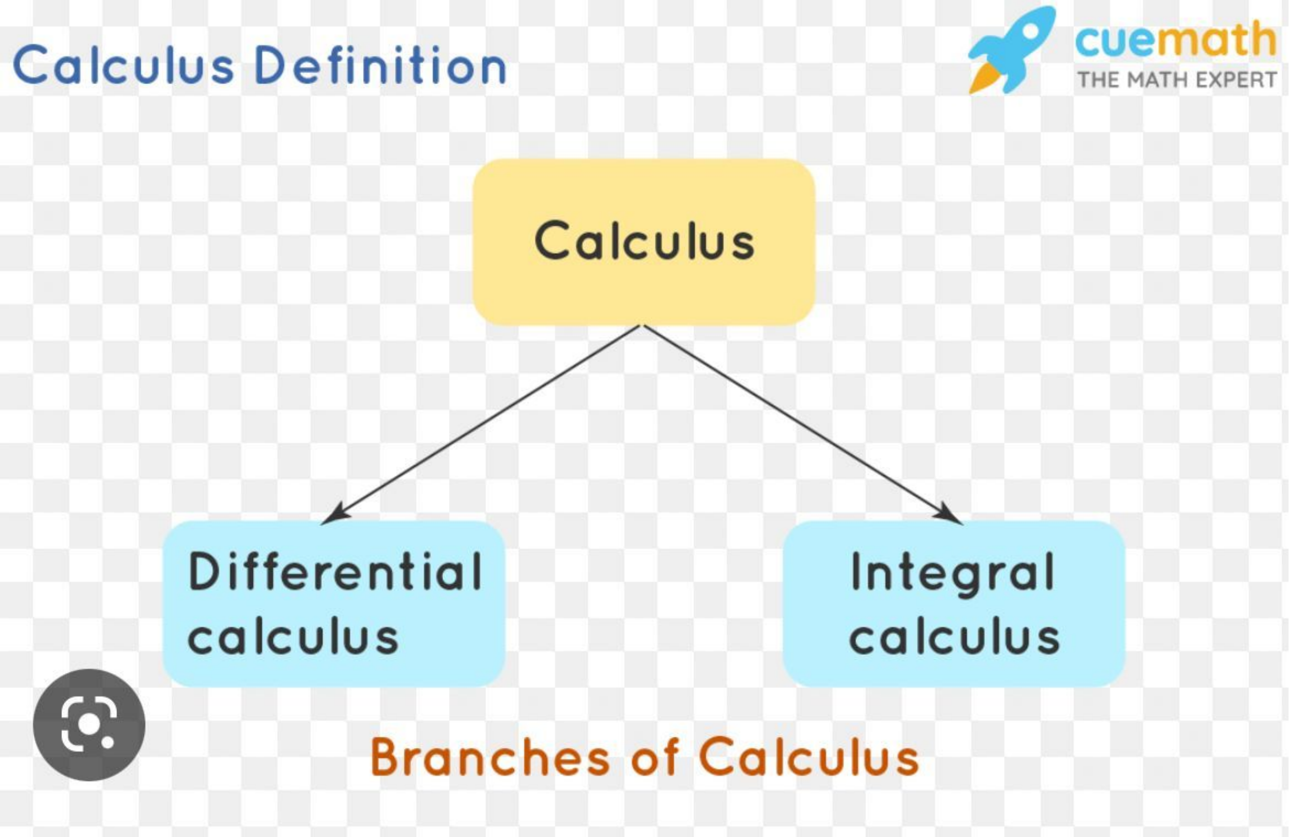
matics for mashine learning

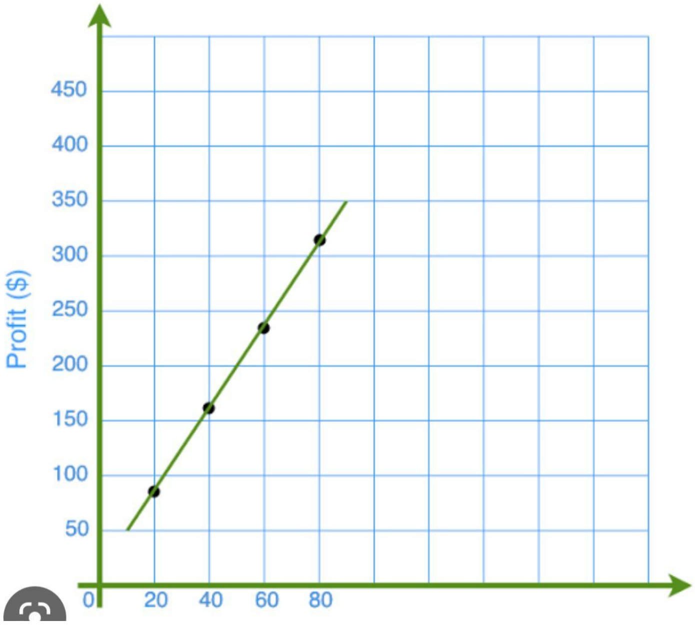
-It helps to finding out the co-reation between two variables by measuring how one variable changes when their is a change in another variable.

D



- TYPES OF CHANGE:

1)CONSTANT CHANGE (Linear): A linear function is a function that has a constant rate of change.

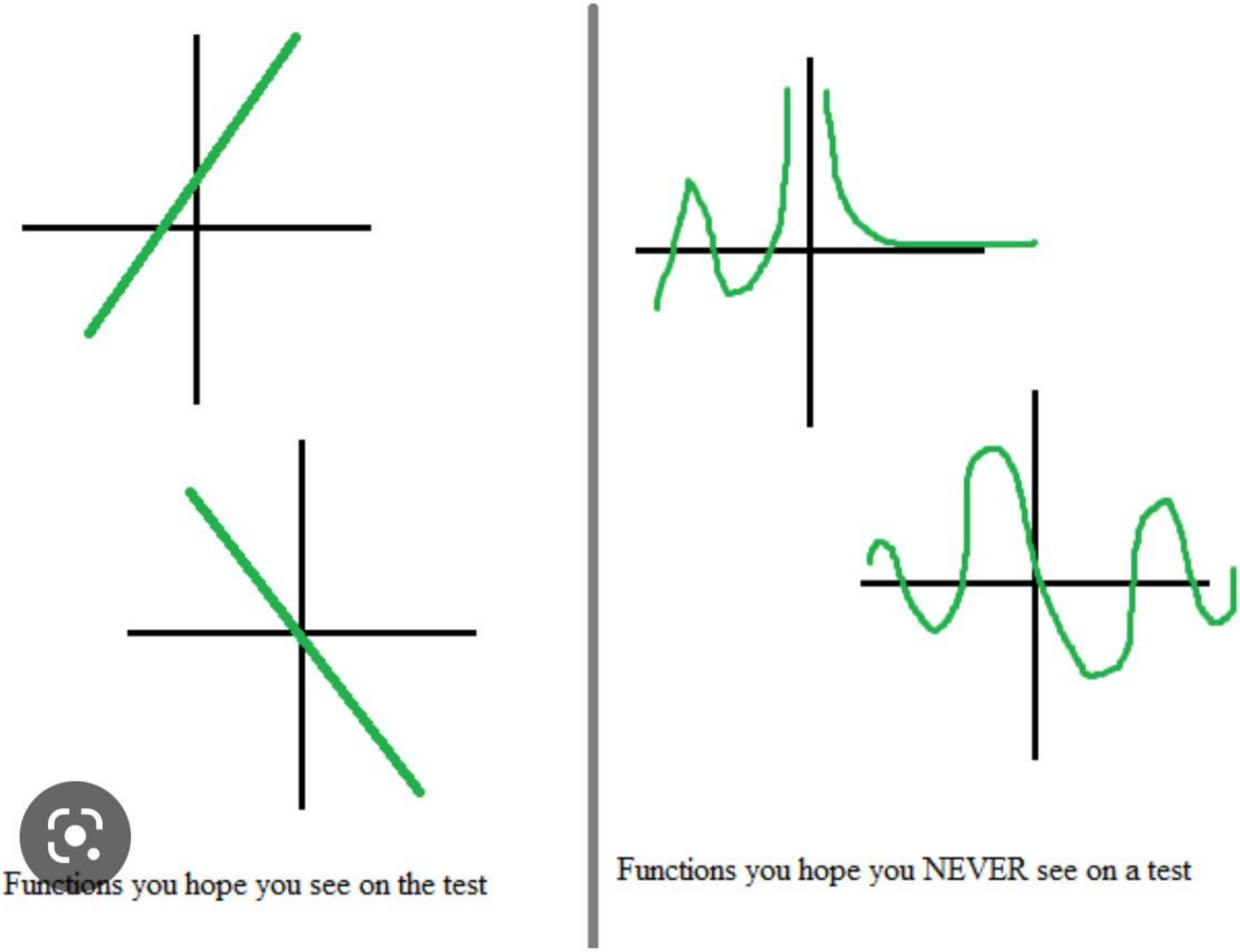


D

FORMULA= y=mx+b
where m= slope
b= intercept

2) NON-CONSTANT CHANGE (Non-linear): A non-linear function does not have a constant rate of change.
So, its graph is not a line.

D



Functions you hope you see on the test

Functions you hope you NEVER see on a test

-Calculus is used to finding the reation of non-linear relations.

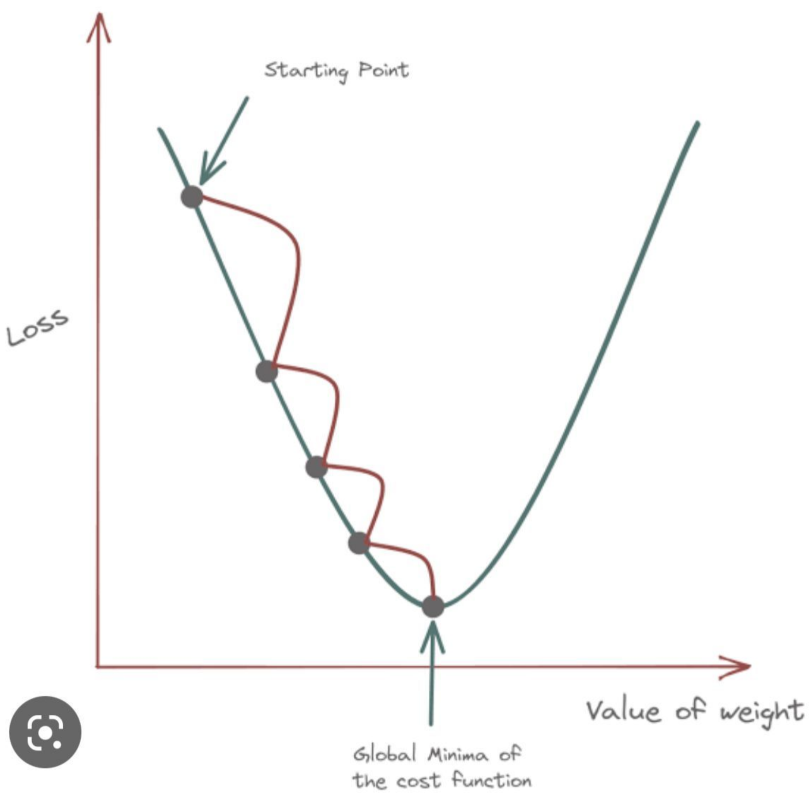
-SLOPE or Gradient-The rate of change is called as slope.

FORMULA=

* DIFFERENTATION:Dividing the non-linear lines into small small peices.The process of finding out the rate of change of a variables with respect to another is known as "Differentation"

-Gobal-minima:- The point at which a function takes the minimum value is called global minimam.
-Learning Rate:Learning rate is one such hyper-parameter that defines the adjustment in the weights of our network with respect to the loss gradient descent.

D



*GRADIENT DESCENT:In Gradient Descent loss function is minimized to reduce error, or we can say, it i s an iterative alogrithm,
that starts from a random point on a
funtion & travels downs its slope in steps,until it reaches the lowest point of that surface.
-where Gradient is mean slope and,
-Gradient Descent menas descending a slope to reach lowest point on the surface.

