Estimation

The purpose of statistical inference is to obtain information about a population from information contained in a sample. Making decisions regarding population based on sample. The techniques of statistical inference can be divided into two major areas: **Parameter estimation** and **Hypothesis testing**.

There are two major approaches in parameter estimation.

1. **Classical approach or frequentist approach**: Inferences are based strictly on information obtained from a random sample selected from the population.

2. **Bayesian approach**: This method utilizes prior subjective knowledge about the probability distribution of the unknown parameter in conjunction with the information provided by the sample data.

Key Differences:

- **Prior Information:** The classical approach assumes no prior knowledge about the parameter, while the Bayesian approach explicitly incorporates it.
- **Parameter Treatment:** The classical approach treats the parameter as a fixed but unknown quantity, while the Bayesian approach treats it as a random variable with a probability distribution.
- **Inference:** The classical approach makes inferences about the parameter based solely on the sample data, while the Bayesian approach updates the prior belief with the sample data to obtain a posterior distribution.
- **Uncertainty Quantification:** The classical approach provides point estimates and confidence intervals to quantify uncertainty, while the Bayesian approach provides a full probability distribution (posterior distribution) for the parameter.

Classical method of estimation

The classical method of parameter estimation refers to techniques based on the frequentist approach, where parameters are treated as fixed but unknown constants, and the goal is to estimate these parameters using sample data. Classical methods rely on the properties of the data and sampling distributions without incorporating prior knowledge.

We can compute two types of estimates about a population parameter.

1. Point Estimation
2. Interval Estimation

Point Estimation

Point estimation is a simpler approach that provides a quick and easy estimate of unknown parameter. It involves using a single value calculated from sample data to estimate an unknown population parameter. This single value is called a point estimate.

For a population parameter θ, it is a single numerical value of a statistic $\hat{\theta}$ that corresponds to that parameter. Any statistic that is used to estimate population parameter is called estimator. A parameter may be estimated by more than one statistic or estimator. An estimate is the specific observed value of the estimator. For repeated sampling we get more than one estimates.

Example: If we want to estimate the average height of all bachelor level students at NCCS college, we might take a sample of 50 adults, calculate their average height, and use that as a point estimate for the population's average height.

Estimation problems occur frequently in many sciences. We often need to estimate the following parameter.

- The mean of a variable in certain population (μ)
  The point estimate for this parameter is $\overline{X}$, the sample mean.
- The variance of a variable in a certain population (σ^2)
  The point estimate for this parameter is s^2, the sample standard deviation.
- The proportion of items in a population that belong to a class of interest (π)
  The reasonable point estimate for π is p or$\hat{\pi} = \frac{X}{n}$, where X = number of items in a random sample of size n that belongs to a category of interest.
- The correlation coefficient of two variables in a population (ρ)
  The point estimate for this parameter is r, the sample correlation coefficient.

Interval Estimation

Interval estimation_is a more robust approach that provides a more accurate and informative estimate by considering the uncertainty in the sample data. It provides a range of values within which the true population parameter is likely to lie, with a specified level of confidence. This range is called a confidence interval.

In many situations, a point estimate does not provide enough information about the parameter of interest. If we are estimating the parameter θ, a single number or value of estimator $\hat{\theta}$ may not be meaningful. It is because random sampling inherently involves chances so $\hat{\theta}$can't be expected to be equal to θ. An interval estimation of the form$\hat{\theta}_L \leq \theta \leq \hat{\theta}_U$ might be more useful. The end points $\hat{\theta}_L$ and $\hat{\theta}_U$ of the interval will be random variables, since they are functions of sample data.

