MEASURES OF DISPERSION

If everything were the same, we would have no need of statistics. But, people's heights, ages, etc., do vary. We often need to measure the extent to which scores in a dataset differ from each other. Such a measure is called the dispersion of a distribution. The measures are:

Range (R)	Mean Deviation (MD)	Variance	Standard Deviation (SD)

Example

The average scores of the class tests of two BBA groups are

Groups	Scores					Average
Section E	46	48	50	52	54	50
Section F	30	40	50	60	70	50

In both groups average scores are equal. But in Section E, the observations are concentrated on the center. All students have almost the same level of performance. We say that there is consistence in the observations. In Section F, the observations are not closed to the center. One observation is as small as 30 and one observation is as large as 70. Thus there is greater dispersion in Section F.

Objectives of Dispersion

- **To know the average variation of different values from the average of a series**
- **To know the range of values**
- **To compare between two or more series expressed in different units**
- **To know whether the Central Tendency truly represent the series or not**

Range

The range is the difference between the highest and lowest values of a dataset.

Example

For the dataset {4, 6, 9, 3, 7} the lowest value is 3, highest is 9, so the range is 9-3=6.

Mean Deviation

The mean deviation is the mean of the absolute deviations of a set of data about the mean. For a sample size N, the mean deviation is defined by

$$\mathrm{MD} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}|,$$

Example

Sonia took five exams in a class and had scores of 92, 75, 95, 90, and 98. Find the mean deviation for her test scores.

$$\begin{aligned}\bar{x} &= \frac{\sum x}{n} \\ &= \frac{92+75+95+90+98}{5} \\ &= \frac{450}{5} \\ &= 90\end{aligned}$$

x	$x - \bar{x}$	$\lvert x - \bar{x} \rvert$
92	2	2
75	-15	15
95	5	5
90	0	0
98	8	8
Total		30

$$\mathrm{MD} = \frac{\sum|x - \bar{x}|}{n} = \frac{30}{5} = 6$$

We can say that on the average, Sonia's test scores deviated by 6 points from the mean.

Variance

The variance (σ^2) is a measure of how far each value in the data set is from the mean.

$$\sigma^2 = \frac{\sum(X - \bar{X})^2}{N}$$

Example

Shimmy took ten exams in STA 240 and had scores of 44, 50, 38, 96, 42, 47, 40, 39, 46, and 50. Find the variance for her test scores.

Mean = (44 + 50 +38 +96 + 42 +47 +40+ 39 + 46+ 50) / 10 = 49.2

x	x - 49.2	$(x - 49.2)^2$
44	-5.2	27.04
50	0.8	0.64
38	11.2	125.44
96	46.8	2190.24
42	-7.2	51.84
47	-2.2	4.84
40	-9.2	84.64
39	-10.2	104.04
46	-3.2	10.24
50	0.8	0.64
Total		2600.4

$$\sigma^2 = 2600.4 / 10 = 260.04$$

Standard Deviation

Standard Deviation it is the square root of the Variance defined as

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

Example

For the above example: Standard Deviation, $\sigma = \sqrt{260.04} = 16.12$.

We can say that on the average, Sonia's test scores vary by 16.12 points from the mean.

Standard Deviation is the most important, reliable, widely used measure of dispersion. It is the most flexible in terms of variety of applications of all measures of variation. It is used in many other statistical operations like sampling techniques, correlation and regression analysis, finding co-efficient of variation, skewness, kurtosis, etc.

Coefficient of Variation

The coefficient of variation (CV) is the ratio of the standard deviation to the mean.

$$CV = \frac{\text{standard deviation}}{\text{mean}} \times 100$$

CV should be computed only for data measured on a ratio scale. It may not have any meaning for data on an interval scale.

Why Coefficient of Variation

The coefficient of variation (CV) is used to compare different sets of data having the units of measurement. The wages of workers may be in dollars and the consumption of meat in their families may be in kilograms. The standard deviation of wages in dollars cannot be compared with the standard deviation of amounts of meat in kilograms. Both the standard deviations need to be converted into coefficient of variation for comparison. Suppose the value of CV for wages is 10% and the value of CV for kilograms of meat is 25%. This means that the wages of workers are consistent.

Example

A company has two sections with 40 and 65 employees respectively. Their average weekly wages are $450 and $350. The standard deviations are 7 and 9. (i) Which section has a larger wage bill? (ii) Which section has larger variability in wages?

