Data play an important role in any statistical investigation. Sampling is a popular method to collect the data. The fundamental assumption behind the sampling method is that if the units of a sample are selected at random, its characteristics will almost be same as they exists in the universe.

Sample

Sample is a small part representing universe and its salient features .

" A sample is that part of the universe which we select for the purpose of investigation."

Importance of sampling

Samples are devices for learning about large masses by observing a few individuals. In fact is that we are living in the age of sampling.

Merits

1. **Economic method** We dont need to investigate the whole population.
2. **Saving of time and Labour** Save both not only in conducting the sampling enquiry but also in the processing, editing and analysing the data.
3. **Testing of accuracy** Sampling methods allow us to investigate the accuracy by comparing the result of two or more samples.
4. **Detailed and intensive enquirey** As the number of units investigated are limited hence it is possible to study them in detail and intensively.
5. **Only method in many cases** If the population is too large or if the testing of units is destructive then we are left with no other way nut to use sampling.

Demerits

1. **misleading results** If a sample survey is not properly planned and carefully executed.
2. **Need of specialised knowledge**
3. **Heterogeneous units** If the units of the population are too heterogeneous.

Sampling theory

It is a study of relationship existing between a population and samples drawn from the population.

Parameter and statistic

A statistic is a characteristic of a sample while a parameter is a characteristic of a population.

Suppose, we have a town whose population is 50,000. The statistical measures based on data of all these persons will be a parameter.

On the other hand, if we draw a sample of 5,000 persons and compute various statistical measures such as mean, SD etc., they will be statistic.

Thus, a parameter is a statistical measure which is related to the population and is based on population, whereas a statistic is a statistical measure which relates to to the sample and is based on sample data.

Objectives of sampling theory

1. Estimation of parameters
   (a) point estimate
   The estimate of a population parameter by a single number.
   (b) Interval estimate
   It is a statement of two values between which the parameter is expected to exist.
2. Testing of hypothesis

Standard error

It is the average amount of variability of the observation of a sampling distribution is computed it is called as standard error. e.g. standard deviation of statistic means is called as standard error.

Parameters	**Statistics**
Population size N	Sample size n
Population mean μ	Sample mean $\overline{x}$
Population SD σ	Sample SD S

Large and small samples A sample with size > 30 is known as a large sample while any sample with size ≤ 30 is known as a small sample.

Hypothesis testing and errors A hypothesis is a statement about the population parameter.

1. Null Hypothesis (H_0)
2. Alternate Hypothesis (H_1)

Errors

1. Type I
   When a Null hypothesis is true but rejected

2. Type II error

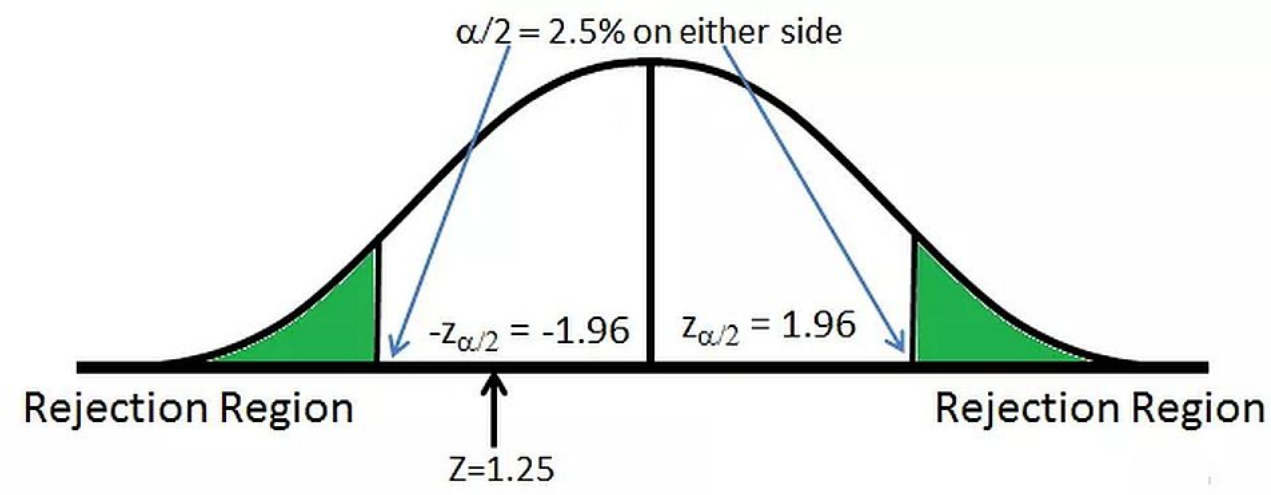
Accepting a null hypothesis when it is false.