Differences between point estimation and interval estimation:

Point Estimation	Interval Estimation
Offers less information about the precision of the estimate.	Offers more information about the precision of the estimate by specifying a confidence level.
Provides a single best-guess value for an unknown parameter.	Provides a range of plausible values for an unknown parameter
Does not account for variability or uncertainty	Explicitly quantifies the uncertainty in the estimate

Assumes the parameter is exactly the given value	Indicates the parameter is likely to lie within the specified range.
When a precise single value is needed for quick decisions	When understanding the reliability or variability of estimates is important
Simple to compute	More complex, often requiring knowledge of the sampling distribution

Why interval estimation more useful than point estimation?
As interval estimation provides a range of values, known as a confidence interval, within which the true population parameter is likely to lie, quantifies the uncertainty associated with the estimate by providing a range of plausible values and offers more information about the precision of the estimate by specifying a confidence level, it is more useful than the point estimation.

Construction of confidence interval
In order to construct interval, estimate of the unknown parameter the sampling distribution of $\hat{\theta}$ should be known. Once the sampling distribution of $\hat{\theta}$ is known we can construct an interval estimate of the unknown parameter θ using probability distribution of $\hat{\theta}$. We can find two statistics $\hat{\theta}_L$ and $\hat{\theta}_U$, such that

$$P\left(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U\right) = 1 - \alpha$$

The interval $\hat{\theta}_L \leq \theta \leq \hat{\theta}_U$ constructed from selected sample is called (1-α) x 100 % confidence interval, 0 ≤ α ≤1, and the fraction 1 – α is called the confidence coefficient or the degree of confidence. The end points $\hat{\theta}_L$and $\hat{\theta}_U$ are called lower and upper limits. The difference between the upper limit and lower limit i.e. $\hat{\theta}_U - \hat{\theta}_L$ is called interval length.

A confidence interval is a specific interval estimate of parameter determined by using data obtained from a sample and by using the specific confidence level of the estimate.

In general, two-sided confidence interval may be expressed as follows

Estimator ± Margin of error
= Estimator ± (Reliability coefficient) x (Standard error of estimator)

Significance level: It is the chance that the population parameter θ we are estimating will not be within our confidence limit i.e. it will lie outside it. It is denoted by α. We can set this probability 0 but only in the expense of much wider confidence interval. It is meaningless to estimate the parameter if the interval length is higher. So, it has to be as shorter as possible but there is some chance that our parameter will be missed out, what we called significance probability. Our confidence interval must be as shorter as possible and significance probability as low as possible. The typical value of α we choose is 0.05 or lower.

Confidence level: The confidence level of an interval estimate of a parameter is the probability that the interval estimate will contain the parameter θ and it is denoted by 1-α. The significance level and confidence level are complementary probabilities.

Interpretation of (1-α) x 100 % confidence interval

If many random samples are collected and a (1-α) x 100 % confidence interval on θ is compute for each sample, then (1-α) x 100 % of these intervals will contain the true value of θ and α x 100 % of these intervals will not contain θ.

The appropriate statement would be that θ lies in the observed interval $[\hat{\theta}_L, \hat{\theta}_U]$ with confidence level (1-α) x 100 %.

Relationship between confidence interval and confidence level

There is close connection between confidence interval and confidence level. Interval length $\hat{\theta}_U - \hat{\theta}_L$ increases with confidence level (1-α) x 100 %. We can increase the confidence interval by increasing the confidence level. Wider the confidence interval is, the more confident we can be that the given interval contains the unknown parameter θ. It is more meaningful to predict the unknown parameter θ in close interval than in wider interval but in doing so we are losing confidence level. It is better to be 95 % confident that average duration illness of certain disease D is between 7 to 11 days than to be 99 % confident that it is between 3 to 15 days. Ideally, we prefer a short interval with a high degree of confidence. Most of the time we use 95 % confidence level for estimation purpose by setting α = 5%. If we have to be more precise, we use 99 % confidence level. But we should not take it as rule; we can use any confidence level to construct confidence interval for parameter θ.

Choosing an appropriate estimator

Any sample statistic that is used to estimate a population parameter is called an estimator. An estimator is not expected to estimate the population parameter without error.

