Probability and Statistics

Bimal Shrestha
Lecturer
Nepal College of Information Technology (NCIT)
Balkumari, Lalitpur

Outlines

- Log-Normal Distribution
- Uniform (or Rectangular) Distribution
- Gamma Distribution
- Beta Distribution
- Exponential Distribution
- Sampling (concept, terminology used, definitions, etc.)
- Sampling Distribution of Mean
- Estimation (Point Estimation, properties of good estimator, and Interval Estimation)
- Determine of Sample size

Log-Normal Distribution

- **Concept**

If a continuous RV X follows log normal distribution then the natural logarithm of X (i.e. lnX) follows normal distribution.

In other words, if a continuous RV X follows normal distribution then the lnX follows log normal distribution.

Symbolically, $X \sim logN(\mu, \sigma^2)$, then

$lnX \sim N(\mu, \sigma^2)$

Here, $\mu \; and \; \sigma^2$ are two parameters of the distribution.

Definition:

A continuous RV X is said to follow log-normal distribution with parameters $\mu \; and \; \sigma^2$, if its probability density function (pdf) is given by

$$f(x) = \frac{1}{x.\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{lnx-\mu}{\boldsymbol{\sigma}}\right)^2}, \qquad x > 0$$

<u>Mean and Variance of log-normal distribution</u>

$\boldsymbol{mean = E(X) = e^{\mu+\frac{\sigma^2}{2}}}$, and

$\boldsymbol{variance = var(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)}$

Examples:

Question 1:

The life time of semi-conductor laser has a log normal with $\mu = 10hrs\ and\ \sigma = 1.5hrs$.

i. What is the probability that the life time exceeds 10000 hrs?

ii. What life time is exceeded by 99% of lasers?

iii. Find the mean and standard deviation of life time.

Solution:

Given, $\mu = 10hrs\ and\ \sigma = 1.5hrs.$

Let X be the life time of semi-conductor laser.

Here, $X \sim logN(\mu, \sigma^2)$

Then, let $Y = lnX \sim N(\mu, \sigma^2)$

$X = e^Y \sim logN(\mu, \sigma^2)$

i. Here, we have find P(X > 10000) = ?

When, $X = 10000$

$\rightarrow Y = lnX = ln10000$

$\rightarrow X = e^Y = 10000$

Now ,

$$P(X > 10000) = P(e^Y > 10000)$$
$$= P(lne^Y > ln10000)$$
$$= P(Y > ln10000)$$
$$= P\left(\frac{Y-\mu}{\sigma} > \frac{ln1000-\mu}{\sigma}\right)$$
$$= P\left(Z > \frac{ln10000-10}{1.5}\right)$$
$$= P(Z > -0.53)$$

Solve it as previous like in Normal Distribution.

ii. Let x be the life time of laser that is exceeded by 99% of lasers.

Therefore, P(X > x) = 0.99

i.e. P(e^Y > x) = 0.99

When, e^Y = x

This implies, Y = lnx

Now, define a variable, $Z = \frac{Y-\mu}{\sigma} = \frac{lnx-10}{1.5} = -z_1(say) \ldots\ldots\ldots\ldots (i)$

Therefore, P(e^Y > x) = P(Y > lnx) = 0.99

Or, P(Z > $-z_1$) = 0.99

Or, 0.5 + P($-z_1$ < Z < 0) = 0.99

Or, P($-z_1$ < Z < 0) = 0.49

Or, P(0 < Z < z_1) = 0.49

Or z_1 = 2.33

Now, equation (i) becomes

$$\frac{lnx-10}{1.5} = -2.33$$

Or, lnx = 10 – 3.495

Or, lnx = 6.505 , from here you can find the value of x.

Practice

Question 2:

The length of time (in seconds) that a user views a page on website before moving to another page is a log normal random variable with parameters $\mu = 0.5 \, and \, \sigma = 1$.

i. What is the probability that a page is viewed for more than 10 seconds?

ii. What is the length of time that 50% of users view the page?

iii. What is the mean and standard deviation of the time until a user moves from the page?

Question 3:

Suppose that ‘X’ has a log normal distribution and that the mean and variance of ‘X’ are 50 and 4000 respectively. Determined the following

a. The parameters $\mu \; and \; \sigma^2$ of the log normal distribution.
b. The probability that ‘X’ is less than 150.

Given, Mean E(X) = 50 , and
Variance, Var(X) = 4000
Let X follows log normal distribution. Then we have

$$\boldsymbol{E(X) = e^{\mu+\frac{\sigma^2}{2}} = 50}$$

Taking ln on both side, we get

$$\boldsymbol{ln e^{\mu+\frac{\sigma^2}{2}} = ln 50}$$

$$\boldsymbol{\mu + \frac{\sigma^2}{2} = 3.912}$$

$$\boldsymbol{2\mu + \sigma^2 = 7.824 \ldots\ldots\ldots 1}$$

and

$$\boldsymbol{variance = var(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)}$$

$$\boldsymbol{variance = var(X) = e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right) = 4000}$$

Taking ln on both, we get

$$\boldsymbol{lnvar(X) = \ln[e^{2\mu+\sigma^2}(e^{\sigma^2}-1)] = ln4000}$$

$$\ln \boldsymbol{e^{2\mu+\sigma^2} + ln\left(e^{\sigma^2}-1\right) = 8.294}$$

$$2\mu + \sigma^2 + ln(\sigma^2 - 1) = 8.294$$

7.824 + $ln(\sigma^2 - 1) = 8.294$ (from equation 1)

$ln(\sigma^2 - 1) = 8.294 - 7.824 = 0.470$

Now, taking antilog on both sides, we get

$$e^{\sigma^2} - 1 = e^{0.470}$$

$$e^{\sigma^2} - 1 = 1.60$$

$$e^{\sigma^2} = 2.60$$

Again, take ln on both sides we get,

$$lne^{\sigma^2} = ln2.60$$

$$\sigma^2 = 0.9555$$

Putting this in equation 1

$$\boldsymbol{2\mu + 0.960 = 7.824}$$

u=3.432

P(X < 150) = ?

If X follows log normal then lnX follows normal distribution

Let Y = lnX

Then define a variable,

$$Z = \frac{lnX - \mu}{\sigma}$$

When X = 150

Then we have lnX = ln150

$$Z = \frac{ln150 - 3.432}{0.98} = 1.61$$

P(X<150) = P(lnX<ln150)

$$= P\left(\frac{lnX-\mu}{\sigma} < \frac{ln150-3.432}{0.98}\right) = P(Z < 1.61)$$

= 0.5 + P(0 < Z < 1.61)

=0.5 + 0.4463

=0.9463

Uniform (or Rectangular) Distribution

A continuous random variable X is said to have a uniform distribution defined on [a, b], if its pdf is given by

$$f(x) = \frac{1}{b-a}\ , a \leq x \leq b$$

Mean and Variance

$$mean = E(X) = \frac{b+a}{2}$$

And, $variance = var(X) = \frac{(b-a)^2}{12}$

Example

1. For rectangular distribution

$$F(x) = \begin{cases} \frac{x}{100}; & 100 < x < 200 \\ 0 \ ; & otherwise \end{cases}$$