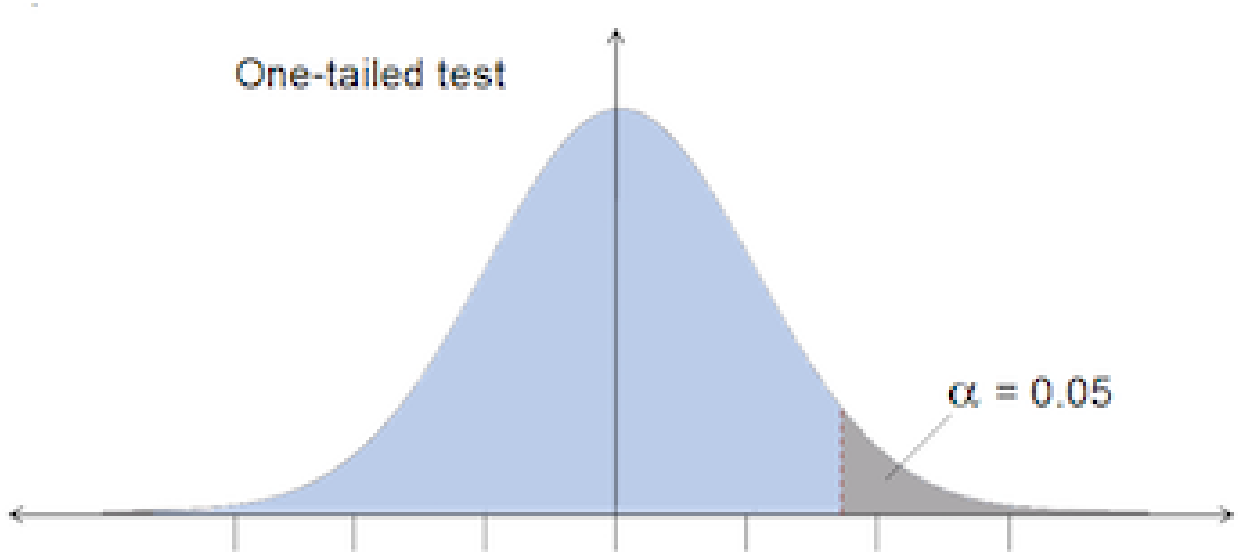
True position	H_0) **Accepted**	H_0 **rejected**
H_0 is true	Correct decision	Type I error
H_0 is not true	Type II error	Correct decision

The maximum possibility of type I error is known as level of significance and it is determined in advance. e.g if level of significance is fixed at 5 %, it means that there is a possibility of making a type I error in 5 out of 100 cases (rejection of a true null hypothesis). We can minimize the type I error by reducing the level of significance. However, controlling the type error, the chances of type II error (acceptance of null hypothesis) increases.

Critical value The value is obtained from a standard table at a particular level of significance.

Two tailed test and One tailed test





Critical value Z_α	1 %	5 %
Two tailed test	$\|z_\alpha\| = 2.58$	$\|z_\alpha\| = 1.96$
Right tailes test	$\|z_\alpha\| = 2.33$	$\|z_\alpha\| = 1.64$
Left tailed test	$z_\alpha = -2.33$	$z_\alpha = -1.64$

Example

In the study of mean, the null hypothesis $H_0 = \mu = \mu_0$

Now, the possible alternate hypothesis be

1. $H_1 : \mu \neq \mu_0$ (i.e. $\mu > \mu_0$ or $\mu < \mu_0$). It is a two tailed test.
2. $H_1 : \mu < \mu_0$ (one tailed test or left sided test).
3. $H_1 : \mu > \mu_0$ (one tailed test or right sided test)

Test of Significance for Single Mean

Under the null hypothesis (H_0): the sample has been drawn from a population with mean μ and variance σ^2, i.e., there is no significant difference between the sample mean ($\overline{x}$) and population mean (μ), the test statistic (for large samples), is

$$Z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

If the SD of the population is not known, then we use

$$Z = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Here, s is the SD of the sample.

Confidence limits for μ: 95% confidence interval for μ is given by

$$|z| \leq 1.96 \Rightarrow |\frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}| \leq 1.96$$

$$\Rightarrow \overline{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

Similarly, we can obtain the 99 % confidence limit for μ as $\overline{x} \pm 2.58\frac{\sigma}{\sqrt{n}}$

Prob: A sample of 900 members has a mean 3.4 cms. Is the sample from a large population of mean 3.25 cms. and s.d. 2.61 cms. ?

