

# Probability and statistics

---

Pavan Sai Ram

# Probability and Statistics

**Mean:** If  $x = \{x_1, x_2, x_3, \dots, x_n\}$  are the set of all 'n' values of a variate, then the Mean is given by

$$\mu = \bar{x} = \frac{\sum x}{n} \text{ and } \bar{x} = \frac{\sum fx}{\sum f}$$

**Median:** The median is that value of the variable which divides the group into two equal parts, one part comprising all values greater and the other, all values less than median.

- If n is odd, the midpoint of n dataset is  $\frac{n+1}{2}$ th position in the dataset.
- If n is even, the midpoint is the sum of the two middle point divided by 2.

$$\text{Median} = L + \left( \frac{\frac{N}{2} - C}{f} \right) h$$

Where, L = lower limit of the median class

N = total frequency

f = frequency of the median class

C = cumulative frequency up to the class preceding the median

h = width of the median class

**Mode:** The mode is defined as that value of the variable which occurs most frequently, i.e. the value of the maximum frequency.

For a grouped distribution,

$$\text{Mode} = L + \left( \frac{\Delta_1}{\Delta_1 + \Delta_2} \right) h = L + \left( \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) h$$

Where, L = lower limit of the class containing the mode.

$$\Delta_1 = f_1 - f_0, \quad \Delta_2 = f_1 - f_2$$

$f_1$  = frequency of the modal class

$f_0$  = frequency of the class preceding to modal class

$f_2$  = frequency of the class succeeding to modal class

**Standard Deviation:** It is defined as the positive square-root of the arithmetic mean of the square of the deviations of the given deviation from their arithmetic mean.

The standard deviation is denoted by the Greek letter  $\sigma$ .

$$\sigma = \sqrt{\frac{\sum x^2}{n}} = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

## Discrete Probability Distribution

- Mean,  $\mu = \sum_i x_i P(x_i)$
- Variance,  $V = \sigma^2 = \sum_i (x_i - \mu)^2 P(x_i)$

### Binomial Distribution:

Definition: If  $p$  is the probability of success and  $q$  is the probability of failure, the probability of  $x$  success out of  $n$  trials is given by,

$$P(x) = {}^nC_x p^x q^{n-x} \text{ or } {}^nC_x p^x (1-p)^{n-x} \quad \text{where, } {}^nC_x = \frac{n!}{x!(n-x)!}$$

- $n$  = Number of trials or observations.
- Mean,  $\mu = np$
- Variance,  $V = \sigma^2 = npq$
- Standard deviation,  $\sigma = \sqrt{npq}$

### Poisson Distribution:

- Binomial distribution where  $n$  is very large or  $p$  is very small.

Definition: A random variable  $x$  is said to have a Poisson distribution with parameter  $m$  if its density  $f$  is given by,

$$f(x) = \frac{e^{-m} m^x}{x!}, \quad x = 0, 1, 2, \dots, m > 0 \text{ and } m = np$$

- Mean,  $\mu = m$
- Variance,  $V = m$
- Standard deviation,  $\sigma = \sqrt{m}$

## Continuous Probability Distribution

### Continuous Random Variable:

A random variable  $x$  is continuous if it can assume any value in some interval or intervals of real numbers and the probability that it assumes any specific value is 0.

### Continuous Density Function:

Let  $X$  be a continuous random variable. A function  $f$  such that,

- $f(x) \geq 0$  for  $x$  real.
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $p[a \leq x \leq b] = \int_a^b f(x)dx$  for  $a$  and  $b$  real.

is called Density Function for  $x$ .

### Cumulative Distribution:

Let  $X$  be continuous with density function  $f$ . The cumulative distribution function for  $X$  denoted by  $F$  is defined as,

$$F(x) = P[X \leq x], x \text{ real}$$

$$\text{Computation: } F(x) = P[X \leq x] = \int_{-\infty}^x f(x)dx$$

### Exponential Distribution:

The continuous probability distributions having the probability density function  $f(x)$  given by,

$$f(x) = \begin{cases} \alpha e^{-\alpha x} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{where } \alpha > 0$$

Or

$$f(x) = \alpha e^{-\alpha x}, x > 0, \alpha > 0$$

Is known as Exponential Distribution.

