**SAMPLING THEORY**

**Sample:** A small section selected from the population is called a sample.

**Sampling:** The process of drawing a sample is called sampling.

**Random sampling:** A sample must be a random selection so that each member of the population has the same chance of being included in the sample. Thus the fundamental assumption underlying theory of sampling is Random sampling.

**Simple sampling:** A special case of random sampling in which each event has the same probability p of success and the chance of success of different events are independent whether previous trials have been made or not, is known as simple sampling.

**Parameters:**The statistical constants of the population such as mean (µ), standard deviation etc. are called the parameters.

**Statistic:** The constants for the sample drawn from the given population i.e., mean (x), standard deviation (S) etc. are called the statistic.

**Objectives of sampling:**

1). Sampling aims at gathering the maximum information about the population with the minimum effort, cost and time.

2). the object of sampling studies is to obtain the best possible values of the parameters under specific conditions.

3). Sampling determines the reliability of these estimates.

**Statistical Inference:**

The logic of the sampling theory is the logic of induction in which we pass from a particular (sample) to general (population). Such a generalisation from sample to population is called **Statistical Inference.**

**Sampling distribution:**

Consider all possible samples of size n which can be drawn from a given population at random. For each sample, we can compute the mean. The means of the samples will not be identical. If we group these different means according to their frequencies, the frequency distribution so formed is known as sampling distribution of the mean. Similarly we can have sampling distribution of the standard deviation etc.

While drawing each sample, we put back the previous sample so that the parent population remains the same. This is called sampling with replacement and all the subsequent formulae will pertain to sampling with replacement.

**Standard error:**

The standard deviation of the sampling distribution is called the standard error (S.E.). Thus the standard error of the sampling distribution of means is called standard error of means. The standard error is used to assess the difference between the expected and observed values.

**Small and Large samples:**

If the size of the sample is greater than or equal to 30 the sample is called large sample. Otherwise the sample is called as Small sample

**Precision:**The reciprocal of the standard error is called precision.

**Statistical hypothesis:**

To reach decisions about populations on the basis of sample information, we make certain assumptions or guesses about the populations involved. Such assumptions, which may or may not be true, are called statistical hypothesis.

**Testing a Hypothesis:**

Testing a hypothesis is meant a process for deciding whether to accept or reject the hypothesis.

The method consists in assuming the hypothesis as correct and then computing the probability of getting the observed sample. If this probability is less than a certain pre-assigned value the hypothesis is rejected.

**Type-I, Type-II Errors:**

If a hypothesis is rejected while it should have been accepted, we say that a Type I error has been committed. This is similar to a good product is rejected by the consumer and hence Type-I error is also known as **producer’s risk**

If a hypothesis is accepted while it should have been rejected, we say that a Type II error has been made. This is similar to of accepting a product of inferior quality and hence Type-II error is also known as **consumer’s risk**

The statistical testing of hypothesis aims at limiting the Type I error to a pre-assigned value (say 5 % or 1%) and to minimize the Type II error. The only way to reduce both types of errors is to increase the sample size, if possible.

**Null hypothesis:**

The hypothesis formulated for the sake of rejecting it, under the assumption that it is true, is called the null hypothesis and is denoted by H0.

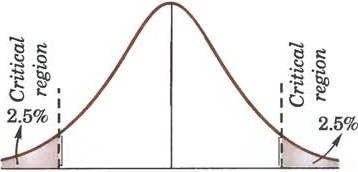
We set up a hypothesis which assumes that there is no significance difference between the sample statistic and corresponding population parameter of between two samples.

To test whether one procedure is better than another, we assume that there is no difference between the procedures.

Similarly to test whether there is a relationship between two variates, we take H0 that there is no relationship. By accepting a null hypothesis, we mean that on the basis of the statist calculated from the sample, we do not reject the hypothesis. It however, does not imply that the hypothesis is proved to be true. Nor its rejection implies that it is disproved.

**Level of Significance:**

The probability level below which we reject the hypothesis is known as the level of significance.