Then, find

i. E(X) = (b+a)/2

ii. Var(X) = (b-a)^2/12

iii. P(X ≥ 150)

iv. P(125 ≤ X ≤ 160)

Given, a = 100 and b = 200

$$F(x) = \begin{cases} \frac{x}{100}; & 100 < x < 200 \\ 0 \ ; & otherwise \end{cases}$$

Change this given distribution function into probability density function.

$$f(x) = \frac{1}{100} \ ; 100 < x < 200$$

iii. P(X≥150) = $\int_{150}^{200} f(x)dx$

$= \int_{150}^{200} \frac{1}{100} dx$

= 0.5

iv. P(125 < X < 160) = $\int_{125}^{160} f(x)dx$

= $\int_{125}^{160} 1/100 dx$

= 0.35

Gamma Distribution

Gamma Function (Properties)

1. $\Gamma n = \int_0^\infty e^{-x} x^{n-1} dx$
2. $\Gamma n = (n-1)\Gamma(n-1)$
3. $\Gamma n = (n-1)!$
4. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$
5. $\int_0^\infty e^{-\beta x} x^{\alpha-1} dx = \frac{\Gamma\alpha}{\beta^\alpha}$
6. $\Gamma 1 = 1$

Probability density function of Gamma Distribution

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}} \; ; x > 0, \alpha \; and \; \beta > 0$$

Here, $\alpha \; and \; \beta$ are two parameters of Gamma distribution.

Sometimes, there is single parameter gamma distribution with α only (i.e. $\beta = 1$) and its pdf is given by

$$f(x) = \frac{1}{\Gamma\alpha} x^{\alpha-1} e^{-x} \; ; x > 0, \alpha > 0$$

Mean and Variance of Gamma Distribution with two parameters $\alpha \; and \; \beta$

$$Mean = E(X) = \alpha\beta$$

And, $Variance = Var(X) = \alpha\beta^2$

Example:

1. Suppose that the life time of a certain kind of computer is random variable X having gamma distribution with α = 2 and β = 6. Find

a. Mean life time of computers.

b. The probability that such a computer will last more than 10 years. P(X > 10) = $\int_{10}^{\infty} f(x)dx$

Solution:

Let X be the life time of a certain kind of computer having gamma distribution with α = 2 and β = 6. Then, its pdf is given by

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}} \; ; x > 0$$

Continue…

or, $f(x) = \frac{1}{6^2\Gamma 2} x^{2-1} e^{-x/6}$

or, $f(x) = \frac{1}{36} x e^{-x/6}$ since $\Gamma n = (n-1)!$

Now,

b) The probability that such a computer will last more than 10 years is given by

$$P(X > 10) = 1 - P(X \leq 10)$$

$$= 1 - \int_0^{10} f(x)dx = 1 - \int_0^{10} \frac{1}{36} x e^{-x/6} dx$$

Integrate yourself using Product Rule.

Practice

The daily consumption of milk in Kathmandu city in excess 2,00,000 liters, is approximately distributed as gamma distribution with α = 2 and $\beta = 10^5$. The city has a daily stock of milk of 3,00,000 liters. What is the probability that the stock is insufficient on a particular day?

Solution:

Let X be the daily consumption of milk.

Let Y = X – 200000 has gamma distribution with α = 2 and $\beta = 10^5$.

Now the pdf is given by

$$f(y) = \frac{1}{\beta^{\alpha}\Gamma\alpha} y^{\alpha-1} e^{-\frac{y}{\beta}} \; ; y > 0$$

$$f(y) = \frac{1}{(10^5)^2 \Gamma 2} y^{2-1} e^{-\frac{y}{10^5}}$$

P(X > 300000) = P (Y+200000 > 300000) = P(Y>100000)

= 1 – P(X ≤ 100000)

= 1 - $\int_0^{100000} f(y)dy$

= 1- $\int_0^{100000} \frac{1}{(10^5)^2 \Gamma 2} y^{2-1} e^{-\frac{y}{10^5}} dy$

= ?

Beta Distribution

Beta Function (Properties)

1. $\beta(m,n) = \int_0^1 x^{m-1}(1-x)^{n-1}dx$
2. $\beta(m,n) = \frac{\Gamma m\, \Gamma n}{\Gamma(m+n)}$

Pdf of Beta Distribution with parameter $\alpha \; and \; \beta$.

$$f(x) = \frac{1}{\beta(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} \; ; \;\; 0 < x < 1 \; and \; \alpha, \beta > 0$$

Here, $\alpha \; and \; \beta$ are the parameters of the Beta distribution.

Mean and Variance of Beta Distribution

$$Mean = E(X) = \frac{\alpha}{\alpha+\beta} \text{ and}$$

$$Variance = Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Example

1. If the annual proportion of erroneous income tax returns filed with the Inland Revenue Department (IRD) can be looked upon as a random variable having beta distribution with $\alpha = 2 \; and \; \beta = 3.$ What is the probability that in any given year there will be fewer than 10% erroneous returns? P(X < 0.10) =?

Solution :

Let X be the erroneous income tax returns having bets distribution with $\alpha = 2 \; and \; \beta = 3.$ then its pdf is

$$f(x) = \frac{1}{\beta(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \text{ , } 0 < x < 1$$

$$f(x) = \frac{1}{\beta(2,3)} x^{2-1}(1-x)^{3-1}$$

$$= \frac{\Gamma 5}{\Gamma 2 \Gamma 3} x(1-x)^2 \quad \text{since}, \beta(\alpha,\beta) = \frac{\Gamma\alpha\Gamma\beta}{\Gamma(\alpha+\beta)}$$

$$f(x) = \frac{24}{2} x\,(1-x)^2$$

$$P(X < 0.10) = \int_0^{0.10} f(x)\,dx$$

$$= \int_0^{0.10} \frac{24}{2} x\,(1-x)^2\,dx$$

= ?

Exponential Distribution

Exponential distribution is a special case of Gamma distribution. When $\alpha = 1 \; and \; \frac{1}{\beta} = \lambda$, then the gamma distribution became

$$f(x) = \lambda e^{-\lambda x} \; ; \; x > 0 \; and \; \lambda > 0$$

Which is the pdf of exponential distribution with parameter λ.

Mean and Variance of exponential distribution

$mean = E(X) = \frac{1}{\lambda}$, and

$Variance = Var(X) = \frac{1}{\lambda^2}$

Example:

The life time of mechanical assembly in a vibration test is exponential distributed with a mean of 400hrs. Then

a. What is the probability that an assembly on test fails in less than 100hrs?
b. What is the probability that operates for more than 500 hrs before failure?
c. If an assembly has been on test for 400 hrs without failure, then what is the probability of a failure in the next 100 hrs?

Solution:

let X be the life time of mechanical assembly having exponential distribution with mean of 400 hrs.