-Steps:
1) Find the slope
2) Randomly initialize the objective function(this becomes our starting point)
3)Putting values of starting pointsb in gradient(e.g x=0)
4) Find step size (Gradient)*(learning rate)
Continue
5)Finding new values of x
new value (of x)=old value (of x)-(step size)

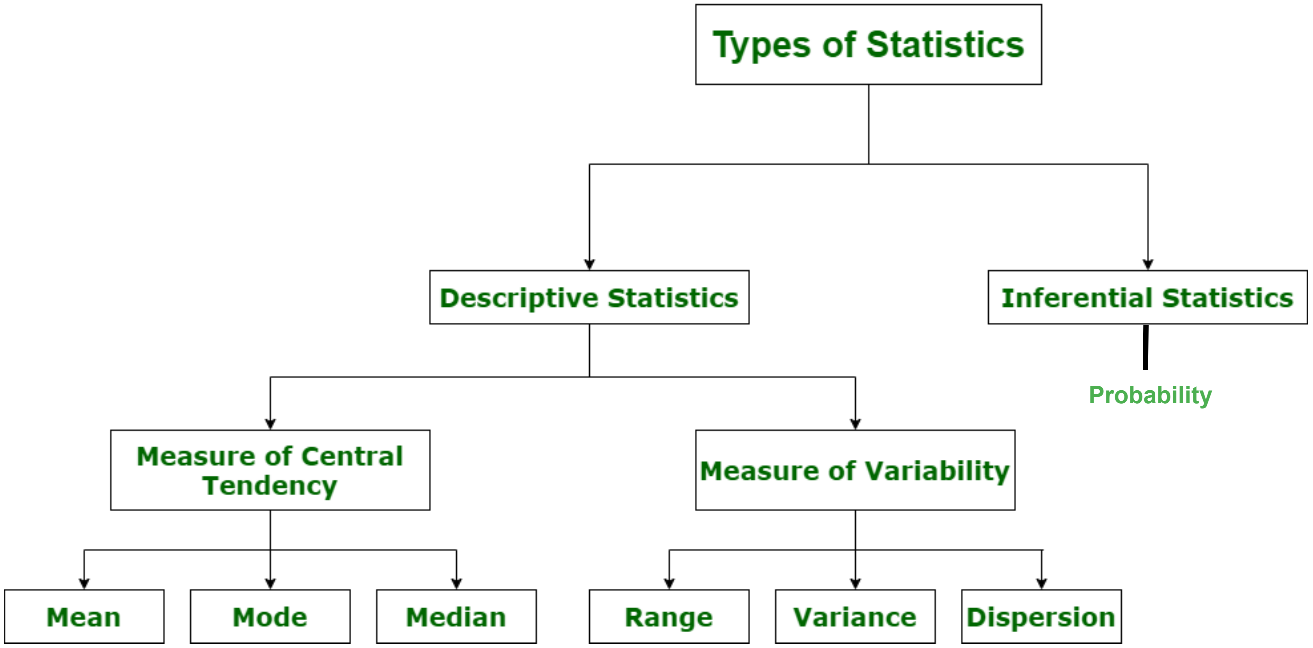
Contiune- repeat same 3,4,5 steps until we reach the minima where slope/gradient is 0(zero)

1. (Statistics)

Statistics is the science concerned with developing and studying methods for collecting, analyzing, interpreting and presenting empirical data.

Basically, there are two types of statistics.

(1). Descriptive Statistics
(2). Inferential Statistics



(1). Descriptive Statistics-
Descriptive statistics is a way of organising, representing, and explaining a set of data using charts, graphs, and summary measures. Histograms, pie charts, bars, and scatter plots are common

ways to summarise data and present it in tables or graphs.

(2). Inferential Statistics-
inferential statistics to convey the meaning of the collected data after it has been collected, evaluated, and summarised. The probability principle is used in inferential statistics to determine if patterns found in a study sample may be extrapolated to the wider population from which the sample was drawn. Inferential statistics are used to test hypotheses and study correlations between variables, and they can also be used to predict population sizes. Inferential statistics are used to derive conclusions and inferences from samples, i.e. to create accurate generalisations.

(1). [Descriptive Statistics]

What is Data in Statistics?

Data is a collection of facts, such as numbers, words, measurements, observations etc.

(Types of Data)

(1).Qualitative data- it is descriptive data.
Example- She can run fast, He is thin.

(2).Quantitative data- it is numerical information.
Example- An Octopus is an Eight legged creature.

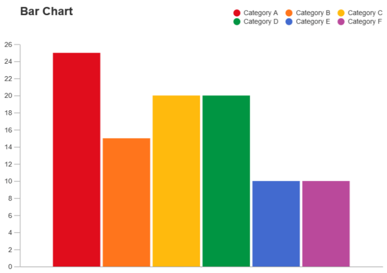
Types of quantitative data-

(1).Discrete data- has a particular fixed value. It can be counted

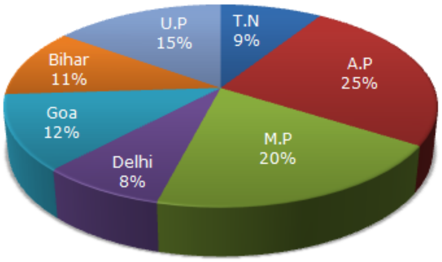
(2).Continuous data- is not fixed but has a range of data. It can be measured.Representation of Dat
a

(Representation of Data)
There are different ways to represent data such as through graphs, charts or tables.
The general representation of statistical data are:

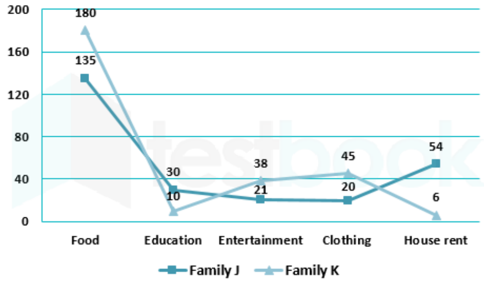
(1) Bar Graph



(2) Pie Chart

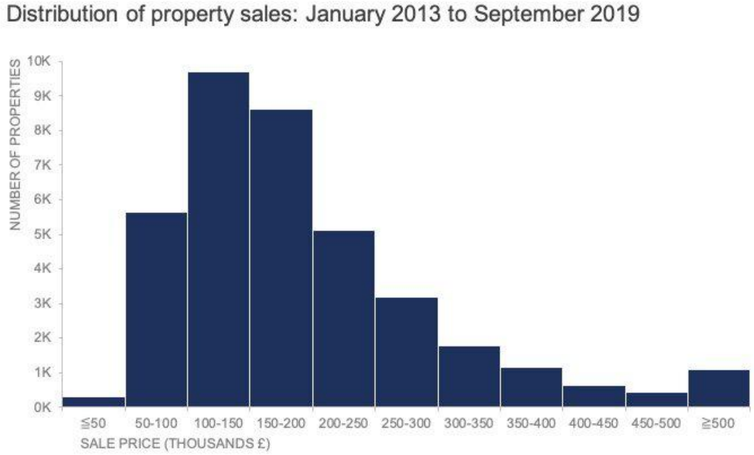


(3) Line Graph



(4) Pictograph

(5) Histogram



(6) box plot

(7) KDE (kernal dencity estimation)

(Measures of Central Tendency)

In Mathematics, statistics are used to describe the central tendencies of the grouped and ungrouped data.

The three measures of central tendency are:

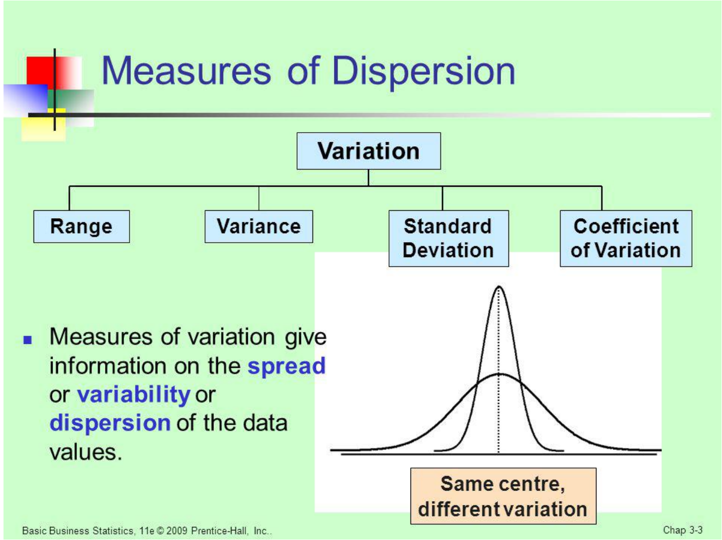
(1). Mean (denote by “mu”) avarage
(2). Median (even and odd)
(3). Mode (most frequent)

(All three measures of central tendency are used to find the central value of the set of data.)

(Measures of Dispersion)-

In statistics, the dispersion measures help interpret data variability, i.e. to understand how homogenous or heterogeneous the data is.In simple words, it indicates how squeezed or scattered the variable is. However, there are two types of dispersion measures, absolute and relative. They are tabulated as below:

(Absolute measures of dispersion)



Measures of Dispersion

Variation

Range

Variance

Standard Deviation

Coefficient of Variation

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

Same centre, different variation

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc..

Chap 3-3

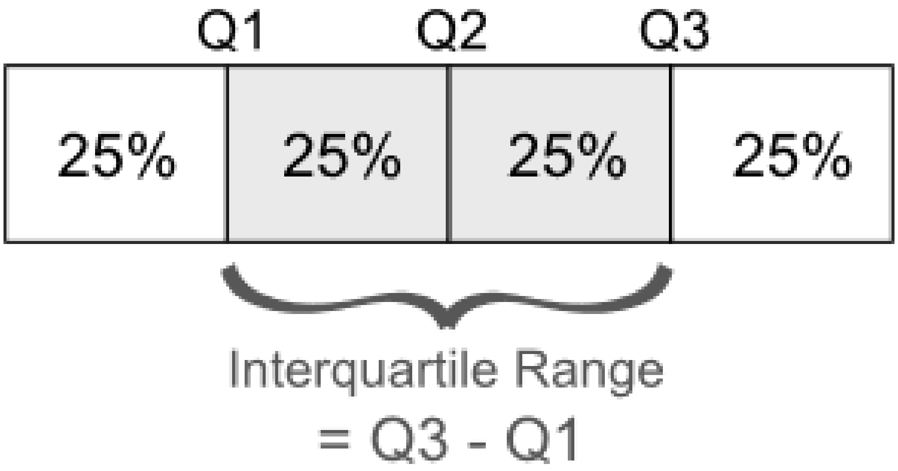
(1).Variance-

Variance is the measure of how notably a collection of data is spread out. If all the data values are identical, then it indicates the variance is zero.There can be two types of variances in statistics, namely, sample variance and population variance. The symbol of variance is given by σ2. Variance is widely used in hypothesis testing,

(2) Standard deviation- denote by σ “sigma” or σx formola is underroot varriance

Standard Deviation is a measure which shows how much variation (such as spread, dispersion, spread,) from the mean exists.

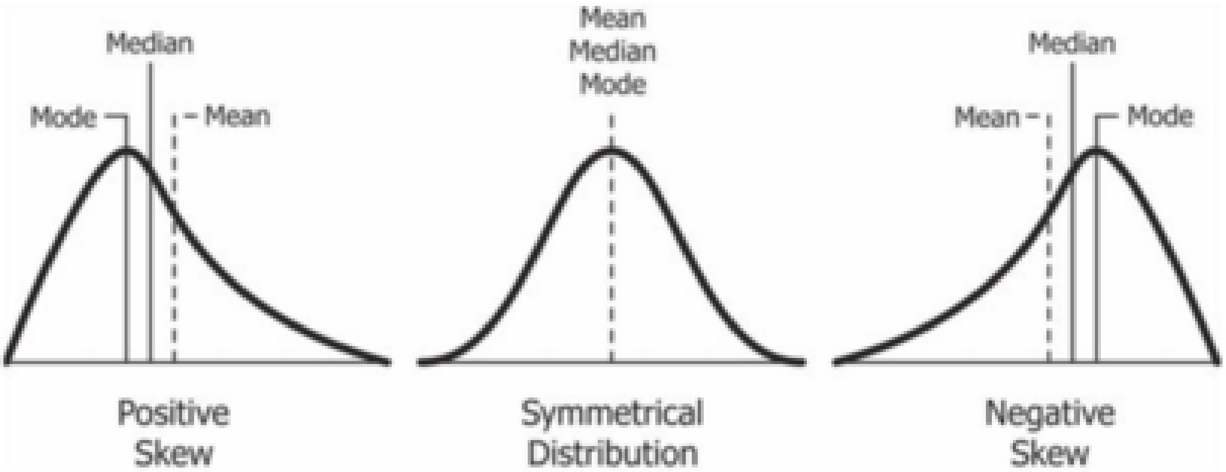
Quartiles and Quartile



(1) Q1 - 25%
(2) Q2 - 50%
(3) Q3 - 75%
(4) IQR (INTER QUATILE RANGE)

The interquartile range tells you the spread of the middle half of your distribution.
IQR=Q3-Q1

(Skewness in Statistics)-



Skewness, in statistics, is a measure of the asymmetry in a probability distribution.
It measures the deviation of the curve of the normal distribution for a given set of data.

