Expected value of mean:

$$E[x] = \sum_{i=1}^{K} x_i P_i \quad \text{or} \quad E[x] = \int_R x f(x)\,dx$$

(continous probability distribution
$f(x)$ = Probability Density function)

Discrete Probability distr:

mean, $\boxed{\mu = \sum x_i P(x_i)}$

variance, $\boxed{V = \sigma^2 = \sum_i (x_i - \mu)^2 p(x_i)}$

Binomial Distribution:

n: no. of trails or Observations

Def: If P is the probability of success and q is probability of failure, the probability of x success out of n trails is given by,

$$\boxed{P(x) = {}^nC_x\, p^x q^{n-x}} \text{ or } {}^nC_x\, p^x (1-P)^{n-x}$$

$$\boxed{{}^nC_x = \frac{n!}{x!(n-x)!}}$$

mean, $\mu = np$

Variance, $V = \sigma^2 = npq$

Standard deviation, $\sigma = \sqrt{npq}$

Poisson Distribution:

Def: A random variable x is said to have a poisson distribution with parameter m if its density f is given by

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

where, $m = np$

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

Continuous Probability Distribution: continuous random variable.

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.
- $\int_{-\alpha}^{\alpha} f(x)\, dx = 1$
- $P[a \leq x \leq b] = \int_{a}^{b} f(x)\, dx$ for a and b real is

called a 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 by,

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

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

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

$$-\infty < \mu < \infty$$

$$\sigma > 0$$

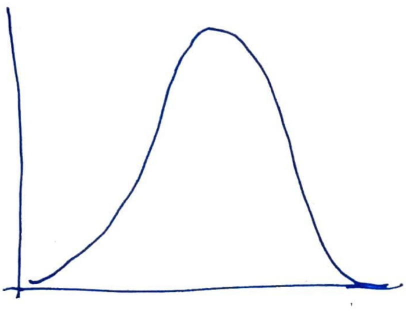
is said to have normal distribution with parameter μ and σ.

Also, $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}[(x-\mu)/\sigma]^2} = 1$

where, μ = mean

σ = standard deviation

The graph of this normal density is a symmetric bell curve centered at its mean μ.



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

- $f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}[(x-\mu)/\sigma]^2}$
- $P[a < x < b] = \int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}[(x-\mu)/\sigma]^2} dx$ is not possible
- Put $z = \frac{x-\mu}{\sigma}$ in the integral

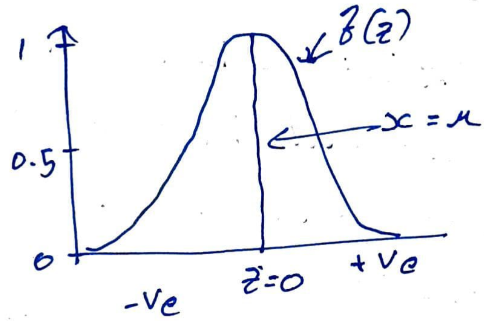
Now your transformed probability is,

$$P\left[\frac{a-\mu}{\sigma} < \frac{x-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right] = \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}z^2} \, dz$$

The new variable z is called the standard normal variable.

Standardization Theorem: let x be normal with mean μ and standard deviation σ. The variable $z = \frac{x-\mu}{\sigma}$ is standard normal.

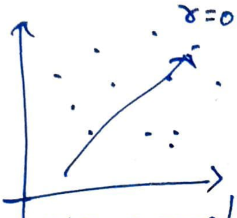
This transformation yields the random variable mean 0 and standard deviation 1.



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.

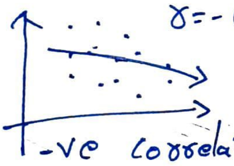
$r = 0.4$



+ve correlation

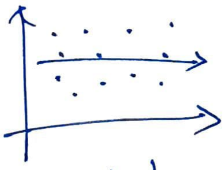
- **Negatively correlated**: Two variables x and y are related in such a way that increase or decrease in one of them corresponds to decrease or increase in the other.

$r = -0.4$



-ve correlation

$r = 0$



No correlation

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

σ_x = standard deviation of x

σ_y = standard deviation of y

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

where, $\sigma_x^2 = \dfrac{\Sigma x^2}{n} - \bar{x}^2$

$\sigma_y^2 = \dfrac{\Sigma y^2}{n} - \bar{y}^2$

$\sigma_{x-y}^2 = \dfrac{\Sigma (x-y)^2}{n} - (\bar{x} - \bar{y})^2$

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

Then the regression formula can be written as,

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

Here, $r \dfrac{\sigma_y}{\sigma_x}$ and $r \dfrac{\sigma_x}{\sigma_y}$ are known as Regression coefficient.

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 obtaind 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 ~~sampling~~ of a large samples is assumed to be a normal distribution

- The standard deviation of a sampling distribution is also called standard Error (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 → μ

Population variance → σ^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 next draw.

step (ii) If N is size of population and n is the 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 next draw.

step (ii): Number of ways in which n sample 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

$c \to 1$ as $N \to \infty$, since $\lim_{N\to\infty} \frac{N-n}{N-1} = 1$

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 a hyothesis - a definite statement about the population parameter, such a hypothesis is of no difference is called the Null Hypothesis is 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.