- Exponential Distribution is a probability density function.

$$f(x) > 0$$

$$\int_{-\infty}^{\infty} f(x) = 1$$

- Mean,  $\mu = \frac{1}{\alpha}$
- Variance,  $V = \frac{1}{\alpha^2}$
- Standard deviation,  $\sigma = \frac{1}{\alpha}$

### Normal Distribution:

A random variable X with the density,

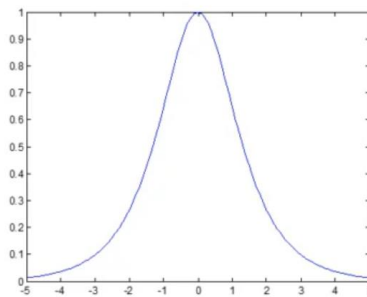
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2}, -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$$

Is said to have normal distribution with parameter  $\mu$  and  $\sigma$ .

$$\text{Also, } \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2} = 1$$

Where,  $\mu$  = mean and  $\sigma$  = standard deviation

The graph of this normal distribution is a symmetric bell curve at its mean  $\mu$ .



To calculate the probability associated with the normal curve requires the integration of the density function.

- $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2}$
- $p[a < x < b] = \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2} dx$  is not possible.
- Put  $Z = \frac{x-\mu}{\sigma}$  in the integral.

Now your transformed probability is,

$$p\left[\frac{a-\mu}{\sigma} < \frac{x-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right] = \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}Z^2} dZ$$

The new variable Z is called the standard normal variable.

**Standardization Theorem:** Let X be normal with mean  $\mu$  and standard deviation  $\sigma$ . The variable  $Z = \frac{(x-\mu)}{\sigma}$  is standard normal.

This transformation yields the random variable mean 0 and standard deviation 1.

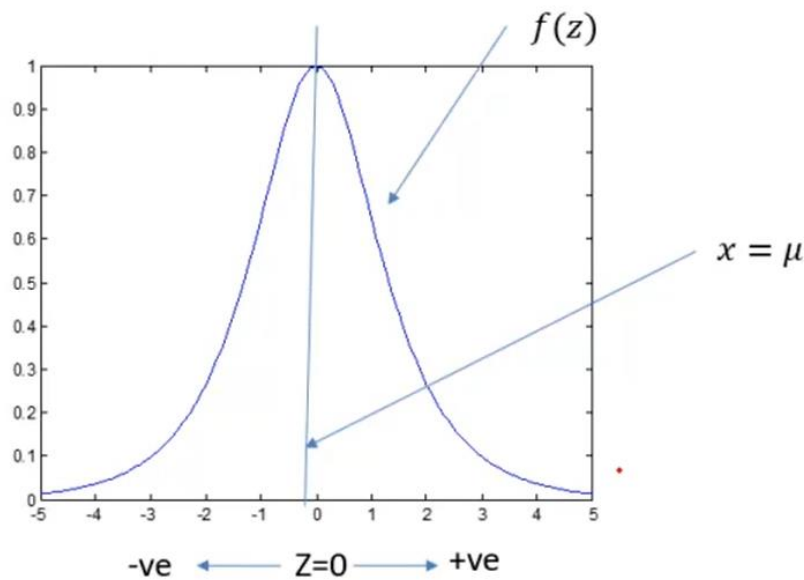
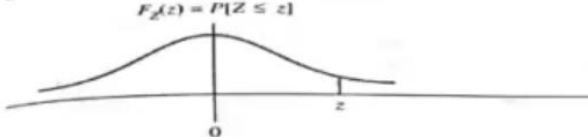


TABLE V  
Cumulative distribution: Standard normal

$F_Z(z) = P[Z \leq z]$



$F_Z(z) = P[Z \leq z]$

$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1137	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1921	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

TABLE V  
Cumulative distribution: Standard normal (concluded)

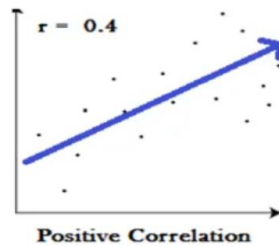
$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	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
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

$P(Z \leq 0.67) =$

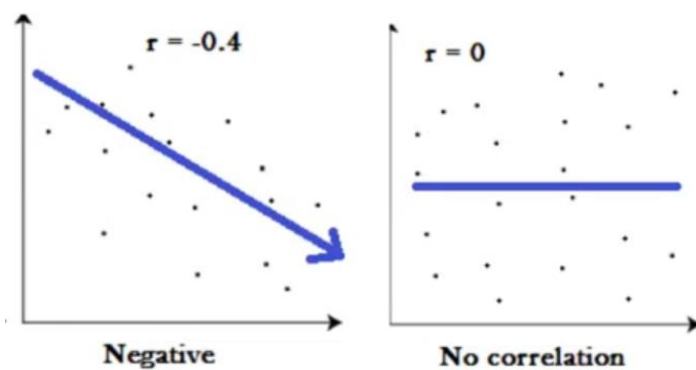
## Correlation and Correlation Coefficient

Co-variation of two independent magnitude is known as Correlation.

