Chapter 1

SAMPLING DISTRIBUTION AND ESTIMATION

Sampling Distribution

A sample statistic is a numerical summary measure calculated from sample data. The mean, median, mode, standard deviation, sample proportion and correlation calculated for sample data are called sample statistics. On the other hand, the same numerical summary measures calculated for population data are called population parameters. A population parameter is always a constant, whereas a sample statistic is a random variable. Because every random variable most possesses a probability distribution, each sample statistic possesses a probability distribution. The probability distribution of a sample statistic is called sampling distribution.

If we take a sample of size n from a population of size N, then there are ${}^{N}C_{n} = k$ (say), possible samples. We can compute the sample statistic, say, T for each of these samples. Let $T_1, T_2, \ldots, T_k$ be the values of the k possible samples. Thus, the statistic T may be regarded as a random variable which can take any one of the values $T_1, T_2, \ldots, T_k$. The probability distribution of the statistic T is called the sampling distribution. The sampling distribution of a statistic depends on the distribution of the population, the size of the sample, and the method of sample selection. The average value and standard deviation of sampling distribution plays vital role in statistics. The standard deviation of the sampling distribution of a statistic T is known as standard error of the statistic. Here we will discuss about the most important sampling distribution i.e. Sampling distribution of mean and proportion.

Sampling Distribution of Mean

Let us consider the sampling distribution of sample mean $\bar{X}$. Let us suppose that a random sample of size n is taken from a normal population with mean $\bar{X}$ and variance σ^2. Let $x_1, x_2, \ldots x_n$

is a random sample (without replacement) of size n from a finite population of size N, then $E(\bar{X}) = \mu$ mean value of the sampling distribution of the sample mean is equal to Population mean and its variance is given by $V(\bar{X}) = \frac{\sigma^2(N-n)}{n(N-1)}$ which measures the variability of sample mean.

Where each observation in this sample, say, $x_1, x_2, \ldots x_n$, is a normally and independently distributed random variable with mean $\bar{X}$ and variance σ^2. Again

If sample is from an infinite (very large) population so that the sampling fraction n/N can be neglected, or if the sampling is done with replacement then $V(\bar{X}) = \frac{\sigma^2}{n}$ is the standard deviation (i.e. square root of variance is known as standard error of the sampling distribution.

Sampling distribution of mean possesses the following properties:

(i) Sample mean $\bar{x}$ is an unbiased estimate of the population mean μ.

(ii) The variance of $\bar{x}$ depends on the sample size (n) and is equal to σ^2/n.

Example 1

A sample of size 36 is drawn from a population consisting of 196 units. If the population standard deviation is 7, find the standard error of the sample mean when the sample drawn (i) without replacement, (ii) with replacement.

Solution

Here,

Population size (N) = 196,

Sample size (n) = 36

Population SD (σ) = 7

Sampling without replacement: The standard error of the sample mean is given by

$$S.E.(\bar{X}) = \frac{s}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} = \frac{7}{\sqrt{36}} \times \sqrt{\frac{196-36}{196-1}} = 1.0568$$

Sampling with replacement: The standard error of the sample mean is given by

$$S.E.(\bar{X}) = \frac{s}{\sqrt{n}} = \frac{7}{\sqrt{36}} = 1.1667$$

Example 2

An electronics company manufactures resistors that have a mean resistance of 100 ohms standard deviation of 10 ohms. The distribution of resistance is normal. Find the prob that a random sample of n = 25 resistors will have an average resistance less than 95 ohms

Solution

Here,

Sample size (n) = 36

Population mean (μ) = 100 ohms.

Population SD (σ) = 10 ohms.

Sample mean ($\bar{X}$) = 95 ohms

Now sampling distribution of sample mean X is normally distributed with mean μ = 100 ohms and variance $\frac{s}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$

Now the probability that average resistance is less than 95 ohms is given by

$$P(\bar{x} < 95) = P(z < \frac{95 - 100}{2}) = p(z < -2.5) = 0.0062$$

Sampling Distribution of Proportion

A proportion is the number of elements with a given characteristic divided by the total number of elements in the group. Let X be the number of characteristics obtained in the population having size N. Then population proportion is given by $P = \frac{X}{N}$.

The sample proportion, $p = \frac{x}{n}$, is the point estimate of the population proportion, P. and the variance of the sample proportion is given by the formula $V(p) = \frac{PQ}{n} \times (\frac{N - n}{N - 1})$

Where $Q = 1 - P$ and x=number of characteristics in the sample and n = sample size.