****

This “standard normal score” or “*z*-score” expresses the value of a variable in terms of its distance from the mean, measured in standard deviations. (Thus if μ = 1000 and » = 50, then the value *x* = 850 has a *z*-score of −3.0: it lies 3 standard deviations below the mean.) We can subscript *z* to indicate the proportion of the standard normal distribution that lies to its right. For instance, since the normal distribution is symmetrical, *z*0.5 = 0. It follows from points made earlier that *z*0.025 = 1.96 and *z*0.005 = 2.58. A picture may help to make this obvious.

*z*.975 = −1.96 *z*.025 = 1.96

**Critical region:**

The region, in which a sample value falling is rejected, is known as the critical region.

We generally take two critical regions which cover 5% and 1% areas of the normal curve. The shaded portion in the figure to 5% level of significance.

Thus the probability of the value of the variate falling in the critical region is the level of significance.

Depending on the nature of the problem, we use a single-tail test double-tail test to estimate the significance of a result. In a double-tail test, the areas of both the tails of the curve representing the sampling distribution are taken into account whereas in the single tail test, only the area on the right of an ordinate is taken into consideration. For instance, to test whether a coin is biased or not, double-tail test should be used, since a biased coin gives either more number of heads than tails (which corresponds to right tail), or more number of tails than heads (which corresponds to left tail only).

**Test of significance:**

The procedure which enables us to decide whether to accept or reject the hypothesis is called the test of significance.

Here we test whether the differences between the sample values and the population values (or the values given by two samples) are so large that they signify evidence against the hypothesis or these differences are so small as to account for fluctuations of sampling.

**Confidence Limits:**

Suppose that the sampling distribution of a statistic S is normal with mean µ and standard deviation. The sample statistic S can be expected to lie in the interval (µ - l.96, µ + l.96) for 95% times i.e., we can be confident of finding µ in the interval (S - 1.96, S + 1.96 ) in 95 % cases. Because of this, we call(S - l.96, S + 1.96) the 95% confidence interval for estimation of µ. The ends of this interval (i.e. S ± 1.96) are called 95% confidence limits (or fiducial limits) for S.

Similarly S ± 2.58 are 99% confidence limits. The numbers 1.96, 2.58 etc. are called confidence coefficients. The values of confidence coefficients corresponding to various levels of significance can be found from the normal curve area table.

**Simple Sampling of Attributes:**

The sampling of attributes may be regarded as the selection of samples from a population whose members possess the attribute K or not K. The presence of K may be called a success and its absence a failure.

Suppose we draw a simple sample of n items. Clearly it is same as a series of n independent trials with the same probability p of success. The probabilities of 0, **1,** 2, ..., n successes are the terms in the binomial expansion of  where q = 1 - p.

We know that the mean of this distribution is np and standard deviation is i.e., the expected value of success in a sample of size n is np and the standard error is 

If we consider the proportion of successes, then

* 1. mean proportion of successes = np/n = p.
  2. standard error of the proportion of successes



and (iii) precision of the proportion of successes= , which varies as √n,since p and q are constants.

### Test of Significance for Large Samples:

If x is the observed number of successes in the sample and z is the standard normal variate then



If <1.96, difference between the observed and expected number of successes is not significant.

If  > 1.96, difference is significant at 5% le1Jel of significance.

If  > 2.58, difference is significant at 1% level of significance.

**Procedure for Testing of Hypothesis:**

**Step1:** Define Null Hypothesis H0

**Step2:** Define alternative Hypothesis. After a care full of study of the given problem decide the nature of the test is decided (whether one -tailed or two-tailed).

**Step3:** Level of significance (LOS) is fixed

**Step4:** The test –statistic  is computed.

**Step5:** Comparison is made between 

1. If  then H0 is accepted or H1 is rejected, it is conclude that the difference between ‘x’ and E(x) is not significant at 
2. If  then H0 is rejected or H1 is accepted, it is conclude that the difference between ‘x’ and E(x) is significant at 

**TEST OF SIGNIFICANCE OF LARGE SAMPLES:**

Test Statistic for Mean of single sample is 

Test Statistic for Means of two large samples is 

Confidence interval for large samples is  and (single mean)

1. A sample of 400 items is taken from a normal population whose mean is 4 and variance is 4. If the sample mean is 4.45, can the samples be regarded as a simple sample.

Soln.: Null hypothesis :- 

Alternative hypothesis :- 

Level of Significance (LOS):- 

Test Statistics:- 

Given, 



Conclusion:



, i.e. is rejected, hence the sample is not simple.