(i) **Wage bill for section A = 40 x 450 = 18000**
Wage bill for section B = 65 x 350 = 22750
Section B is larger in wage bill.

(ii) **Coefficient of variance for Section A = 7/450 x 100 =1.56 %**
Coefficient of variance for Section B = 9/350 x 100 = 2.57%
Section B is more consistent so there is greater variability in the wages of section A.

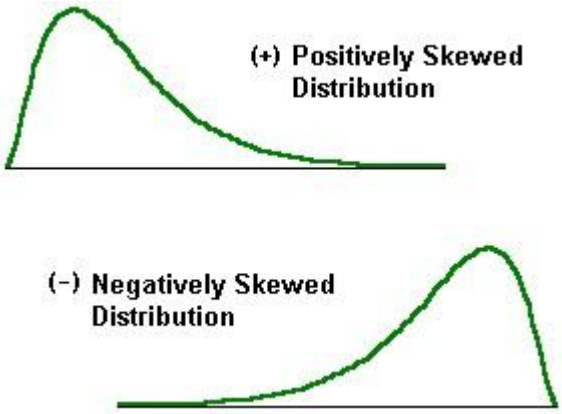
Skewness

It is the degree of departure from symmetry of a distribution.

A positively skewed distribution has a "tail" which is pulled in the positive direction.

A negatively skewed distribution has a "tail" which is pulled in the negative direction.

Example



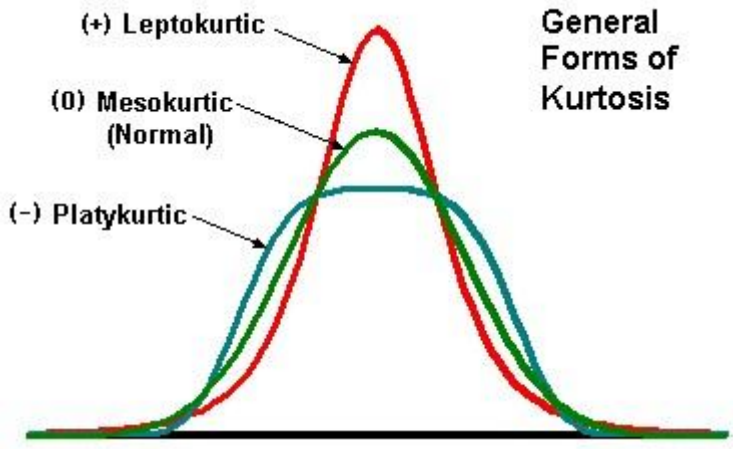
Kurtosis

Kurtosis is the degree of peakedness of a distribution.

A normal distribution is a mesokurtic distribution.

A leptokurtic distribution has higher peak than normal distribution and has heavier tails.

A platykurtic distribution has a lower peak than a normal distribution and lighter tails.



Correlation

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.

Example

Height and weight are related; taller people tend to be heavier than shorter people. The relationship isn't perfect. People of the same height vary in weight, and we can easily think of two people we know where the shorter one is heavier than the taller one. Nonetheless, the average weight of people 5'5" is less than the average weight of people 5'6", and their average weight is less than that of people 5'7", etc. Correlation can tell us just how much of the variation in peoples' weights is related to their heights.

Types of correlation

Positive correlation

Here, as the values a variable increase, the values of the other variable also increase and as the value of a variable decreases, the value of the other variable also decreases.

Example

Relation between training and performance of employees in a company

Relation between price and supply of a product

Negative correlation

Here, as the values a variable increase, the values of the other variable also decrease and as the value of a variable decreases, the value of the other variable also increases.

Example

Relation between television viewing and exam grades

Relation between price and demand of a product

Zero correlation

Here, change in one variable has no effect on the other variable.