Statistic = Parameter + Systematic Error (Bias) + Chance error (Random error)

When taking a sample in survey or running an experiment or doing observational study, randomness is important for minimizing bias. If we assume that the samples are random and that there is no bias, then

Statistic = Parameter + Chance error

Chance error still exists, because when the samples are random, the statistics will be random and will vary from sample to sample. The sampling distribution, is an important concept in inferential statistics, which when known, one has a good idea of the likely size of the chance error.

In many cases, a parameter θ may be estimated by more than one estimator, say $\hat{\theta}_1$, $\hat{\theta}_2$, …etc. For example, sample mean or sample median can be used to estimate the population mean.

Which estimator to use to estimate a given parameter?

The following are the desirable properties of a ‘good’ estimator.

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

Unbiasedness: An estimator is unbiased if its expected value equals the true value of the parameter being estimated. A good estimator should be unbiased i.e. it should be close to the true value of the unknown parameter. This ensures that the estimator, on average, is correct.

Mathematically,

An estimator $\hat{\theta}$ is called an unbiased estimator of the population parameter θ if,

$$E(\hat{\theta}) = \theta.$$

It means that the average values of $\hat{\theta}$ taken over all samples it is equal to the population parameter θ. The term $E(\hat{\theta})$ is read as expected value of $\hat{\theta}$.

Parameter	**Unbiased Estimator**
Population mean (μ)	Sample mean ($\overline{X}$) (for both SRSWR and SRWOR)
Population proportion of success (π)	Sample proportion of success (p) (for both SRSWR and SRSWOR)
Population variance (σ^2)	Sample variance (s^2) (for SRSWR)
Modified population variance (S^2)	Sample variance (s^2) (for SRSWOR)

Note: Sample variance (s^2) is biased for (σ^2) if sampling plan is SRSWOR because $E(s^2) = \frac{N}{N-1}\,\sigma^2$.

The amount of biased is given by,

$$\text{Bias} = E\,(s^2) - \sigma^2 = \frac{N}{N-1}$$

For large population size N, the bias is negligible.

Efficiency:

An efficient estimator has the smallest possible variance among all unbiased estimators of a parameter. Lower variance means more reliable estimates with less spread. Efficiency refers to the size of the standard error of the statistic.

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimator of population parameter θ, then $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ if,

$$Var\ (\theta_1) < Var\ (\theta_2)$$

$$\Rightarrow S.E.(\theta_1) < S.E.(\theta_2)$$

If we consider all possible unbiased estimators of some parameter θ, the one with the smallest variance is called the most efficient estimator of θ. Thus, an estimator with a smaller variance will have more chance of producing an estimate that is close to the true value of the unknown parameter.

Consistency:

An estimator is consistent if it converges to the true parameter value as the sample size increases.

A statistic $\hat{\theta}$ is a consistent estimator of a population parameter θ if its variance tends to zero as n tends to infinity.

i.e. $Var\ (\theta) \to 0\ as\ n \to \infty$

It means that as the sample size increases, it becomes almost certain that the value of the statistic comes very close to the value of the population parameter. Thus, consistent estimator is more reliable with large sample. This guarantees that with a large enough sample, the estimate will be arbitrarily close to the true value.

Sufficiency: An estimator is sufficient if it uses all the information in the data relevant to the parameter. A sufficient statistic captures all the information about the parameter contained in the sample data. For example, the sample mean is more sufficient estimator than sample mean in estimating population mean because it makes use of more information contained in sample than sample median because sample mean considers the magnitude of individual data in the sample while median only considers its rank order.