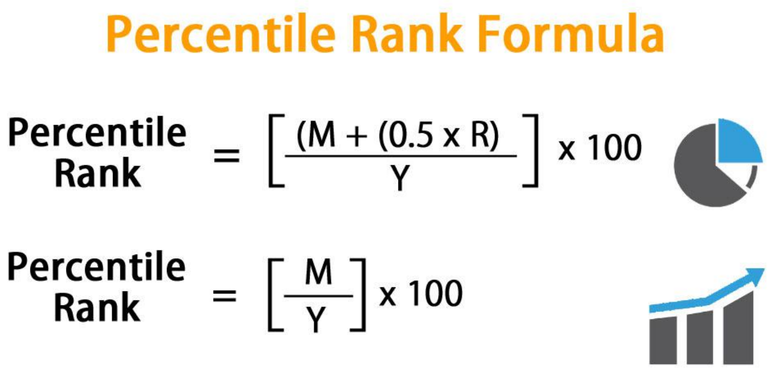
(Percentage and percentile)

percentage and percentile-

(1)- percentage-

The percentage is a mathematical value presented out of 100 and percentile is the per cent of values below a specific value.

Percentage = (Numerator Denominator) × 100 or (X Y) × 10



Percentile Rank Formula

$$\text{Percentile Rank} = \left[\frac{(M + (0.5 \times R)}{Y}\right] \times 100$$

$$\text{Percentile Rank} = \left[\frac{M}{Y}\right] \times 100$$

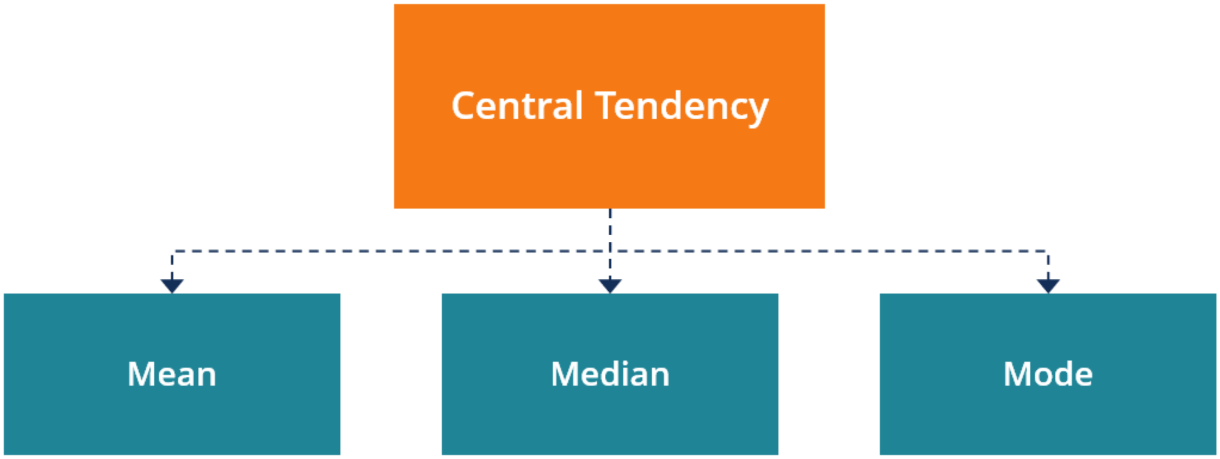
(2)-Percentile-

A percentile is a comparison score between a particular score and the scores of the rest of a group.

It shows the percentage of scores that a particular score surpassed.

(P) percentile = (nth percentile/100) × Total number of values in the list .

(Measures of central tendency)



A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.

There are three main measures of central tendency:

(1) mean

The mean is the sum of the value of each observation in a dataset divided by the number of observations.

This is also known as the arithmetic average.

Looking at the retirement age distribution again:
54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

The mean is calculated by adding together all the values (54+54+54+55+56+57+57+58+58+60+60 = 623) and dividing by the number of observations (11) which equals 56.6 years.

(2) median

The median is the middle value in distribution when the values are arranged in ascending or descending order.
The median divides the distribution in half (there are 50% of observations on either side of the median value).
Looking at the retirement age distribution (which has 11 observations), the median is the middle value, which is 57 years:

54, 54, 54, 55, 56, (57), 57, 58, 58, 60, 60
median

(3) mode-

The mode is the most commonly occurring value in a distribution.(most freqvant value)

(Consider this dataset showing the retirement age of 11 people, in whole years:) Example
54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60 mode = 54 is your mode

This table shows a simple frequency distribution of the retirement age data.

.(Measure of Variability) –

Measure of Variability is also known as measure of dispersion and used to describe variability in a sample or population.
In statistics there are three common measures of variability as shown below:

(i) Range :

It is given measure of how to spread apart values in sample set or data set.
Range = Maximum value - Minimum value

(ii) Variance :

It simply describes how much a random variable defers from expected value and it is also computed as square of deviation.

$$S^2 = \sum_{i=1}^{n} [(x_i - x)^2 \div n]$$

In these formula, n represent total data points, x represent mean of data points and x_i represent individual data points.

(iii) Dispersion :

It is measure of dispersion of set of data from its mean.

σ= √ (1÷n) ∑ni=1 (xi -)2

What’s a Z-Score?

Z-score is also known as standard score gives us an idea of how far a data point is from the mean.

It indicates how many standard deviations an element is from the mean. Hence, Z-Score is measured

in terms of standard deviation from the mean. For example, a standard deviation of 2 indicates the

value is 2 standard deviations away from the mean. In order to use a z-score, we need to know the

population mean ( ) and also the population standard deviation (σ).

Formula for Z-Score

A z-score can be calculated using the following formula.

z = (X –) / σ

z = Z-Score,
X = The value of the element,
  = The population mean, and
σ = The population standard deviation

Example 1:

Question:
You take the GATE examination and score 500. The mean score for the GATE is 390 and the standard

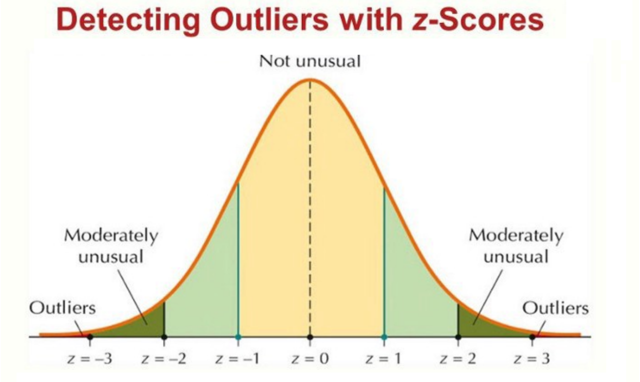
deviation is 45. How well did you score on the test compared to the average test taker?

Solution:
The following data is readily available in the above question statement
Raw score/observed value = X = 500
Mean score =   = 390
Standard deviation = σ = 45

By applying the formula of z-score,

z = (X –) / σ
z = (500 – 390) / 45
z = 110 / 45 = 2.44

This means that your z-score is 2.44.



(Since the Z-Score is positive 2.44, we will make use of the positive Z-Table.)

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89
32	15.13	16.36	18.29	20.07	22.27	42.58	46.19	49.48	53.49
34	16.50	17.79	19.81	21.66	23.95	44.90	48.60	51.97	56.06
38	19.29	20.69	22.88	24.88	27.34	49.51	53.38	56.90	61.16
42	22.14	23.65	26.00	28.14	30.77	54.09	58.12	61.78	66.21
46	25.04	26.66	29.16	31.44	34.22	58.64	62.83	66.62	71.20
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15
55	31.73	33.57	36.40	38.96	42.06	68.80	73.31	77.38	82.29
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38
65	39.38	41.44	44.60	47.45	50.88	79.97	84.82	89.18	94.42
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43
75	47.21	49.48	52.94	56.05	59.79	91.06	96.22	100.84	106.39
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33
85	55.17	57.63	61.39	64.75	68.78	102.08	107.52	112.39	118.24
90	59.20	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12
95	63.25	65.90	69.92	73.52	77.82	113.04	118.75	123.86	129.97
100	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81

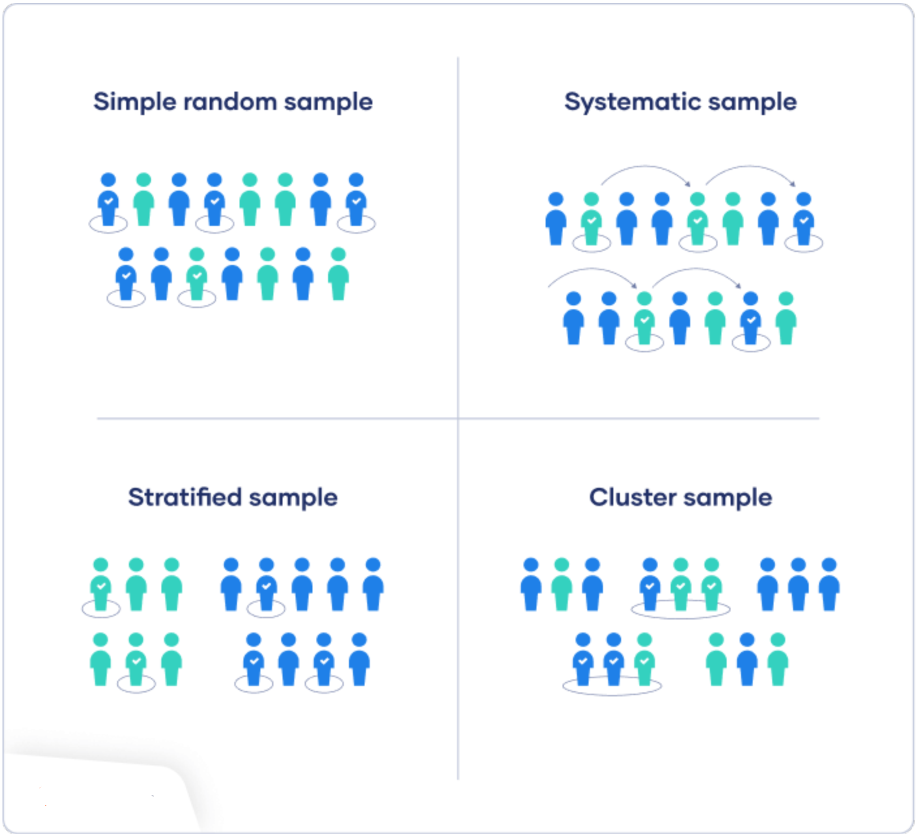
STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
-0.0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414

(Probability sampling methods)

Probability sampling means that every member of the population has a chance of being selected. It is mainly used in quantitative research .If you want to produce results that are representative of the whole population, probability sampling techniques are the most valid choice.

There are four main types of probability sample.



1. Simple random sampling
In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include
the whole population.

To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.

Example: Simple random sampling
You want to select a simple random sample of 1000 employees of a social media marketing company. You assign a number to every employee
in the company database from 1 to 1000, and use a random number generator to select 100 numbers.
2. Systematic sampling
Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population
is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

Example: Systematic sampling
All employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number
6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.