Example

Relation between height and exam grades

Correlation coefficient

It measures the strength and the direction of the relationship between two variables. Its value always lies between - 1 and + 1. It is defined as

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}}$$

Interpretation of correlation coefficient

r = 0 indicates no relation

r = + 1 indicates a perfect positive relation

r = - 1 indicates a perfect negative relation

Values of r between 0 and 0.3 (0 and - 0.3) indicate a weak positive (negative) relation

Values of r between .3 and .7 (.3 and - .7) indicate a moderate positive (negative) relation

Values of r between 0.7 and 1(- 0.7 and -1) indicate a strong positive (negative) relation

Example

Compute correlation coefficient and interpret the result from the following table

Age (x)	43	21	25	42	57	59
Glucose Level (y)	99	65	79	75	87	81

	x	y	xy	x^2	y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Total	247	486	20485	11409	40022

So, Σ x = 247, Σ y = 486, Σ x y = 20485, Σ x^2 = 11409, Σ y^2 = 40,022, n = 6

Putting these values in the equation of r, we get: r = 0.5298 which means the variables have a moderate positive correlation.

Regression

It is a statistical measure that attempts to model the relationship between a dependent variable (denoted by Y) and few other independent variables (denoted by X's).

$$y = a + bx$$

where:

$$a = \frac{\sum y - b\sum x}{n}$$

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

A linear regression line has an equation of the form Y = a + b X

a gives expected amount of change in Y for X=0

b gives expected amount of change in Y for 1 unit change in X

Example

For the previous example, fit a linear regression line and interpret the result.

[Try yourself]

Correlation vs. Regression

Correlation	Regression
It cannot predict	It can predict
It cannot express cause and effect	It can express cause and effect
r ranges from - 1 to + 1	a and b ranges from - ∞ to + ∞

Probability

The probability of an event A is the number of ways event A can occur divided by the total number of possible outcomes. [0 ≤ P (A) ≤ 1]

$$P(A) = \frac{\text{number of outcomes favorable to event A}}{\text{total number of outcomes}}$$

Example

A glass jar contains 6 red, 5 green, 8 blue and 3 yellow marbles. If a single marble is chosen at random from the jar, what is the probability of choosing a red marble?

Number of ways it can happen: 6 (there are 6 reds)

Total number of outcomes: 22 (there are 22 marbles in total)

So the required probability = $\frac{6}{22}$

Experiment

An action where the result is uncertain is called an experiment.

Example

Tossing a coin, throwing dice etc. are all examples of experiments.

Sample Space

All the possible outcomes of an experiment is called a sample space.

Example

A die is rolled, the sample space S of the experiment is S = {1, 2, 3, 4, 5, 6}.

Event

A single result of an experiment is called an event.

Example

Getting a Tail when tossing a coin is an event.

Probability Example

A total of five cards are chosen at random from a standard deck of 52 playing cards. What is the probability of choosing 5 aces?

$$P(\text{5 aces}) = \frac{0}{30} = 0$$

A die is rolled, find the probability that an even number is obtained.

Sample space, S = {1, 2, 3, 4, 5, 6}

The event "an even number is obtained", E = {2, 4, 6}

P (E) = n (E) / n(S) = 3 / 6

A teacher chooses a student at random from a class of 30 girls. What is the probability that the student chosen is a girl?

$$\text{P (girl)} = \frac{30}{30} = 1$$

In a lottery, there are 10 prizes and 25 blanks. A lottery is drawn at random. What is the probability of getting a prize?

$$\text{P (getting a prize)} = \frac{10}{(10 + 25)} = \frac{10}{35} = \frac{2}{7}$$

At a car park there are 60 cars, 30 vans and 10 Lorries. If every vehicle is equally likely to leave, find the probability of: a) Van leaving first b) Lorry leaving first.

a) Let S be the sample space and A be the event of a van leaving first.

So, n (S) = 100 and n (A) = 30

$$P(A) = \frac{30}{100} = \frac{3}{10}$$

b) Let B be the event of a lorry leaving first. So, n (B) = 10.

$$P(B) = \frac{10}{100} = \frac{1}{10}$$

In a box, there are 8 black, 7 blue and 6 green balls. One ball is picked up randomly. What is the probability that ball is neither black nor green?

Total number of balls = (8 + 7 + 6) = 21

Let E = event that the ball drawn is neither black nor green

= event that the ball drawn is blue.

$$P(E) = \frac{n(E)}{n(S)} = \frac{7}{21} = \frac{1}{3}$$

Two coins are tossed, find the probability that two heads are obtained.