**Positively Correlated:** Two variables X and Y are related in such a way that increase or decrease in one of them corresponds to increase or decrease in the other.



**Negatively Correlated:** Two variables X and Y are related in such a way that increase or decrease in one of them corresponds to decrease or increase in the other.



The numerical measure of correlation between two variables X and Y is known as **Pearson's Coefficient of Correlation**, denoted by  $r$  and is defined as,

$$r = \frac{\sum_i^n (x - \bar{x})(y - \bar{y})}{n\sigma_x\sigma_y}$$

Where,  $\bar{x}$  = Mean of  $x$ .

$\bar{y}$  = Mean of  $y$ .

$\sigma_x$  = Standard deviation of  $x$ .

$\sigma_y$  = Standard deviation of  $y$ .

Or



$$r = \frac{\sigma_x^2 + \sigma_y^2 + \sigma_{x-y}^2}{2\sigma_x\sigma_y}$$

Where,  $\sigma_x^2 = \frac{\sum x^2}{n} - \bar{x}^2$ ,

$$\sigma_y^2 = \frac{\sum y^2}{n} - \bar{y}^2,$$

$$\sigma_{x-y}^2 = \frac{\sum (x - y)^2}{n} - (\bar{x} - \bar{y})^2$$

## Regression

**Regression:** Regression means to study the relationship between X and Y.

- The best fitting straight line of the form  $Y = aX + b$ , is called the regression line of Y on X.
- The best fitting straight line of the form  $X = aY + b$ , is called the regression line of X on Y.

Regression line of Y on X is,

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

Regression line of X on Y is,

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

If  $Y = y - \bar{y}$  and  $X = x - \bar{x}$ ,

Then the regression formula can be written as,

$$Y = r \frac{\sigma_y}{\sigma_x} X \text{ and } X = r \frac{\sigma_x}{\sigma_y} Y$$

Here,  $r \frac{\sigma_y}{\sigma_x}$  and  $r \frac{\sigma_x}{\sigma_y}$  are known as Regression Coefficient.

**Note:** r should always lie in the interval [-1,1]

## SAMPLING THEORY

**Population:** Population in statistics means all members of a defined group that we are studying or collecting information for making decision. It is a group of phenomena that has something in common.

**Sampling theory:** It is the field of statistics that is involved with the collection, analysis and interpretation of data gathered from random samples of a population to estimate the characteristic of the population.

- In our day to day life we need to draw some information from population.
- It is impossible to examine every individual.
- We examine a small part of the population known as **Sample**.
- From the sample we draw conclusion about the entire population based on the information from the sample.
- Predict about a population from sample also known as **Statistical Inference**.

**Sample:** A finite subset of universe (or population) is called a sample. The process of selecting a sample from the population is called **Sampling**.

**Random sampling:** The selection of an individual or items from the population in such a way that each has the same chance of being selected is called Random sampling.

**There are two different ways of selecting a random sample:**

- **Sampling with replacement:** Sampling where a member of the population may be selected more than once.
- **Sampling without replacement:** Sampling where a member is not chosen more than once.

**Sampling distribution:** A sampling distribution is a probability distribution of a statistic obtained through a large number of samples drawn from a specific population. The sampling distribution of a given population is the distribution of frequencies of a range of different outcomes that could possibly occur for a statistic of a population.

- Sampling distribution of a large samples is assumed to be a normal distribution.

The standard deviation of a sampling distribution is also called **Standard deviation** (SE).

Reciprocal of the standard error is called **Precision**.

**Statistics:** A statistical measure of sample observation and as such it is a function of sample observations.

Statistical inferences are drawn about population values i.e., parameter based on the sample observations i.e., statistics.

Population mean,  $\mu$

Population variance,  $\sigma^2$

Sample mean,  $\mu_{\bar{x}}$  or  $\bar{X}$

Sample variance,  $\sigma_{\bar{x}}^2$

### Sampling distribution of means:

#### Case 1: Random sampling with replacement

Step(i): Items are drawn one by one and are put back to population before the next draw.