Interval Estimation of population mean μ (population standard deviation σ is known)

The sampling distribution of $\overline{x}$ is a normal probability distribution with mean μ and standard deviation of $\frac{\sigma}{\sqrt{n}}$ provided that the population standard deviation σ is known, irrespective of sample size.
So,

$$Pr\left\{-Z_{\frac{\alpha}{2}} \leq \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq +Z_{\frac{\alpha}{2}}\right\} = 1 - \alpha$$

or, $$Pr\left\{-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \overline{x} - \mu \leq +Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

or, $$Pr\left\{\overline{x} - Z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + Z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

The term $Z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}$is called the margin of error or maximum error of estimate. It is the maximum likely difference between the point estimate of a parameter and the actual value of the parameter.

Thus, (1 – α) x 100 % confidence interval for population mean μ is given by,

$$\overline{x} \pm z_c \frac{\sigma}{\sqrt{n}}$$

Where,

$\overline{x}$ = sample mean

Z_c = critical value of z or tabulated value of z for α level of significance

$\frac{\sigma}{\sqrt{n}}$ = standard error of sample mean $\overline{x}$

Note: If population is finite and sampling fraction is more that 5 % use finite correction multiplier in the standard error formula.

Example: Aircrew escape systems are powered by a solid propellant. The burning rate of this propellant is an important product characteristic. Specifications require that the mean burning rate must be 50 cm/s. It is known that the standard deviation of burning rate is σ = 2 cm/s. The experimenter selects a random sample of n = 25 and obtains a sample average burning rate of $\bar{x}$ = 51.3 cm/s. Find a 95% CI on the mean burning rate.

Solution:

R.V. X = Burning rate of the propellant (cm/s)
Sample size (n) = 25
Sample mean burning rate ($\bar{x}$) = 51.3 cm/s
Known population SD (σ) = 2 cm/s
Significance probability (α) = 0.05
Confidence probability (1 – α) = 0.95
Critical Z for 5 % significance probability i.e. Zc i.e. $Z_{0.025}$= 1.96

The standard error of mean is given by,

$$S.E.(\bar{x}) = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{25}} = 0.4$$

The lower confidence limit is,

L = $\bar{x}$ - $z_{\alpha/2}$ * $\frac{\sigma}{\sqrt{n}}$ = 51.3 – 1.96 * $\frac{2}{\sqrt{25}}$ = 51.6 – 0.78 = 50.52 cm/s

The upper confidence limit is,

U = $\bar{x}$ + $z_{\alpha/2}$ * $\frac{\sigma}{\sqrt{n}}$ = 51.3 + 1.96 * $\frac{2}{\sqrt{25}}$ = 51.6 + 0.78 = 52.08 cm/s

Conclusion:

There is a 95 % chance that the mean burning rate of propellent is between 50.52 and 52.08. The reason for this very confidence interval (50.52, 52.08) is that the standard deviation is very small i.e. 0.4 cm/s

Exercise:

1. An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed with a standard deviation of 40 hours. If a sample of 30 bulbs has an average life of 780 hours.
   (a) Compute standard error of mean
   (b) Find 95 % confidence interval for the population mean of all bulbs produced by this firm.

Interval Estimation of population mean μ (population standard deviation σ is unknown)

If population standard deviation is not known the sampling distribution of sample mean $\overline{x}$ is Student's t distribution with (n – 1) degrees of freedom.

Thus, (1 – α) x 100 % confidence interval for population mean μ is given by,

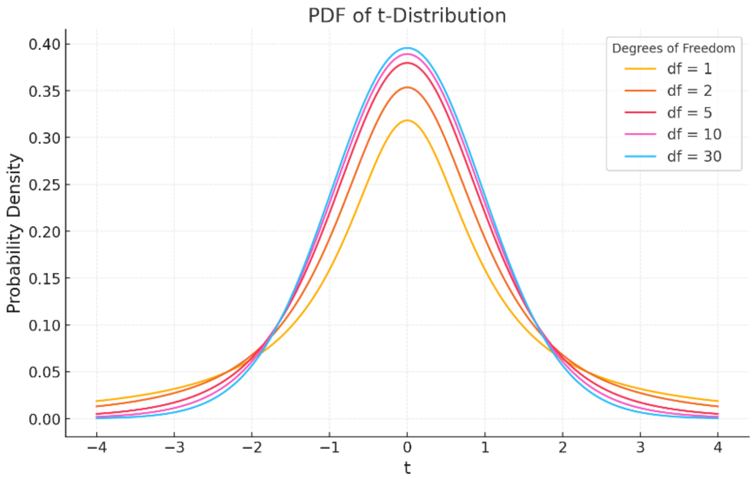
$$\overline{x} \pm t_c \cdot \frac{s}{\sqrt{n}}$$