1. The mean of two large samples of 1000 and 2000 members are 168.75 cms and 170 cms respectively . Can the samples be regarded as drawn from the same population of S.D 6.25 cms?

Soln.: Null hypothesis :- 

Alternative hypothesis :- 

Level of Significance (LOS):- 

Test Statistics:- 

Given, 



Conclusion:



, i.e. is rejected, hence the samples are not drawn from the same population.

1. A random sample of 1000 men from North India shows that their mean wage is Rs. 5 per day with a S.D of Rs. 1.50. A sample of 1500 men from South India gives a mean wage of Rs. 4.50 per day with S.D of Rs. 2. Does the mean rate of wages varies as between the two regions.

Soln.

Null hypothesis :- 

Alternative hypothesis :- 

Level of Significance (LOS):- 

Test Statistics:- 

Given, 



Conclusion:



, i.e. is rejected, hence the rate of wages varies between the regions.

1. A normal population has a mean of 0.1 & S.D of 2.1 . Find the probability that the mean of a sample of size 900 drawn from this population will be negative.

Soln. 



**TEST OF SIGNIFICANCE OF SMALL SAMPLES:**

**\*\*** Test Statistic for Large samples is  with degree of freedom = n – 1,

where  and 

\*\* Comparison is made between 

1. If  then H0 is accepted or H1 is rejected, it is conclude that the difference between andis not significant at 
2. If  then H0 is rejected or H1 is accepted, it is conclude that the difference between and  is significant at 

\* \* Confidence interval for large samples is  .

\*Test statistic for two small samples is  where and 

and 

\*\*If the two samples are of the same size and the data are paired,

Then ‘ t ‘ is defined by  where . Here  is the difference of the members of the samples and  is the mean of the differences with degree of freedom = n – 1.

1. Find the student’s t for the following variable values in a sample of eight, -4,-2,-2,0,2,2,3,3 taking the mean of the universe to be zero.

Soln. 

Given that 









|  |  |  |
| --- | --- | --- |
|  | Boys | Girls |
| Mean | 124 | 121 |
| S.D | 12 | 10 |
| n | 18 | 14 |

1. A group of boys & girls were given an intelligence test. The mean score, S.D and numbers in each group are follows.

Is the mean score of boys significantly differ from girls?

Soln.

Null hypothesis :- 

Alternative hypothesis :- 

Level of Significance (LOS):- 

Test Statistics:- 

Where , 



Given that 



Now 

Conclusion:



, i.e. is accepted, hence the man scores are not significantly differ.

1. A sample of 10measurements of the diameter of a sphere gave a mean of 12 cm & S.D 015 cm, find 95% confidence limits for the actual diameter.

Soln. Confidence limits =

given that , 

Confidence limits =

1. A random sample of size 25 from a normal population has the mean & S.D=8.4. Does this information refuse the cliam that the mean of the population is .

Soln. Null hypothesis :- 

Alternative hypothesis :- 

Level of Significance (LOS):- 

Test Statistics:- 



Conclusion:



, i.e. is rejected, hence the information refuse the claim.

**CHI-SQUARE () TEST:**

If be a set of observed frequencies and  be the corresponding set of expected frequencies, the is defined by the relation,

 With n-1 degree of freedom.

**GOODNESS OF FIT:**

The value of is used to test whether the deviations of the observed frequencies from the expected frequencies are significant or not. It provides a test of goodness of fit and may be used to examine the validity of some hypothesis about an observed frequency distribution. Also it is non-parametric test.

**PROCEDURE TO TEST SIGNIFICANCE AND GOODNESS OF FIT:**

(i) Set up a null hypothesis and calculate 

(II) Find the degree of freedom and read the corresponding values of  at prescribed significance level from the table.

(iii) If < or < with given degree of freedom then the distribution is a good fit otherwise not a good fit.

1. Genetic theory states that children having one parent of blood type M and other blood type N will always be one of the three types M, MN, N and that the proportions of these types will on average be 1: 2: 1. A report states that out of 300 children having one M parent and one N parent, 30% were found to be of type M, 45%of type MN and remainder of type N. Test the hypothesis by Chi-square test.