Each coin has 2 possible outcomes: H (heads) and T (Tails)

Sample space, S = {(H, T), (H, H), (T, H), (T, T)}

The event "two heads are obtained", E = {(H, H)}

P (E) = n (E) / n (S) = 1 / 4

Two dice are rolled; find the probability that the sum of the values is

a) equal to 1 b) equal to 4 c) less than 13

a) The sample space, S = { (1,1),(1,2),(1,3),(1,4),(1,5),(1,6)
(2,1),(2,2),(2,3),(2,4),(2,5),(2,6)
(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)
(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)
(5,1),(5,2),(5,3),(5,4),(5,5),(5,6)
(6,1),(6,2),(6,3),(6,4),(6,5),(6,6) }

Let E be the event "sum equal to 1". There are no such outcomes.
So, P (E) = n (E) / n (S) = 0 / 36 = 0

b) Let E be the event "sum equal to 4". E = {(1, 3), (2, 2), (3, 1)}.
So, P (E) = n (E) / n (S) = 3 / 36

c) Let E be the event "sum is less than 13". E = S.
So, P (E) = n (E) / n (S) = 36 / 36 = 1

Sampling

It is the selection process of a sample from a population.

Example

Selection of class monitors from the entire class.

Simple random sampling

A sampling procedure that assures that each element in the population has an equal probability of being selected in the sample is called simple random sampling.

Example

There are 50 students in the class. We are to select 2 class monitors. Each student has a probability of 1/50 to be selected. So, selection of class monitors from the class is an example of simple random sampling.

Stratified Sampling

It is a sampling technique where we divide the entire population into different groups and then randomly select the objects from those different groups.

Example

There are 50 students in the class. We are to select a group combining both male and female students. We divide the 50 students into 2 groups: male and female. Then we select students randomly from these 2 groups. Hence, our selected group will be a stratified sample and the selection process will be called stratified sampling.

Cluster Sampling

It is a sampling technique where the entire population is divided into clusters and a random sample of these clusters are selected. All observations in the selected clusters are included in the sample.

Example

In a study of homeless people across Dhaka, all the wards are selected and a significant number of homeless people are interviewed in each one. Here, the selected wards are the clusters. So, the selected sample is a cluster sample and the selection process is cluster sampling.

Systematic Sampling

It is a sampling technique involving the selection of elements from an ordered sampling frame. Here, every k^{th} element in the frame is selected, where *k*, the sampling interval, is calculated as:

$$k = \frac{N}{n}$$

Here, *n* is the sample size and *N* is the population size.

Example

Suppose we want to sample 8 houses from a street of 120 houses. 120/8=15, so every 15th house is chosen after a random starting point between 1 and 15. If the random starting point is 11, then the houses selected are 11, 26, 41, 56, 71, 86, 101, and 116.

Lottery method

Here, sampling units are represented by small chits of paper which are folded and mixed together. From this the required numbers are picked out blind folded.

Example

ID no. of 60 students is written on small chits of papers which can be folded in such a way that they are indistinguishable from each other. Then 10 folded chits are drawn from this lot at random. This selection method of 10 students is called lottery method.

Expectation of random variables

For discrete random variables, E (X) = $\sum x\, P\,(x)$

Example

What is the expected value when we roll a fair die?

There are 6 possible outcomes: 1, 2, 3, 4, 5, 6, each of these has a probability of 1/6 of occurring. Let X represents the outcome of the experiment.

X	1	2	3	4	5	6
P(X)	1/6	1/6	1/6	1/6	1/6	1/6

E (X) = 1 x 1/6 + 2 x 2/6 + 3 x 3/6 + 4 x 4/6 + 5 x 5/6 + 6 x 6/6 = 7/2

The probability distribution of X, the number of red cars Tanya meets on his way to work each morning, is given by the following table:

X	0	1	2	3	4
P(X)	.41	.37	.16	.05	.05

Find the number of red cars that Tanya expects to run into each morning.

Since X is a discrete random variable,

E (X) = 0 x .41 + 1 x .37 + 2 x .16 + 3 x .05 + 4 x .05 = .88

For continuous random variables, E (X) = $\int x\, P\,(x)dx$

Example

A company uses certain software to check errors on any program. The number of errors found is represented by a random variable X whose density function is given by

$$f(x) = \begin{cases} \frac{2(x+2)}{5} & 0 < x < 4 \\ 0, & otherwise \end{cases}$$

Find the average number of errors the company expects to find in a given program.