If you use this technique, it is important to make sure that there is no hidden pattern in the list that might skew the sample. For example, if the HR database groups employees by team, and team members are listed in order of seniority, there is a risk that your interval might skip over people in junior roles, resulting in a sample that is skewed towards senior employees.

3. Stratified sampling

Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.

To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g., gender identity, age range, income bracket, job role).

Based on the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup.

Example: Stratified sampling

The company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.

4. Cluster sampling

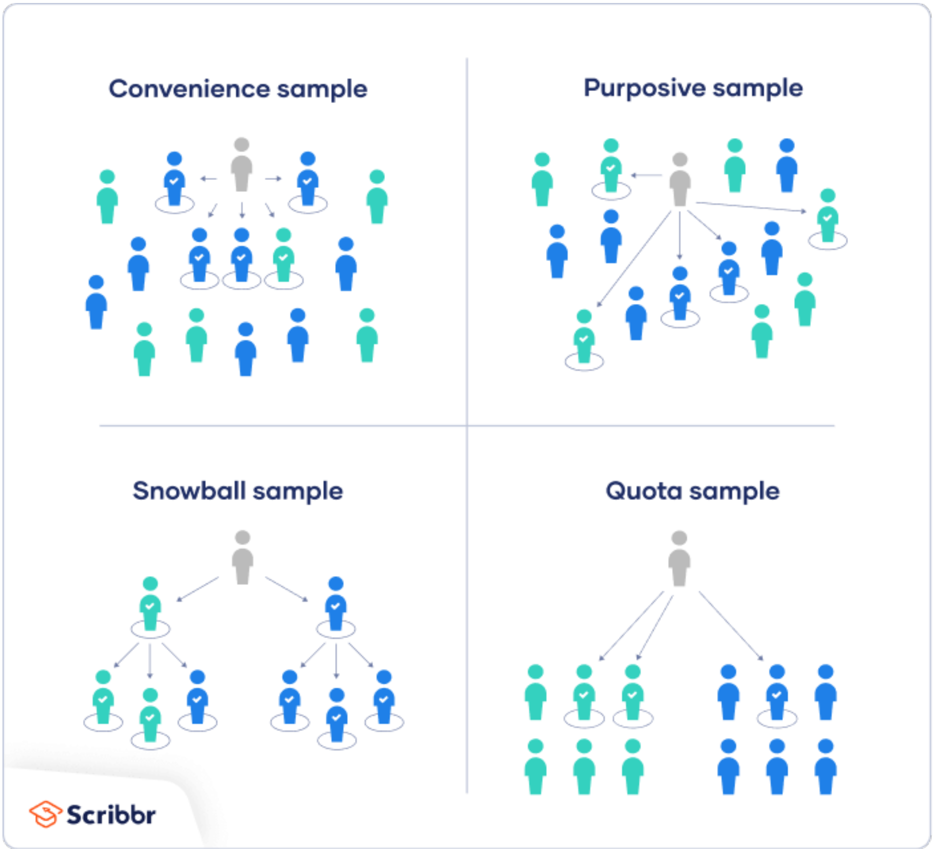
Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.

If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above. This is called multistage sampling.

This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It’s difficult to guarantee that the sampled clusters are really representative of the whole population.

Example: Cluster sampling

The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don’t have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.



Non-probability sampling methods

In a non-probability sample, individuals are selected based on non-random criteria, and not every individual has a chance of being included.

This type of sample is easier and cheaper to access, but it has a higher risk of sampling bias. That means the inferences you can make about the population are weaker than with probability samples, and your conclusions may be more limited. If you use a non-probability sample, you should still aim to

make it as representative of the population as possible.

Non-probability sampling techniques are often used in exploratory and qualitative research. In these types of research, the aim is not to test a hypothesis about a broad population, but to develop an initial understanding of a small or under-researched population.

1. Convenience sampling

A convenience sample simply includes the individuals who happen to be most accessible to the researcher.

This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population, so it can’t produce generalizable results. Convenience samples are at risk for both sampling bias and selection bias.

Example: Convenience sampling

You are researching opinions about student support services in your university, so after each of your classes, you ask your fellow students to complete a survey on the topic. This is a convenient way to gather data, but as you only surveyed students taking the same classes as you at the same level, the sample is not representative of all the students at your university.

2. Voluntary response sampling

Similar to a convenience sample, a voluntary response sample is mainly based on ease of access. Instead of the researcher choosing participants and directly contacting them, people volunteer themselves (e.g. by responding to a public online survey).

Voluntary response samples are always at least somewhat biased, as some people will inherently be more likely to volunteer than others, leading to self-selection bias.

Example: Voluntary response sampling

You send out the survey to all students at your university and a lot of students decide to complete it. This can certainly give you some insight into the topic, but the people who responded are more likely to be those who have strong opinions about the student support services, so you can’t be sure that their opinions are representative of all students.

3. Purposive sampling

This type of sampling, also known as judgement sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research.

It is often used in qualitative research, where the researcher wants to gain detailed knowledge about a specific phenomenon rather than make statistical inferences, or where the population is very small and specific. An effective purposive sample must have clear criteria and rationale for inclusion. Always make sure to describe your inclusion and exclusion criteria and beware of observer bias affecting your arguments.

Example: Purposive sampling

You want to know more about the opinions and experiences of disabled students at your university, so you purposefully select a number of students with different support needs in order to gather a varied range of data on their experiences with student se rvices.

4. Snowball sampling

If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people you have access to “snowballs” as you get in contact with more people. The downside here is also representativeness, as you have no way of knowing how representative your sample is due to the reliance on participants recruiting others. This can lead to sampling bias.

Example: Snowball sampling

You are researching experiences of homelessness in your city. Since there is no list of all homeless people in the city, probability sampling isn’t possible. You meet one person who agrees to participate in the research, and she puts you in contact with other homeless people that she knows in the area.

(Frequently asked questions about sampling)

(1)What is sampling?

A sample is a subset of individuals from a larger population. Sampling means selecting the group that you will actually collect data from in your research. For example, if you are researching the opinions of students in your university, you could survey a sample of 100 students.

In statistics, sampling allows you to test a hypothesis about the characteristics of a population.

(2)Why are samples used in research?

Samples are used to make inferences about populations. Samples are easier to collect data from because they are practical, cost-effective, convenient, and manageable.

(3) What is probability sampling?

Probability sampling means that every member of the target population has a known chance of being included in the sample.
Probability sampling methods include simple random sampling, systematic sampling, stratified sampling, and cluster sampling.

(4) What is non-probability sampling?

In non-probability sampling, the sample is selected based on non-random criteria, and not every member of the population has a chance of being included.Common non-probability sampling methods include convenience sampling, voluntary response sampling, purposive sampling, snowball sampling, and quota sampling.

(5) What is multistage sampling?

In multistage sampling, or multistage cluster sampling, you draw a sample from a population using smaller and smaller groups at each stage.
This method is often used to collect data from a large, geographically spread group of people in national surveys, for example. You take advantage of hierarchical groupings (e.g., from state to city to neighborhood) to create a sample that’s less expensive and time-consuming

to collect data from.

(6)What is sampling bias?
Sampling bias occurs when some members of a population are systematically more likely to be selected in a sample than others

1 BACKGROUND

Definitions and Terms

Null Hypothesis (H_0): A statement of no change and is 0 assumed true until evidence indicates otherwise

Alternate Hypothesis (H_a): A statement that the researcher is trying to find evidence to support

Type I Error: Reject the null hypothesis when the null hypothesis is true

Type II Error: Do not reject the null hypothesis when the alternative hypothesis is true

Test Statistics (*t*): A single number that summarizes the sample data used to conduct the test hypothesis

Standard Error: How far sample statistics (e.g., mean) deviates from the actual population mean

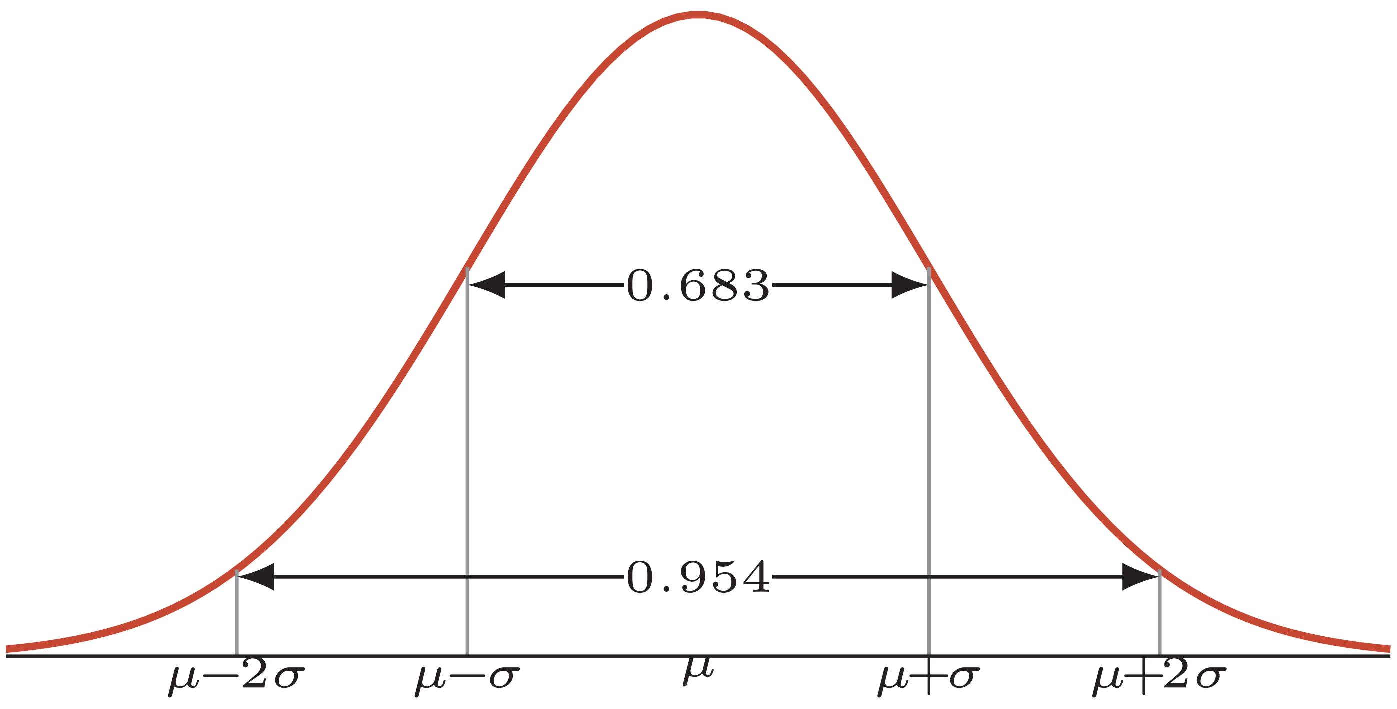
***p*-value:** Probability of observing a test statistics