Soln. Null hypothesis :- 

Alternative hypothesis :- 

Level of Significance (LOS):- 

Test Statistics:- 

The Ratio of M,MN, N are 1:2:1

1+2+1 =4

Total no. Of children = 300

Expected value of type M =

,, ,, of type MN =

,, ,, of type N =

Observed value of type M= 30% of 300



Observed value of type MN= 45% of 300



Observed value of type N= 25% of 300





for n-1 d.o.f ( is 5.99

Conclusion:- , hence is accepted.

The given data supports the genetic theory.

1. Fit a binomial distribution for the following Data & also test the goodness of fit.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **X** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **Total** |
| **F** | **5** | **18** | **28** | **12** | **7** | **6** | **4** | **80** |

Soln. 



Null hypothesis :- Binomial is a good fit

Alternative hypothesis :- Binomial is not a good fit

Level of Significance (LOS):- 

Test Statistics:- 

Using 















|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Oi | 5 | 18 | 28 | 12 | 7 | 6 | 4 |
| Ei | 4 | 15 | 25 | 22 | 11 | 3 | 0 |

The first class is combined with the second & the last classes are combined with the last, but second class in order to make the expected frequency in each class >10, we have

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Oi | 23 | 28 | 12 | 17 |
| Ei | 19 | 25 | 22 | 14 |



for 5%, d.o.f =4-2=2 is given by 5.99

Conclusion:-

, hence is rejected.

Hence Binomial is not a good fit.

**\**

**ASSIGNMENT QUESTIONS**

1. A sample of 900 members is found to have a mean of 3.4 cm. Can it be reasonably regarded as a truly random sample from a large population with mean 3.25 cm and S.D. 1.61 cm?
2. An unbiased coin is thrown n times. It is desired that the relative frequency of the appearance of heads should lie between 0.49 and 0.51. Find the smallest value of n that will ensure this result with 90% confidence.
3. A sample of height of 6400 soldiers has a mean of 67.85inches and a standard deviation of 2.56 inches while a simple sample of h-eights of 1600 sailors has a mean of 68.55 inches and a standard deviation of 2.52 inches. Do the data indicate that the sailors are on the average taller than soldiers?
4. A sample of 400 items is taken from a normal population whose mean is 4 and variance 4. If the sample mean is 4.45, can the samples be regarded as a simple sample?
5. To know the mean weights of all 10-year old boys in Delhi, a sample of 225 is taken. The mean weight of the sample is found to be 67 pounds with a S.D. of 12 pounds. Can you draw any inference from it about the mean weight of thepopulation?
6. If 60 new entrants in a given university are found to have a mean height of 68.60 inches and 50 seniors a mean height of 69.51 inches; is the evidence conclusive that the mean height of the seniors is greater than that of the new entrants? Assume the standard deviation of height to be 2.48 inches.
7. Asample of 100 electric bulbs produced by manufacturer A showed a mean life time of 1190 hours and a standard deviation of 90 hours. Asample of 75 bulbs produced by manufacturer B showed a mean life time of 1230 hours**,** with a standard deviation of 120 hours. Is there a difference between the mean life time of two brands at a significance level of (i) 0.05 (ii) 0.01.
8. A random sample of 1000 men from North India shows that their mean wage is Rs. 5 per day with a S.D. of Rs. 1.50. Asample of 1500 men from South India gives a mean wage of Rs. 4.50 per day with a standard deviation of Rs. 2. Does the mean rate of wages varies as between the two regions?
9. A certain stimulus administered to each of 12 patients resulted in the following increases of blood pressure: 5, 2, 8, -1, 3, 0, -2, 1, 5, 0, 4, 6. Can it be concluded that the stimulus will in general be accompanied by an increase in blood pressure.
10. From a random sample of 10 pigs fed on diet A, the increases in weight in a certain period were 10, 6, 16, 17, 13, 12, 8, 14, 15, 9 lbs. For another random sample of 12 pigs fed on diet B, the increases in the same period were 7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17 lbs. Test whether diets A and B differ significantly as regard s their effect on increases in weight?
11. Find the student's t for the following variable values in a sample of eight: -4**,** -2,-2, 0, 2, 2, 3, 3; taking the mean of the universe to be zero.
12. Asample of 10 measurements of the diameter of a sphere gave a mean of 12 cm and standard deviation 0.15 cm. Find 95% confidence limits for the actual diameter.
13. Arandom sample of size 25 from a normal population has the mean 47.5 and standard deviation 8.4. Does this information refute the claim that the mean of the population is µ = 42.1.
14. Find out the reliability of the sample mean of the following data: Breaking strength of 10 specimens of 1.04 cms diameter hard-drawn copper wire:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Specimen l | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Breaking Strength (kgs) : 578 | 572 | 570 | 568 | 572 | 570 | 570 | 572 | 526 | 584 |