Step(ii): If  $N$  is size of the population and  $n$  is size of sample then we have  $N^n$  possible samples.

Prediction of sample mean and variance

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

#### Case 2: Random sampling without replacement

Step(i): Items are drawn one by one and are not put back to the population before the next draw.

Step(ii): Number of ways in which  $n$  samples can be drawn from a population of  $N$  is  ${}^N C_n$  ways.

Prediction of sample mean and variance

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}}^2 = \left[ \frac{N-n}{N-1} \right] \frac{\sigma^2}{n} = C \frac{\sigma^2}{n}$$

$C = \frac{N-n}{N-1}$ , called finite population correction factor.

$C \rightarrow 1$  as  $N \rightarrow \infty$ , since  $\lim_{N \rightarrow \infty} \frac{N-n}{N-1} = 1$

## HYPOTHESIS TESTING

**Hypothesis:** A statistical hypothesis is a statement about a population or more populations which we should verify on the basis of information available from a sample.

**There are two types of hypothesis:**

- i. Null Hypothesis
- ii. Alternate Hypothesis

**Null Hypothesis:** For applying the tests of significance we first set up hypothesis- a definite statement about the population parameter, such a hypothesis of no difference is called the Null Hypothesis denoted by  $H_0$ .

**Example:**

- To test whether one procedure is better than another.  
 $H_0$ : There is no difference between the procedure.
- To test whether there is a relationship between two variates.  
 $H_0$ : There is no difference between the two variates.

**Alternate Hypothesis:** Any hypothesis which is complementary to the null hypothesis is called an Alternate Hypothesis. It is denoted by  $H_1$ .

**Example:**

- If we want to test the null hypothesis,  

$$H_0: \mu = \mu_0$$
 Then the alternate hypothesis would be,
  - I.  $H_1: \mu \neq \mu_0 \leftarrow$  Two-Tailed Test
  - II.  $H_1: \mu < \mu_0 \leftarrow$  Left-tailed Test
  - III.  $H_1: \mu > \mu_0 \leftarrow$  Right-tailed Test

**Types of Error:**

- i. Type-I error: The error made when the null hypothesis  $H_0$  is True but it is rejected by the test procedure.
- ii. Type-II error: The error made when the null hypothesis  $H_0$  is False but it is accepted by the test procedure.

**There are four possibilities of a statistical hypothesis:**

- i. Hypothesis is True but test rejects it. (Type-I error)
- ii. Hypothesis is False but test accepts it. (Type-II error)

- iii. Hypothesis is True and our test accepts it. (Correct decision)
- iv. Hypothesis is False and our test rejects it. (Correct decision)

### Significance level:

The maximum probability of committing type-I error is known as the significance level. This probability is conventionally fixed at 0.05 (5%) or 0.01 (1%). These are called level of significance. It is denoted by  $\alpha$ .

**Critical Region:** A region of rejecting  $H_0$  when it is true.

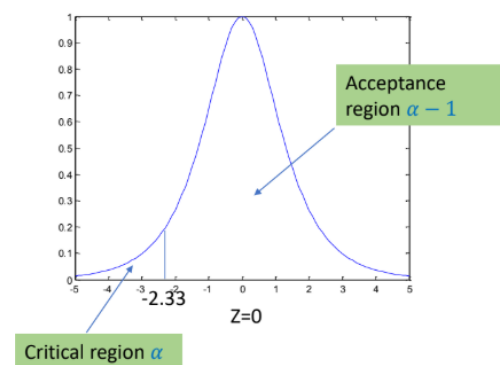
**Acceptance Region:** A region of accepting  $H_0$  when it is true.