It is denoted by H_1 to test the null hypothesis.

Example: If we want to test the null hypothesis

$$H_0: \mu = \mu_0$$

Then the alternate hypothesis would be,

I. $H_1: \mu \neq \mu_0 \longleftarrow$ Two-tailed test

II. $H_1: \mu < \mu_0 \longleftarrow$ left-tailed test

III. $H_1: \mu > \mu_0 \longleftarrow$ 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%). It is denoted by α. These are called level of significance.

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

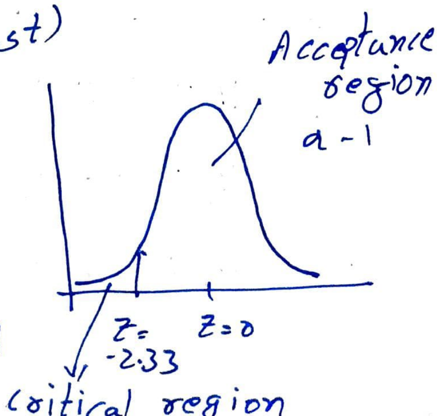
Acceptance Region: A region of accepting H_0 when it is true.

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

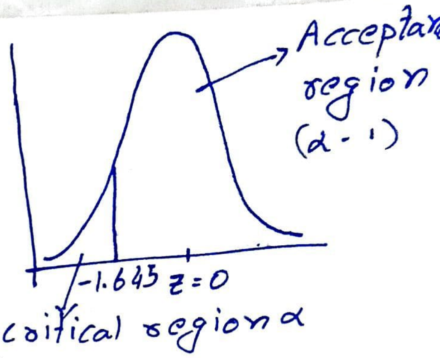
If $\alpha = 1\%$ or 0.01

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)



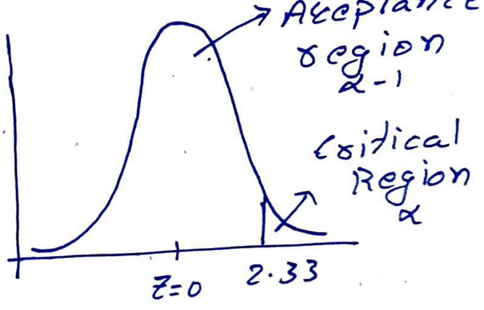
If $\alpha = 5\%$ or 0.05

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)



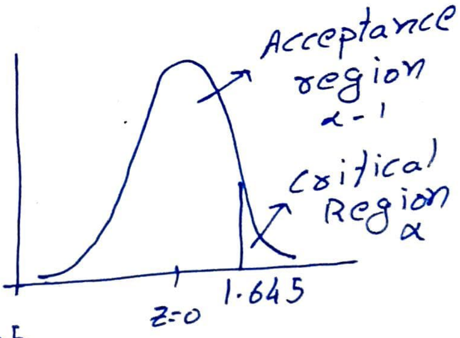
If $\alpha = 1\%$ or 0.01

Then $P(z > z_\alpha) = 0.01$

$\Rightarrow z_\alpha = 2.33$

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

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



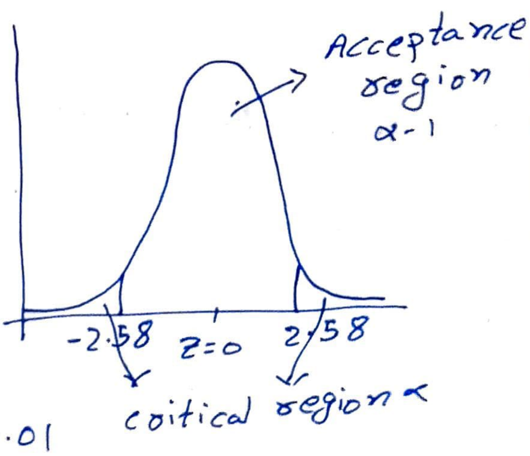
If $\alpha = 5\%$ or 0.05

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 $P(z < z_{\alpha/2}) = 0.01$

$\Rightarrow z_{\alpha/2} = -2.58$

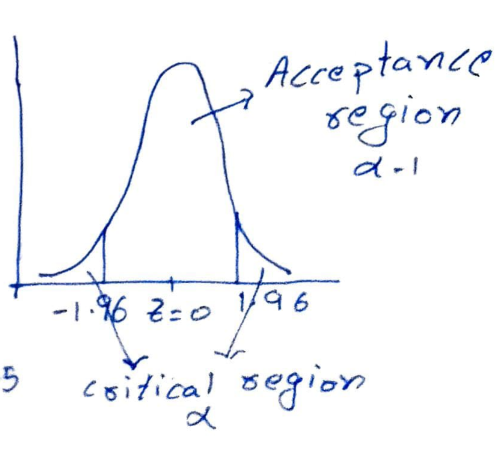
and

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

$z_{\alpha/2} = 2.58$

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 $P(z < z_{\alpha/2}) = 0.05$