1. Two horses A and B were tested according to the time (in seconds) to run a particular race with the following results:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Horse A : | 28 | 30 | 32 | 33 | 33 | 29 | and 34 |
| Horse B : | 29 | 30 | 30 | 24 | 27 | and 29 |  |

Test whether you can discriminate between two horses?

1. Agroup of 10 rats fed on a diet A and another group of 8 rats fed on a different diet B, recorded the following increase in weights:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Diet A | 5 | 6 | 8 | 1 | 12 | 4 | 3 | 9 | 6 | 10 | gm |
| Diet B | 2 | 3 | 6 | 8 | 10 | 1 | 2 | 8 | gm. |  |  |

Does it show that superiority of diet A over that of B?

1. The means of two random samples of sizes 9 and 7 are 196.42 and 198.82 respectively. The sum of the squares of the deviations from the mean is 26.94 and 18.73 respectively. Can the sample be considered to have been drawn from the same normal population?
2. A set of five similar coins is tossed 320 times and the result is

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. of heads : | 0 | 1 | 2 | 3 | 4 | 5 |
| Frequency | 6 | 27 | 72 | 112 | 71 | 32 |

Test the hypothesis that the data follow a binomial distribution.

1. Fita Poisson distribution to the following data and test for its goodness of fit at level of significance 0.05.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| x: | 0 | 1 | 2 | 3 | 4 |  | |
| f: | 419 | 352 | 154 | 56 | 19 |  |  |

1. Five dice were thrown 96 times and the number of times 4, 5 or 6 were thrown were:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. of dice showing 4, 5 or 6: | 5 | 4 | 3 | 2 | 1 | 0 |
| Frequency | 8 | 18 | 35 | 24 | 10 | 1 |

Find the probability of getting this result by chance?

1. The frequencies of localities according to the number of deaths due to cholera during eight years in 1000 localities is as follows:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No. of deaths | 0 | 1 | 2 | 3 | **4** | 5 | 6 | 7 |
| No. Of localities | 314 | 355 | 204 | 86 | 29 | 9 | 3 | 0 |

Fit a suitable distribution to the data andt.est the goodness of fit.

1. Two samples of sizes **9** and 8 give the sum of squares of deviations from their respective means equal to 160 sq. inches and 91 sq. inches respectively. Can these be regarded as drawn from the same normal population?
2. Measurements on the length of a copper wire were taken in 2 experiments A and B as under:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A's measurements (mm) | 12.29 | 12.25 | 11.86 | 12.13 | 12.44 | 12.78 | 12.77 | 11.90 | 12.47 |
| B's measurements (mm) | 12.39 | 12.46 | 12.34 | 12.22 | 11.98 | 12.463 | 12.23 | 12.06 |  |

Test whether B's measurements are more accurate than A's. (The readings taken in both cases being unbiased)

1. Two samples of 9 and 7 individuals have variances 4.8 and 9.6 respectively. Is the variance 9.6 significantly greater than the variance 4.6?
2. Test for breaking strength were carried out on two lots of 5 and 9 steel wires. The variance of one lot was 230 and that of other was 492. Is there a significant difference in their variability?
3. Show how you would use Fisher's z-test to decide whether the two sets of observations 17, 27, 18**,** 25, 27, 29, 27, 23, 17 and 16, 16, 20, 16, 20, 17, 15, 21, indicate samples from the same universe.
4. In two groups of ten children each, the increase in weight due to different diets during the same period. were in pounds

3, 7, 5, 6, 5, 4**,** 4, 5, 3, 6

8, 5, 7, 8, 3, 2, 7, 6, 5, 7

Is there a significant difference in their variability?