Statistical Inference

Statistical inference is mainly concerned with making inferences regarding the unknown aspects of the distribution of the population based on samples taken from it. The unknown aspect may be the form of the distribution or values of the parameters involved or both. Statistical inference is broadly classified into two

1. Estimation of parameters
2. Testing of hypotheses

Estimation deals with methods of determining numbers which may be taken as the values of the unknown parameters (called the point Estimation) as well as with the determination of intervals which will contain the unknown parameters with a specified probability (known as interval estimation), based on samples taken from the population.

Testing of hypotheses deals with the methods for deciding either to accept or reject the hypotheses based on samples taken from the population, with the degree of validity of the decision indicated in terms of probability.

Statistics

Any function of the sample values is known as a statistics.

Point Estimation

Any value or function of the sample values suggested as the value of the parameters is known as the ***Estimator***.

The following are the most important desirable properties of a good estimate.

1. Unbiasedness
2. Consistency
3. Efficiency
4. Sufficiency

UNBIASEDNESS

Let 't' be a statistic suggested as an estimate of the parameter θ. 't ' is said to be unbiased estimate of θ if $E(t)= \theta$.

CONSISTANCY

Let ' t_n' be a statistic where n is the sample size. t_n is said to be a consistent estimate of parameter θ , if for any two positive numbers ' ϵ' and ' η' (however small they may be) a N can be found out such that $n \geq N$.

$P(|t_n - \theta| < \epsilon) > 1 - \eta$

We can say that t_n is said to be consistent estimate of θ if it tends 'in probability' to θ as value of $n \to \infty$

EFFICIENCY

Let t_1 and t_2 two unbiased estimates of a parameter θ. Then t_1 is said to be more efficient than t_2 if V (t_1) is less than V (t_2)

SUFFICIENCY

An estimate of a parameter θ is called a sufficient estimate if it contains all the information about θ contained in the sample.

METHODS OF ESTIMATION

1. **Method of maximum likelihood**

   Let f(x, θ_1, θ_2… θ_k) be the p.d.f of the population where θ_1, θ_2… θ_k are the parameters. Let x_1, x_2,…x_n be a random sample taken the population . The likelihood function of the sample is,

   L(x_1,x_2,…x_n : θ_1, θ_2… θ_k) = f(x_1, θ_1, θ_2… θ_k) f(x_2, θ_1, θ_2… θ_k)…f(x_n, θ_1, θ_2… θ_k)

   For any given sample this may regarded as a function of unknown parameters θ_1, θ_2… θ_k . Those values of θ_1, θ_2… θ_k which maximizes the likelihood function are called maximum likelihood estimates of θ_1, θ_2… θ_k).

2. **Method of moments**

   Let f(θ_1, θ_2… θ_k)be the p.d.f of the population and let x_1,x_2,…x_n be a random sample taken the population. In this method of moments we find the first k moments of the population and equate them to the corresponding moments of the sample and the values of θ_1, θ_2… θ_k which are obtained as the solutions of these equations are taken as their estimates.

   Of these two methods suggested, the M.L method is superior because of the following desirable properties

   1. M.L estimates are asymptotically unbiased
   2. M.L estimates are consistent
   3. M.L estimates are most efficient
   4. M.L estimates are sufficient if sufficient estimates exist.
   5. M.L estimates are asymptotically normally distributed

Interval Estimation

In point estimation a number calculated from the sample is suggested as the estimate of the unknown parameter. In interval estimation we find out two statistics t_1 and t_2 ($t_1<t_2$) such that the probability that the interval (t_1,t_2) contains the true value of the unknown parameter has a pre-assigned value α called the confidence coefficient of the interval. Such an interval is called a confidence interval with confidence coefficient α.

The Confidence interval for the mean of a normal population

Let $x_1, x_2, \ldots, x_n$ be a sample chosen from Normal population N(μ,σ) Let $\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n}$ and $s^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$.

(a) Confidence interval for μ when σ known

We know that $\bar{x}$ has normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. So t= $\frac{(\bar{x}-\mu)\sqrt{n}}{\sigma}$ has N(0,1). From the standard normal table we can able to find $t_{\frac{\alpha}{2}}$ such that $P\left(|t| \le t_{\frac{\alpha}{2}}\right) = \alpha$, where α is the given confidence coefficient. This gives the confidence interval for μ as

$$\left[\bar{x} - \frac{\sigma}{\sqrt{n}} t_{\frac{\alpha}{2}}, \bar{x} + \frac{\sigma}{\sqrt{n}} t_{\frac{\alpha}{2}}\right]$$