$H_1: \mu < \mu_0$  (Left-tailed Test)

If  $\alpha = 1\%$  or 0.01, then

$$P(z < z_\alpha) = 0.01 \Rightarrow z_\alpha = -2.33$$

Critical value of  $z$  for  $\alpha = 1\%$  or 0.01

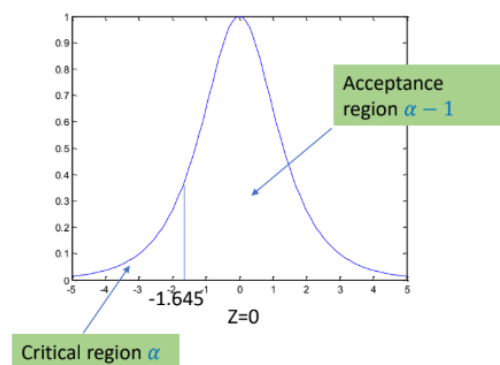


$H_1: \mu < \mu_0$  (Left-tailed Test)

If  $\alpha = 5\%$  or 0.05

$$\text{Then, } P(z < z_\alpha) = 0.05 \Rightarrow z_\alpha = -1.645$$

Critical value of  $z$  for  $\alpha = 5\%$  or 0.05

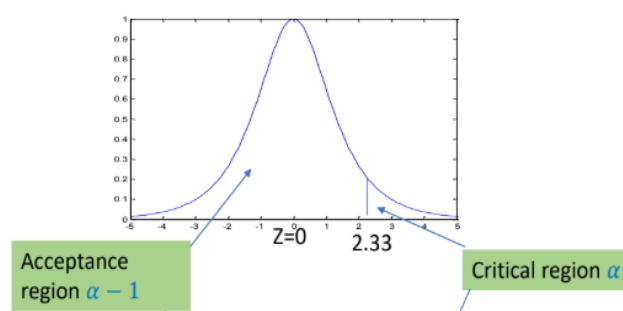


$H_1: \mu > \mu_0$  (Right-tailed Test)

If  $\alpha = 1\%$  or 0.01

$$\text{Then, } P(z > z_\alpha) = 0.01 \Rightarrow z_\alpha = 2.33$$

Critical value of  $z$  for  $\alpha = 1\%$  or 0.01

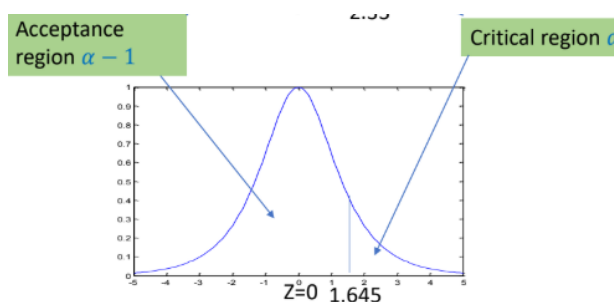


$H_1: \mu > \mu_0$  (Right-tailed Test)

If  $\alpha = 5\%$  or 0.05

Then,  $P(z > z_\alpha) = 0.05 \Rightarrow z_\alpha = 1.645$

Critical value of  $z$  for  $\alpha = 5\%$  or 0.05



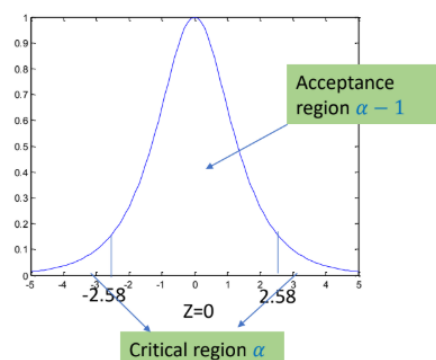
$H_1: \mu \neq \mu_0$  (Two-tailed Test)

If  $\alpha = 1\%$  or 0.01

Then,  $P(z < z_{\alpha/2}) = 0.01 \Rightarrow z_{\alpha/2} = -2.58$

And  $P(z > z_{\alpha/2}) = 0.01 \Rightarrow z_{\alpha/2} = 2.58$

Critical value of  $z$  for  $\alpha = 1\%$  or 0.01



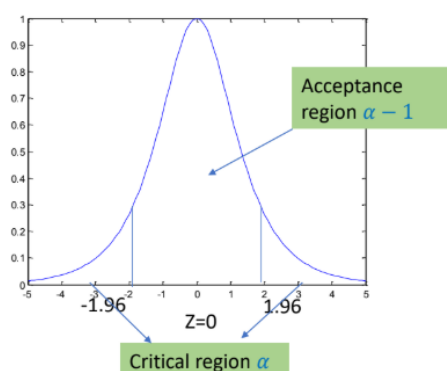
$H_1: \mu \neq \mu_0$  (Two-tailed Test)

If  $\alpha = 5\%$  or 0.05

Then,  $P(z < z_{\alpha/2}) = 0.05 \Rightarrow z_{\alpha/2} = -1.96$

And  $P(z > z_{\alpha/2}) = 0.05 \Rightarrow z_{\alpha/2} = 1.96$

Critical value of  $z$  for  $\alpha = 5\%$  or 0.05



$\alpha$	1%	5%
Left-tailed Test	-2.33	-1.645
Right-tailed Test	2.33	1.645
Two-tailed Test	2.58	1.96

**Test of significance:** The process which helps us to decide about the acceptance or rejection of the hypothesis is called the test of significance.