Where,

$\overline{x}$ = sample mean
t_c or $t_{\alpha/2}(n-1)$ = Critical value of t for α level of significance and n – 1

Properties of t distribution

1. The t-distribution is symmetric and bell-shaped, similar to the normal distribution
2. The t-distribution has a single peak, or mode. This peak is located at the mean of the distribution
3. Compared to the normal distribution, the t-distribution has heavier tails. This means that there is a higher probability of observing values far from the mean in the t-distribution than in the normal distribution. In other words, it gives more probability to extreme values.
4. The shape of the t-distribution depends on the degrees of freedom (d.f.). The degrees of freedom are related to sample size and generally calculate as d.f. = n – 1, where n is the sample size. There is a specific t distribution for specific degrees of freedom.



5. For smaller d.f. the tails of t distribution are heavier and as the degrees of freedom increases the t-distribution approaches the standard normal distribution. In fact, with infinite degrees of freedom, the t-distribution is identical to the normal distribution.
6. The mean of t distribution is zero i.e. E(t) = 0
7. The variance of t distribution is slightly greater than 1 which is given by,

$$V(t) = \frac{\nu}{\upsilon-2} = \frac{d.f.}{d.f.-2} \qquad \text{for } \nu > 2$$

   If $\nu < 2$, the variance is undefined.
8. The t distribution ranges from - ∞ $to + \infty$
9. The t-distribution is specially used when the population standard deviation is unknown and the sample size is small.
10. The t-distribution was invented by William Sealy Gosset, a statistician working for the Guinness Brewery in Dublin, Ireland, in the early 20th century.
11. Probability density function of t distribution is given by,

$$f(t;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \cdot \left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Where,

t : The variable of the t-distribution.
ν : Degrees of freedom (a positive integer or real number).
$\Gamma(x)$: The gamma function, which generalizes the factorial

Example:
Interval Estimation of population mean using t distribution
The following measurements were recorded for the dying time, in hours, of a certain brand of latex paint:

3.4 2.5 4.8 2.9 3.6 2.8 3.3 5.6 3.7 2.8
4.4 4.0 5.2 3.0 4.8

Construct a 95 % confidence interval for the mean dying time for the paint.

Solution:
Here, the variable X = dying time of certain brand of latex paint (in hours)
Sample size (n) = 15
The (1- α) x 100 % confidence interval for population mean (μ) is given by, (when pop sd σ is unknown)

$$\overline{X} \pm t_c \frac{s}{\sqrt{n}}$$

Given,

Significance level (α) = 5 % = 0.05
Confidence level (1 – α) = 95 % = 0.95
Degrees of freedom = n – 1 = 15 -1 = 14
Critical value of t i.e. t_c = $t_{\alpha/2}(n-1)$ = $t_{0.025}(14)$ = 2.145

Now,

$\sum X$ = 3.4 + … + 4.8 = 56.8
$\sum X^2$ = 3.4^2 + … + 4.8^2 = 228.28

Now,

$$\overline{X} = \frac{\sum X}{n} = \frac{56.8}{15} = 3.7867 \text{ hrs}$$

$$S = \sqrt{\frac{1}{n-1}\left\{\sum X^2 - n.\overline{X}^2\right\}} = 0.9709 \text{ hrs}$$