If population size is large or the sampling is done with replacement then the variance term becomes $V(p) = \frac{PQ}{n}$

$$S.E.(p) = \sqrt{\frac{PQ}{n}} \times \sqrt{\frac{N - n}{N - 1}}$$

is the standard error of proportion deviation (i.e. square root of variance is known as standard error of the sampling distribution)

Sampling distribution p of proportion possesses the following properties:

(i) Sample proportion is an unbiased estimate of the population proportion p.

(ii) The variance of *p* depends on the sample size *(n)* and is equal to $\frac{PQ}{n}$

Note: The rule of thumb is if $\frac{n}{N} < 0.05$, then the population is said to be large and is used the formula for the variance as $\frac{PQ}{n}$.

Example 3

From the consignment of 100 apples 20 apples is drawn by simple random sampling method without replacement. If out of the 20 apples 15 apples found defective, what should be the standard error of the sample proportion?

Solution

We have,

Population size (N) = 100

Sample size (n) = 20

No. of defectives (x) = 15

SE (p) = ?

We know,

Sample proportion of defectives $(p) = \frac{x}{n} = \frac{15}{20} = 0.75$

Sample proportion of non-defectives $(q) = 1 - 0.75 = 0.25$

Checking ratio

$\frac{n}{N} = \frac{20}{100} = 0.20 > 0.05$, the adjustment is needed.

Now,

$$SE(p) = \sqrt{\frac{pq}{n-1}}\sqrt{\frac{N-n}{N}} = \sqrt{\frac{0.75\times0.25}{20-1}} = \sqrt{\frac{100-20}{100}} = 0.088$$

Hence, the required standard error of sample proportion is 0.088.

Standard error of some well known statistic are

Statistic	Standard error
Mean (when σ known and population size infinite)	$S.E.(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
Mean (when σ known and population size finite i.e. N)	$S.E.(\bar{X}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$
Mean (when σ unknown and population size infinite)	$S.E.(\bar{X}) = \frac{s}{\sqrt{n}}$
Mean (when σ unknown and population size finite i.e. N)	$S.E.(\bar{X}) = \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$
Difference of means(when σ's are known)	$S.E.(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
Difference of means (when σ's are unknown)	$S.E.(\bar{X}_1 - \bar{X}_2) = \sqrt{\left(s^2\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}\right)}$
Proportion (when population size is infinite)	$S.E.(p) = \sqrt{\frac{PQ}{n}}$
Proportion (when population size is finite i.e. N)	$S.E.(p) = \sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}}$
Difference of proportions	$S.E.(p_1 - p_2) = \sqrt{\left(PQ\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}\right)}$

Central Limit Theorem

The central limit theorem states that as the sample size gets large enough, the samp distribution of the mean is approximately normally distributed. This statement is t regardless of the shape of the distribution of the individual values in the population. If x_1, x x_n is a random sample from normal population with mean μ and variance σ^2, then the sam mean $\bar{x}$ is also normally distributed with mean μ and variance σ^2/n, i.e., $X \sim N(\mu, \sigma^2/n)$. result is true even if the population from which the samples are drawn is not normal, prov

the sample size is sufficiently large as stated in the following Central Limit Theorem.

"If $X_1, X_2, \ldots X_n$ is a random sample of size n from any population, then the sample mean $(\bar{x})$ is normally distributed with mean μ and variance σ^2/n provided n is sufficiently large."

"If $X_1, X_2, \ldots X_n$ are independent random variables following any distribution, then under certain very general conditions, their sum $\Sigma X = X_1 + X_2 + \ldots + X_n$ is asymptotically normally distributed, i.e., ΣX follows normal distribution as $n \to \infty$.

By using this theorem, it has been proved that the sampling distributions of most of the statistics, like sample proportion (p), difference of sample proportions (P_1-P_2). Difference of sample means $(\mu_1 - \mu_2)$ difference of sample standard deviation (s_1-s_2) are asymptotically normal, i.e., the standardized variates corresponding to any one of these statistics is N(0, 1) for large samples. Thus, if t is any statistic, then by central limit theorem,

$$z = \frac{t - E(t)}{S.E.(t)} = \frac{\bar{x} - \mu}{\sigma} \sim N(0, 1)$$

asymptotically, i.e., as $n \to \infty$.