Here, we have $E(X) = \frac{1}{\lambda} = 400$

$\lambda = 0.0025$

Now, the pdf is given by

$f(x) = \lambda e^{-\lambda x}$

$f(x) = 0.0025e^{-0.0025x}$

i. $P(X < 100) = \int_0^{100} f(x)dx$

$= \int_0^{100} 0.0025e^{-0.0025x}dx$

= ?

ii. $P(X > 500) = \int_{500}^{\infty} f(x)dx$

= 1 – P (X ≤ 500)

iii. Lack of memory property of exponential distribution

P (400 < X < 500/X>400) =P (X < 100) =

The time between arrivals of taxi at a busy intersection is exponentially distributed with a mean of 10 minutes.

a. What is the probability that you wait longer than one hour for a taxi?
b. Suppose you have already been waiting for one hour for a taxi, what is the probability that one arrives within the next 10 minutes?
c. Determine ‘x’ such that the probability that you wait less than ‘x’ minute is 0.10.

a. $P(X > 60) = ?$

b. $P(60 < X < 70 / X > 60) = P(X < 10) = ?$

P(X < x) = 0.10

Find x = ?

P(X<x) =F(x) =0.10

Or, 1- $e^{-\lambda x}$ = 0.10

Or, $1 - e^{-0.10x} = 0.10$

Or, 0.90 = $0.1e^{-0.10x}$

$or, e^{-0.10x} = 0.9$

X = 1.05

$$F(x) = \int_0^x f(x)dx$$

$$= \int_0^x \lambda e^{-\lambda x} dx$$

$$= \lambda \left[\frac{e^{-\lambda x}}{-\lambda}\right]$$

$$= - \left[e^{-\lambda x}\right]_0^x$$

$$= 1 - e^{-\lambda x}$$

Sampling and Estimation

- A sampling is a method of selecting a fraction of the population in a way that the selected sample represents the population

- Sampling is a statistical procedure that is concerned with the selection of the individual observation; it helps us to make statistical inferences about the population.

- Sampling is the selection of a subset of the individuals from the study population to estimate the characteristics of the whole population.

Terminology Used

- **Population**: It is a group of individuals, objects, or items from which sample are drawn for the purpose of the study. For example; the inhabitants of a country or regions, number of trees in the jungle, books in the library, etc.

- Finite population: Individuals in the population are countable i.e. limited or finite number of individuals or objects. For example, number of students in NCIT, number of NMC registered doctors, etc.

- Infinite population: Individuals in the population are not countable. For example, number of fishes in the sea, stars in the sky, etc.

- Study population: The population of interest in which research will be done.

- Target population: The study population from which samples are drawn.

- Population size: The number of individuals in the population is called population size. It is denoted by N.

- **Sample**: A finite subset of a population selected from the population for the purpose of investigation is called sample. A sample is selected for the purpose of reaching conclusion about the population from the sample information.

▪**Sample size**: The number of individuals in the sample is called sample size. It is denoted by n.

▪**Parameter**: The numerical value calculated from the population is called parameter. A population has several characteristics or statistical measures such as mean, standard deviation, variance etc. which describe or characterize the population. These statistical measures computed from the population are known as population parameters. Usually population parameters are denoted by Greek letter (or capital letter) such as μ for population mean, ρ for population correlation coefficient, σ for population standard deviation, **P** for population proportion, etc.

▪**Statistic :** The numerical value calculated from the sample is called statistic i.e. the statistical measure computed from the sample observations. They are denoted by simply small letters such as sample mean as $\bar{x}$, sample standard deviation as 's', sample correlation coefficient as 'r', etc. The statistic is a function of the sample observations. Thus, any function of the observed sample values calculated to estimate some parameters of the population is called statistic. So it is also regarded as estimator of the parameter. Statistic is a random variable.

▪**Sampling unit**: The population units selected in sample is called sampling units.

▪**<u>Sampling frame or Sampling list</u>**: The list of population units from which sample units are selected is called sampling list or sampling frame. If the list is not available, it should be prepared before conducting the main survey.

Objectives of the Sampling

1. To estimate the population parameters from the sample statistics.
2. To minimize the cost, time, and labor, etc.
3. To update the data record.
4. To get information immediately.
5. To find the accurate results for the infinite population
6. To test the hypothesis about the population from which sample(s) are drawn.

Types of sampling

Non-Probability Sampling

- Non-probability sampling is a sampling technique where the samples are gathered in a process that does not give all the individuals in the population equal chances of being selected.
- Most researchers are bounded by time, money and workforce and because of these limitations, it is almost impossible to randomly sample the entire population and it is often necessary to employ another sampling technique, the non-probability sampling technique.
- Subjects in a non-probability sample are usually selected on the basis of their accessibility or by the purposive personal judgment of the researcher.
- In this sampling technique, selected sample may or may not represent the entire population. Thus, results cannot be generalized.

Types of Non-probability sampling

- Purposive sampling
- Convenience sampling
- Quota sampling
- Snowball sampling

Purposive sampling

- It is a common non-probability sampling
- In this method, the sample selection is entirely based on the judgement or purposive of the researcher.
- In this method the researcher selects the sample based on the personal judgement that may represent the entire population.
- This method is widely used and more popular for the small sample from small population.
- When using this method the researcher must be confident that the selected sample is truly representative of the entire population
- It is commonly known as judgmental sampling or subjective sampling

If the researcher makes a wise decision and with having a great expertise, this sampling technique provides an accurate, adequate and more representative sample of the population

It helps the researcher for quick decision in any problem arising in collection of data.

However, sometimes the results can be useless due to bias.

Merits

- It is less costly and less time consuming.
- It ensures proper representation of the population when the researcher has full knowledge about the composition of the population.
- It prevents unnecessary and irrelevant items entering into the sample by chance.
- It gives better results if the investigator is unbiased and has the capacity of keen observation and sound judgement.
- It ensures intensive study of the selected items.

Demerits

- It is not suitable for large samples.
- Sampling error can not be calculated.
- There is enough scope for biasness in the selection of sample.
- There is no equal chance for all the observations of the population being selected in the sample.

Convenience sampling

- Convenience sampling is a specific type of non-probability sampling method that relies on data collection from population members who are conveniently available to participate in study.
- This sampling method involves getting participants wherever you can find them and typically wherever is convenient. In convenience sampling no inclusion criteria identified prior to the selection of subjects. All subjects are invited to participate.
- Convenience sampling is used in exploratory research where the researcher is interested in getting an inexpensive approximation of the truth.
- It is often used in preliminary research

- This method is not scientific and may not represent the population.
- But this type of sampling is most useful for pilot survey.

Merits

- It is very useful in pilot study
- It saves money and time

Demerits

- Sampling errors increases
- Difficult to generalize the population from sampling results
- No evidence of representation of population on the sample

Quota sampling

- In quota sampling, a population is first segmented into mutually exclusive sub-groups, just as in stratified sampling. Then judgment is used to select the subjects or units from each segment based on a specified proportion.
- The selection of sample items within the quota is entirely depends on the judgments principles.
- This sampling is generally used in public opinion studies survey.

Merits

- It saves money and times.
- Easy to conduct and easy to get information in short period of time.
- It does not need any sampling.

Demerits

- Selection of the sample is non random so the samples may be biased.
- Within quota the sampling may be unrepresentative

Snowball sampling

- Snowball sampling is a technique for developing the research sample where existing study subjects recruit the future subjects from among their acquaintances.
- Thus the sample group appears to grow like rolling snowball.
- This sampling technique is often used in hidden populations which are difficult for researchers to access. For example, populations would be drug users, or prostitutes, etc.
- Since sample members are not selected from a sampling frame, snowball samples are subject to numerous biases.