$z_{\alpha/2} = -1.96$

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

critical ~~ve~~ value of z for $\alpha = 5\%$ or 0.05

α	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,

$$\bar{x} \pm z_{0.05} \frac{\sigma}{\sqrt{n}}$$

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

For 99% confidence the mean interval is,

$$\bar{x} \pm z_{0.01} \frac{\sigma}{\sqrt{n}}$$

$$\bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}}$$

If μ is the 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 l.o.s (α)

step 4: Compute the statistics

$$z_{cal} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \longrightarrow \text{Test statistic for mean}$$

step 5: If $|z_{cal}| \leq |z_{tab}|$, H_0 is accepted, otherwise it is rejected let μ_1 and μ_2 be the mean of two population.

let $(\bar{x}_1, \sigma_1)$; $(\bar{x}_2, \sigma_2)$ be the mean and S.D of two 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 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 no. of successes. Then the test statistics is

$$Z = \frac{x-\mu}{\sigma} = \frac{x-np}{\sqrt{npq}}, \quad \begin{array}{l} p = \text{probability of success} \\ q = \text{probability of failure} \end{array}$$

The standard Error,

$SE = \sqrt{pq/n}$

Significance Table

α	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. Set up the Null Hypothesis, H_0: a statement of

   $H_0 : P = P_0$

2. Set up 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 $|z_{cal}| \le |z_{tab}|$, H_0 is accepted, else is rejected.

Testing of hypothesis on Difference of proportion:
let P_1 & P_2 be the sample proportion in respect of an attribute corresponding to two large samples of size n_1 & 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 Set up the Null Hypothesis, H_0: a statement of

$H_0: P_1 = P_2$

② Set up the Alternate hypothesis, H_1: a statement of

i. $H_1: P_1 < P_2$, left tailed test

ii. $H_1: P_1 > P_2$, right tailed test

iii. $H_1: P_1 \neq P_2$, 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 $|z_{cal}| \le |z_{tab}|$, 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 refers 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 $(i=1,2,3\ldots n)$ be a random sample of size n drawn from a normal population with mean μ and variance σ^2. The statistic t is defined as,

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

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 $(n-1)$ d.f

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

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,

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

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 for μ t for the mean population μ:

- If $t_{0.05}$ is the tabulated value of t for n-1 degrees of freedom at 95% of level of significance, then

$$\bar{x} \pm \frac{s}{\sqrt{n}} t_{0.05}$$ is the confidence limits

for 95% confidence.

- If $t_{0.01}$ 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 $(i = 1, 2, 3 \ldots n_1)$ and y_j $(j = 1, 2, 3 \ldots n_2)$ drawn from a normal population

let $(\bar{x}, \sigma_x), (\bar{y}, \sigma_y)$ be the mean and S.D of two large samples of size n_1 and n_2 respectively.

let μ be the population mean & σ be the population variance.

To test, $H_0 : \bar{x} = \bar{y}$

Test statistics, $t_{cal} = \dfrac{(\bar{x} - \bar{y})}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

where, $s^2 = \dfrac{1}{n_1 + n_2 - 2} \left\{ \sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \right\}$

or

$s^2 = \dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$, 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,

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

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

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

If $O_i\,(i=1,2,\ldots n)$ and $E_i\,(i=1,2,\ldots n)$ respectively denotes a set of observed & estimated frequencies the quantity chi-squared denoted by χ^2 is defined as follows,

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

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

If the calculated value of χ^2 is less than the table value of χ^2 at a specified level of significance the hypothesis is accepted, otherwise the hypothesis is rejected.

Steps for testing:

- Set up the Null Hypothesis, H_0: a statement of $H_0: O_i = E_i$
- Set up the Alternate Hypothesis, H_1: a statement of

  i) $H_1: O_i < E_i$ Left tailed test

  ii) $H_1: O_i > E_i$ Right tailed test

  iii) $H_1: O_i \neq E_i$ Two tailed test

- Test statistics, $\chi^2_{(cal)} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$
- If $|\chi^2_{cal}| \leq |\chi^2_{tab}|$, H_0 is accepted, else rejected.

Joint Distribution

Consider a two dimensional random variable of a bivariate random variable.

Definition Let X and Y be discrete random variables. The order pair (X 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. $\sum_{\text{all } x} \sum_{\text{all } y} f_{XY}(x, y) = 1$

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

- Each entry is in the table is a number b/w 0 & 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 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 give 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

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

<u>Covariance</u>: let X and Y be random variables with mean μ_x and μ_y respectively. The covariance between X and Y, denoted by $Cov(X, Y)$ or σ_{xy}^2 is given by

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

Computational formula,

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

<u>Pearson Coefficient of Correlation</u>: let X and Y be random variables with mean μ_x and μ_y respectively and variance σ_x^2 and σ_y^2 respectively. The correlation ρ_{xy} between X and Y is given by,

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

where, $\sigma_x^2 = E[x^2] - [E[x]]^2$

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

<u>Independent random variable</u>: 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

$E[XY] =$

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

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

<u>Note</u>: If X and Y are independent,

$$Cov(X, Y) = 0$$

$$\rho_{xy} = 0$$

THE END