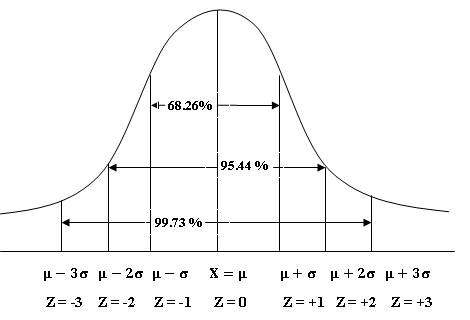
**Normal distribution and area under the normal curve**

The area under the normal probability curve between the ordinates at X= μ-σ and X= μ+σ is 0.6826.In other words, the range X = μσ covers 68.26% of the observations as shown in the figure below. This is known as 1σ limit of normal distribution



**Area under Normal Probability Curve**

**2.** The probability that random variable X lies in the interval (μ-2σ, μ+2σ) is given by

http://ecoursesonline.iasri.res.in/pluginfile.php/5354/mod_resource/content/1/Lesson%2012%20_files/image065.gif

http://ecoursesonline.iasri.res.in/pluginfile.php/5354/mod_resource/content/1/Lesson%2012%20_files/image066.gif

Hence the area under the normal probability curve between the ordinates at X= μ-2σ and X= μ+2σ is 0.95445. in other words, the range X = μ2σ covers 95.445% of the observations.

Similarly, the probability that random variable X lies in the interval (μ−3σ, μ+3σ) is given by

http://ecoursesonline.iasri.res.in/pluginfile.php/5354/mod_resource/content/1/Lesson%2012%20_files/image070.gif

http://ecoursesonline.iasri.res.in/pluginfile.php/5354/mod_resource/content/1/Lesson%2012%20_files/image071.gif

The area under the normal probability curve between the ordinates at X=μ−3σ and X= μ+3σ is 0.9973.In other words, the range X = μ2σ covers 99.73% of the observations.

Thus, the probability that a normal variate X lies outside the range  3σ is given as

http://ecoursesonline.iasri.res.in/pluginfile.php/5354/mod_resource/content/1/Lesson%2012%20_files/image074.gif

**Examples of Normal Distribution**

(i)     The age at first calving of cows belonging to the same breed and living under similar environmental conditions tend to normal frequency distribution.

(ii)   The milk yield of cows in a large herd tends to follow a normal frequency distribution.

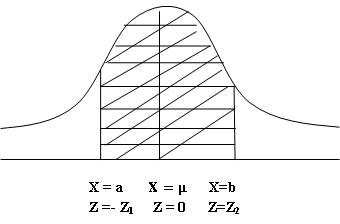
(iii) The chemical constituents of milk like fat, SNF, protein etc. for large samples follow normal distribution.

**Computation of Area Under Normal Probability Curve**

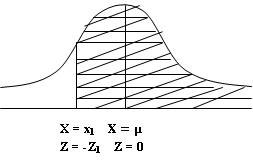
Probability that a continuous random variable X in any value between a and b is

http://ecoursesonline.iasri.res.in/pluginfile.php/5354/mod_resource/content/1/Lesson%2012%20_files/image075.gif

which is the area bounded by the curve p(x), X-axis and the ordinates at X=a and X=b and is shown



Similarly the area to the right of the ordinates at X=x1 and x1is less than mean

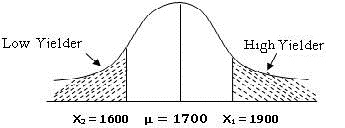


At X=X1 Z=-Z1 so P(X>X1)=0.5+P(-Z1<Z<0)=0.5+P(0<Z<Z1) , where P(0<Z<Z1)can be read from the normal tables. The application of this distribution in solving problems is illustrated through following examples.

**Example 1.** Average lactation yield for 1000 cows maintained at a farm is 1700 kg and their standard deviation is 85 kg. A cow is considered as high yielder if it has a lactation yield greater than 1900 kg and poor yielder if it has lactation yield less than 1600 kg. Find the number of high yielding and poor yielding cows.