- Snowball sampling is a special type of non-probability sampling used when the desired sample characteristics is rare.
- Snowball sampling relies on referrals from initial subjects to generate the additional subjects.
- In snowball sampling researcher begin by identifying someone that meets the criteria for inclusion in their study. Researchers then ask them to recommend others who also meet the criteria.
- Snowball sampling is especially useful when researchers are trying to reach populations that are inaccessible or hard to find.

Merits

- Snowball sampling has been found to be economical, efficient and effective in various studies.
- It can produce in-depth results and can produce relatively quickly.
- It enabled to access to hidden population.

Demerits

- It is quite difficult to generalize from a particular case because samples are not randomly drawn but dependent on the subjective choices of the respondents
- Quality of data and selection bias that limits the validity of the sample.
- Do not represent the population

Probability sampling

- It is a sampling technique in which the probability of getting any particular sample drawn from the population is calculated and each units in the population has an equal chance of being selected.
- Probability sampling is also called random sampling.
- The advantage of probability sampling is that sampling error can be calculated.
- This type of sampling has to be used to make inferences of the study valid and reliable.

Types of Probability sampling

- Simple random sampling
- Systematic sampling
- Stratified random sampling
- Cluster sampling

Simple Random Sampling

- Simple random sampling (SRS) is a method of selection of a sample comprising of n number of sampling units out of the population having N number of sampling units such that every sampling unit has an equal chance of being chosen.
- In this method of sampling, the units are to be selected in such a way that each and every unit of the population has an equal chance of being selected.
- It is very simplest method of probability sampling and it is free from sampling bias because of randomness.
- It is applicable when the population is small, population with similar characteristics and readily available. But difficult and tedious for large population.

The samples can be drawn in two possible ways.

1. The sampling units are chosen without replacement in the sense that the units once chosen are not placed back in the population.
2. The sampling units are chosen with replacement in the sense that the chosen units are placed back in the population.

Simple random sampling with replacement(SRSWR)

SRSWR is a method of selection of n units out of the N units one by one such that at each stage of selection each unit has equal chance of being selected, i.e., 1/N.

If the size of population is 'N' and sample size is 'n', in sampling with replacement, the probability of selection of a unit at each draw remains $\frac{1}{N}$ and the sample of size 'n' can be drawn in N^n ways so that the probability of selecting each sample from the population is $\frac{1}{N^n}$.

Simple random sampling without replacement(SRSWOR)

SRSWOR is a method of selection of n units out of the N units one by one such that at any stage of selection, anyone of the remaining units have same chance of being selected.

If the size of population is 'N' and sample size is 'n', in sampling without replacement, the probability of selecting a unit at $r^{th}\ draw\ is\ \frac{1}{N-r+1}$ and the sample of size 'n' can be drawn in ${}^{N}C_{n}$ ways so that the probability of selecting each sample from the population is $\frac{1}{{}^{N}C_{n}}$.

Merits

- It is a scientific method so there is less personal bias.
- Simple random sampling is representative of the population.
- Reliable and accurate results is obtained at the limited cost, time and labor.
- It is more ideal for statistical purposes.
- It is very easy to assess the sampling error in this method.

Demerits

- It needs complete updated lists of units of population, which may not available.
- It cannot be employed where the units of the population are heterogeneous in nature.
- Numbering the population is too much tedious and time consuming and hard to achieve in practice.

Systematic sampling

- This method needs a complete list of population frame in which the desired sample is to be selected.
- The list of the population is to be arranged in any manner such as alphabetical or numerical or regional or any other ways.
- There are some steps that we need to follow in order to achieve a systematic random sample.
- Firstly, we need to assign that the units are numbered serially in the population from 1 to N, if the population size is N.
- Then we need to decide the sample size n, so that the required sample size is n.

- After then we need to compute the sampling interval size, and the required sampling interval size is calculated as, k= N/n.
- Now the first sample unit is selected randomly an integer between 1 to k.
- Finally we need to take every kth unit from the sampling frame.
- This method is quite popular and widely used if a complete list of the population is available.
- Systematic sampling is to be applied only if the given population is logically homogeneous because systematic sample units are uniformly distributed over the population.
- It is more often applied to field studies when the population is large and scattered.

Merits

- It is very simple and more suitable as well as convenient method for the purpose of study.
- It saves time, money and human resources as well.
- It is much more efficient than the simple random sampling.
- It gives greater accuracy when the population is large but homogeneous.

Demerits

- It can fare very badly if the list has periodic arrangement of the population.
- Updated list of the population may not be available.

Stratified Sampling

- This method is followed when the population is not homogeneous.
- In this method, population first divide into groups on the basis of certain characteristics of the population. These groups are called strata. And each and every stratum made as homogeneous as possible.
- Then the sample units are selected randomly from each stratum using simple random sampling.
- So stratification is the process of grouping the units of the population into relatively homogeneous groups before sampling.

- These groups are based on some predetermined criteria such as geographical location, or demographical characteristic.
- Then random sampling or systematic sampling is applied within each stratum.
- If the sampling from the strata is simple random sampling then whole procedure is called stratified random sampling.

Merits

- Improves the accuracy of the estimation
- The sample is adequate as well as more representative since the samples comes from the homogeneous stratum.
- It provides accurate, reliable and more consistent sample.
- Allows the use of different sampling techniques for different groups (or subpopulation).

Demerits

- It can be difficult to select relevant stratification variables.
- It is not useful when there is no homogeneous subgroups.
- It can be expensive to implement.
- Analysis is quite complicated.
- Problems occurs when strata is not clearly defined.

Cluster sampling

- In cluster sampling, cluster i.e., a group of population elements, constitutes the sampling unit, instead of a single element of the population.
- In this method, a sample of areas is chosen in the first and a sample respondents within those areas is selected in the second. So it is called two stage or multistage sampling.
- In this method, the total population is divided into clusters such as villages, wards, blocks, schools, and a sample of the clusters is selected. The sample comprises a census of each random cluster selected. This means that all elements are taken from selected clusters.

Merits

- Reduced field cost
- Applicable where no complete list of units is available
- It increases the levels of the efficiency of sampling.

Demerits

- Requires group-level information to be known.
- Commonly has higher sampling error than other sampling techniques.
- Cluster sampling may fail to reflect the diversity in the sampling frame.
- Analysis is more complicated than SRS.
- Other elements in the same cluster may share similar characteristics.

Sampling Distribution of Mean (Normal Population)

Firstly, sampling distribution of statistic is defined as the distribution of the sample statistic in terms of frequency distribution or probability distribution.

The frequency distribution or probability distribution formed from the values of sample means of all possible samples selected from a given population is called sampling distribution of sample mean.

Here, sample mean is considered as a random variable. And while selecting samples from the population then we can draw a random sample with and without replacement.