Find the 95 % confidence limits of true mean.

Solution

Null hypothesis (H_0): The sample has been drawn from the population with mean(μ) = 3.25 cms. and S.D. $\sigma = 2.61$ cms.

Here, we are given

$\overline{x} = 3.4cms., n = 900, \mu = 3.25cms$ and $\sigma = 2.61cms.$

$$Z = \frac{3.40 - 3.25}{\frac{2.61}{\sqrt{900}}} = \frac{0.15 \times 30}{2.61} = 1.73$$

Since $|Z| < 1.96$, therefore H_0 can be accepted at 5 % level of significance.

95 % confidence limits are $\overline{x} \pm 1.96\frac{\sigma}{\sqrt{n}} \Rightarrow 3.40 \pm 1.96 \times \frac{2.61}{\sqrt{900}} \Rightarrow 3.40 \pm 0.1705$

Hence the limits are 3.5705 and 3.2295.

Prob: A sample of 1600 units is found to have a mean of 3.4 cms. Can it be reasonably regarded as a simple sample from a large population with mean 3.2 cms and SD 2.3 cms.

Solution

Here, $n = 1600, \mu = 3.2, \overline{x} = 3.4$ and $\sigma = 2.3$

H_0 : The sample is drawn from a population with mean 3.2 cms.

Now,

$$|Z| = \frac{|\overline{x} - \mu|}{\frac{\sigma}{\sqrt{n}}} = \frac{|3.4 - 3.2|}{\frac{2.3}{\sqrt{1600}}} = 3.478$$

Since $|Z| > 3$, therefore we will reject H_0.

Problem A population has a mean of 159.7 cms and SD 4.5 cms. How large a sample would be necessary to make the standard error of the mean less than or equal to 0.5 cm.

Soultion

Given, $n = ?, \overline{x} = 159.7$ and $\sigma = 4.5$ and $SE < 0.5$

$$SE = \frac{\sigma}{\sqrt{n}} \Rightarrow 0.5 = \frac{4.5}{\sqrt{n}} \Rightarrow 0.5 \times \sqrt{n} = 4.5 \Rightarrow \sqrt{n} = 9$$

We get $n = 81$.

Therefore, the size of the sample is 81 at least.

Problem An automatic machine was designed to pack 2.0 Kg of Vanaspati. A sample of 100 tins was examined to test the machine. The average weight was found to be 1.94 Kg with SD 0.10 Kg. Is the machine working properly?

Solution

H_0 : Machine is working properly.

Given, $n = 100, \mu = 2Kg, \overline{x} = 1.94Kg$ and $s = 0.10Kg$

Now,

$$|Z| = \frac{|\overline{x} - \mu|}{\frac{s}{\sqrt{n}}} = \frac{|1.94 - 2|}{\frac{0.10}{\sqrt{100}}} = 6$$

Since the calculated value of Z is greater than the tabular value, therefore we reject H_0. Hence, machine is not working properly.

Problem In the past the average length of an outgoing telephone call from a business office has been 143 seconds. A manager wishes to check whether that average has decreased after the introduction of policy changes. A sample of 100 telephone calls produced a mean of 133 seconds, with a standard deviation of 35 seconds. Perform the relevant test at the 1 % level of significance.

Solution

$H_0 : \quad \mu = 143$ $H_1 : \quad \mu < 143$

Given $n = 100, \overline{x} = 133, s = 35$

Now,

$$|Z| = \frac{|\overline{x} - \mu|}{\frac{s}{\sqrt{n}}} = \frac{|133 - 143|}{\frac{35}{\sqrt{100}}} = 2.85$$

The calculated value of Z is greater than the tabular value hence we reject H_0. OR

$$Z = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{133 - 143}{\frac{35}{\sqrt{100}}} = -2.85$$

The calculated value of Z is lesser than the tabular value hence we reject H_0.

Problem

The average household size in a certain region several years ago was 3.14 persons. A sociologist wishes to test, at the 5 % level of significance, whether it is different now. Perform the test using the information collected by the sociologist: in a random sample of 75 households, the average size was 2.98 persons, with sample standard deviation 0.82 person.

Solution

$H_0 : \quad \mu = 3.14$ $H_1 : \quad \mu \neq 3.14$

Given $n = 75, \overline{x} = 2.98, s = 0.82$

Now,

$$|Z| = \frac{|\overline{x} - \mu|}{\frac{s}{\sqrt{n}}} = \frac{|2.98 - 3.14|}{\frac{0.82}{\sqrt{75}}} = 1.68$$

The calculated value of Z is less than the tabular value hence we accept H_0.