**Solution:**

Here, =1700 kg and σ = 85 kg and let X denote the lactation milk yield



**Fig. 1**

a)   To find number of high yielder cows we first find the probability of cows yielding more than 1900 kg. i.e. P(X > 1900 kg) fig.1.  So, we first compute the value standard normal variate i.e.  Z1and then find area under shaded region using normaltables

At X1 = 1900 kg.http://ecoursesonline.iasri.res.in/pluginfile.php/5354/mod_resource/content/1/Lesson%2012%20_files/image081.gif

P(X1 > 1900 kg) = P (Z1 >2.353) =0.5 - P (0≤ Z1 ≤ 2.353) = 0.5-0.49069=0.00931

Number of high yielder cows = N xP(z1 > 2.353) = 0.00931 X 1000 = 9.31 = 9 cows

b)  To find number of low yielder cows, we first find the probability of cows yielding less than 1600 kg i.e. P(X2 < 1600 kg). So, we first compute the value standard normal variate i.e. Z2 and then find area under shaded region using normaltables

http://ecoursesonline.iasri.res.in/pluginfile.php/5354/mod_resource/content/1/Lesson_12__files/image010.gif

 P(X2 < 1600) = P (Z2 < -1.18)=0.5 - P(0≤ Z2 ≤ 1.18) = 0.5-0.38109=0.119

Number of low yielder cows = N xP(X2 < 1600) = 0.119 X 1000 = 119 cows

**Conclusion:**Total number of high yielding & low yielding cows are 9 and 119 respectively.

**Example 2.** An Intelligence test was administrated to 1000 students. The average score of students was 42 with standard deviation of 24. Find

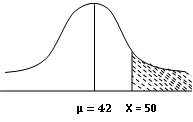
(a)  Number of students exceeding a score of 50

(b)  Number of students scoring between 30 & 58

(c)  Value of score exceeded by top 100 students.

**Solution:**In this problem****=42 and σ = 24 and let X denote the score obtained

(a)       Number of students exceeding score 50



As shown in figure 12.6 we want to find P(X>50) i.e. probability of shaded portion

At X=50,

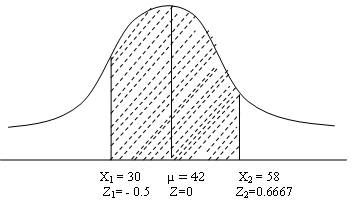
http://ecoursesonline.iasri.res.in/pluginfile.php/5354/mod_resource/content/1/Lesson%2012%20_files/image090.gif

P(X>50) =P(Z > 0.334) = 0.5 - P(0≤ Z ≤ 0.334)= 0.5 - 0.1308== 0.3692

No of students = 1000 X 0.3692= 369.2 ~ 369 students

(b) Number of students scoring between 30 and 58

As shown in figure 12.7 we want to find P(30<X<58) i.e. probability of shaded portion



http://ecoursesonline.iasri.res.in/pluginfile.php/5354/mod_resource/content/1/Lesson%2012%20_files/image092.gif

P( Z1> -0.5) = P(0≤Z1 ≤ 0.5)= 0.1915

http://ecoursesonline.iasri.res.in/pluginfile.php/5354/mod_resource/content/1/Lesson%2012%20_files/image093.gif

P(Z2<0.6667)= P(0 ≤  Z2≤ 0.6667)=0.2476

P(30<X<58)=P(-0.5≤ Z≤0.6667) =0.1915+0.2476=0.4391

 No of students = 1000 X .4391 = 439.1 ~ 439 students

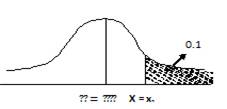
(c) Value of score exceeded by top 100 students. Let x1 be the value of score exceeded by top 100 students, the probability of top 100 students = 100/N = 100/1000 = 0.1 such that P(X>x1) = 0.1

At X= x1,

http://ecoursesonline.iasri.res.in/pluginfile.php/5354/mod_resource/content/1/Lesson_12__files/image012.gif= Z1. From Fig below the P(X>x1) shown as shaded region

P(X>x1)=P(Z>Z1)=0.1 ⇒P(0 ≤ Z ≤ Z1)=0.4 http://ecoursesonline.iasri.res.in/pluginfile.php/5354/mod_resource/content/1/Lesson_12__files/image014.gif = 1.286

x1= 72.86 ~73

****

**Conclusion**

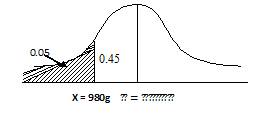
(a) 369 students scored more than 50.

(b) 439 students scored between 30 & 58.

(c) Minimum score of top 100 students is 73.

**Example 3.**  Tins are filled by an automatic filling machine with ghee. Average quantity filled in tin is 1000g. It is found that 5% of tins had ghee less than 980 grams. Find the standard deviation.