(b) Confidence interval for μ when σ unknown

In this case we make use of the result that So t= $\frac{(\bar{x}-\mu)\sqrt{n-1}}{s}$ has students t distribution with n-1 degrees of freedom. From the t table we can able to find $t_{\frac{\alpha}{2}}$ such that $P\left(|t| \le t_{\frac{\alpha}{2}}\right) = \alpha$, where α is the given confidence coefficient. This gives the confidence interval for μ as

$$\left[\bar{x} - \frac{s}{\sqrt{n-1}} t_{\frac{\alpha}{2}}, \bar{x} + \frac{s}{\sqrt{n-1}} t_{\frac{\alpha}{2}}\right].$$

The Confidence interval for the proportion of a binomial population

In this case we make use of the result that So t= $\frac{\bar{p}-p}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}}$ has N(0,1). From the standard normal table we can able to find $t_{\frac{\alpha}{2}}$ such that $P\left(|t| \le t_{\frac{\alpha}{2}}\right) = \alpha$, where α is the given confidence coefficient. This gives the confidence interval for p as

$$\left[\bar{p} - t_{\frac{\alpha}{2}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}, \bar{p} + t_{\frac{\alpha}{2}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\right]$$

where $\bar{p} = \frac{x}{n}$, x is the number of favourable case in n repetition.

The Confidence interval for the variance of a normal population

Let $x_1, x_2, \ldots, x_n$ be a sample chosen from Normal population N(μ,σ) and let $s^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$. Define $u = \frac{ns^2}{\sigma^2}$ then u has chi-square variables with n -1 degrees of freedom. So for a given α from chi-square table with n-1 degrees of freedom we can able to find two values $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ Such that $P\left\{\frac{ns^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{ns^2}{\chi^2_{1-\alpha/2}}\right\} = \alpha$. So confidence interval for the variance is $\left[\frac{ns^2}{\chi^2_{\alpha/2}}, \frac{ns^2}{\chi^2_{1-\alpha/2}}\right]$.

Testing of Hypothesis

A hypothesis is an assertion about the form of the distribution or the value of the parameters of statistical populations.

Examples

1. The weight of the students follows the normal distribution
2. Height of the college students follows the normal distribution with mean 5 feet
3. The given data follows normal distribution with mean 10 and standard deviation 4.

Simple and Composite Hypothesis

If the population specifies the population completely then the hypothesis is called a simple hypothesis and otherwise it is called composite hypothesis.

In the above example 1 and 2 are composite and 3 is simple.

Statistical test

A procedure by which we may accept or reject the hypothesis based on sample taken from the population is called a statistical test.

Null hypothesis and Alternative hypothesis

The hypothesis that is tested is called the '*null hypothesis*' and is usually denoted by H_0. The hypothesis which we will accept or reject according as we reject or accept H_0. This hypothesis is called the '*alternative hypothesis*' and is usually denoted by H_1.

Two Types of Errors

It is impossible to assert whether a hypothesis is correct or wrong by a statistical test, as the decision is based on a sample only. A true hypothesis may be rejected and a false hypothesis accepted I a test. In short we admit that our procedure may result in committing one or the other of the following two types of errors

1. Rejecting H_0 when it is true
2. Accepting H_0 when it is false

The first is called Type I error and the second is called Type II error.

Action taken based on sample data	State of Nature	
	H_0 is true	H_0 is false
Reject H_0	Type I error	No error
Accept H_0	No error	Type II error

Test Statistics The function of the sample observations is chosen to take the decision either to accept or reject the hypothesis is called the test statistics.

Significance level and Power of the test

The probability of the test statistics falling in the critical region when the hypothesis is true is called the significance level or size of the test. That is the significance level is the probability of the first type of error and is denoted by α.

ie significance level = α = Prob.{Rejecting $H_0 \mid H_0$}

Probability of correct decision is called the power of the test. That is power of the test is the probability of rejecting the null hypothesis when the alternative hypothesis is true and is denoted by β.

$$\begin{aligned} \text{Power } &= \beta = \text{Prob.}\{\text{Rejecting } H_0 \mid H_1\} \\ &= 1 - \text{Prob.}\{\text{Accepting } H_0 \mid H_1\} \\ &= 1 - \text{Prob.}\{\text{second type of error}\}. \end{aligned}$$

Acceptance Region and Critical Region

In a test procedure we calculate the test statistics based on which we take the decision to accept or reject the null hypothesis. For this we divide the range of variation of the test statistics into two regions, acceptance regions and rejection region or critical region such that the probabilities of the two types of errors are not very large. If the computed value of the test statistics falls in the rejection region we reject the null hypothesis.

Size of the critical Region

Probability for a selected sample belong to the critical region is called the size of the critical region.