Significance level (α): Probability of making Type I error

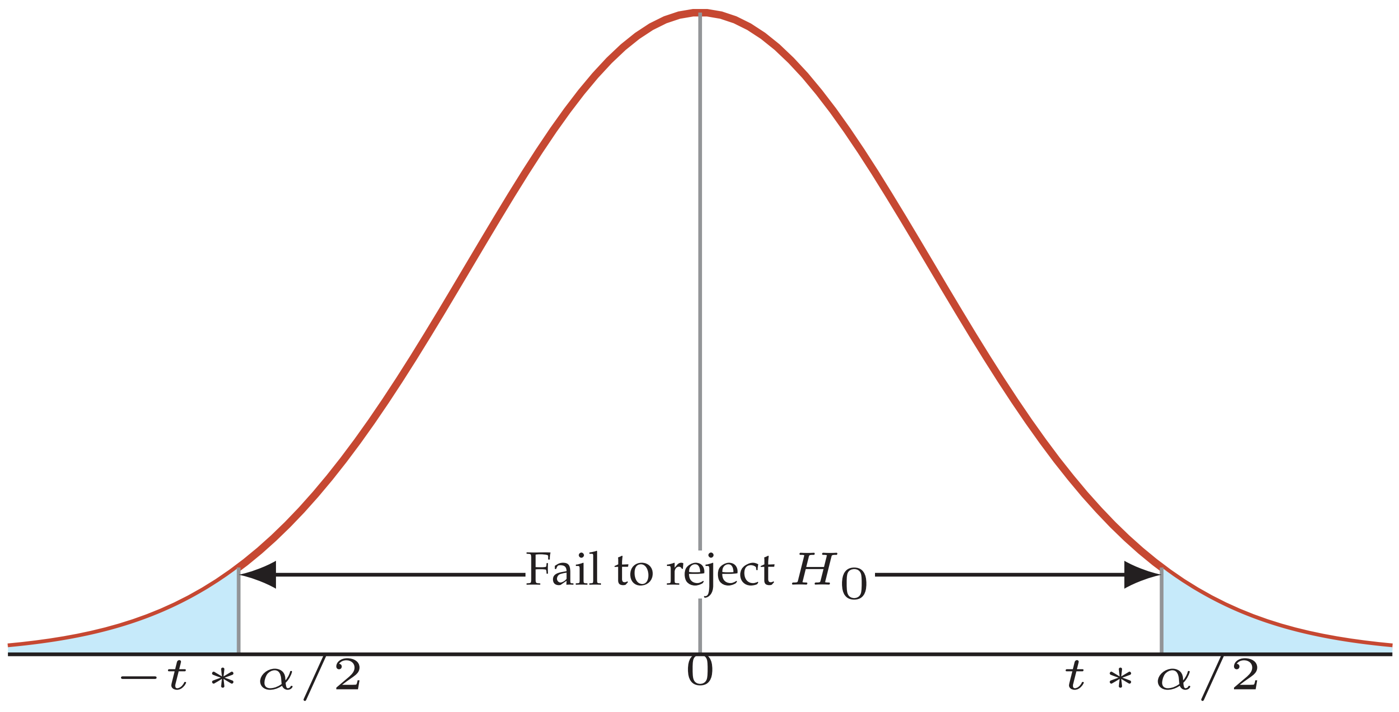
One tailed test: Test statistics falls into one specified tail of its sampling distribution

Two tailed test: Test statistics can falling into either tail of its sampling distribution

Normal curve:



Acceptance/Rejection regions:
Two tailed test



2 Hypothesis Testing

Steps to Significance Testing

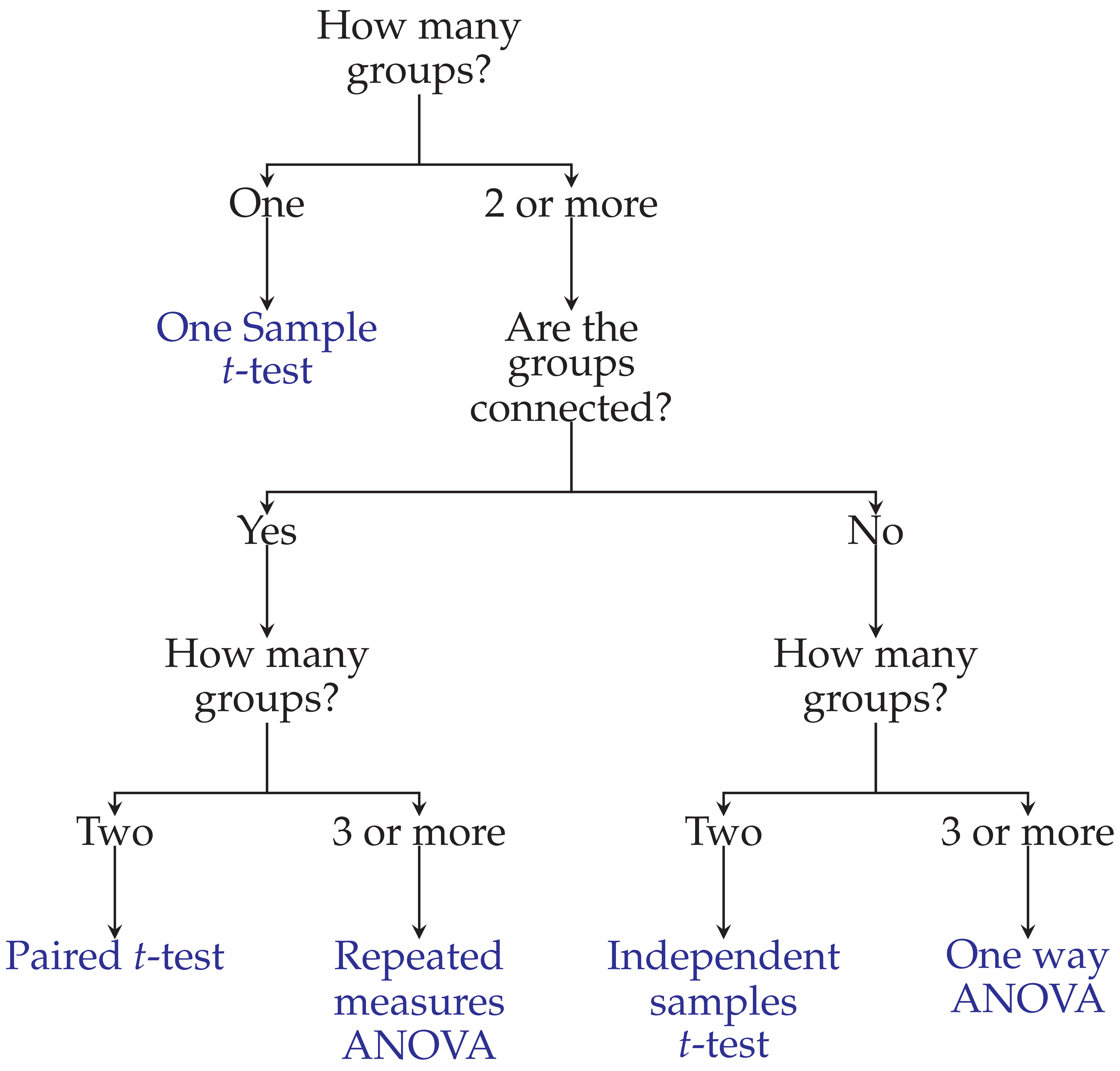
1. Define H_0 and H_a
2. Identify test, α, find critical value, test statistics
3. Construct acceptance/rejection regions
4. Calculate test statistics

   Critical value approach: *Determine critical region*

   p-value approach: *Calculate p-value*
5. Retain or reject the hypothesis

3 Choosing a Statistical Test

Decision Tree



Decision Tree - by data structure

Categorical Data: Use Chi Square

Sample size (n):

n < 30 and Population Variance is unknown - *t-test*

n < 30 and Population Variance is known - *z-test*

n > 30 - *z-test* or *t-test*

4 Examples

Chi Square test for independence:

Checks whether two categorical variables are related or not (independence)

E.g., Is the distribution of sex and voting behavior due to chance or is there a difference between sexes on voting behavior?

T-Test:

Looks at the difference between two groups (e.g., undergrad/grad)

E.g., Do undergrad and grad students differ in the amount of hours they spend studying in a given month?

ANOVA (Analysis of Variance):

Tests the significance of group differences between two or more groups

Only determines that there is a difference between groups, but does not tell which is different

E.g., Do GRE scores differ for low-, middle, and high-income students?

ANCOVA (Analysis of Covariance):

Same as ANOVA, but adds control of one or more covariates that my influence dependent variable

E.g., Do SAT scores differ for low-, middle-, and high-income students after controlling for single/dual parenting?

1. (Statistics)

Statistics is the science concerned with developing and studying methods for collecting, analyzing, interpreting and presenting empirical data.

Basically, there are **four** types of statistics.

(1). Mathmatics for mashine learning

(2).Descriptive statistics

(3).Inferential statistics

(4).Probability

(1). Descriptive Statistics-

Descriptive statistics is a way of organising, representing, and explaining a set of data using charts, graphs, and summary measures. Histograms, pie charts, bars, and scatter plots are common

ways to summarise data and present it in tables or graphs.

(2). Inferential Statistics-

inferential statistics to convey the meaning of the collected data after it has been collected, evaluated, and summarised. The probability principle is used in inferential statistics to determine if patterns found in a study sample may be extrapolated to the wider population from which the sample was drawn. Inferential statistics are used to test hypotheses and study correlations between variables, and they can also be used to predict population sizes. Inferential statistics are used to derive conclusions and inferences from samples, i.e. to create accurate generalisations.

(1). [Descriptive Statistics]

What is Data in Statistics?

Data is a collection of facts, such as numbers, words, measurements, observations etc.

(Types of Data)

(1).Qualitative data- it is descriptive data.
Example- She can run fast, He is thin.

(2).Quantitative data- it is numerical information.
Example- An Octopus is an Eight legged creature.

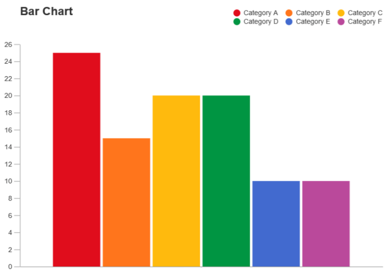
Types of quantitative data-

(1).Discrete data- has a particular fixed value. It can be counted

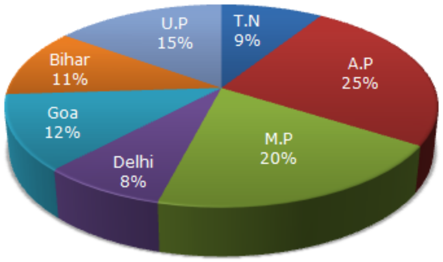
(2).Continuous data- is not fixed but has a range of data. It can be measured.Representation of Data

(Representation of Data)
There are different ways to represent data such as through graphs, charts or tables.
The general representation of statistical data are:

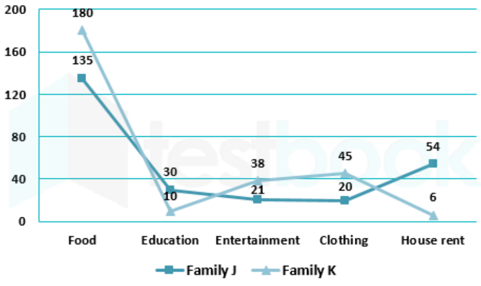
(1) Bar Graph



(2) Pie Chart

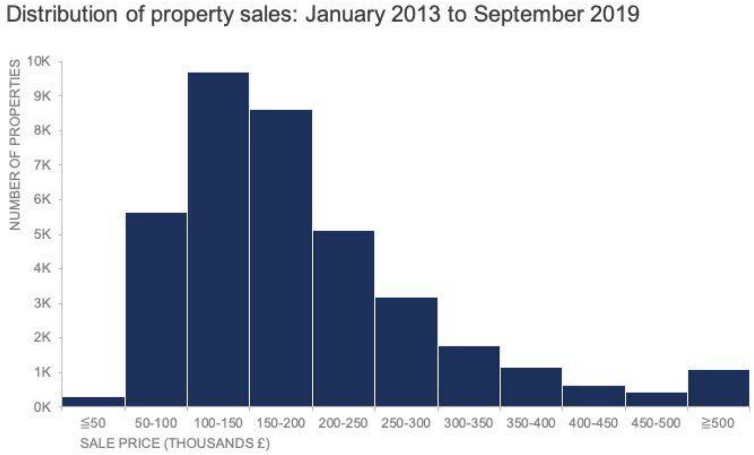


(3) Line Graph



(4) Pictograph

(5) Histogram



(6) box plot

(7) KDE (kernal dencity estimation)

(Measures of Central Tendency)

In Mathematics, statistics are used to describe the central tendencies of the grouped and ungrouped data.