**Solution**: Here =1000g and let X be quantity of ghee filled in a tin



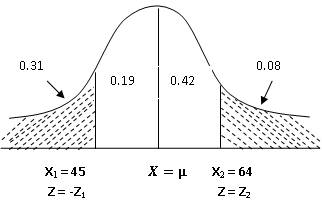
From figure above, P(X <980)=0.05 ⇒ P(Z < -Z1) = -1.645

  ⇒ (980-1000/σ) -1.645 ⇒ -20 = -1.645 σ , Hence σ =12.1840~12.18 gm

**Conclusion :**

Standard derivation of the ghee filled in tin is 12.18 gm.

**Example 4.**  In a normal distribution, 31% of items are under 45 and 8% are over 64. Find the mean & standard deviation.



**Solution:**

Let X denotes the variable under consideration. We are given that P (X1< 45) = 0.31 and P (X2 > 64) = 0.08. If X has normal distribution with mean  and standard deviation σ. Then standard normal variates corresponding to X1= 45 and X2 = 64 (from figure above) are

When X1 = 45,   http://ecoursesonline.iasri.res.in/pluginfile.php/5354/mod_resource/content/1/Lesson%2012%20_files/image103.gif

When X2 = 64,   http://ecoursesonline.iasri.res.in/pluginfile.php/5354/mod_resource/content/1/Lesson%2012%20_files/image104.gif

From the fig. 12.10, P (0 <Z < Z2) = 0.42 ⇒ Z2=  1.405 (from normal tables)

http://ecoursesonline.iasri.res.in/pluginfile.php/5354/mod_resource/content/1/Lesson%2012%20_files/image105.gif (1)

P (-Z1 <Z < 0) = 0.19 ⇒ P (0 <Z < Z1) = -0.496(from normal tables)

http://ecoursesonline.iasri.res.in/pluginfile.php/5354/mod_resource/content/1/Lesson%2012%20_files/image106.gif (2)

Solving equations (1) & (2) we get

μ= 49.95 ~ 50 and σ = 9.99 ~ 10

**Conclusion**

Mean & standard deviation of given distribution are 50 & 10 respectively.

**Importance of Normal Distribution**

Normal distribution plays a very important role in statistics because

(i)      Most of the discrete probability distributions occurring in practice e.g., Binomial and Poisson can be approximated to normal distribution as n number of trials tends to increase.

(ii)    Even if a variable is not normally distributed, it can be sometimes be brought to normal by a simple mathematical transformation, if the distribution of X is skewed, the distribution log x might come out to be normal.

(iii)    If X~N(μ,σ2) then P [μ−3σ < x < μ+3σ] =0.9973 ⇒ [|Z|˃3] = 1− 0.9973=0.0027. Thus the probability of standard normal variate going outside the limits 3 is practically zero. This property of normal distribution forms the basis of entire large sample theory.

(iv)    Many of the sampling distribution e.g., student’s t, Snedecor’s F, Chi square distributions etc tend to normality for large samples. Further, the proof of all the tests of significance in the sample is based upon the fundamental assumptions that the populations from which the samples have been drawn are normal.

(v)      The whole theory of exact sample (small sample) tests viz. t , χ2, F etc, is based on the fundamental assumption that the parent population from which the samples have been drawn follows normal distribution.

(vi)    Normal distribution finds large applications in statistical quality control in industry for setting up of control limits.

(vii)  Theory of normal curves can be applied to the graduation of the curve which is not normal.

**SAMPLING THEORY AND SAMPLING DISTRIBUTION**

**Introduction**

As earlier stated, statistical methodology is broadly studied under two heads, that is, descriptive and inferential. Descriptive statistics help in describing the characteristics of numerical data while inferential statistics or statistical inference helps in drawing valid conclusions about population in any statistical investigation on the basis of information contained in the sample which has been drawn from the same population.

**Some Basic Concepts**

**Universe or population**

A universe or population means the entire field under investigation about which knowledge is sought. It is the totality of persons, objects, items or anything conceivable pertaining to certain characteristics. In statistical usage, the term population is applied to any finite or infinite collection of individuals as per the statistical dictionary definition given by Kendall and Buckland. It is obvious that for any statistical investigation, complete enumeration of the population is rather impracticable. For example if we want to have an idea of the average per capita monthly income of the people in Kenya, we will have to enumerate all the earning individuals in the country which is rather a very difficult task because of administrative and financial implications. A population can be of two kinds (i) Finite and (ii) Infinite. In a finite population, number of items is definite such as, number of students or teachers in a college, daily milk yield of 500 milking animals in a livestock farm. On the other hand, an infinite population has infinite number of items e.g. the population of pressures at various points in the atmosphere, the population of real numbers between 0 and 1, the population of all integers, number of water drops in an ocean, number of leaves on a tree or number of hairs on the head etc.