**Confidence Interval (limits):** Confidence interval of mean given variance of a normal distribution is,

$$\bar{X} \pm Z_c \frac{\sigma}{\sqrt{n}}$$

For 95% confidence, the mean interval is,

$$\bar{X} \pm Z_{0.05} \frac{\sigma}{\sqrt{n}} \text{ or } \bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

For 99% confidence, the mean interval is,

$$\bar{X} \pm Z_{0.01} \frac{\sigma}{\sqrt{n}} \text{ or } \bar{X} \pm 2.58 \frac{\sigma}{\sqrt{n}}$$

If  $\mu$  is mean of the population,

Procedure for testing of hypothesis of mean for large n:

- |         |   |
|---------|---|
| Step 1: | Set up the null hypothesis $H_0$ .  |
| Step 2: | Set up the alternate hypothesis $H_1$ .   |
| Step 3: | Choose the level of significance ( $\alpha$ ).  |
| Step 4: | Compute the statistics, $Z_{cal} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightarrow$<br>Test statistic for mean. |
| Step 5: | If $ Z_{cal}  \leq  Z_{tab} $ , $H_0$ is accepted, otherwise it is rejected.                                      |

Let  $\mu_1$  and  $\mu_2$  be the means of two populations.

Let  $(\bar{X}_1, \sigma_1)$ ;  $(\bar{X}_2, \sigma_2)$  be the mean and standard deviation of two large samples of size  $n_1$  and  $n_2$  respectively.

We wish to test the null hypothesis  $H_0$  that there is no difference between the population means.

$$H_0: \bar{X}_1 = \bar{X}_2$$



The test statistic is,

$$Z_{cal} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The confidence limits for the difference of means of the population are,

$$(\bar{X}_1 - \bar{X}_2) \pm Z_c \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Note:

If the samples are drawn from the same population, then  $\sigma_1 = \sigma_2$ , and

$$Z_{cal} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

### Testing of hypothesis on proportion:

- Standard Normal Distribution can be used to test the hypothesis on a proportion P with large sample size.
- Construct confidence interval on P.
- Population proportion can be defined as,  $P = \frac{X}{N}$

Where, X = Count of successes in the population.

N = Size of population

Let X be the observed number of successes in a sample size of n and  $\mu = np$  be the expected number of successes. Then the test statistics is,

$$Z = \frac{X - \mu}{\sigma} = \frac{X - np}{\sqrt{npq}}$$

Where, p = Probability of success.

q = Probability of failure.

The Standard Error,  $S.E = \sqrt{\frac{pq}{n}}$

$\alpha$	1%	5%
Left-tailed Test	-2.33	-1.645
Right-tailed Test	2.33	1.645
Two-tailed Test	2.58	1.96

### Steps for testing:

1. Setup the Null Hypothesis,  $H_0$  : a statement (or)  $H_0 : P = P_0$  .
2. Setup the Alternate Hypothesis,  
 $H_1$  : a statement or
  - i)  $H_1 : P < P_0$  Left tailed test
  - ii)  $H_1 : P > P_0$  Right tailed test
  - iii)  $H_1 : P \neq P_0$  Two tailed test

3. Test statistics,

$$Z_{cal} = \frac{X - \mu}{\sigma} = \frac{X - np}{\sqrt{npq}}$$

4. If  $|Z_{cal}| \leq |Z_{tab}|$ ,  $H_0$  is accepted, else is rejected.

### **Testing of hypothesis on difference of proportion:**

Let  $P_1$  and  $P_2$  be the sample proportions in respect of an attribute corresponding to the large samples of size  $n_1$  and  $n_2$  drawn from two population.