The mean of sampling distribution of a sample mean $(\bar{x})$ is given by

$$E(\bar{x}) = \sum \bar{x} \, . \, p(\bar{x})$$

And also the variance of sampling distribution of a sample mean is given by

$$Var(\bar{x}) = \mathrm{E}\,[\bar{x} - \mathrm{E}(\bar{x})\,]^2 = \sum [\bar{x} - E(\bar{x})]^2 . p(\bar{x})$$

Example

A population consists of 4 units A, B, C, and D with values 8, 2, 6, and 4.

i. Construct a sampling distribution of sample mean by selecting samples of size 2 in random sampling without replacement.

ii. Find the mean and variance of the sampling distribution of the sample mean

iii. Also examine whether the sample mean is an unbiased estimator of population mean.

Solution:

i. Given,

Population size (N) = 4, and sample size (n) = 2

In random sampling without replacement, the total number of possible samples is ${}^{4}C_{2} = 6$ and each sample has equal chance of being selected and that is $\frac{1}{6}$.

The possible samples with their corresponding sample means and probability of selecting each sample are given in the following table:

Sample Unit	Possible samples	Sample mean ($\bar{x}$)	Probability p($\bar{x}$)
AB	(8, 2)	5	$1/6$
AC	(8, 6)	7	$1/6$
AD	(8, 4)	6	$1/6$
BC	(2, 6)	4	$1/6$
BD	(2, 4)	3	$1/6$
CD	(6, 4)	5	$1/6$

The sampling distribution of sample mean is constructed as follows:

Sample mean ($\bar{x}$)	3	4	5	6	7
Probability p($\bar{x}$)	$1/6$	$1/6$	$2/6$	$1/6$	$1/6$

The mean and variance of the sampling distribution of sample mean are calculated as follows:

Sample mean ($\bar{x}$)	Probability p($\bar{x}$)	$\bar{x}$.p($\bar{x}$)	$\bar{x}$ - E($\bar{x}$)	$[\bar{x} - \mathrm{E}(\bar{x})]^2$	$[\bar{x} - \mathrm{E}(\bar{x})]^2$. p($\bar{x}$)
3	$1/6$	$3/6$	-2	4	4/6
4	$1/6$	$4/6$	-1	1	1/6
5	$2/6$	$10/6$	0	0	0
6	$1/6$	$6/6$	1	1	1/6
7	$1/6$	$7/6$	2	4	4/6
Total	1	$\Sigma\bar{x}$.p($\bar{x}$)=5		18	$\Sigma[\bar{x} - \mathrm{E}(\bar{x})]^2$. p($\bar{x}$) =10/6

ii. The mean of the sampling distribution of sample mean is given by

$$E(\bar{x}) = \sum \bar{x}.p(\bar{x}) = \frac{30}{6} = 5$$

And the variance of the sampling distribution of sample mean is given by

$$Var(\bar{x})_{WOR} = \mathrm{E}\,[\bar{x} - \mathrm{E}(\bar{x})\,]^2$$

$$= \sum [\bar{x} - \mathrm{E}(\bar{x})\,]^2 \,.p(\bar{x})$$

$$= 4 \times \frac{1}{6} + 1 \times \frac{1}{6} + 0 \times \frac{2}{6} + 1 \times \frac{1}{6} + 4 \times \frac{1}{6} = \frac{10}{6}$$

$$= 1.667$$

iii. Since, $E(\bar{x}) = 5$ and the population mean is given by

$$\mu = \frac{\sum X}{N} = \frac{8+2+6+4}{4} = 5$$

Here, we get $E(\bar{x}) = \mu = 5$

Hence, the sample mean is an unbiased estimator of population mean.

Example

Suppose a population created by recording the number of days with temperature more than 30℃ per month in 4 months of a summer season in Kathmandu valley is 2, 4, 6, and 8.

i. Construct a sampling distribution of sample mean by selecting he samples of size 2 with replacement.

ii. Find the mean of the sampling distribution of sample mean and show that the sample mean is equal to the population mean.

iii. Find the variance of the sampling distribution of sample mean.

Given, N = 4

‘n’ = 2

In sampling with replacement, the number of possible samples is

$N^n = 4^2 = 16$

The each sample has equal chance of being selected i.e. 1/16.

The possible samples and their corresponding sample means and also the probability of selecting each sample is given by

Possible samples	Sample means ($\bar{x}$)	Probability p($\bar{x}$)
(2, 2)	2	1/16
(2, 4)	3	1/16
(2, 6)	4	1/16
(2, 8)	5	1/16
(4, 2)	3	1/16
(4,4)	4	1/16
(4,6)	5	1/16
(4,8)	6	1/16
(6,2)	4	1/16
(6, 4)	5	1/16
(6, 6)	6	1/16
(6,8)	7	1/16
(8,2)	5	1/16
(8,4)	6	1/16
(8, 6)	7	1/16
(8, 8)	8	1/16

The sampling distribution of sample mean is given by

Sample mean $(\bar{x})$	probability $p(\bar{x})$
2	1/16
3	2/16
4	3/16
5	4/16
6	3/16
7	2/16
8	1/16
Total	1

- The mean of the sampling distribution of sample mean is given by

Sample means $(\bar{x})$	Probability $p(\bar{x})$	$(\bar{x}).\, p(\bar{x})$
2	1/16	2/16
3	2/16	6/16
4	3/16	12/16
5	4/16	20/16
6	3/16	18/16
7	2/16	14/16
8	1/16	8/16
Total	1	80/16

Now, the mean of the sample mean is given by

$$E(\bar{x}) = \sum \bar{x}.p(\bar{x}) = \frac{80}{16} = 5$$

The variance of the sampling distribution of sample mean

Sample mean ($\bar{x}$)	Prob. $p(\overline{x})$	$\bar{x} - E(\bar{x})$	$[\bar{x} - E(\bar{x})]^2$	$[\bar{x} - E(\bar{x})]^2 . p(\overline{x})$
2	1/16	-3	9	9/16
3	2/16	-2	4	8/16
4	3/16	-1	1	3/16
5	4/16	0	0	0
6	3/16	1	1	3/16
7	2/16	2	4	8/16
8	1/16	3	9	9/16
Total	1			40/16

The variance of sample mean is calculated as

$$Var(\bar{x})_{WR} = \mathrm{E}\left[\bar{x} - \mathrm{E}(\bar{x})\right]^2$$

$$= \sum[\bar{x} - \mathrm{E}(\bar{x})]^2 \,.\, p(\bar{x}) = \frac{40}{16} = 2.5$$

Example

Following data represent the amount of water (in liter) taken by 4 members of a family (or population) 1, 2, 3, and 4. Then

i. Construct a sampling distribution of mean of samples of size 2 using simple random sampling with replacement.

ii. Find mean and standard error of the sampling distribution of sample mean.

Standard Error (with replacement i.e. infinite population)

Statistic	Standard Error (S.E)
sample mean($\bar{x}$)	$\sigma/\sqrt{n}$ or $\frac{s}{\sqrt{n}}$
Sample proportion (p)	$\sqrt{\frac{PQ}{n}}$ or $\sqrt{\frac{pq}{n}}$
$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$ or $\sqrt{\frac{s_1^2}{n_1} + \frac{{s_2}^2}{n_2}}$
$p_1 - p_2$	$\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$ or $\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$

Standard Error (without replacement i.e. finite population)

Statistics	S.E
Sample mean ($\bar{x}$)	$\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$
Sample proportion (p)	$\sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}}$

Note:

1. $standard\ error, S.E\ (\bar{x}) = \sqrt{Var(\bar{x})}$

2. $Sample\ variance, Var(x) = s^2 = \frac{1}{n-1}\sum(x - \bar{x})^2$

Here, x is the value of sample observation.