Now the issue here is what sample size to be taken as a large sample. As a general rule, statisticians have found that for many population distributions, when the sample size is at least 30, the sampling distribution of the mean is approximately normal. However, we can apply the central limit theorem for even smaller sample size if the population distribution is approximately bell shape (normal shape). In the case when the distribution is extremely skewed or has more than one mode, we have to take sample size larger than 30 to ensure the normality.

Concept of Inferential Statistics

The inductive inference may be termed as the logic of drawing statistically valid conclusions about the population characteristics on the basis of a sample drawn from it in a scientific manner. We shall develop the technique which enables us to generalize the results of the sample to the population; to find how far these generalizations are valid, and also to estimate the population parameters along with the degree of confidence. The answers to these and many other related problems are provided by a very important branch of Statistics, known as the *Statistical Inference*.

Estimation

Estimation of population parameters like mean, variance, proportion, correlation coefficient, etc., from the corresponding sample statistics is one of the very important problems of statistical inference. The theory of estimation was founded by Prof. R.A. Fisher in a series of fundamental papers about 1930 and is divided into two groups.

(i) Point Estimation

(ii) Interval Estimation.

In Point Estimation, a sample statistic (numerical value) is used to provide an estimate of population parameter whereas in Interval Estimation, probable range is specified within which the value of the parameter might be expected to lie.

Point Estimation

A particular value of a statistic which is used to estimate a given parameter is known as a point *estimate* or *estimator* of the parameter. A good estimator is one which is as close I true value of the parameter as possible. The following are some of the criteria which should satisfy to be a good estimator.

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

Unbiasedness

A statistic $t = t(x_1, x_2, \ldots, x_n)$, a function of the sample observations $x_1, x_2, \ldots x_n$ is said to be an unbiased estimate of the corresponding population parameter θ, if $E(t) = \theta$ i.e., if the mean value of the sampling distribution of the statistic is equal to the parameter. For example, the sample mean x is an unbiased estimate of the population mean μ; the sample proportion p is an unbiased estimate of the population proportion P, S^2 is unbiased estimate of variance.

i.e. $E(\bar{X}) = \mu$ and $E(p) = P$, $E(S^2) = \sigma^2$ where, $S^2 = \frac{1}{n-1}\Sigma(X - \bar{X})^2$

If $E(t) > \theta$, *then the statistic* t *is said to be a* biased estimator of θ.

Let $E(t) = \theta + Q$ then 'Q' is called the *'amount of bias'* in the estimate. If $Q > 0$, i.e., $E(t) > \theta$, *then* t *is said to be* positively biased *and if* $Q < 0$, i.e., $E\{t\} < 0$, *it is said to be* negatively biased.

Consistency

A statistic t based on a sample of size n is said to be a consistent estimator of the parameter θ if it converges in probability to 0, *i.e.*, if $t_n \to 0$ as n. Symbolically, $\lim_{n\to\infty} P(t_n \to \theta) = 1$.

For any distribution, sample mean x is a consistent estimator of the population mean, sample proportion 'p' is a consistent estimator of population proportion P and sample variance s^2 is a consistent estimator of the population variance σ^2. The variance of sampling distribution of the estimator enables us to determine if the statistic is a consistent estimator of the parameter or not. The result is contained in the following theorem.

Theorem. A statistic $t = t_n = t(x_1, x_2, \ldots, x_n)$ is a consistent estimator of the parameter θ if

$$\text{Var}(t) \to 0 \text{ as } n \to \infty.$$

Efficiency

If we have more than one consistent estimators of a parameter θ, then *efficiency* is the criterion which enables us to choose between them by considering the variances of the sampling distributions of the estimators. Thus, if t_1 and t_2 are consistent estimators of a parameter θ such that

$$\text{Var}(t_1) < \text{Var}(t_2), \text{ for all } n$$

then t_1 is said to be more efficient than t_2. *In other words,* an estimator with lesser variability is said to be more efficient and consequently more reliable than the other.

If there exist more than two consistent estimators for the parameter θ, then considering the class of all such possible estimators we can choose the one whose sampling variance is minimum. Such an estimator is known as the *most efficient estimator* and provides a measure of the efficiency of the other estimators.

Definition. If t is the most efficient estimator of a parameter θ with variance v and t_1 is any other estimator with variance v_1, then the efficiency E of t_1 is defined as:

$$E = \frac{v}{v_1}$$

Since t is the most efficient estimator, its sampling variance v is minimum, i.e. $v < v_1$

The efficiency of any estimator cannot exceed unity.