**Sample**

A sample is finite subset of the population, selected from it by using scientific procedure with the objective of investigating its properties. More generally, a sample is a subset of a population. Thus, sample means some units selected out of a population which represent it. For example, if an investigator selects 100 animals from 2000 animals in a herd then these 100 animals will be termed as a sample and number of the individuals in the sample is called sample size.

**Sampling**

The process of selecting a sample is called sampling. It is a tool which enables us to draw conclusions about the characteristics of the population after studying only those items which are included in the sample. The main objective of sampling is

�      To obtain the maximum information about the characteristics of population with the available sources e.g. time, money, manpower etc.

�      To obtain best estimates of population parameter

**Parameter and statistic**

The statistical constants of the population like mean (μ), variance (σ2), skewness (β1), kurtosis (β2), correlation coefficient (ρ) etc. are known as parameters. Similar statistical measures computed from the sample observations e. g. mean  variance (s2), skewness (b1), kurtosis (b2), correlation coefficient(r) etc. have been termed by Prof. R. A. Fisher as statistics. Let us consider a finite population of *n* units and let y1, y2, …, ynbe the observations on the *n* units in the population. Then we have

 Mean (μ) = 1/n

http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image003.png

B v Variance =

http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image004.png

Suppose we draw a sample of size n from this population. Let y1, y2, …, yn be theobservations on the sample units. Then we can compute sample mean  and sample variance (s2) as given below:

 Mean = 

Variance = 

Usually the parameter values are not known and their estimates based on sample values are used. Thus statistics which may be regarded as an estimate of the parameter obtained from the sample is a function of sample values only and vary from sample to sample. If t is any general statistic which is a function of the sample observations y1, y2, …, ynthen  a statistic t = f (y1, y2, …, yn) is said to be unbiased estimate of population parameter θ if E(t) = θ.

**Sampling distribution**

If we draw a sample of size n from a given finite population of size N, then the total number of possible samples is NCn = k (say). For each of these samples we can compute some statistic t = t (x1, x2, …, xn) of say mean , then we have,

|  |  |  |  |
| --- | --- | --- | --- |
| **Sample number** | **Statistic** | | |
| **T** | http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image012.png | **s2** |
| 1  2  3  -  -  -  K | t1  t2  t3  -  -  -  tk | http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image013.png  http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image014.png  http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image015.png  -  -  -  http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image016.png | http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image017.png  http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image018.png  http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image019.png  -  -  -  http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image020.png |

The set of values of the statistic so obtained one for each sample constitutes what is called the sampling distribution of the statistic. For example, the values t1, t2, …, tk determine the sampling distribution of the statistic t. Since the statistic t takes different values, then it can be regarded as a random variable which can take values t1, t2, …, tk and various statistical constants like mean, variance, skewness, kurtosis etc. can be computed for the distribution.

**Standard Error**

Standard error is the standard deviation of the sampling distribution of a statistic and is

abbreviated as S.E. The standard errors of some of the well known statistics for *large samples* are given

below. where n is the sample size, the population variance, and *P* the population proportion, and *Q* = 1 *–P* nl and n2 represent the sizes of two independent random samples respectively drawn from the given population(s).

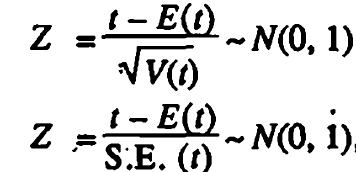
Below are the standard errors of some important statistics for large samples, are given below where n is the sample size, σ2 the population variance

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Statistic** | **Standard Error** |
| 1. | Sample mean http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image005.png | http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image024.png |
| 2. | Sample proportion p | http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image025.png |
| 3. | Sample standard deviation | http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image026.png |
| 4. | Sample variance (s2) | http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image027.png |
| 5. | Sample correlation coefficient( r) | http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image029.png |
| 6. | Difference between two sample means http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image030.png | http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image031.png |
| 7. | Difference between two sample Standard deviation  (s1-s2) | http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image032.png |
| 8. | Difference between two sample proportion (p1-p2) | http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson%2013_files/image033.png |

***Importance of Standard Error***

1. It plays a very important role in large sample theory and forms the basis of the testing of hypothesis. Thus, if the discrepancy between the observed and expected (hypothetical value of a statistic) is greater than or equal to Zα times S.E., the hypothesis is rejected at α level of significance otherwise the deviation is not regarded as significant and is considered as due to fluctuations of sampling or chance causes.

If t is a statistic, then for large samples,



Thus, if the discrepancy the observed and the expected (hypothetical)' value of statistic than  times its S.E

The null hypothesis is rejected at  level of significance. Similarly if



the deviation is not regarded significant at 5% level of significance. In other words, the deviation *t-E(t)* could have arisen due to fluctuations of sampling due to fluctuations of sampling and the data do not provide us any evidence against the null hypothesis which may therefore. be accepted at level of significance.