The test statistic is,  $Z = \frac{P_1 - P_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

where,  $p = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$ ,  $q = 1 - p$

### Steps for testing:

1. Setup the Null Hypothesis,  $H_0$  : a statement (or)  $H_0 : P = P_0$  .
2. Setup the Alternate Hypothesis,  
 $H_1$  : a statement or
  - iv)  $H_1 : P < P_0$  Left tailed test
  - v)  $H_1 : P > P_0$  Right tailed test
  - vi)  $H_1 : P \neq P_0$  Two tailed test

3. Test statistics,

$$Z_{cal} = \frac{P_1 - P_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, p = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}, q = 1 - p$$

4. If  $|Z_{cal}| \leq |Z_{tab}|$ ,  $H_0$  is accepted, else is rejected.

## Student t Distribution

The name 'student' derived a theoretical distribution to test the significance of a sample mean where the small sample is drawn from a normal population.

It is similar to the normal distribution with its bell shape but has heavier tails.

The t distribution (Student's t distribution) is a probability distribution that is used to estimate population parameters when the sample size is small and or when the population variance is unknown.

T distribution is determined by its degrees of freedom. The degrees of freedom refer to the number of independent observations in a set of data.

The number of independent observations is equal to sample size minus one.

The distribution of the t-statistic from samples of size 8 would be described by a t-distribution having 8-1 or 7 degrees of freedom.

### Student-t distribution:

Let  $x_i$  ( $i = 1, 2, 3, \dots, n$ ) be a random sample of size  $n$  drawn from a normal population with mean  $\mu$  and variance  $\sigma^2$ . The statistic  $t$  is defined as,

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - \mu}{s} \sqrt{n}$$

Here,  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

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

$v = n - 1$ , denote the degrees of freedom of  $t$ .

Probability density function for student's t distribution with  $(n-1)$  degrees of freedom is,

$$y = f(t) = \frac{y_0}{\left(1 + \frac{t^2}{v}\right)^{\frac{v+1}{2}}}$$

Where,  $y_0$  is a constant such that the area under the curve is unity.

Note: If  $v$  is large ( $v \geq 30$ ) the graph of  $f(t)$  closely approximates standard normal.

### **Student t test:**

1. Set up the null hypothesis,  $H_0: \bar{x} = \mu$
2. Set up the alternate hypothesis,
  - i.  $H_1: \bar{x} < \mu$ , Left Tailed Test
  - ii.  $H_1: \bar{x} > \mu$ , Right Tailed Test
  - iii.  $H_1: \bar{x} \neq \mu$ , Two Tailed Test
3. The test statistic is,

$$t_{cal} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

4. If  $|t_{cal}| \leq |t_{tab}|$ , accept  $H_0$ , otherwise reject  $H_0$ .

Note: Significance level will be taken 1% and 5%.

### **Confidence limit of t for the mean population $\mu$ :**

If  $t_{0.05}$  is the tabulated value of t for  $n-1$  degrees of freedom at 99% of level of significance, then

$\bar{X} \pm \frac{s}{\sqrt{n}} t_{0.01}$  is the confidence limits for 99% confidence.

### **Test of hypothesis of difference between sample means:**

Consider 2 independent sample  $x_i (i = 1, 2, 3 \dots n_1)$  and  $y_i (i = 1, 2, 3 \dots n_2)$  drawn from a normal population.

Let  $(\bar{X}, \sigma_x)$ ,  $(\bar{Y}, \sigma_y)$  be the population mean and  $\sigma$  be the population variance.

To test,  $H_0: \bar{X} = \bar{Y}$

$$\text{Test statistics, } t_{cal} = \frac{\bar{X} - \bar{Y}}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{Where, } s^2 = \frac{1}{n_1 + n_2 - 2} \{ \sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \}$$

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}, \text{ degrees of freedom } v = n_1 + n_2 - 2$$

## Chi-Square Distribution

It is a continuous probability distribution.

The probability density function of Chi-Square distribution with  $v$  degrees of freedom is,

$$f(x) = \frac{x^{\frac{v}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{v}{2}} \Gamma\left(\frac{v}{2}\right)}$$

Mean,  $\mu = v, \sigma^2 = 2v$

Chi-Squared distribution provides a measure of correspondence between the theoretical frequencies and observed frequencies.

If  $O_i (i = 1, 2, 3 \dots n)$  and  $\varepsilon_i (i = 1, 2, 3 \dots n)$  respectively denotes a set of observed and estimate frequencies the quality Chi-Squared denoted by  $\chi^2$  is defined as follows,

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - \varepsilon_i)^2}{\varepsilon_i}$$

It helps us to test the goodness of fit of these distributions.