S.E. $(\bar{X}) = \frac{s}{\sqrt{n}}$ = 0.2507 hrs

Now the 95 % confidence interval for population mean is given by,

$$\bar{X} \pm t_c \frac{s}{\sqrt{n}}$$

= 3.7867 ± 2.145 X 0.9709/$\sqrt{15}$
= 3.7867 ± 2.145 x 0.2507
= 3.7867 ± 0.5378
= 4.3245, 3.2489

Conclusion:

The 95 % confidence interval for population mean dying time of certain latex paint is 3.2509 hrs to 4.3265 hrs). The mean dying time of certain latex paint is between 3.2509 hrs and 4.3265 hrs., unless there is a unlucky chance of 5 %.

Interval length = Upper limit – lower limit = 4.3245 – 3.2489 = 1.0756 hrs

Exercise:

1. It is desired to estimate the average age of students who graduate with an MBA degree in the university system. A random sample of 64 graduating students showed that the average age was 27 years with a standard deviation of 4 years.
   (a) Estimate a 95 % confidence interval estimate of the true average age of all such graduating students at the university.
   (b) How would the confidence interval limit change if the confidence level was increased from 95 % to 99%?

Interval estimation of population total

The (1 – α) x 100 % confidence interval for population total is given by,

$$N\,\overline{x} \; \pm \; t_c \, . \, \frac{N\,s}{\sqrt{n}}$$

Where,

$\overline{x}$ = sample mean

s = Sample standard deviation

t_c = Critical value of t for α level of significance and n – 1

Interval Estimation of population proportion of success π

Consider sampling from a population which is divided into two mutually exclusive classes.

- One class possessing particular attribute (success)
- Other class not possessing that attribute (failure)

Then, random variable X = No. of successes follows binomial probability distribution with parameters n and p i.e. X ~ B (n, p).

The proportion of success in sample or estimated proportion of success is given by,

$$p = \frac{X}{n}$$

If we take large sample, and p is not close to ‘0’ or ‘1’, we can approximate the sampling distribution of X by using normal probability distribution.

The mean of binomial random variable X is

$$\mu = n\,P$$

$$\sigma = \sqrt{nPQ}$$

The distribution of *X* is considered to be approximately normal if *n* is greater than 20 and if *np* and *nq* are both greater than 5 i.e. X ~ N (np, $\sqrt{npq}$) if above conditions are met.

Then dividing each side by n, we get

$$p = \frac{X}{n} \sim N\left(P, \sqrt{\frac{PQ}{n}}\right)$$

If a random sample of size n is selected from a large population with probability of success 'p' then the sampling distribution of p then:

- Sample proportion of success p has normal distribution if n is large
- The mean of p is P i.e. $\mu_p = P$ or π
- The standard deviation of p is $\sqrt{\frac{PQ}{n}}$ i.e. $\sigma_p = \sqrt{\frac{PQ}{n}}\ or \sqrt{\frac{\pi(1-\pi)}{n}}$
- The estimated standard error of p is $\sqrt{\frac{p\,q}{n}}$ or $\sqrt{\frac{\hat{\pi}\,(1-\hat{\pi})}{n}}$

The (1 – α) x 100 % confidence interval for population proportion of success is given by,

$p \pm Z_c\sqrt{\frac{p.q}{n}}$ (if f < 5 % or infinite population or sampling with replacement)

$p \pm Z_c\sqrt{\frac{p\,.\,q}{n}}\sqrt{\frac{N-n}{N-1}}$ (if f ≥ 5 %, or finite population or sampling without replacement)

We have to replace parameters P and Q by estimates from prior study or related study or preliminary sample.

Example:

In a recent poll of 200 households, it was found that 152 households had at least one computer. Estimate the proportion of households in the population that have at least one computer.