Sufficiency

A statistic $t = t(x_1, x_2, \ldots x_n)$ is said to be a sufficient estimator of parameter θ if it contains all the information in the sample regarding the parameter. In other words, a sufficient statistic utilises all the information that a given sample can furnish about the parameter. If $t = t(x_1, x_2, \ldots x_n)$ is a statistic based on a random sampleof size n from a population with probability function or pdf $p(x, \theta)$, then it is a sufficient estimator of θ if the conditional distribution of $x_1\ x_2, \ldots, x_n$ for given value of t is independent of θ, *i.e.,* if the conditional probability $P(x_1 \cap x_2 \ldots x_k t = k)$ does not depend on θ.

The sample mean $\bar{X}$ is sufficient estimator of population mean and sample proportion p is a sufficient estimator of population proportion *P*.

Properties of **Sufficient Estimators**

1. If a sufficient estimator exists for some parameter then, it is also the most efficient estimator.
2. It is always consistent.
3. It may or may not be unbiased.

Interval Estimation

In point estimation, a single value of a statistic (t) is used as an estimate of the population parameter (θ). But even the best possible point estimate may deviate enough from the true parameter value to make the estimate unsatisfactory. The answer is provided by the technique of *interval estimation*. This consists in the determination of two constants c_1 and c_2, say, such that

$$P(c^1 < \theta < c^2, \text{ for given value of t}) = 1 - \alpha.$$

where α is the level of significance. The interval $[c_1, c_2]$, within which the unknown value of the parameter θ is expected to lie is known as *Confidence Interval* (Neyman) or *Fiducial Interval* (R.A. Fisher); the limits c_1 and c_2 so determined are known as *Confidence Limits or Fiducial Limits* and $1-\alpha$, is called the *confidence coefficient,* depending upon the desired precision of the estimate.

For example, α = 0.5 (or 0.01) gives the 95% (or 99%) confidence limits.

If t is the statistic used to estimate the parameter θ, then (1-α)% *confidence limits for* $\theta = t \pm S.E.(t) \times t_{\alpha/2}$

where α is the significant or critical value of *t* at level of significance α for a two-tailed test. Thus, the computation of confidence limits for a parameter, involves the following three steps:

(i) Compute the appropriate sample statistic t

(ii) Obtain the Standard Error of the sampling distribution of the statistic *t i.e.*, S.E. (*t*).

(iii) Choose appropriate confidence coefficient (1 - α), depending on the precision of the estimate.

Interval Estimation of population mean for Large Samples

For large samples, the underlying distribution of the standardized variate corresponding to the sampling distribution of the statistic *t* will be asymptotically normally distributed i.e.

For practical purposes, samples may be regarded as large if the sample size is greater than 30 *i.e.*> 30.

From areas under normal probability curve, we have

$$P(-1.96 < Z < 1.96) = 0.95$$

Thus 95% *confidence limits for population mean are* $\bar{X} \pm 1.\frac{96\sigma}{\sqrt{n}}$ where σ is assumed to be known, and the interval is the 95% confidence interval for estimating μ. Then, (1 - α)% confidence limits are given by $\bar{X} \pm \frac{Z_{\alpha/2}\,\sigma}{\sqrt{n}}$ for large N.

And (100 – α)% confidence limits are given by $\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{(N-n)}{(N-1)}}$ for large n is assumed and simple random sampling without replacement for finite N.

Two-Tailed Significant (Critical) Values of Z

Confidence coefficient (1-α)	50%	68.27%	90%	95%	95.45%	98%	99%	99.73%
Significant value	0.6745	1	1.645	1.96	2	2.33	2.58	3

For small sample i.e. n < 30 the student t statistics is followed and given by $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ where s is given by $S = \sqrt{\frac{1}{n-1}\Sigma(x - \bar{x})^2}$

Then, (100-α)% confidence limits are given by

$\bar{x} \pm t_{\alpha/2,\,n-1}\frac{S}{\sqrt{n}}$ for unknown population size

And (1-α) confidence limits are given by

$\bar{x} \pm t_{\alpha/2,\,n-1}\frac{S}{\sqrt{n}}\sqrt{\frac{(N-n)}{N-1}}$ for simple random sampling without replacement for finite N.