Small sample test or t-test

The t-test used to test the significance of

1. The mean of a small sample
2. The difference between the means of two small samples or to compare two small samples

Test of significance of the mean of small sample
Steps involved
To calculate the significance of sample mean at 5 % level of significance

- H_0 : The population mean (μ) is equal to the given value of the mean (i.e. $\mu = \mu_0$).
- calculate $t = \dfrac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$ or $t = \frac{(\overline{x}-\mu)\sqrt{n}}{s}$.
- Compare the calculated value with the tabular value of t at $(n-1)$ degree of freedom and 5 % level of significance.($dof = (n-1)$)

Problem A random sample of size 20 drawn from a normal population yielded the following results: $\overline{x} = 49.2, s = 1.33$. Test $H_0 : \mu = 50 vs. H_1 : \mu \neq 50$ at $\alpha = 0.01$. (The tabular value of t at 19 dof and 1 % level of significance is 2.86)

Using the formula

$$t = \frac{|\overline{x} - \mu|}{\frac{s}{\sqrt{n}}} = \frac{|49.2 - 50|}{\frac{1.33}{\sqrt{20}}} = 2.690$$

Since the calculated value is less than the tabular value, we accept H_0.

Problem

The height of 9 children selected at random from a given colony had a mean 63.5 cms. and variance 6.25 cms. Test, at 5 % level of significance, the hypothesis that the children of the given colony are on average 65 cms long and not less than 65 cm. in all. (The value of t for 8 d.f. at 5 % level of significance is 2.262)

Solution

H_0 : The average height of the children is 65 cms. or $\mu = 65$. $H_1 : \mu < 65$

$n = 9, \overline{x} = 63.5 cms., variance = 6.25$ (or SD = 2.5) and $\mu = 65$

Using the formula

$$t = \frac{|\overline{x} - \mu|}{\frac{s}{\sqrt{n}}} = \frac{|63.5 - 65|}{\frac{2.5}{\sqrt{9}}} = 1.8$$

Since the calculated value is less than the tabular value, we accept H_0.

Problem Six boys are selected at random from a school and their marks in Mathematics found to be 63,63,64,66,60 and 68 out of 100. In the light of

these marks discuss the general observations that the mean in Mathematics in the school were 66. (The value of t for 5 d.f. at 5 % level of significance is 2.571)

Solution

$H_0 : \mu = 66$

Marks	$d_i = (x_i - 64)$	d^2
63	-1	1
63	-1	1
64	0	0
66	2	4
60	-4	16
68	4	16
$\sum x = 384$	$\sum d = 0$	$\sum d^2 = 38$

$$\overline{x} = \frac{\sum x_i}{n} = \frac{384}{6} = 64$$

$$s = \sqrt{\frac{\sum d^2}{(n-1)}} = \sqrt{\frac{38}{(6-1)}} = 2.756$$

Using the formula

$$t = \frac{|\overline{x} - \mu|}{\frac{s}{\sqrt{n}}} = \frac{|64 - 66|}{\frac{2.756}{\sqrt{6}}} = 1.777$$

Since the calculated value is less than the tabular value, we accept H_0.

χ^2 test

Chi-square test is a measurement which

- tell about magnitude of difference between actual or observed frequencies (f_o) and corresponding theoretical or expected frequencies (f_e).
- explains that whether difference is significant or due to sample fluctuations?

$$\chi^2 = \sum \left[\frac{(f_0 - f_e)^2}{f_e} \right]$$

Use of Chi-square test

- Test of independence
- Test of goodness of fit

Problem The following figures show the distribution of digits in numbers chosen at random from a telephone directory) :

Digits	0	1	2	3	4	5	6	7	8	9	Total
frequency	1026	1107	997	966	1075	933	1107	972	964	853	10,000

Test whether the digits may be taken to occur equally frequently in the directory. (The tabular value of χ^2 at 5 % level of significance for 9 degree of freedom is 16.919.)

Solution H_0 : Digits are equally frequently
Expected frequency of each digit (f_e) $\frac{10{,}000}{10} = 1000$

Digits	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
0	1026	1000	26	676	0.676
1	1107	1000	107	11,449	11.449
2	997	1000	-3	9	0.009
3	966	1000	-34	1156	1.156
4	1075	1000	75	5625	5.625
5	933	1000	-67	4489	4.489
6	1107	1000	107	11449	11.449
7	972	1000	-28	784	0.784
8	964	1000	-36	1296	1.296
9	853	1000	-147	21609	21.609
Total		10,000			58.542

$$\chi^2 = \sum \left[\frac{(f_0 - f_e)^2}{f_e} \right] = 58.542$$

The calculated value of χ^2 is much greater than the tabular value, hence we reject our null hypothesis.