3. Standard Error is the measure of the accuracy of a mean and an estimate.

4. A small SE is an indication that the sample mean is a more accurate reflection of the actual population mean. A larger sample size will normally result in a smaller SE (while SD is not directly affected by sample size).

5. SE helps us to understand how far a sample mean is from the true population mean, then we can use this to understand how accurate any individual sample mean is in relation to the true population mean.

Note:

Population proportion (P) = $\frac{X}{N}$ = Proportion of success in Population.

Where, ‘X’ is the number of success in the population and ‘N’ is the size of population.

And $Q = 1 - P$, known as Proportion of failure in population.

Similarly, sample proportion (p) = $\frac{x}{n}$ = Proportion of success in sample.

Where, ‘x’ is the number of success in sample and ‘n’ is the size of sample

And q = 1 – p , known as Proportion of failure in sample.

Example

1. A random sample of 25 ballet papers are drawn without replacement from a box of 700ballet papers of a election booth. If the number of invalid ballet papers in the box is 140. find the standard error of the sample proportion of the invalid ballet papers in the sample.
2. Assume that the life time of 1000 electric bulbs manufactured by a certain company are normally distributed with mean life of 1600 hrs and standard deviation of 100 hrs. If the sample of size 49 are randomly selected from the population,

a. With replacement (infinite population or population size does not specified)

b. Without replacement (finite and population size is specified)

Find the standard error of the sampling distribution of sample mean.

1. Population size N =700

Sample size , n = 25

No. of invalid ballet papers , X = 140

Population Proportion , P = X/N= 140 / 700 =0.2

Q = 1 – P = 1- 0.2 = 0.8

Now the standard error of sample proportion (w/o replacement) is given by $\mathrm{S}.E(p) = \sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}}$

Estimation

The theory of estimation was found by Professor R. A. Fisher.

Definition

The statistical method of estimating the population parameter (s) from the sample observations drawn from the population is called estimation.

The estimation of the population parameter (s) can be done in either of the following two ways:

1. Point Estimation
2. Interval Estimation

Point Estimation

The procedure of estimating the population parameter (s) in a single value using the sample observations which is drawn from the population is called Point estimation.

A statistic t is said to be a point estimator of the parameter θ, if the estimate falls nearest to the true value of the parameter θ . Thus , the point estimation provides nearly exact or true value of the unknown parameter to be estimated under investigation.

Properties of a Good Estimator

The estimator is said to be good estimator of the parameter if it satisfies the following properties.

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

Unbiasedness

A statistic t, a function of sample observations $x_1, x_2, \ldots\ldots\ldots, x_n$ is said to be an unbiased estimator of population parameter θ if

$$E(t) = \theta$$

In other words, a statistic t is said to be an unbiased estimator of population parameter θ if the mean of the sampling distribution of t is equal to θ .

This unbiasedness is a property of examining a good estimator through average or expectation.

Consistency

An estimator t, based on a sample size of n, is said to be a consistent estimator of population parameter θ if t converges in terms of probability to the parameter θ as n $\rightarrow \infty$ i.e. $t \xrightarrow{P} \theta \; as \; n \; \rightarrow \infty$.

That is if $\lim_{n \rightarrow \infty} P(|t - \theta| \geq \varepsilon) = 0, for \; every \; \varepsilon > 0$

Or equivalently, if $\lim_{n \rightarrow \infty} P(|t - \theta| \leq \varepsilon) = 1, for \; every \; \varepsilon > 0$

Sufficiency

A statistic t is said to be sufficient estimator of θ if it contains all the information about the parameter θ.

This definition gives the concept that if we have a random sample $x_1, x_2, \ldots\ldots\ldots, x_n$ from a population with a probability function $f(x, \theta)$, a statistic t which provides as much information as the random sample $x_1, x_2, \ldots\ldots\ldots, x_n$ could reveal about the population parameter θ.

In other words, the sufficient statistic may be defined as follows:

Let X = $(x_1, x_2, \ldots\ldots\ldots, x_n)$ be a sample of size n from $\{f(x, \theta); \theta \epsilon \Theta\}$. Then a statistic $t = t_n(x)$ is said to be a sufficient estimator of θ, if the conditional distribution of X given $t = t_n(x)$ is independent of θ.

Efficiency

If $t_1 and\ t_2$ are two consistent estimators of a parameter θ. Then, the estimator t_1 is said to be more efficient estimator than t_2, if variance of t_1 is less than variance of t_2 i.e. var(t_1) < var(t_2).

If t_1 be most efficient estimator with variance var(t_1) and t_2 is any other estimator with variance var(t_2) , then the relative efficiency E of t_2 as compared to t_1 is defined as

$$E = \frac{var(t_1)}{var(t_2)}$$

Here, it is obvious that 0 < E < 1.

Interval Estimation

Since the point estimator does not coincide with a true value of a population parameter θ and also fails to provide degree of uncertainty which bound to occur. So it is preferred to another estimator as an interval which provides the uncertainty in estimation. A procedure of finding such interval that includes the true value of population parameter θ with a certain probability is called interval estimation.

The values of t_1 and t_2 can be obtain when any probability, say (1- α) is given such that

$$\int_{t_1}^{t_2} f(t)dt = 1 - \alpha$$

$$\rightarrow \quad P(t_1 < \theta < t_2) = 1 - \alpha \,,$$

where $P(\theta < t_1) = {}^{\alpha}/_{2}$ and $P(\theta > t_2) = {}^{\alpha}/_{2}$

Confidence Interval

The interval (t_1, t_2) within which the true value of unknown population parameter θ is expected to lie, with a certain probability or confidence $(1-\alpha)$ is called confidence interval. The confidence interval is also known as confidence limits or fiducial limits. The probability $1-\alpha$ is called confidence level. A confidence interval of confidence level $1-\alpha$ is often referred to as $100(1-\alpha)\%$ confidence interval.

Confidence Interval for Mean of Normal Population

Case I : Large sample case:

i. when population variance σ^2 is known.

Here large sample means the sample size is greater than 30 i.e. n > 30.

100(1−α)% confidence interval for mean μ is given by

$$\left(\bar{x} - z\alpha/_2 \frac{\sigma}{\sqrt{n}} , \quad \bar{x} + z\alpha/_2 \frac{\sigma}{\sqrt{n}}\right)$$

Here, lower limit = $\bar{x} - z\alpha/_2 \frac{\sigma}{\sqrt{n}}$ and

Upper limit = $\bar{x} + z\alpha/_2 \frac{\sigma}{\sqrt{n}}$

Here, $\bar{x}$ is the sample mean

$z_{\alpha/2}$ is the value of Z (obtained from SNT)

σ is the population standard deviation

'n' is the sample size.