case of sample SD (s) = $\sqrt{\frac{1}{n}\Sigma(x - \bar{x})^2}$

00-α)% confidence limits are given by

$\pm t_{\alpha/2, n-1}\frac{s}{\sqrt{n-1}}$ for unknown population size

$\pm t_{\alpha/2, n-1}\frac{s}{\sqrt{n-1}}\sqrt{\frac{(N-n)}{N}}$ for simple random sampling without replacement for finite N

ample 4

random sample of 500 have the average of 68.6 with standard deviation 2.5. Find 95% and % confidence interval population average.

lution

Here, sample size n = 500

Sample mean $\bar{x}$ = 68.6

Sample standard deviation s = 2.5

% confidence limits population mean μ is given by

$$\bar{x} \pm \frac{Z_{\alpha/2}\,\sigma}{\sqrt{n}} = \bar{x} \pm \frac{1.96 \times s}{\sqrt{n}} = 68.6 \pm 1.92 \times \frac{5}{\sqrt{500}} = 68.6 \pm 0.219$$

king (-) sign 68.6 – 0.219 = 68.38

king (+) sign 69.8+0.219 = 68.819

nce 99% confidence limits population mean μ is 68.38 to 68.819.

mple 5

andom sample of 12 records gave the average length of 163.99 minutes with standard iation of 3.043 minutes. Find the 95% confidence limits for population mean and population l length if population consists of 100 units.

ution

Here, sample size (n) = 12

Sample mean ($\bar{x}$) = 163.99

Sample standard deviation (s) = 3.043

Population size (N) = 100

Confidence limit (α) = 5%

$t_{\alpha/2, n-1}$ = 2.20

confidence limits population mean μ is

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n-1}}\sqrt{\frac{(N-n)}{N}}$$

Est. [S.E. (p)] = $\sqrt{\frac{0.13 \times 0.87}{500}} = 0.015$.

The most probable limits or the population proportion of bad pineapples are:

$p \pm 3$ S.E. (p) = $0.130 \pm 3 \times 0.015 = 0.130 \pm 0.045$ = (0.085 and 0.175)

Hence the percentage of bad pineapples in the consignment lies almost surely between the limits 8.5 and 17.5.

(b) If no estimate of P is available, then 95% confidence limits for P are given by

$$p \pm Z_{\alpha/2} \text{ S.E. } (p) = p \pm 1.96 \text{ S.E. } (p) = p \pm 1.96 \sqrt{\frac{pq}{n}}$$

$$= 0.130 \pm 1.96 \times 0.015 = 0.130 \pm 0.044 = (0.086 \text{ and } 0.174).$$

Example 8

A random sample of 700 units from a large consignment showed that 200 were damaged. Find (i) 95% and (ii) 99% confidence limits for the proportion of damaged units in the consignment.

Solution

We are given sample size (n) = 700.

Proportion of damaged units in the sample (p) = 0.28 then q = 1-p

= 1- 0.28 = 0.72

Hence an estimate of Standard Error of p is given by

S.E.(p) = $\sqrt{\frac{pq}{n}} = \sqrt{\frac{0.28 \times 0.72}{700}} = 0.017$

95% confidence Limits for P, i.e., proportion of damaged units in the consignment are given by

$= p \pm Z_{\alpha/2} \sqrt{\frac{pq}{n}} = p \pm 1.96 = 0.286 + 1.96 \times 0.017$

$= 0.286 \pm 0.033 = (0.253, 0.319)$.

Hence 95% confidence limit for proportion of damaged is 25.3% to 31.9%

(i) 99% Confidence Limits for P are

$$p \pm Z_{\alpha/2} \sqrt{\frac{pq}{n}} = p \pm 2.58 \sqrt{\frac{pq}{n}}$$

$= 0.286 \pm 2.58 \times 0.017 = 0.286 \pm 0.044$ = (0.242 and 0.330).

Hence 99% confidence limit for proportion of damaged is 24.2% to 33%.

Example 9

Out of 20,000 customers' ledger accounts, a sample of 600 accounts was taken to test the accuracy of posting and balancing where in 45 mistakes were found. Assign limits within which the number of defective cases can be expected at 95% level.

Solution

Here we are given

Sample size (n) = 600,

Population size (N) = 20,000.