The three measures of central tendency are:

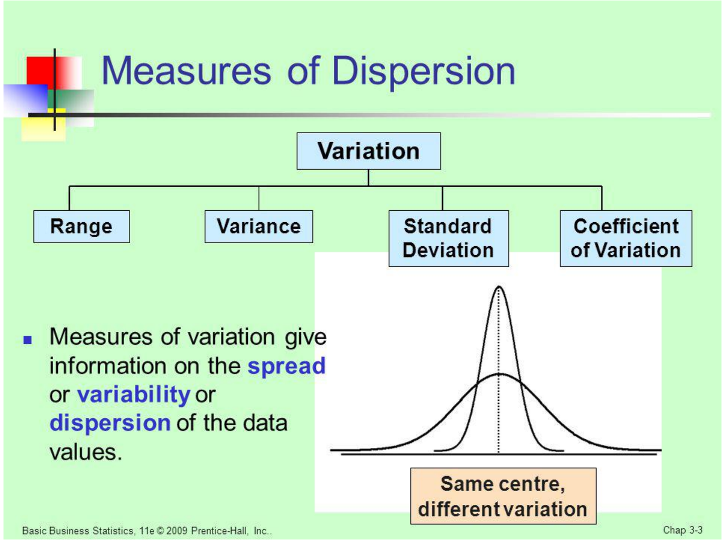
(1). Mean (denote by “mu”) avarage
(2). Median (even and odd)
(3). Mode (most frequent)

(All three measures of central tendency are used to find the central value of the set of data.)

(Measures of Dispersion)-

In statistics, the dispersion measures help interpret data variability, i.e. to understand how homogenous or heterogeneous the data is.In simple words, it indicates how squeezed or scattered the variable is. However, there are two types of dispersion measures, absolute and relative. They are tabulated as below:

(Absolute measures of dispersion)



Measures of Dispersion

Variation

Range

Variance

Standard Deviation

Coefficient of Variation

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

Same centre, different variation

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc..

Chap 3-3

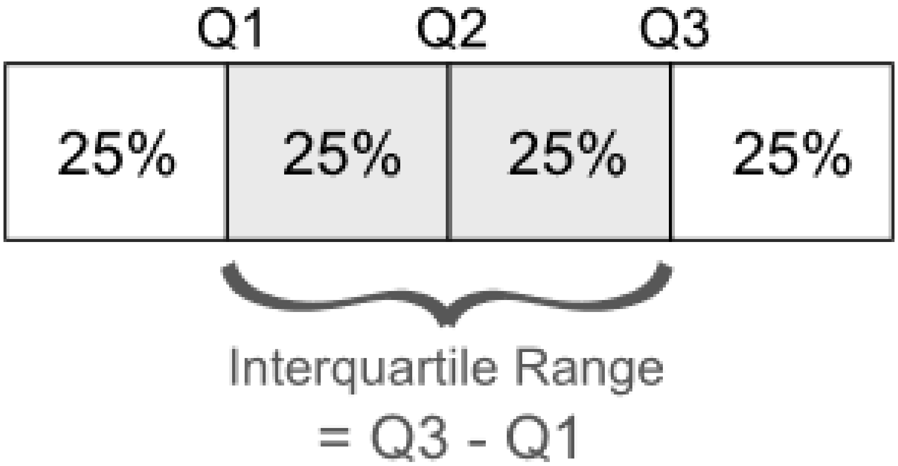
(1).Variance-

Variance is the measure of how notably a collection of data is spread out. If all the data values are identical, then it indicates the variance is zero.There can be two types of variances in statistics, namely, sample variance and population variance. The symbol of variance is given by σ2. Variance is widely used in hypothesis testing,

(2) Standard deviation- denote by σ “sigma” or σx formola is underroot varriance

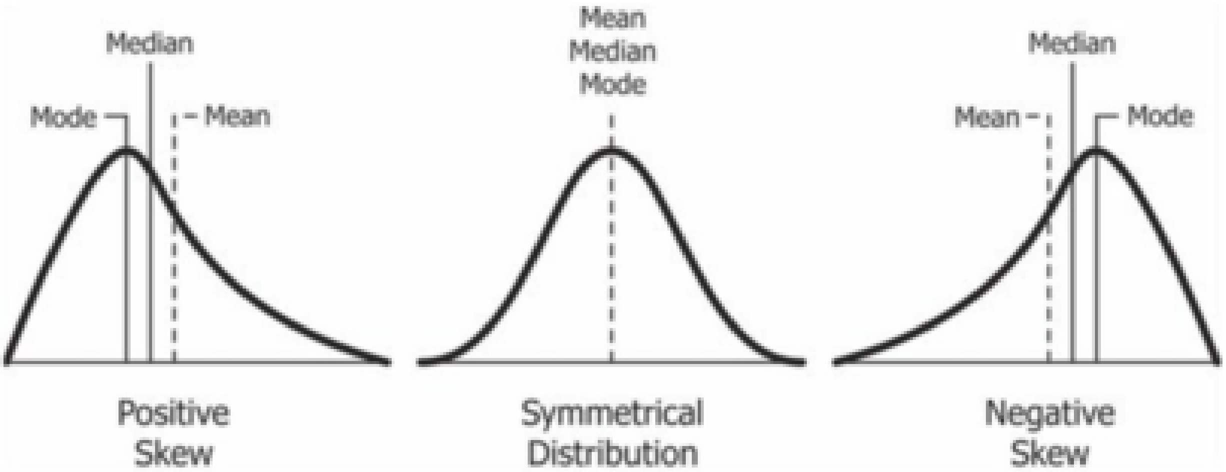
Standard Deviation is a measure which shows how much variation (such as spread, dispersion, spread,) from the mean exists.

Quartiles and Quartile



(1) Q1 - 25%
(2) Q2 - 50%
(3) Q3 - 75%
(4) IQR (INTER QUATILE RANGE)
The interquartile range tells you the spread of the middle half of your distribution.
IQR=Q3-Q1

(Skewness in Statistics)-



Skewness, in statistics, is a measure of the asymmetry in a probability distribution.
It measures the deviation of the curve of the normal distribution for a given set of data.

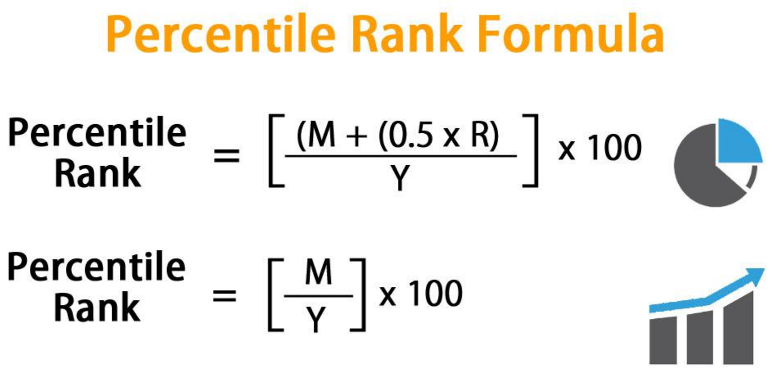
(Percentage and percentile)

percentage and percentile-

(1)- percentage-

The percentage is a mathematical value presented out of 100 and percentile is the per cent of values below a specific value.

Percentage = (Numerator Denominator) × 100 or (X Y) × 10



Percentile Rank Formula

$$\text{Percentile Rank} = \left[\frac{(M + (0.5 \text{ x } R)}{Y}\right] \text{ x } 100$$

$$\text{Percentile Rank} = \left[\frac{M}{Y}\right] \text{ x } 100$$

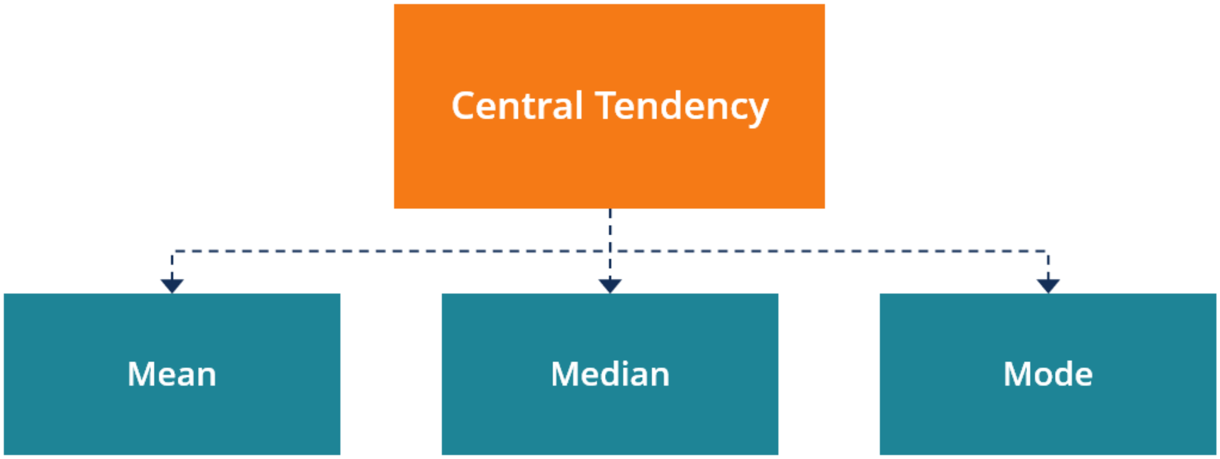
(2)-Percentile-

A percentile is a comparison score between a particular score and the scores of the rest of a group.

It shows the percentage of scores that a particular score surpassed.

(P) percentile = (nth percentile/100) × Total number of values in the list .

(Measures of central tendency)



A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.

There are three main measures of central tendency:

(1) mean

The mean is the sum of the value of each observation in a dataset divided by the number of observations.

This is also known as the arithmetic average.

Looking at the retirement age distribution again:
54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

The mean is calculated by adding together all the values (54+54+54+55+56+57+57+58+58+60+60 = 623) and dividing by the number of observations (11) which equals 56.6 years.

(2) median

The median is the middle value in distribution when the values are arranged in ascending or descending order.

The median divides the distribution in half (there are 50% of observations on either side of the median value).

Looking at the retirement age distribution (which has 11 observations), the median is the middle value, which is 57 years:

54, 54, 54, 55, 56, (57), 57, 58, 58, 60, 60
median

(3) mode-

The mode is the most commonly occurring value in a distribution.(most freqvant value)

(Consider this dataset showing the retirement age of 11 people, in whole years:) Example
54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60 mode = 54 is your mode

This table shows a simple frequency distribution of the retirement age data.

.(Measure of Variability) –

Measure of Variability is also known as measure of dispersion and used to describe variability in a sample or population.

In statistics there are three common measures of variability as shown below:

(i) Range :

It is given measure of how to spread apart values in sample set or data set.
Range = Maximum value - Minimum value

(ii) Variance :

It simply describes how much a random variable defers from expected value and it is also computed as square of deviation.

$$S^2= \sum_{i=1}^{n} [(x_i - x)^2 \div n]$$

In these formula, n represent total data points, x represent mean of data points and x_i represent individual data points.

(iii) Dispersion :

It is measure of dispersion of set of data from its mean.

$$\sigma= \surd\ (1\div n) \sum ni=1\ (xi -\ \ )2$$

What’s a Z-Score?

Z-score is also known as standard score gives us an idea of how far a data point is from the mean. It indicates how many standard deviations an element is from the mean. Hence, Z-Score is measured in terms of standard deviation from the mean. For example, a standard deviation of 2 indicates the value is 2 standard deviations away from the mean. In order to use a z-score, we need to know the population mean ( ) and also the population standard deviation (σ).

Formula for Z-Score

A z-score can be calculated using the following formula.

$$z = (X - \ \ ) / \sigma$$

z = Z-Score,
X = The value of the element,
  = The population mean, and
σ = The population standard deviation

Example 1:

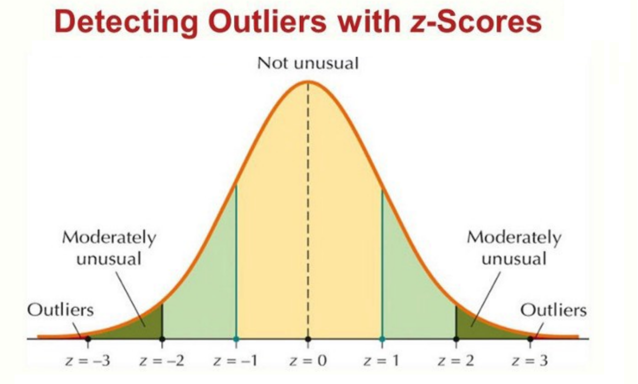
Question:
You take the GATE examination and score 500. The mean score for the GATE is 390 and the standard deviation is 45. How well did you score on the test compared to the average test taker?