The random variable X is given as a continuous random variable, so

$$\mathrm{E}(x) = \int_0^4 x \left(\frac{2(x+2)}{5}\right) \mathrm{dx}$$

$$= \frac{2}{5}\int_0^4 (x^2 + 2x)\mathrm{dx}$$

$$= \frac{2}{5}\left[\frac{x^3}{3} + x^2\right]_0^4$$

$$= \frac{2}{5}\left(\frac{4^3}{3} + 4^2\right) - 0$$

$$= \frac{2}{5}\left(\frac{112}{3}\right)$$

$$= \frac{224}{15}$$

Variance of random variables

For both discrete and continuous random variables, $V(X) = E(X^2) - [E(X)]^2$

Example

What will be the variance when we roll a fair die?

There are 6 possible outcomes: 1, 2, 3, 4, 5, 6, each of these has a probability of 1/6 of occurring. Let X represents the outcome of the experiment.

X	1	2	3	4	5	6
P(X)	1/6	1/6	1/6	1/6	1/6	1/6

$E(X) = 1 \times 1/6 + 2 \times 2/6 + 3 \times 3/6 + 4 \times 4/6 + 5 \times 5/6 + 6 \times 6/6 = 7/2$

$E(X^2) = 1^2 \times 1/6 + 2^2 \times 2/6 + 3^2 \times 3/6 + 4^2 \times 4/6 + 5^2 \times 5/6 + 6^2 \times 6/6 = 147/2$

$V(X) = E(X^2) - [E(X)]^2 = 147/2 - (7/2)^2$

The probability distribution of X, the number of red cars Tanya meets on his way to work each morning, is given by the following table:

X	0	1	2	3	4
P(X)	.41	.37	.16	.05	.05

Find the variance of the number of red cars that Tanya runs into each morning.

Since X is a discrete random variable,

$E(X) = 0 \times .41 + 1 \times .37 + 2 \times .16 + 3 \times .05 + 4 \times .05 = .88$

$E(X^2) = 0^2 \times .41 + 1^2 \times .37 + 2^2 \times .16 + 3^2 \times .05 + 4^2 \times .05 = 2.26$

$V(X) = E(X^2) - [E(X)]^2 = 2.26 - (.88)^2 = 1.4856$

$P(x) = x$, $0<x<1$
Calculate the variance.

Probability distribution

It is a table or an equation that links each outcome of an experiment with its probability.

Example

Probability distribution that results from the rolling of a fair die is

X	1	2	3	4	5	6
P(X)	1/6	1/6	1/6	1/6	1/6	1/6

Discrete distributions

binomial distribution, negative binomial distribution, geometric distribution, hyper geometric distribution, discrete uniform distribution, Poisson distribution

Continuous distributions

F-distribution, t-distribution, exponential distribution, beta distribution, gamma distribution, normal distribution, continuous uniform distribution

Binomial Distribution

A random variable X belongs to binomial distribution if it follows the distribution

$$P(x) = {}^{n}c_{x}\, p^{x} q^{n-x}$$

n = number of trials, p = probability of success, q = probability of failure (p + q = 1)

For a binomial distribution, E (X) = np and V (X) = npq = np (1 – p)

Example

A survey found that 30% IUBAT students earn money from tuitions. If 5 students are selected at random, find the probability that at least 3 of them have tuitions.

Here: n = 5 , p = 30% = 0.3 , q = (1 - 0.3) = 0.7 , x = 3, 4, 5

$P(3) = {}^{5}c_{3}(0.3)^{3}(0.7)^{5-3} = 0.132$

$P(4) = {}^{5}c_{4}(0.3)^{4}(0.7)^{5-4} = 0.028$

$P(5) = {}^{5}c_{5}(0.3)^{5}(0.7)^{5-5} = 0.002$

P (at least 3 students have tuitions) = 0.132 + 0.028 + 0.002 = 0.162

A fair coin is tossed 8 times. Find the probability of exactly 3 tails.

Here: n = 8 , x = 3 , p = q = 0.5

$P(3) = {}^{8}c_{3}(0.5)^{3}(0.5)^{8-3} = 0.219$

Rehab randomly guesses 5 questions. Find the probability that he gets exactly 3 correct. Each question has 5 possible choices.

Here: n = 5 , x = 3 , p = 1 / 5 = .2 , q = (1 – .2) = .8

$P(3) = {}^{5}c_{3}(0.2)^{3}(0.8)^{5-3} = 0.05$

Hospital records show that of patients suffering from a certain disease, 75% die of it. What is the probability that of 6 randomly selected patients, 4 will recover?