Number of mistakes in the sample ledger accounts (x) = 45

Proportion of mistakes in the sample (p) $= \frac{x}{n} = \frac{45}{600} = 0.075$

$q = 1 - 0.075 = 0.925$

95% confidence limits for population proportion P are given by:

$$p \pm Z_{\alpha/2}\sqrt{\frac{pq}{n-1}}\sqrt{\frac{N-n}{N}}$$

$$= 0.075 \pm 1.96\sqrt{\frac{0.075 \times 0.925}{600-1}}\sqrt{\frac{20000-600}{20000}}$$

$$= 0.075 \pm 1.96 \times 0.01 \times 0.97$$

$$= 0.075 \pm 0.019 = (0.056, 0.094)$$

Hence, the number of defective cases in a lot of 20,000 are expected to lie between 20,000 × 0.056 and 20,000 × 0.094 i.e., 1,020 and 1,880.

Determination of Sample Size

Estimation of sample size by using mean

Let $\bar{x}$ be the sample mean from a random sample of size n drawn from population with mean μ and standard deviation σ.

Now, $Z = \frac{\bar{x} - E(\bar{x})}{SE(\bar{x})} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

At α level of significance $(1-\alpha)$ confidence limit is

$$P\left(\left|\frac{\bar{x} - \mu}{\sqrt{\frac{\sigma}{\sqrt{n}}}}\right| \leq Z_{\alpha/2}\right) = 1 - \alpha$$

$$\Rightarrow P(|\bar{x} - \mu| \leq \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}) = 1 - \alpha$$

Now, $\bar{x} - \mu = d$ (margin of error) then

$$d = \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}$$

$$\Rightarrow \sqrt{n} = \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}$$

$\Rightarrow$ In case of σ is not known take $\sigma = s$

For the finite population of size N, sample size $= \frac{\sigma^2 Z^2_\alpha}{d^2 + \frac{\sigma^2 Z^2_\alpha}{N}} = \frac{n}{1 + \frac{n}{N}}$

Example 10

In measuring reactions time, a psychologist estimates that the standard deviation is 0.05

seconds. How large a sample of measurement must be taken in order to be 99% confident that the error of his estimate will not exceed 0.01 seconds?

Solution

Here

Sample size (n) = ?

Standard deviation (s) = 0.05

Confidence interval $(1 - \alpha) = 99\% = 0.99$

or $\alpha = 0.01$

$Z_{\alpha/2} = 2.58$

Error (d) = 0.01

Here $\sigma = s$

$$n = \frac{\sigma^2 Z^2_{\alpha/2}}{d^2} = \frac{(0.05)^2 \times (1.96)^2}{(0.01)^2} = 166.4 \approx 167$$

Hence required sample size is 167.

Example 11

The mean systolic blood pressure of a certain group of people was found to be 125 mm of Hg with standard deviation of 15 mm of Hg. Calculate sample size to verify the result at 5% level of significance if error do not exceed 2. Also find sample size if sample is selected from population of size 500.

Solution

Here

Standard deviation (s) = 15

Level of significance (a) = 5%

Sample size (n) = ?

Error (d) = 2

Here $\sigma = s$

Now,

$$n = \frac{\sigma^2 Z^2_{\alpha/2}}{d^2} = \frac{(15)^2 \times (1.96)^2}{(2)^2} = 216.09 \approx 216$$

When $N = 500$

$$\text{Sample size} = \frac{n}{1 + \frac{n}{N}} = \frac{216}{1 + \frac{216}{500}} = 150.83 \approx 151$$

Example 12

Determine the minimum sample size required so that the sample estimate lies within 10% of the true value with 95% level of confidence when coefficient of variation is 60%.

Solution:

Here, C.V. = 60% = 0.6

$$P(|\bar{x} - \mu| \leq 0.1\mu) = 0.95 \quad \text{.....(i)}$$

Confidence level $(1-\alpha) = 95\% = 0.95$ then $\alpha = 0.05$

Now, $P\left(\left|\frac{\bar{x}-\mu}{\sqrt{\frac{\sigma}{\sqrt{n}}}}\right| \le Z_{\alpha/2}\right) = 1-\alpha$

$\Rightarrow P(|\bar{x}-\mu| \le \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}) = 0.95$

$\Rightarrow P(|\bar{x}-\mu| \le 1.96 \times \frac{\sigma}{\sqrt{n}}) = 0.95$ .....(ii)

From equation (i) and (ii)

$0.1\mu = 1.96 \times \frac{\sigma}{\sqrt{n}}$

$\Rightarrow \sqrt{n} = \frac{1.96}{0.1} \times \frac{\sigma}{\mu}$

$\Rightarrow n = (1.96/0.1 \times \sigma/\mu)^2$

$\Rightarrow n = 384.16 \times CV^2$

$\Rightarrow n = 384.16 \times (0.6)^2$

$\Rightarrow n = 138.29 \approx 138$

Hence required sample size is 138.