2). The magnitude of S.E. gives an index of the precision of the estimate of parameter. The reciprocal of the S.E. is taken as the measure of reliability or precision of the sample e.g. S.E. of sample mean and sample proportion are respectively

http://ecoursesonline.iasri.res.in/pluginfile.php/5356/mod_resource/content/1/lesson_13_files/image015.gif

which vary inversely as the square root of the sample size. Therefore reducing the S.E. increases the precision and vice versa

3) S.E. helps in determination of the probable limits/confidence limits within which the population parameter may be expected to lie.

**TESTING OF HYPOTHESIS**

A statistical hypothesis is some assumption or statement, which may or may not be true about a population, which we want to test on the basis of evidence from a random sample. It is a definite statement about population parameter which must be subjected to some test.

**Null hypothesis**

According to Prof. R. A. Fisher a hypothesis which is tested for possible rejection under the assumption that it is true is usually called Null Hypothesis and is denoted by H0.The common way of stating a hypothesis is that there is no difference between the two values, namely the population mean and the sample mean. The term no difference means that the difference, if any, is merely due to sampling fluctuations. Thus, if the statistical test shows that the difference is significant, the hypothesis is rejected. A statistical hypothesis which is stated for the purpose of possible acceptance is called Null Hypothesis. To test whether there is any difference between the two populations we shall assume that there is no difference. Similarly, to test whether there is relationship between two variates, we assume there is no relationship. So a hypothesis is an assumption concerning the parameter of the population. The reason is that a hypothesis can be rejected but cannot be proved. Rejection of no difference will mean a difference, while rejection of no relationship will imply a relationship. For example if we want to test that the average milk production of Karan Swiss cows in a lactation is 3000 litres then  the null hypothesis may be expressed symbolically as  Ho: μ = 3000 litres.

**Alternative hypothesis**

Any hypothesis which is complementary to the null Hypothesis is called an alternative hypothesis. It is usually denoted by H1orHA. For example if we want to test the null hypothesis that the population has a specified mean μoi.e. Ho: μ=μo then the alternative hypothesis could be

(i)          H1: μ ≠ μo (μ > μo or μ < μo)

(ii)        H1: μ > μo

(iii)      H1: μ < μo

The alternative hypothesis in (i) is known as two tailed alternative and the alternatives in (ii) and (iii) are known as right tailed and left tailed alternatives. The setting of alternative hypothesis is very important since it enables us to decide whether to use a single tailed (right or left) or two tailed test. The null hypothesis consists of only a single parameter value and is usually simple while alternative hypothesis is usually composite.

**Simple and composite hypothesis**

If the statistical hypothesis completely specifies the population or distribution, it is called a simple hypothesis, otherwise it is called a composite hypothesis. For example, if we consider a normal population N (, σ2) where σ2is known and we want to test the hypothesis H0: =25 against H1:  =30. From these hypotheses, we know that can take either of the two values, 25 or 30. In this case H0 and H1 are both simple. But generally H1 is composite, i.e., of the form H1: ≠25, viz, H1:  <25 or H1: >25. In sampling from a normal population N (μ, σ2), the hypothesis H: μ=μ0and σ2=σ02 is a simple hypothesis because it completely specifies the distribution. On the other hand (i) μ=μ0(σ2is not specified)(ii) σ2=σ02  (μ is not specified) (iii) μ<μ0, σ2=σ02 etc. are composite hypothesis.

**Types of errors in testing of hypothesis**

The main objective in sampling theory is to draw a valid inference about the population parameters on the basis of the sample results. In practice we decide to accept or reject a null hypothesis (H0) after examining a sample from it.  As such we are liable to commit errors. The four possible situations that arise in testing of hypothesis are expressed in the following dichotomous table:

|  |  |  |
| --- | --- | --- |
| Decision from sample | **True Situation** | |
| Hypothesis is true | Hypothesis is false |
| Accept the hypothesis | No error | Type II error |
| Reject the hypothesis | Type I error | No error |

In testing hypothesis, there are two possible types of errors which can be made. The error of rejection of a hypothesis H0 when H0is true is known as Type I error and error of acceptance of a hypothesis H0 when H0is false is known as type II error. When setting up an experiment to test a hypothesis it is desirable to minimize the probabilities of making these errors. But practically it is not possible to minimize both these errors simultaneously. These two types of errors can be better understood with an example where a patient is given a medicine to cure some disease and his condition is examined for some time. It is just possible that the medicine has a positive effect but it is considered that it has no effect or adverse effect. Therefore it is the Type I error. On the other hand if the medicine has an adverse effect but it is considered to have had a positive effect, it is called Type II error. Now let us consider the implications of these two types of error. If type I error is committed, the patient will be given another medicine, which may or may not be effective. But if type II error is committed i.e., the medicine is continued inspite of an adverse effect, the patient may develop some other complications or may even die. This clearly shows that the type II error is much more serious than the type I error. Hence in drawing inference about the null hypothesis, generally type II error is minimized even at the risk of committing type I error which is usually chosen as 5 per cent or 1 per cent.