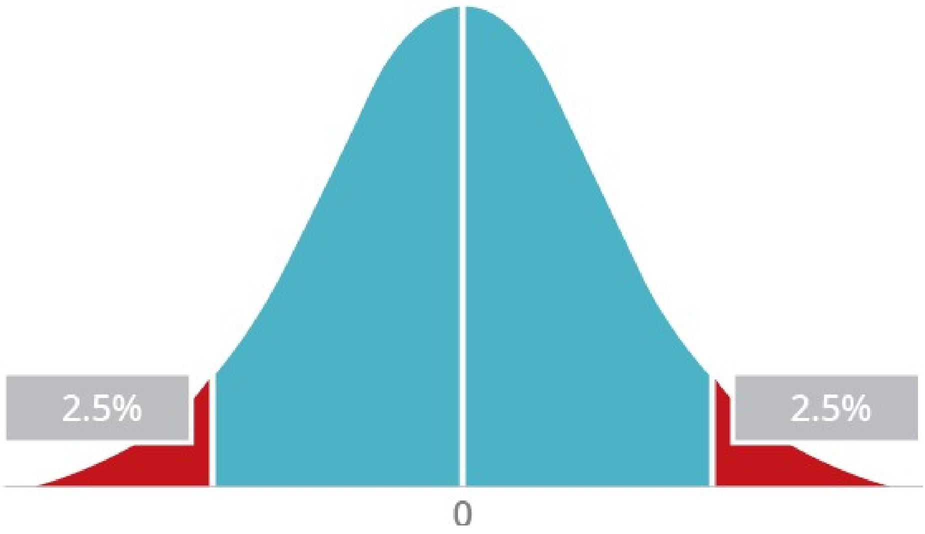
Size of the critical region is also known as significance level

Best Critical Region

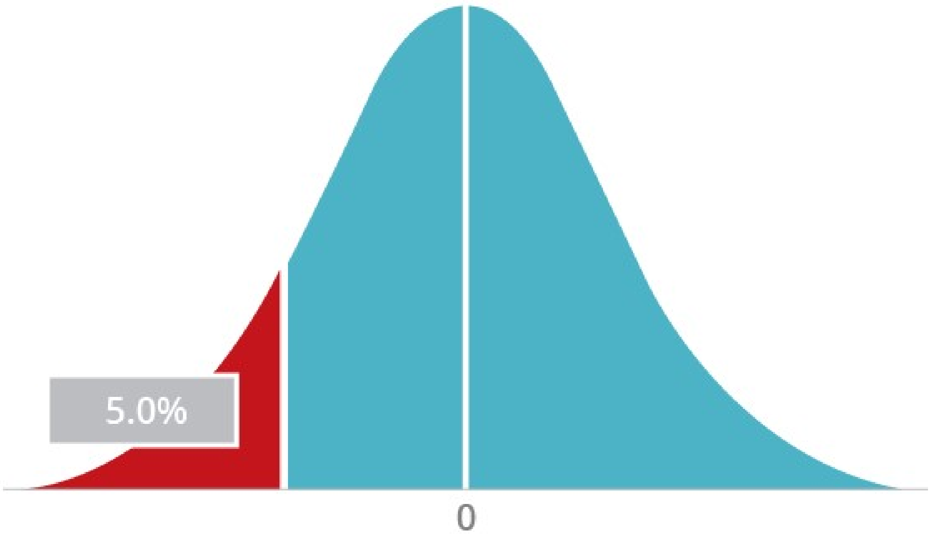
While fixing the size of the critical region we general considered significance level and try to minimize it or type I error. This result in increase in the other type of error. This is undesirable. So we have to keep the critical region such that it has least type II error. Such a region is called best critical region.

Two-tailed and one tailed Test

A two-tailed test is a **statistical test** in which the critical area of a distribution is two sided and tests whether a sample is either greater than or less than a certain range of values. If the sample that is being tested falls into either of the critical areas, the alternative hypothesis will be accepted instead of the null hypothesis. Suppose if the level of significance is 5% (0.05), then the rejection region is 2.5% on left tail and 2.5% on right tail.



A one-tailed test is a statistical test in which the critical area of a distribution is one-sided so that it is either greater than or less than a certain value, but not both. If the sample that is being tested falls into the one-sided critical area, the alternative hypothesis will be accepted instead of the null hypothesis. Suppose if the level of significance is 5% (0.05), then in the case of one tailed test the rejection region is 5% either falling in the left side only or right side only.



Steps in a Statistical Test Procedure

The different steps in testing of hypothesis is as follows

1. Define the population and formulate the hypothesis
2. Choose an appropriate test statistics
3. Divide the range of variation of the test statistics into two regions, acceptance region (*A*) and rejection or critical region (*C*) so that probability of type I error significance level has a pre-assigned value
4. Take a sample, calculate the test statistics and decide whether to accept or reject the hypothesis.