Remarks:

1. If the population is finite of size N, then the $100(1-\alpha)\%$ confidence interval for mean μ is given by

$$\left(\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}\right)$$

2. The maximum error of the estimate is denoted by E and is given by

$$E = Z\alpha_{/2}\frac{\sigma}{\sqrt{n}}$$

ii. When σ^2 is not known:

Usually, the population variance σ^2 is not known.

<u>100(1−α)% confidence interval for mean μ is given by</u>

$$\left(\bar{x} - z\alpha/_2 \frac{s}{\sqrt{n}} , \quad \bar{x} + z\alpha/_2 \frac{s}{\sqrt{n}}\right)$$

Here, s is the sample standard deviation.

Example

A random sample of 40 students is drawn from a certain campus and their height showed a mean of 68.6 inches and a standard deviation of 2.5 inches. Construct a 95% confidence interval for mean height of all the students of the campus.

Solution:

Given, sample size, (n) = 40 (large sample)

Sample mean, $\bar{x}$ = 68.6

Sample standard deviation, s = 2.5

And $100(1 - \alpha)\% = 95\% \Rightarrow 1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$

$\therefore z_{0.025} = 1.96$ from table

Continue…

Hence, the 95% confidence interval for mean μ is given by

$$\left(\bar{x} - z\alpha/_2 \frac{s}{\sqrt{n}} \ , \ \bar{x} + z\alpha/_2 \frac{s}{\sqrt{n}}\right)$$

$$\Rightarrow \left(68.6 - 1.96 \times \frac{2.5}{\sqrt{40}} \ , 68.6 + 1.96 \times \frac{2.5}{\sqrt{40}}\right)$$

$$\Rightarrow (67.726, 69.274)$$

Which is the required confidence interval.

Example

An analysis for PH (a measure of acidity) in a random sample of water from 40 rainfalls showed that mean PH is 3.7 and standard deviation is 0.5. find a 99% confidence interval for the mean PH in rainfalls.

Solution:

Given, sample size (n) = 40 (large sample)

Sample mean, $\bar{x} = 3.7$

Sample standard deviation, s = 0.5

And $100(1 - \alpha) = 99\% \Rightarrow 1 - \alpha = 0.99 \Rightarrow \boldsymbol{\alpha = 0.01}$

$\therefore z_{0.005} = 1.645$

Hence, the 99% confidence interval for mean μ is given by

$$\left(\bar{x} - z\alpha/_2 \frac{s}{\sqrt{n}} \ , \ \bar{x} + z\alpha/_2 \frac{s}{\sqrt{n}}\right)$$

$$\Rightarrow \left(3.7 - 1.645 \times \frac{0.5}{\sqrt{40}} \ , 3.7 + 1.645 \times \frac{0.5}{\sqrt{40}}\right)$$

$$\Rightarrow (3.496, 3.904)$$

Required confidence interval.

Practice

1. The standard deviation for a method of measuring the concentration of nitrate ions in water is known to be 0.05. If 100 measurements give a mean of 1.13, find the 95% and 98% confidence limits for the true mean.
2. A random sample of size 50 drawn from normal population has mean of 68.5 and s.d. of 2.7. Construct 98% fiducial limits for population mean. Also determine the maximum size of error of the estimated value of sample mean.

Case II: Small Sample Case ($n \leq 30$).

$100(1-\alpha)\%$ confidence interval for mean μ is given by

$$\left(\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\ ,\qquad \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right)$$

Here, s is the unbiased sample standard deviation

Since, the unbiased estimator of the population variance is given by

$$s^2 = \frac{1}{n-1}\sum(x-\bar{x})^2$$

Sometimes, the biased sample variance or sample standard deviation is given. In this case,

$100(1-\alpha)\%$ confidence interval for mean μ is given by

$$\left(\bar{x} - t_{\alpha/2,n-1}\frac{s_b}{\sqrt{n-1}}\ ,\qquad \bar{x} + t_{\alpha/2,n-1}\frac{s_b}{\sqrt{n-1}}\right)$$

Here, s_b is the biased sample standard deviation.

And $t_{\alpha/2,n-1}$ is the value of t with (n – 1) degree of freedom.

The maximum error of the estimate is given by

$$E = t_{\alpha/2,n-1}\cdot\frac{s}{\sqrt{n}}$$

Example

The sample data drawn from a normal population is given below:

6.5, 15.2, 6.4, 18.9, 12.7, 5.3, 21.6, 9.4

Construct a 95% confidence interval for the population mean.

Solution:

Here, sample size (n) = 8 (small sample)

Firstly, we calculate sample mean ($\bar{x}$) and sample variance s^2.

Continue…

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
6.5	- 5.5	30.25
15.2	3.2	10.24
6.4	- 5.6	31.36
18.9	6.9	47.61
12.7	0.7	0.49
5.3	- 6.7	44.89
21.6	9.6	92.16
9.4	- 2.6	6.76
$\Sigma x_i = 96$		$\Sigma(x_i - \bar{x})^2 = 263.76$

Continue..

The sample mean is given by

$$\bar{x} = \frac{\sum x_i}{n} = \frac{96}{8} = 12$$

And the sample variance is given by

$$s^2 = \frac{1}{n-1}\sum(x - \bar{x})^2$$

$$= \frac{1}{8-1} \times 263.76 = 37.68$$

$$\Rightarrow s = 6.1384$$

Also, $100(1 - \alpha)\% = 95\% \Rightarrow 1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$ (two – tailed test)

$$\therefore t_{\alpha/2, n-1} = t_{0.025,7} = 2.365$$

Continue..

Hence, the 95% confidence interval for population mean, for small sample, is given by

$$\left(\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\ ,\ \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right)$$

$$= \left(12 - 2.36\ X\ \frac{6.1384}{\sqrt{8}}\ , 12 + 2.36\ X\ \frac{6.1384}{\sqrt{8}}\right)$$

$$= (12 - 5.13,\ \ 12 + 5.13)$$

$$= (6.87,\ \ 17.13)$$

Here, $since, t_{\alpha/2,n-1} = t_{0.025,\,7} = 2.36$

Practice

1. The daily expenditure (in Rs.) of 7 students in a certain college are 102, 98, 104, 102, 98, 96, 100. Find a 95% confidence interval for the true mean expenditure, assuming as approximate normal distribution.
2. A random sample of 25 women of fertile age is drawn from the population of certain Metropolitan City and the sample gave the following age distribution:

Age of Women	15 - 19	20 - 24	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49
No. of Women	4	5	7	5	2	1	1

Construct 98% fiducial interval for the average fertile age of all women of the Metropolitan City. Also find maximum size of error of the estimate of sample average fertile age of women.

Solution: 1

Given, sample size , (n) =7

Calculate sample mean and sample variance

X	$x - \bar{x}$	$[X - \bar{x})$]square
102	2	4
98	-2	4
104	4	16
102	2	4
98	-2	4
96	-4	16
100	0	0
$\Sigma x = 700$		= 48

The sample mean is given by

$$\bar{x} = \frac{\sum x}{n} = \frac{700}{7} = 100$$

Similarly, the sample variance is given by

$$s^2 = \frac{1}{n-1}\sum(x - \bar{x})^2 = \frac{48}{6} = 8$$

S = 2.82

Also 100(1 – α)% = 95%

1 – α = 0.95

‘α ‘= 0.05

$t_{0.025,6} = 2.447$

Now, the 95% confidence interval for true population is given by (small sample)

$$\left(\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\ ,\ \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right)$$

$$\left(100 - 2.447\ X\ \frac{2.82}{\sqrt{7}}\ , 100 + 2.447\ X\ \frac{2.82}{\sqrt{7}}\right)$$

$$(97.385\ , 102.615\ )$$

Which is the required confidence interval.