Probability of committing type I error and type II error are denoted by α and β and are called size of type I and type II error respectively. In Industrial Quality Control, while inspecting the quality of a manufactured lot, the Type I error and type II error amounts to rejecting a good lot and accepting a bad lot respectively. Hence α=P(Rejecting a good lot) and β=P( Accepting a bad lot). The sizes of type I and type II errors are also known as producer’s risk and consumer’s risk respectively. The value of (1-β) is known as the power of the test.

**Level of significance**

It is the amount of risk of the type I error which a researcher is ready to tolerate in making a decision about H0. In other words, it is the maximum size of type I error, which we are prepared to tolerate is called the level of significance. The level of significance denoted by α is conventionally chosen as 0.05 or 0.01. The level of 0.01 is chosen for high precision and the level 0.05 for moderate precision. Sometimes this level of risk is further brought down in medical statistics where the efficiency of life saving drug on the patient is tested. If we adopt 5% level of significance, it means that on 5 out of 100 occasions, we are likely to reject a correct H0 .In other words this implies that we are 95% confident that our decision to reject Hois correct.. That is, we want to make the significance level as small as possible in order to protect the null hypothesis and to prevent, as far as possible, the investigator from inadvertently making false claims. Level of significance is always fixed in advance before collecting the sample information.

**P-value concept**

Another approach followed in testing of hypothesis is to find out the P-value at which H0 is significant i.e., to find the smallest level α at which H0 is rejected. In this situation, it is not inferred whether H0 is accepted or rejected at a level of 0.05 or 0.01 or any other level. But the researcher only gives the smallest level α at which H0 is rejected. This facilitates an individual to decide for himself as to how much significant the research results are. This approach avoids the imposition of a fixed level of significance. About the acceptance or rejection of H0, the experimenter can himself decide the level of α by comparing it with the P-value. The criterion for this is that if the P-value is less than or equal to α, reject H0 otherwise accept H0.

**Degrees of freedom**

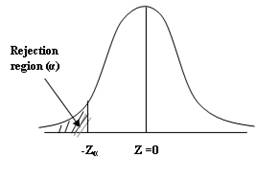
For a given set of conditions, the number of degrees of freedom is the maximum number of variables which can freely be designed (i.e., calculated or assumed) before the rest of the variates are completely determined. In other words, it is the total number of variates minus the number of independent relationships existing among them. It is also known as the number of independent variates that make up the Statistic. In general, degree of freedom is the total number of observations (n) minus the number of independent linear constraints (k) i.e. n-k.

**Critical region**

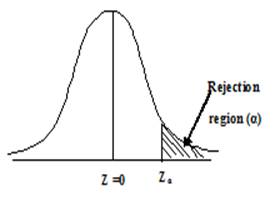
The total area under a standard curve is equal to one representing probability distribution. In test of hypothesis the level of significance is set up in order to know the probability of making type I error of rejecting the hypothesis which is true. A statistic is required to be used to test the null hypothesis H0. This test is assumed to follow some known distribution. In a test, the area under the probability density curve is divided into two regions, viz, the region of acceptance and the region of rejection. If the value of test statistics lies in the region of rejection, the H0 will be rejected. The region of rejection is also known as a critical region. The critical region is always on the tail of the distribution curve. It may be on both sides of the tails or on one side of the tail depending upon alternative hypothesis H1.