Problem

In an anti malaria campaign in a certain area, quinine was administered to 812 persons out of the total population of 3,248. The number of fever cases is given below

Treatment	Fever	No fever
Quinine	20	792
No Quinine	220	2216

Discuss the usefulness of quinine in checking malaria. The tabular value of χ^2 at 5 % level of significance for 1 degree of freedom is 3.841.

Solution The given data can be represented as

	Quinine	No quinine	**Total**
No fever	792	2216	3008
Fever	20	220	240
Total	812	2436	3248

H_0 : Quinine is not effective in treating malaria

	Quinine	No quinine	**Total**
No fever	$\frac{812\times3008}{3248} = 752$	$\frac{2436\times3008}{3248} = 2256$	3008
Fever	$\frac{812\times240}{3248} = 60$	$\frac{2436\times240}{3248} = 180$	240
Total	812	2436	3248

$$\chi^2 = \sum\left[\frac{(f_0-f_e)^2}{f_e}\right] = \frac{(792-752)^2}{752}+\frac{(20-60)^2}{60}+\frac{(2216-2256)^2}{2256}+\frac{(220-180)^2}{180}$$
$$= 2.128 + 26.667 + 0.709 + 8.889 = 38.393$$

The calculated value of χ^2 is much greater than the tabular value, hence we reject our null hypothesis.

Problems A sample of 400 under-graduate students and 400 students of post graduate was taken to know their opinion about autonomous college. 290 of the undergraduate and 310 of the post graduate students favoured the autonomous status. Present these facts in the form of a table and test at 5 % level of significance, that the opinion regrading autonomous status of colleges are independent of the level of classes of students. (The tabular value of χ^2 at 5 % level of significance for 1 degree of freedom is 3.841.)

Solution The given data can be represented as

	Undergraduate	post graduate	**Total**
favored	290	310	600
against	110	90	200
Total	400	400	800

H_0 : The opinions are independent of the level of classes of students

	Undergraduate	post graduate	**Total**
favored	$\frac{400\times600}{800} = 300$	$\frac{400\times600}{800} = 300$	600
against	$\frac{400\times200}{800} = 100$	$\frac{400\times200}{800} = 100$	200
Total	400	400	800

$$\chi^2 = \sum\left[\frac{(f_0-f_e)^2}{f_e}\right] = \frac{(290-300)^2}{300}+\frac{(310-300)^2}{300}+\frac{(110-100)^2}{100}+\frac{(90-100)^2}{100}$$
$$= 0.33 + 0.33 + 1 + 1 = 2.66$$

The calculated value of χ^2 is less than the tabular value, hence we accept our null hypothesis.

Problem A set of five similar coins is tossed 320 times and the result is given in the following table

No. of heads	0	1	2	3	4	5
frequency	6	27	72	112	71	32

Test the hypothesis that data followed a binomial distribution. (The tabular value of χ^2 at 5 % level of significance for 5 degree of freedom is 11.07.)

Solution H_0 : data followed the binomial distribution

$$P(H) = \tfrac{1}{2}, P(T) = \tfrac{1}{2}.$$

Let r represents the number of heads then theoretical frequencies are obtained as

$$P(r=0) = 320 \binom{5}{0} (\tfrac{1}{2})^5 = 10$$

$$P(r=1) = 320 \binom{5}{1} (\tfrac{1}{2})^5 = 50$$

$$P(r=2) = 320 \binom{5}{2} (\tfrac{1}{2})^5 = 100$$

$$P(r=3) = 320 \binom{5}{3} (\tfrac{1}{2})^5 = 100$$

$$P(r=4) = 320 \binom{5}{4} (\tfrac{1}{2})^5 = 50$$

$$P(r=5) = 320 \binom{5}{5} (\tfrac{1}{2})^5 = 10$$

$$\chi^2 = \sum \left[\frac{(f_0 - f_e)^2}{f_e}\right] = \frac{(6-10)^2}{10} + \frac{(27-50)^2}{50} + \frac{(72-100)^2}{100} + \frac{(112-100)^2}{100} + \frac{(71-50)^2}{50} + \frac{(32-10)^2}{10} = 78.68$$

The calculated value of χ^2 is greater than the tabular value, hence we reject our null hypothesis.