Here: n = 6 , x = 4 , p = 25% = .25 , q = 75% = .75

$P(4) = {}^{6}c_{4}(0.25)^{4}(0.75)^{6-4} = 0.033$

For a binomial distribution, mean is 2 and variance is 1. Find the constants.

E (X) = np = 2	npq / np = .5	np = 2
V (X) = npq = 1	q = .5	n x .5 = 2
	So, p = .5	n = 4

Hence the constants are: n = 4, p = .5 and q = .5

Poisson distribution

A random variable X belongs to binomial distribution if it follows the distribution

$P(x) = e^{-m}m^{x} / x!$

m = mean, e = 2.72 (a constant)

For a Poisson distribution, E (X) = V (X) = m

Example

Vehicles pass through a junction on a busy road at an average rate of 300 per hour. Find the probability that none passes in a given minute.

Here: m = 300 / 60 = 5 , x = 0

$P(0) = (2.72)^{-5}(5)^{0} / 0! = (2.72)^{-5}$

A company makes electric motors. The probability that a motor is defective is 0.01. What is the probability that a sample of 300 motors will contain exactly 5 defective motors?

Here: m = 300 x 0.01 = 3 , x = 5

$P(5) = (2.72)^{-3}(3)^{5} / 5! = 0.101$

Electricity fails according to Poisson distribution with average of 3 failures per 20 weeks, calculate the probability that there will not be more than 1 failure during a specific week.

Here: m = 3 / 20 = 0.15 , x = 0, 1

$P(0) = (2.72)^{-0.15}(0.15)^{0} / 0!$

$P(1) = (2.72)^{-0.15}(0.15)^{1} / 1!$

P (there will not be more than 1 failure) = P (0) + P (1) = 0.99

Normal distribution

A random variable X belongs to binomial distribution if it follows the distribution

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

u = mean, σ = standard deviation, e = 2.72 (a constant), π = 3.414 (a constant)

If we have mean μ and standard deviation σ, then the variable

$$Z = \frac{X-\mu}{\sigma}$$ **is said to follow a standard normal distribution.**

Z table for necessary calculation help

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Example

P (Z < 1.4) = Ψ (1.4) = 0.9192	P (Z < - 1.4) = 1 - P (Z < 1.4) = 1 – Ψ (1.4) = 1 – 0.9192 = 0.0808
P (Z > 1.4) = 1 - P (Z < 1.4) = 1 – Ψ (1.4) = 1 – 0.9192 = 0.0808	P (0.67 < Z< 2.33) = P (Z < 2.33) - P (Z < 0.67) = Ψ (2.33) - Ψ (0.67) = 0.9901 – 0.7480 = 0.2421

Importance of normal distribution

The normal distribution is the most used statistical distribution. The principal reasons are:

a) Normality arises naturally in many physical, biological, and social measurement situations.
b) Normality is important in statistical inference.

Hypothesis Test

A statistical hypothesis is a statement about a population which we want to verify on the basis of information contained in a sample. Hypothesis Test is an attempt to arrive at a correct decision about a pre-stated statistical hypothesis.

Example

Internet server claims that computer users in IUBAT spend on the average 15 hours per week on browsing. We conduct a survey based on a sample of 250 users to arrive at a correct decision. Here, the server's claim is referred to as a statistical hypothesis and we are doing a hypothesis test.

Null hypothesis

It is a statement which tells us that no difference exits between the parameter and the statistic being compared to it.

Example

Given the test scores of two random samples of men and women, does one group differ from the other? A possible null hypothesis is

$H_0 : \mu_1 = \mu_2$

μ_1 = mean of population 1 and μ_2 = mean of population 2

Alternative hypothesis

The alternative hypothesis is the logical opposite of the null hypothesis. The rejection of a null hypothesis leads to the acceptance of the alternative hypothesis.

Example

Given the test scores of two random samples of men and women, does one group differ from the other? A possible alternative hypothesis is

$H_1 : \mu_1 > \mu_2$

μ_1 = mean of population 1 and μ_2 = mean of population 2

One tailed test

A hypothesis test where the alternative is one sided is called a one tailed test.

Example

$H_1: \mu_1 > \mu_2$

μ_1 = mean of population 1 and μ_2 = mean of population 2

Two tailed test

A hypothesis test where the alternative is two sided is called a two tailed test.