***One - tailed test***

A test of any statistical hypothesis where the alternative hypothesis is one tailed (right-tailed or left- tailed) is called a one tailed test. For example, a test for testing the mean of a population H0: μ=μ0 against the alternative hypothesis H1: μ>μ0 (Right - tailed) or H1: μ<μ0 (Left- tailed) is a single-tailed test. If the critical region is represented by only one tail, the test is called one-tailed test or one-sided test. In right tailed test (H1: μ>μ0) the critical region lies entirely on the right tail of the sampling distribution of http://ecoursesonline.iasri.res.in/pluginfile.php/5362/mod_resource/content/1/Lesson_16_files/image002.gifas shown in Fig, 2 , while for the left tail test (H1: μ<μ0 ), the critical region is entirely in the left tail of the distribution of as shown in Fig. 1

******

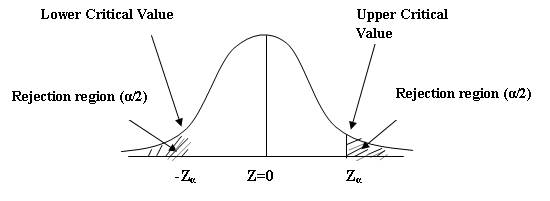
**Fig. 1 Left tailed Test**

******

**Fig. 2 Right tailed Test**

***Two-tailed test***

A test of statistical hypothesis where the alternative hypothesis is two-sided such as: H0: μ=μ0 against the alternative hypothesis H1: μ≠μ0(μ>μ0 or μ<μ0) is known as a two-tailed test and in such a case the critical region is given by the portion of the area lying on both the tails of the probability curve of the test statistic as shown in Fig. 3



**Fig. 3 Two Tailed Test**

In a particular problem, whether one-tailed or two-tailed test is to be applied depends entirely on the nature of the alternative hypothesis.

**Critical values or significant values**

The value of a test statistic which separates the critical (or rejection) region and the acceptance region is called the critical value or significant value. It depends upon

* the level of significance and
* the alternative hypothesis, whether it is two tailed or single tailed.

The critical value of the test statistic at α level of significance for a two tailed test is given by Zα where Zα is determined by the equation P(|Z| > Zα)*,*where Zαis the value so that the total area of the critical region on both tails is α. Since normal probability curve is symmetric curve, we get P[Z>Zα] = α/2 i.e., the area of each tail is α/2. Thus Zα is the value such that area to the right of Zα is α/2 and to the left of - Zα is α/2 as shown in fig. 16.3 .In case of single tail alternative, the critical value of Zα is determined so that total area to the right of it (for right tailed test) is α (as shown in fig. 2) and for left tailed test the total area to the left of - Zα is α (as shown in Fig. 1).Thus the significant or critical value of Z for a single tailed test (left or right) at level of significance α is same as the critical value of Z for a two tailed test at level of significance 2α.

**Table 2 Critical values of (Z𝛂) of Z**

|  |  |  |  |
| --- | --- | --- | --- |
| **Critical values**  **(Zα)** | **Level of significance** | | |
| **1%** | **5%** | **10%** |
| Two tailed test  Right tailed test  Left tailed test | http://ecoursesonline.iasri.res.in/pluginfile.php/5362/mod_resource/content/1/Lesson%2016_files/image009.gif  Zα = 2.33  Zα = -2.33 | http://ecoursesonline.iasri.res.in/pluginfile.php/5362/mod_resource/content/1/Lesson%2016_files/image010.gif  Zα = 1.645  Zα = -1.645 | http://ecoursesonline.iasri.res.in/pluginfile.php/5362/mod_resource/content/1/Lesson%2016_files/image011.gif  Zα = 1.28  Zα = -1.28 |

**Test of Significance**

The tests of significance which are dealt hereafter pertain to parametric tests. A statistical test is defined as a procedure governed by certain rules, which leads to take a decision about the hypothesis for its acceptance or rejection on the basis of sample observations. Test of significance enables us to decide on the basis of sample results if

(i) deviation between observed sample statistic and the hypothetical parameter value or

 (ii) the deviation between two sample statistics is significant.

Test of significance is a procedure of either accepting or rejecting a Null hypothesis. The tests are usually called tests of significance since here we test whether the difference between the sample values and the population values or between the values given by two samples are so large that they signify evidence against the hypothesis or these differences are small enough to be accounted for as due to fluctuations of sampling, i.e. they may be regarded as due only to the fact that we are dealing with a sample and not with the whole population. Statistical tests play an important role in biological sciences, dairy industry, social sciences and agricultural sciences etc. The use of these tests is made clear through a number of practical examples:

1.      An automatic machine is filling 500 ml. of milk in the pouch. Now to make sure whether the claim is correct or not one has to take a random sample of the filled in pouches and note the actual quantity of milk in the pouches. From these sample observations it would be decided whether the automatic machine is filling the right quantity of milk in the pouches. This is done by performing test of significance.

2.      There is a process A which produced certain items. It is considered that a new process B is better than process A. Both the processes are put under operation and then items produced by them are sampled and observations are taken on them. A statistical based test on sample observations will help the investigator to decide whether the process B is better than A or not.