Solution:
The following data is readily available in the above question statement
Raw score/observed value = X = 500
Mean score =   = 390
Standard deviation = σ = 45

By applying the formula of z-score,

$z = (X - \ \ ) / \sigma$
$z = (500 - 390) / 45$
$z = 110 / 45 = 2.44$

This means that your z-score is 2.44.



(Since the Z-Score is positive 2.44, we will make use of the positive Z-Table.)

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89
32	15.13	16.36	18.29	20.07	22.27	42.58	46.19	49.48	53.49
34	16.50	17.79	19.81	21.66	23.95	44.90	48.60	51.97	56.06
38	19.29	20.69	22.88	24.88	27.34	49.51	53.38	56.90	61.16
42	22.14	23.65	26.00	28.14	30.77	54.09	58.12	61.78	66.21
46	25.04	26.66	29.16	31.44	34.22	58.64	62.83	66.62	71.20
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15
55	31.73	33.57	36.40	38.96	42.06	68.80	73.31	77.38	82.29
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38
65	39.38	41.44	44.60	47.45	50.88	79.97	84.82	89.18	94.42
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43
75	47.21	49.48	52.94	56.05	59.79	91.06	96.22	100.84	106.39
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33
85	55.17	57.63	61.39	64.75	68.78	102.08	107.52	112.39	118.24
90	59.20	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12
95	63.25	65.90	69.92	73.52	77.82	113.04	118.75	123.86	129.97
100	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81

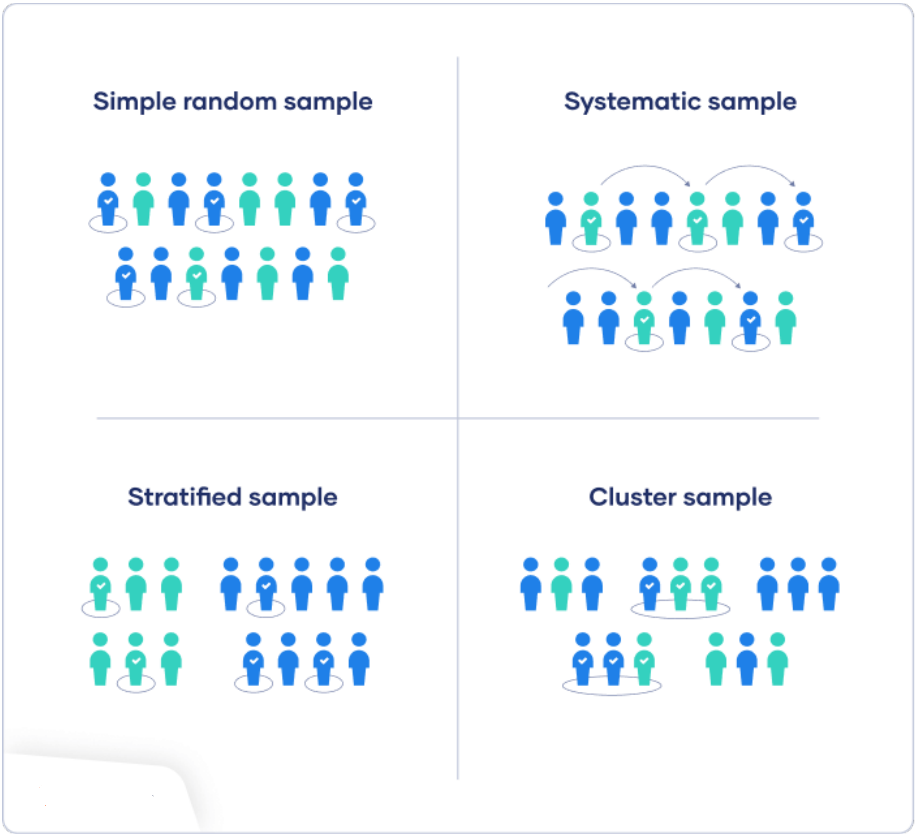
STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
-0.0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414

(Probability sampling methods)

Probability sampling means that every member of the population has a chance of being selected. It is mainly used in quantitative research .If you want to produce results that are representative of the whole population, probability sampling techniques are the most valid choice.

There are four main types of probability sample.



1. Simple random sampling
In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include
the whole population.

To conduct this type of sampling, you can use tools like random number generators or other techniques th at are based entirely on chance.

Example: Simple random sampling
You want to select a simple random sample of 1000 employees of a social media marketing company. Yo u assign a number to every employee
in the company database from 1 to 1000, and use a random number generator to select 100 numbers.
2. Systematic sampling
Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Ever y member of the population
is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular int ervals.

Example: Systematic sampling
All employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly s elect a starting point: number
6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you e nd up with a sample of 100 people.

If you use this technique, it is important to make sure that there is no hidden pattern in the list that might skew the sample. For example, if the HR database groups employees by team, and team members are listed in order of seniority, there is a risk that your interval might skip over people in junior roles, resulting in a sample that is skewed towards senior employees.

3. Stratified sampling

Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.

To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g., gender identity, age range, income bracket, job role).

Based on the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup.

Example: Stratified sampling

The company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.

4. Cluster sampling

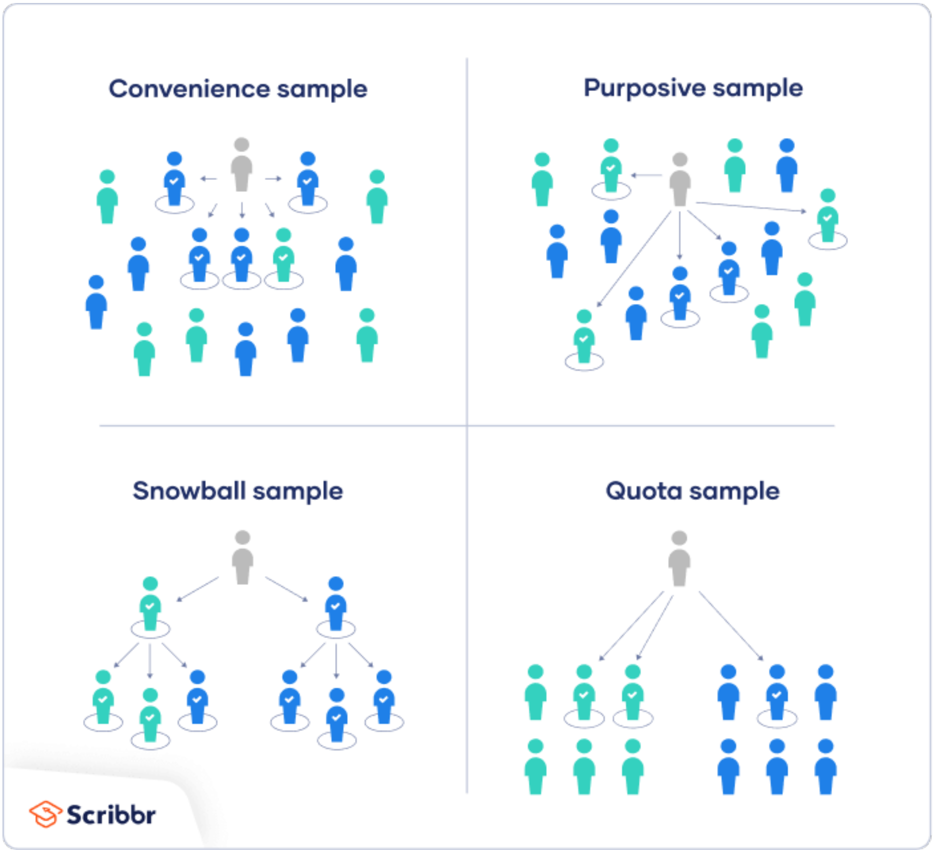
Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.

If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above. This is called multistage sampling.

This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It’s difficult to guarantee that the sampled clusters are really representative of the whole population.

Example: Cluster sampling

The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don’t have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.



Non-probability sampling methods
In a non-probability sample, individuals are selected based on non-random criteria, and not every individual has a chance of being included.

This type of sample is easier and cheaper to access, but it has a higher risk of sampling bias. That means the inferences you can make about
the population are weaker than with probability samples, and your conclusions may be more limited. If you use a non-probability sample, you
should still aim to

make it as representative of the population as possible.
Non-probability sampling techniques are often used in exploratory and qualitative research. In these types of research, the aim is not to
test a hypothesis about a broad population, but to develop an initial understanding of a small or under-researched population.

1. Convenience sampling

A convenience sample simply includes the individuals who happen to be most accessible to the researcher.

This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population, so it can’t produce generalizable results. Convenience samples are at risk for both sampling bias and selection bias.

Example: Convenience sampling

You are researching opinions about student support services in your university, so after each of your classes, you ask your fellow students to complete a survey on the topic. This is a convenient way to gather data, but as you only surveyed students taking the same classes as you at the same level, the sample is not representative of all the students at your university.

2. Voluntary response sampling

Similar to a convenience sample, a voluntary response sample is mainly based on ease of access. Instead of the researcher choosing participants and directly contacting them, people volunteer themselves (e.g. by responding to a public online survey).

Voluntary response samples are always at least somewhat biased, as some people will inherently be more likely to volunteer than others, leading to self-selection bias.

Example: Voluntary response sampling

You send out the survey to all students at your university and a lot of students decide to complete it. This can certainly give you some insight into the topic, but the people who responded are more likely to be those who have strong opinions about the student support services, so you can’t be sure that their opinions are representative of all students.

3. Purposive sampling

This type of sampling, also known as judgement sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research.

It is often used in qualitative research, where the researcher wants to gain detailed knowledge about a specific phenomenon rather than make statistical inferences, or where the population is very small and specific. An effective purposive sample must have clear criteria and rationale for inclusion. Always make sure to describe your inclusion and exclusion criteria and beware of observer bias affecting your arguments.

Example: Purposive sampling

You want to know more about the opinions and experiences of disabled students at your university, so you purposefully select a number of students with different support needs in order to gather a varied range of data on their experiences with student se rvices.

4. Snowball sampling
If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people you have access to “snowballs” as you get in contact with more people. The downside here is also representativeness, as you have no way of knowing how representative your sample is due to the reliance on participants recruiting others. This can lead to sampling bias.

Example: Snowball sampling
You are researching experiences of homelessness in your city. Since there is no list of all homeless people in the city, probability sampling isn’t possible. You meet one person who agrees to participate in the research, and she puts you in contact with other homeless people that she knows in the area.

(Frequently asked questions about sampling)

(1)What is sampling?
A sample is a subset of individuals from a larger population. Sampling means selecting the group that you will actually collect data from in your research. For example, if you are researching the opinions of students in your university, you could survey a sample of 100 students.

In statistics, sampling allows you to test a hypothesis about the characteristics of a population.

(2)Why are samples used in research?
Samples are used to make inferences about populations. Samples are easier to collect data from because they are practical, cost-effective, convenient, and manageable.

(3) What is probability sampling?

Probability sampling means that every member of the target population has a known chance of being included in the sample.
Probability sampling methods include simple random sampling, systematic sampling, stratified sampling, and cluster sampling.

(4) What is non-probability sampling?
In non-probability sampling, the sample is selected based on non-random criteria, and not every member of the population has a chance of being included.Common non-probability sampling methods include convenience sampling, voluntary response sampling, purposive sampling, snowball sampling, and quota sampling.

(5) What is multistage sampling?
In multistage sampling, or multistage cluster sampling, you draw a sample from a population using smaller and smaller groups at each stage.
This method is often used to collect data from a large, geographically spread group of people in national surveys, for example. You take advantage of hierarchical groupings (e.g., from state to city to neighborhood) to create a sample that’s less expensive and time-consuming

to collect data from.