Estimation of sample size by using proportion

Let p be sample proportion from random sample of size n drawn from population with proportion P

Now,

$$Z = \frac{p - E(p)}{SE(p)} = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$$

At α level of significance $(1-\alpha)$ confidence limit is

$$P\left(\left|\frac{p-P}{\sqrt{\frac{PQ}{n}}}\right| \le Z_{\alpha/2}\right) = 1 - \alpha$$

$$\Rightarrow P(|p-P| \le \sqrt{\frac{PQ}{n}} Z_{\alpha/2} = 1 - \alpha$$

Now, $p - P = d$ (margin of error) then

$$d = \sqrt{\frac{PQ}{n}} Z_{\alpha/2}$$

$$\Rightarrow \sqrt{n} = \frac{Z_{\alpha/2}\sqrt{PQ}}{d}$$

$$\Rightarrow n = \frac{PQ\, Z^2_{\alpha/2}}{d^2}$$

In case of P is not known take P = p

For the finite population of size N, sample size $= \frac{PQ\,Z^2_{\alpha/2}}{d^2 + \frac{PQ\,Z^2_{\alpha/2}}{N}} = \frac{n}{1 + \frac{n}{N}}$

Example 13

A researcher wants to conduct a survey of disabled at Kathmandu valley. What should be the sample size of the prior estimate of population of disables in the population is 10% and the desired error is estimation is 2% and level of significance is 5%.

Solution

Here

Sample size (n) = ?

Population proportion (p) = 10% = 0.1

q = 1 – p = 0.9

Error (d) = 2% = 0.02

Level of significance (α) = 5%

Here P = p

$$n = \frac{Z_{a/2}{}^2\,P.Q.}{d^2} = \frac{(1.96)^2 \times 0.1 \times 0.9}{(0.02)^2} = 864.36 \square 865$$

Hence required sample size is 865.

Example 14

For p = 0.2, d = 0.05 and z = 2 find n. Also find n if N = 1000.

Solution

Here P = p

Now,

$$n = \frac{Z^2_{\alpha/2}\,PQ}{d^2} = \frac{4 \times 0.2 \times (1 - 0.2)}{0.05^2} = 256$$

When N = 1000

$$\text{Sample size} = \frac{n}{1 + \frac{n}{N}} = \frac{256}{1 + \frac{256}{1000}} = 203.82 \approx 204.$$

Relationship of sample size with desired level of error

For estimation of unknown parameters population mean μ based on the sample statistic ($\bar{X}$ = sample mean), we wish to consider the relationship between the error, risk and sample size. Since the sampling distribution of sample means for large samples are normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, the standard normal variate is defined by,

$$Z_\alpha = \frac{\bar{X} - \mu}{S.E(\bar{X})}$$

$$= \frac{E}{S.E(\bar{X})} = \frac{E}{\frac{\sigma}{\sqrt{n}}}$$

Where,

$\bar{X}$ = sample mean
μ = population mean
σ = population standard deviation
n = sample size
Z = standard normal variate
α = the risk that sample mean ($\bar{X}$) differ by population mean μ.

Example 15

If the population proportion of success is 0.65 and n = 100, what will be the value of sampling error when acceptance region is 0.95?

Solution

Here,

Population proportion of success (P) = 0.65
$\therefore$ Q = 1 - P = 0.35
Sample size (n) = 100
Sampling Error (E) = ?
Significant level (α) = 1 - 0.95 = 0.05
z_α = 1.96 [Two tailed]

We know,

$$n = \left(\frac{Z_\alpha}{E}\right)^2 PQ$$

or, $$100 = \left(\frac{1.96}{E}\right)^2 0.65 \times 0.35$$

Taking root both sides, we get;

or, $$10 = \frac{1.96}{E}\sqrt{0.65 \times 0.35}$$

or, $$10E = 1.96\sqrt{0.65 \times 0.35}$$

or, $$E = \frac{0.935}{10} = 0.0935$$

$\therefore$ The value of sampling error = 0.0935.