3.      Psychologists are often interested in knowing whether the level of IQ of a group of students is up to a certain standard or not. In this case some students are selected and an intelligence test is conducted. The scores obtained by them are subjected to certain statistical test and a decision is made whether their IQ is up to the standard or not.

There is no end to such types of practical problems where statistical tests can be applied. Here one important point may be noted. Whatever conclusions are drawn about the population (s), they are always subjected to some error.

**Steps in hypothesis testing**

Various steps in test of significance are as follows:

         (i)   Set up the Null hypothesis Ho.

         (ii) Set up the alternative hypothesis H1 .This will decide whether to go for single tailed test or two tailed test.

         (iii) Choose the appropriate level of significance depending upon the reliability of the estimates and permissible risk. This is to be decided before sample is drawn.

       (iv) Compute the test statistic http://ecoursesonline.iasri.res.in/pluginfile.php/5362/mod_resource/content/1/Lesson_16_files/image004.gif

          (v)  Compare the computed value of Z in previous step with the significant value Zα at a given level of significance.

         (vi)  Conclusion :

a)      If |Z| < Zα i.e. if calculated value of Z (test statistic) is less than Zα,we say it is not significant, null hypothesis is accepted at level of significance α.

b)      If |Z| > Zα i.e. if calculated value of Z (test statistic) is greater than Zα, we say it is significant and null hypothesis is rejected at level of significance α.

**Z-Test**

As discussed before in the importance of normal distribution, for large sample size n, say n >30 almost all distributions for example, Binomial, Poisson are approximated by normal distribution and hence the standard normal variate or Z-test can be applied. The distribution of Z is always normal with mean zero and variance one.

**Test of Significance for Large Samples**

If X ~ N (μ, σ2), and its standard variate Z having a mean zero and variance one i.e. Z ~ N (0, 1),

http://ecoursesonline.iasri.res.in/pluginfile.php/5364/mod_resource/content/1/Lesson%2017_files/image001.png

From normal probability tables, it was seen that

P[-3≤Z≤3]=P[|Z| ≤ 3] = 0.9973 ⇒ P[|Z| ≤ 3] = 1-P[|Z| ≤ 3] = 0.0027 which gives 99.73% acceptance region. In hypothesis testing, the commonly used levels of significance are 10%, 5% and 1% with corresponding critical values of Z as shown in table 2 above.

**Applications of Z -Test**

1. **Test for single proportion**

Consider two binomial populations and where observations on various items or objects are made on them. Remember in binomial distributions, there only two possible outcomes viz. defective or not defective item, and we may be interested in testing the hypothesis, of whether the proportion of say defective items in a population is P0.

If in a sample of size n, X be the number of persons possessing the given attributes then observed proportion of successes

http://ecoursesonline.iasri.res.in/pluginfile.php/5364/mod_resource/content/1/Lesson%2017_files/image007.png

http://ecoursesonline.iasri.res.in/pluginfile.php/5364/mod_resource/content/1/Lesson%2017_files/image008.png

http://ecoursesonline.iasri.res.in/pluginfile.php/5364/mod_resource/content/1/Lesson%2017_files/image009.png

http://ecoursesonline.iasri.res.in/pluginfile.php/5364/mod_resource/content/1/Lesson%2017_files/image010.png

For large n, the normal deviate test for the proportion of success becomes.

http://ecoursesonline.iasri.res.in/pluginfile.php/5364/mod_resource/content/1/Lesson%2017_files/image012.png

**Example**

 In a large consignment of baby food packets, a random sample of 100 packets revealed that 5 packets were leaking. Test whether the sample comes from the population (large consignment) containing 3 percent leaked packets.

**Solution:** In this example, we have n=100, X=5, P=0.03,

http://ecoursesonline.iasri.res.in/pluginfile.php/5364/mod_resource/content/1/Lesson%2017_files/image013.png

1. Statement of hypothesis

H0: P = 0.03 .i.e., the proportion of the leaked pouches in the population is 3 per cent

H1: P ≠ 0.03. (two tailed test)

1. Since we are not given the level of significance, we can choose at our discretion. Assuming we choose to test at 5% level of significance, then the critical value of Z at 5% for two tailed test is



Test criterion

1. Since the sample size is large, we use the standard normal deviate (Z) test for single proportion

http://ecoursesonline.iasri.res.in/pluginfile.php/5364/mod_resource/content/1/Lesson%2017_files/image014.png

1. Decision making

Since calculated value of Z=1.17 is less than  therefore we fail to reject H0 at 5% level of significance implying that the sample is from the population (large consignment) of packets containing 3% leaked packets.