Practice

3. A random sample of 12 value from a normal population showed a mean of 31.9 inches and sum of square of deviation from this mean equal to 125 square inches. Obtain on 98% and 99% fiducial limits for population mean.

4. The 10 bearings made by a certain process have a mean diameter of 0.0506 cm and with standard deviation of 0.0040 cm. Assuming that the data may be looked upon as a random sample from a normal population, what can we assert with 95% confidence about maximum error.

Note: here given s.d. is biased s.d

Solution 3 :

Given, sample size, (n) = 12 (small sample)

Sample mean, $\bar{x}$ = 31.9

Sum of squares of deviation from this mean , $\sum(x - \bar{x})^2$= 125

Now, the sample variance is calculated as

$$s^2 = \frac{1}{n-1}\sum(x - \bar{x})^2 = \frac{125}{11} = 11.36$$

S = 3.370

For 98% confidence interval

100(1 – α) = 98%

Or, 1 – α = 0.98

Or α = 0.02

$t_{0.01,11} = 2.718$

Now, the 98% C.I. for population mean is given by

$$\left(\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\ ,\ \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right)$$

Similarly, for 99% C.I

Here, 100(1 –α)% = 99%

1 – α = 0.99

‘α’ = 0.01

$t_{0.005,11} = 3.106$

Now, the 99% C.I for population mean is given by

$$\left(\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\ ,\ \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right)$$

Examples

1. Hotel's manager in Kathmandu wants to know the hotel's average daily registration . The following table presents the numbers of guest registered each of 27 randomly selected days. Calculate the sample mean, standard errors of mean, and construct 95% confidence limits of the population mean.

61, 57, 53, 60 64, 57, 54, 58, 63, 61, 50, 59, 50, 60, 57, 58, 62, 63, 60, 54, 54, 61, 51, 53, 62, 57, 60 .

Confidence Interval for the population proportion (large sample size i.e. n > 30)

Population proportion (P) = $\frac{X}{N}$ = Proportion of success in Population.

Where, ‘X’ is the number of success in the population and ‘N’ is the size of population.

And $Q = 1 - P$, known as Proportion of failure in population.

Similarly, sample proportion (p) = $\frac{x}{n}$ = Proportion of success in sample.

Where, ‘x’ is the number of success in sample and ‘n’ is the size of sample

And q = 1 – p , known as Proportion of failure in sample.

Standard Error, S.E. (p) = $\sqrt{\frac{PQ}{n}}$.

When Population Proportion (P) is unknown,

In this case, S.E. (p) = $\sqrt{\frac{pq}{n}}$

The 100(1 – α)% confidence interval for population proportion (P) is given by

$$\left(p \pm Z\alpha_{/2} \sqrt{\frac{PQ}{n}} \right)$$

Sometimes, P is not known and this is replaced by its unbiased estimate p and in this case, the 100(1 – α)% confidence interval for P is given by

$$\left(p \pm Z\alpha_{/2}\sqrt{\frac{pq}{n}}\right)$$

Remarks:

1. The 100(1 – α)% confidence interval for difference of two population means $(\mu_1 - \mu_2)$, for large sample, is given by

$$\left\{(\bar{x} - \bar{y}) \pm Z\alpha_{/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right\}$$

$\bar{x}$ is the mean of first sample

$\bar{y}$ is the mean of second sample

σ_1^2 is the variance of first sample

σ_2^2 is the variance of second sample

n_1 is the size of first sample

n_2 is the size of second sample

2. The 100(1 – α)% confidence interval for difference of two population means $(\mu_1 - \mu_2)$, for small sample, is given by

$$\left\{(\bar{x} - \bar{y}) \pm t\alpha/_{2,n_1+n_2-2} \cdot s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right\}$$

Here, s is the unbiased estimate of population standard $\boldsymbol{\sigma}$

Where, the pooled or common sample standard deviation is given by.

$$s = \sqrt{\frac{1}{n_1+n_2-2}[\sum(x-\bar{x})^2 + \sum(y-\bar{y})^2]}$$

3. For large sample, the 100(1 – α)% confidence interval for difference of two population proportion $(P_1 - P_2)$ is given by

$$\left[(p_1 - p_2) \pm Z\alpha_{/2}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right]$$

p_1 is the proportion of success in first sample

p_2 is the proportion of success in second sample

And $q_1 = 1 - p_1$

Also $q_2 = 1 - p_2$

Example

1. A biased coin was thrown 400 times and heads were resulted 240 times. Find the 90% confidence interval for population proportion.
2. A sample poll of 100 voters chosen at random from all voters in a given district shows that 55% of them were in favour of a particular candidate. Find the 95% and 99% confidence limits for the proportion of voters in his favour, if a very large number of voters are allowed to cast their voters.
3. A random sample of 500 apples was taken from a large consignment and 60 were found to be bad. Obtain the 98% confidence limit for the percentage number of good apples in the consignment.

4. A machine puts out 16 imperfect articles in a sample of 500 articles. After the machine is overhauled, it puts out 3 imperfect articles in another sample of 100. Construct a 90% confidence interval of true difference between two population proportions.

5. A random sample of 40 tube lights of brand A showed that the average life time of the tube lights is 418 hours of continuous use. Another random sample of 50 tube lights of brand B showed the same figure is 402 hours continuous use. The population standard deviation are known to be 26 hours and 22 hours for two brands A and B respectively. Construct 97% confidence interval for the difference between the mean life times of the two brands of the tube lights.

Determine Sample Size

1. The sample size ‘n’ is determined with 1 – α confidence level when population standard deviation is given as:

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

Here, $Z_{\alpha/2}$ is the value of Z (Use from table)

σ is the population standard deviation

E is the allowable error or margin of error

‘n’ is sample size

The sample size ‘n’ is determined with 1 – α confidence level using population proportion P is given by

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

Where Q = 1 – P

Note: sometimes P is unknown then we use P = Q = 0.5

Example

1. If the proportion of people opposing the death penalty is thought to be 0.52, how many people should have been interviewed by the department of law in order to estimate the proportion, so that the department could be 95% confidence that the population proportion is within ± 0.1 of the sample proportion?
2. A researcher wants to determine the average time it takes a mechanic to rotate the tires of a car, and she want to be able to assert with 95% confidence that mean of her sample is off by at most 0.5 minute. If she can presume from the past experiment that population standard deviation is 1.6 minutes, how large a sample will she have to take?

Practice:

1. Suppose that we want to estimate the true proportion of defectives in a very large shipment of adobe bricks, and that we want to be at least 95% confident that the error is at most 0.04. How large a sample will we need if

a. We have no idea what the true proportion might be;

b. We know that the true proportion does not exceed 0.12?