If the calculates value of  $\chi^2$  is less than the table value of  $\chi^2$  at a specified level of significance the hypothesis is accepted, otherwise the hypothesis is rejected.

### **Steps for testing:**

1. Set up the Null Hypothesis,  $H_0$ : a statement or  $H_0: O_i = \varepsilon_i$
2. Set up the Alternate Hypothesis,  $H_1$ : a statement or
  - i.  $H_1: O_i < \varepsilon_i$ , Left Tailed Test
  - ii.  $H_1: O_i > \varepsilon_i$ , Right Tailed Test
  - iii.  $H_1: O_i \neq \varepsilon_i$ , Two Tailed Test
3. Test statistics,  $\chi_{cal}^2 = \sum_{i=1}^n \frac{(O_i - \varepsilon_i)^2}{\varepsilon_i}$
4. If  $|\chi_{cal}^2| \leq |\chi_{tab}^2|$ ,  $H_0$  is accepted, else rejected.

## Joint Distribution

Consider a two-dimensional random variable or a bivariate random variable.

**Definition:** Let  $X$  and  $Y$  be discrete random variables. The order pair  $(X \text{ and } Y)$  is called a two-dimensional discrete random variable. A function  $f_{xy}$  such that,  $f_{xy}(x, y) = P[X = x \text{ and } Y = y]$  is called the Joint Density for  $(X, Y)$ .

Necessary and sufficient conditions for a function to be a discrete joint Density,

1.  $f_{xy}(x, y) \geq 0$
2.  $\sum_{all\ x} \sum_{all\ y} f_{xy}(x, y) = 1$

It is more common to represent the probability function in the form of a table.

Each entry in the table is a number between 0 and 1.

### Marginal Distribution:

Let  $(X, Y)$  be a two-dimensional discrete random variable with joint density  $f_{xy}$ . The marginal density of  $X$ , denoted by  $f_X$ , is given by,

$$f_X(x) = \sum_y f_{XY}(x, y)$$

The marginal density of  $Y$ , denoted by  $f_Y$  is given by,

$$f_Y(y) = \sum_x f_{XY}(x, y)$$

**Note:** The marginal density for  $X$  is obtained by summing across the rows of the table, that for  $Y$  is obtained by summing down the columns.

### Expectation and Covariance

Expectation of  $X$  denoted by  $E[X]$  is given by,

$$E[X] = \sum_x \sum_y x f_{XY}(x, y) = \sum_x x f_X(x)$$

Expectation of  $Y$  denoted by  $E[Y]$  is given by,

$$E[Y] = \sum_x \sum_y y f_{XY}(x, y) = \sum_y y f_Y(y)$$

Expectation of  $XY$  denoted by  $E[XY]$  is given by,

$$E[XY] = \sum_x \sum_y xy f_{XY}(x, y)$$

### Covariance:

Let  $X$  and  $Y$  be random variables with mean  $\mu_x$  and  $\mu_y$  respectively. The covariance between  $X$  and  $Y$ , denoted by  $Cov(X, Y)$  or  $\sigma_{xy}^2$  is given by,

$$Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

Computational formula,

$$Cov(X, Y) = E[XY] - E[X]E[Y]$$

### Pearson Coefficient of Correlation:

Let  $X$  and  $Y$  be random variables with mean  $\mu_x$  and  $\mu_y$  respectively and variance  $\sigma_x^2$  and  $\sigma_y^2$  respectively. The correlation  $\rho_{xy}$  between  $X$  and  $Y$  is given by,

$$\rho_{xy} = \frac{Cov(X, Y)}{\sqrt{\sigma_x^2 * \sigma_y^2}}$$

Where,  $\sigma_x^2 = E[X^2] - [E[X]]^2$

$$\sigma_y^2 = E[Y^2] - [E[Y]]^2$$

### Independent random variable:

Let  $X$  and  $Y$  be random variables with joint density  $f_{xy}$  and marginal densities  $f_x$  and  $f_y$  respectively.  $X$  and  $Y$  are independent if,

$$f_{XY}(x, y) = f_X(x)f_Y(y)$$

Or

Let  $X$  and  $Y$  be random variables with joint density  $f_{XY}$ . If  $X$  and  $Y$  are independent then,

$$E[XY] = E[X]E[Y]$$

Note: If  $X$  and  $Y$  are independent,  $Cov(X, Y) = 0$  and  $\rho_{xy} = 0$ .

\*\*\* THE END \*\*\*