Example

$H_1: \mu_1 \neq \mu_2$

μ_1 = mean of population 1 and μ_2 = mean of population 2

Level of significance

It is the probability with which we are willing to risk rejecting the null hypothesis even though it is true. We denote it as α.

Type I error

It is the probability of rejecting the null hypothesis when the null hypothesis is true.

P (type I error) = P (reject H_0 I H_0 true) = α

Type II error

It is the probability of accepting the null hypothesis when the null hypothesis is false.

P (type II error) = P (accept H_0 I H_0 false) = β

Example

Consider a defendant in a trial. The null hypothesis is "defendant is not guilty;" the alternate is "defendant is guilty." A Type I error would correspond to convicting an innocent person; a Type II error would correspond to setting a guilty person free.

		Reality	
		Not guilty	Guilty
Verdict	Guilty	Type I Error : Innocent person goes to jail	Correct Decision
	Not guilty	Correct Decision	Type II Error : Guilty person goes free

Test statistic

The statistic used to provide evidence about the null hypothesis is called test statistic.

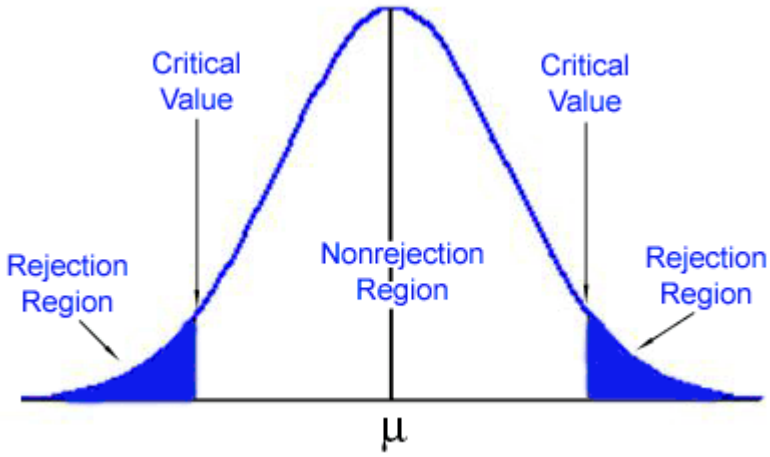
Example

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$ **is a test statistic used for testing sample means.**

Critical / Rejection region

If the value of the test statistic falls into this region, we reject the null hypothesis.

Example



Steps in hypothesis testing

a) State the null hypothesis, H_o
b) State the alternative hypothesis, H_1
c) Choose the level of significance, α
d) Select an appropriate test statistic
e) Calculate the value of the test statistic
f) Determine the critical region
g) reject H_o if the value of the test statistics falls in the critical region; otherwise accept H_o

Example

A firm produces bulbs having a length of life normally distributed with mean 1600 hours and standard deviation 180 hours. Test the hypothesis $\mu = 1600$ vs. $\mu \neq 1600$ if the random sample of 30 bulbs has an average life 1576 hours.

1. H_o : $\mu = 1600$

2. H_1 : $\mu \neq 1600$

3. $\alpha = 0.01$

4. $Z = (1576 - 1600) / (80 / \sqrt{30}) = -1.64$

5. The critical region at $\alpha = 0.05$ for two tailed test is ± 2.58

6. Since our calculated value of Z falls in the acceptance region, so we accept H_o

A sample of 16 observations taken from a normal population has mean 110 and standard deviation 30. Test the hypothesis $\mu = 100$ vs. $\mu > 100$ at 0.05 level of significance.

1. H_o : $\mu = 100$

2. H_1 : $\mu > 100$

3. $\alpha = 0.05$

4. $Z = (110 - 100) / (30 / \sqrt{16}) = 1.33$

5. The critical region at $\alpha = 0.05$ for one tailed test is 1.64

6. since our calculated value of Z falls in the acceptance region, so we accept H_o

A sample of size 20 taken from normal distribution has mean 16.4 and standard deviation 2.255. Does this suggest that the population mean is greater than 15?

1. H_o : $\mu \leq 15$

2. H_1 : $\mu > 15$

3. $\alpha = 0.05$

4. $t = (16.4 - 15) / (2.255 / \sqrt{20}) = 2.776$

5. The critical region at $\alpha = 0.05$ for one tailed test is 1.64

6. since our calculated value of t falls in the critical region, so we reject H_o