(6)What is sampling bias?
Sampling bias occurs when some members of a population are systematically more likely to be selected i n a sample than others

1 BACKGROUND

Definitions and Terms

Null Hypothesis (H_0): A statement of no change and is 0 assumed true until evidence indicates otherwise

Alternate Hypothesis (H_a): A statement that the researcher is trying to find evidence to support

Type I Error: Reject the null hypothesis when the null hypothesis is true

Type II Error: Do not reject the null hypothesis when the alternative hypothesis is true

Test Statistics (*t*): A single number that summarizes the sample data used to conduct the test hypothesis

Standard Error: How far sample statistics (e.g., mean) deviates from the actual population mean

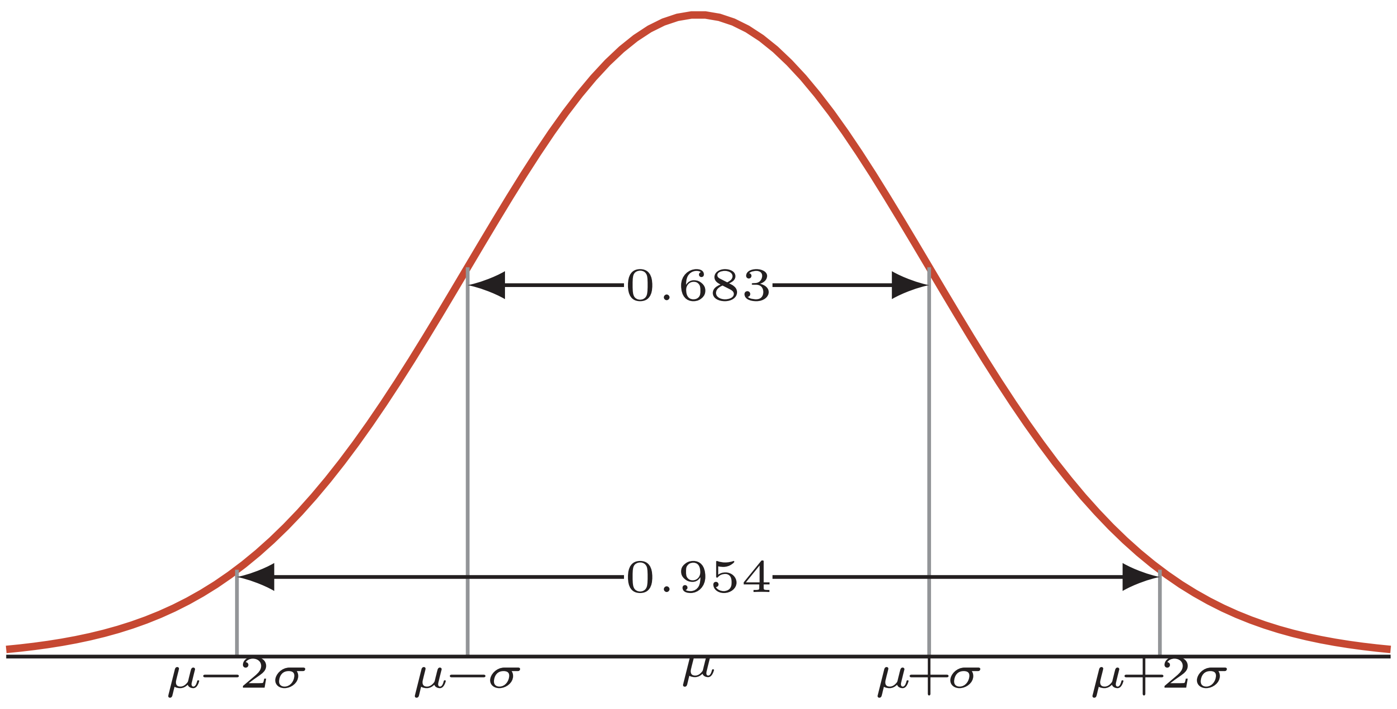
***p*-value:** Probability of observing a test statistics

Significance level (α): Probability of making Type I error

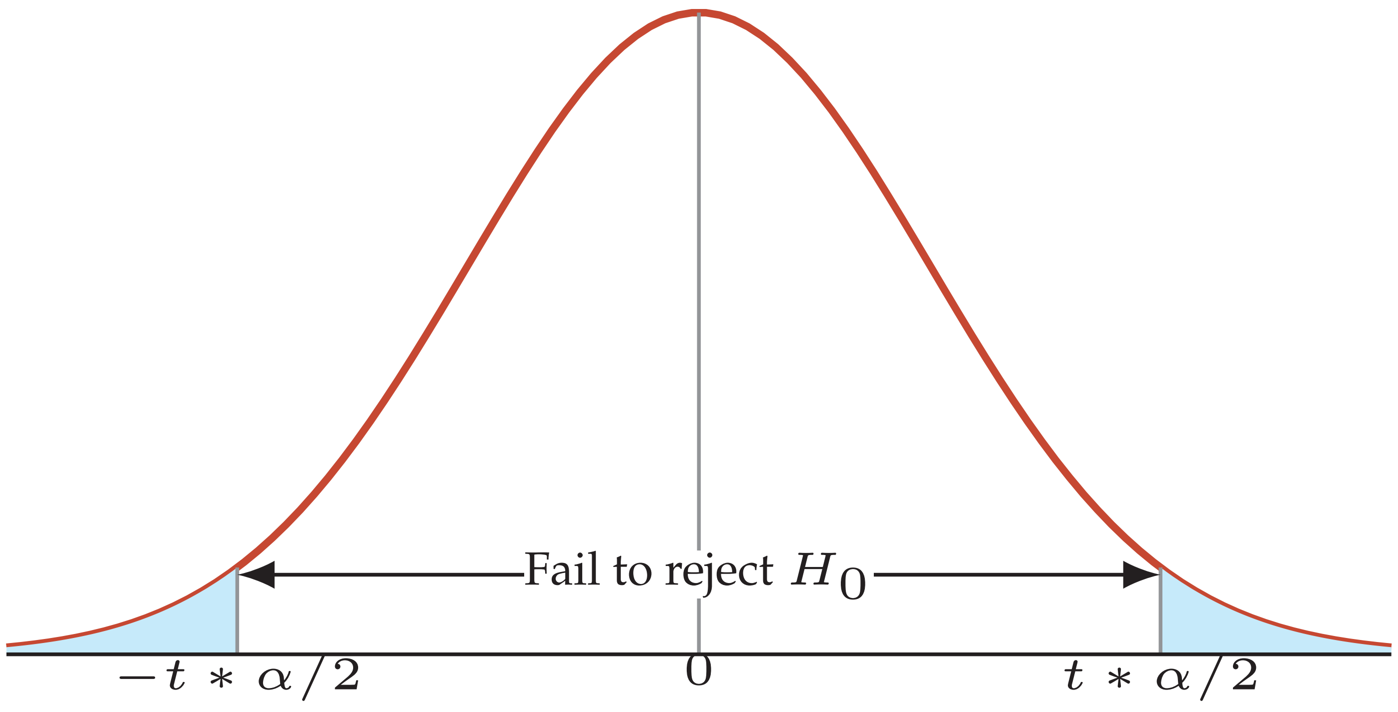
One tailed test: Test statistics falls into one specified tail of its sampling distribution

Two tailed test: Test statistics can falling into either tail of its sampling distribution

Normal curve:



Acceptance/Rejection regions:
Two tailed test



2 Hypothesis Testing

Steps to Significance Testing

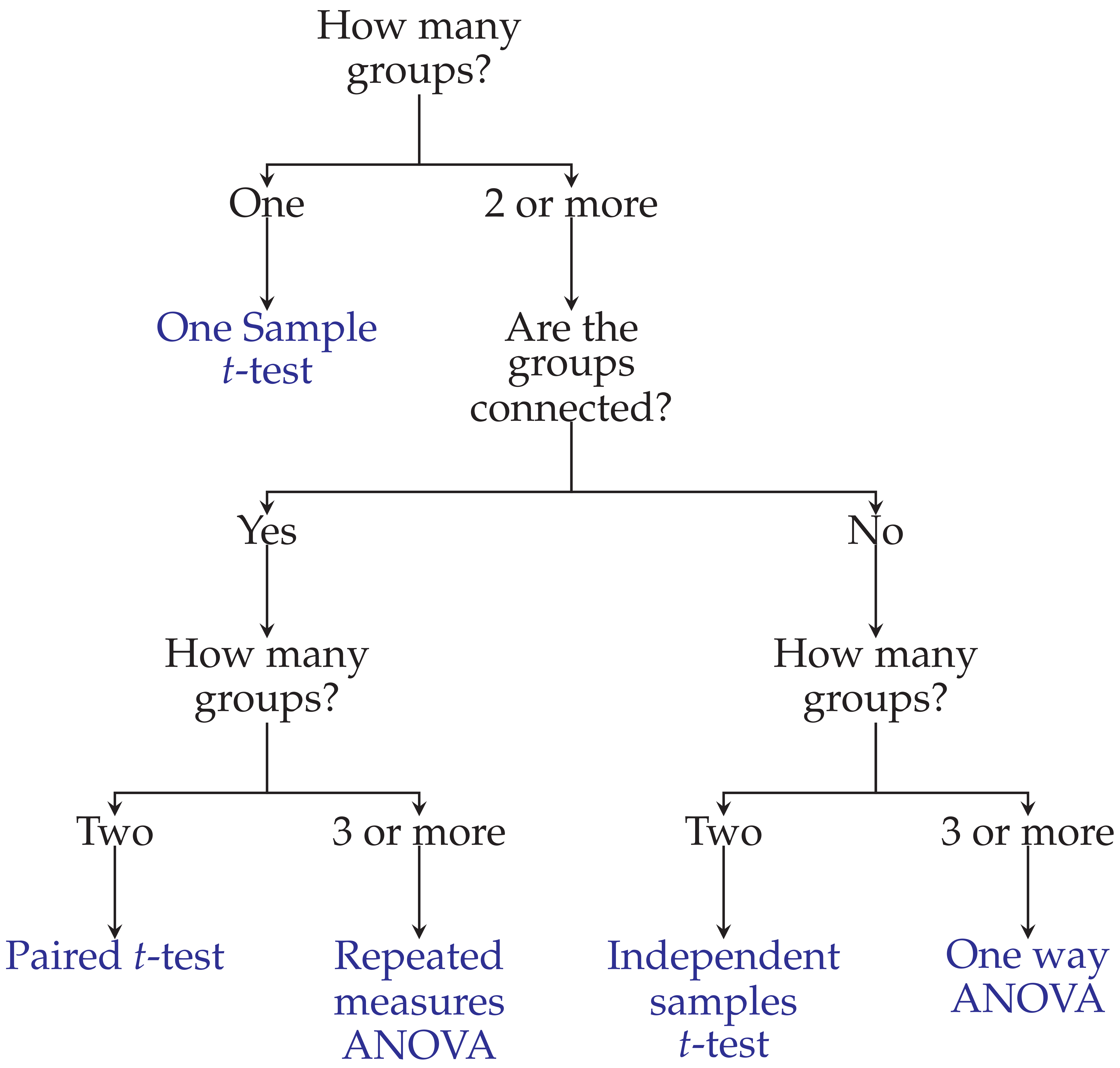
1. Define H_0 and H_a
2. Identify test, α, find critical value, test statistics
3. Construct acceptance/rejection regions
4. Calculate test statistics

   Critical value approach: *Determine critical region*

   p-value approach: *Calculate p-value*
5. Retain or reject the hypothesis

3 Choosing a Statistical Test

Decision Tree



Decision Tree - by data structure

Categorical Data: Use Chi Square

Sample size (n):

n < 30 and Population Variance is unknown - *t-test*

n < 30 and Population Variance is known - *z-test*

n > 30 - *z-test* or *t-test*

4 Examples

Chi Square test for independence:

Checks whether two categorical variables are related or not (independence)

E.g., Is the distribution of sex and voting behavior due to chance or is there a difference between sexes on voting behavior?

T-Test:

Looks at the difference between two groups (e.g., undergrad/grad)

E.g., Do undergrad and grad students differ in the amount of hours they spend studying in a given month?

ANOVA (Analysis of Variance):

Tests the significance of group differences between two or more groups

Only determines that there is a difference between groups, but does not tell which is different

E.g., Do GRE scores differ for low-, middle, and high-income students?

ANCOVA (Analysis of Covariance):

Same as ANOVA, but adds control of one or more covariates that my influence dependent variable

E.g., Do SAT scores differ for low-, middle-, and high-income students after controlling for single/dual parenting?