EXERCISE – 1

1. What do you mean by sampling distribution? Why this study is important in statistics?
2. Elaborate the statement "The central limit theorem forms the basis of inferential statistics".
3. Describe the different criteria of good estimator.
4. Explain how sample size affects the margin of error with examples.
5. A sample of size 25 is drawn from a finite population consisting of 150 units. If the population standard deviation is 10, find the standard error of sample mean.

Ans: 4.17

6. A random sample of 64 projector lamp indicated a sample average life of 3500 hours. The standard deviation of life is 200 hours. Then calculate the standard error of average life of projector lamp.

Ans: 25

7. A simple random sample of size 20 is drawn without replacement from a finite population of 75 units, if the number of defective units in the population is 12. Ascertain standard error of the sample proportion.

Ans: 0.09

8. A manager has sent 300 SMS for a mass meeting and checked 25 of them either they were correctly delivered or not. While checking he found that 15 were not correctly delivered. Find out standard error of sample proportion of SMS incorrectly delivered.

Ans: 0.0892

9. A candidate prepares for a local election. During his campaign, 42 out of 70 randomly selected people in town A and 59 out of 100 randomly selected people in town B showed they would vote for this candidate. Calculate the standard error for the difference in support that this candidate getting in town A and B.

Ans: 0.076

10. Internet connections are often slowed by delays at nodes. Five hundred packets were sent through the same network between 5 to 6 pm and after four hours i.e between 10 to 11 pm 300 packets were sent. The earlier sample showed mean delay time of 0.8 second with standard deviation of 0.1 second whereas the second sample showed mean delay time of 0.5 seconds with a standard deviation of 0.08 second. Calculate the standard error of difference between two sample means.

Ans: 0.006

11. A random sample of size 64 has been drawn from a population with standard deviation 20. The mean of the sample is 80. Calculate 95% confidence limit for the population mean. How does the width of the confidence interval change if sample size is 256 instead?

Ans: 75.1, 84.9, 77.55, 82.5

12. A random sample of size 65 was taken to estimate the mean life of 1000 laptop batteries and the mean and standard deviation were found to be 6300 hours and 9.5 hours respectively. Find a 95% confidence interval for the population mean.

Ans: 6297.75, 6302.25

13. It is observed that 28 successes in 70 independent Bernoulli trial. Compute 90% confidence interval for population proportion.

Ans: 0.304, 0.496

14. In a random sample of 400 chips from a large consignment, 20 items were found to defective. Find 99% confidence limits for the percentage of defective chips in the consignment.

Ans: 2.2%, 7.8%

15. In laboratory experiment, for the test of a material in good condition, a sample of 400 units was drawn. When they were tested, 80 were good. Find 95% confidence limits for the percentage of good.

Ans: 18% to 22%

16. In a sample survey of 100 professionals in a city, 23% preferred a particular brand of laptop. Find 99% confidence limits for percentage of all professionals in the city preferring the brand of laptop.

Ans: 19.57%, 26.43%

17. A factory is producing 50000 CD daily from a sample of 500 CD, 2% were found to be of substandard quality. Estimate the percentage of CD that can be reasonable expected to spoiled in the daily production at 95% confidence level.

Ans: 0.0077, 0.0323

18. A random sample of 100 defective computers of a university, 75 were successfully repaired and 25 were unable to repair due to motherboard problems. Find 95% confidence limits for the percentage of computer which can be repaired in that university.

Ans: 66.5%, 83.5%

19. A sample of size 100 produces the sample mean 16. Assuming population standard deviation 3, compute 95% confidence interval for population mean.

Ans: 15.412, 16.588

20. Assuming population standard deviation 3, how large should a sample be to estimate population mean with margin of error not exceeding 0.5?

Ans: 139

21. The principle of a college wants to estimate the proportion of students who were interested to develop startup. What size of a sample should he select so as to have the difference of proportion of interested students with true mean not to exceed by 10% with almost certainty? It is believed from previous records that the proportion of interested students was 0.30?

Ans: 189

22. Mr. X wants to determine the average time to complete a project the past records show that population standard deviation is 10 days. Determine the sample size so that he may be 95% confident that sample average remains within ± 2 days of the averages.

Ans: 96

23. The average time taken by server to execute an algorithm varies from time to time. From the past experience it is known that the time taken is normally distributed with standard deviation of 6.7 minutes. The IT manager wishes to estimate the average by drawing a random sample such that the probability is 0.95 that the mean of the sample will not deviate by more than 1 minute from the population mean. What should be sample size?

Ans: 173

❑❑❑