Solution:

Population is divided into two categories:

- Households having at least one computer (considered as success)
- Households not having at least one computer (considered as failure)

Sample size (n) = 200
No. of successes in the sample (X) = 152
Sample proportion of success (p) = X / n = 152 / 200 = 0.76
Sample proportion of failure (q) = 1 – 0.76 = 0.24
Significance probability (α) = 0.05
Confidence probability (1 – α) = 0.95

The (1 – α) x 100 % confidence interval for population proportion of success is given by,

$$p \pm Z_c\sqrt{\frac{\hat{P}\,.\,\hat{Q}}{n}} = p \pm Z_C\sqrt{\frac{p.q}{n}}$$

Critical Z for 5 % level of significance i.e. Z_C = 1.96

Standard Error of p = S.E. (p) = $\sqrt{\frac{p.q}{n}}$ = $\sqrt{\frac{0.76 \text{ x } 0.24}{200}}$ =0.0302

Now the 95 % confidence interval for population proportion of success is given by

0.76 ± 1.96 x 0.0302
= 0.76 ± 0.05919
= 0.8192, 0.7008

Conclusion:

We are 95 % sure that the proportion of households having at least one computer is between 0.7008 (70.08 %) to 0.8192 (81.92 %).

Exercise:

1. A student polls his school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation. Compute a 90% confidence interval for the true percent of students who are against the new legislation, and interpret the confidence interval.

Interval Estimation of population variance σ^2 (Not covered in the course)

It can be shown that an unbiased estimator of the population variance σ^2 is sample variance s^2 i.e. E (s^2) = σ^2, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$. Hence sample variance s^2 provides the point estimation of population variance σ^2. In order to find the interval estimation, we have to use the sampling distribution of s^2. If a sample is drawn from a normally distributed population with mean μ and variance σ^2, the statistic $\frac{(n-1)s^2}{\sigma^2}$ is distributed as chi-square with n-1 degrees of freedom. The chi-square distribution of the quantity $\frac{(n-1)s^2}{\sigma^2}$ allows us to construct confidence intervals for the variance and the standard deviation (when the original population of data is normally distributed).

For a confidence level, 1 – α, we have inequality $\chi^2_{1-\alpha/2} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{\alpha/2}$ i.e.

$P\left\{\chi^2_{1-\alpha/2} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{\alpha/2}\right\}$ = 1 - α

On solving the inequality, a (1-α) x 100 % confidence interval for σ^2 is given by,

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

Since chi-square distribution in not symmetric, the confidence intervals obtained by above formula is not symmetric about the point estimate s^2.

Example:

The variability in weight in 2 pounds packets of a guaranteed food is expressed by a standard deviation of 0.05 ounce. To test this sample of 25 packets was picked and weighted giving the following results (in ounces). Obtain the 95 % fiducial limits for the population variance.

32.11 31.97 32.18 32.05 32.10 32.03 32.25 32.07 32.07 32.15
32.05 32.14 32.19 31.96 32.03 31.98 32.07 31.99 32.09 32.08
32.16 32.03 32.18 32.04 31.98.

Solution:

Random variable X = weight of packets of a guaranteed food (pound)
Sample size (n) = 25
Confidence probability (1 – α) = 95 % = 0.95
Significance probability (α) = 0.05

The (1-α) x 100 % confidence interval for σ^2 is given by,

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

Where,

n = sample size
s^2 = sample variance
$\chi^2_{\alpha/2} = \chi^2_U$ = Upper critical value of chi-square
$\chi^2_{1-\alpha/2} = \chi^2_L$ = Lower critical value of chi-square

Here,

Degrees of freedom = n – 1 = 25 – 1= 24
Lower critical chi-square value i.e. $\chi^2_{\alpha/2}$ or χ^2_L = 12.401
Upper critical chi-square value i.e. $\chi^2_{1-\alpha/2}$ or χ^2_U = 39.364

Now, the sample variance is given by,

$$s^2 = \frac{1}{n-1}\left\{\sum X^2 - n\overline{X}^2\right\}$$

Computational result,

$\sum X$ = 801.95 = 32.11 + …. + 31.98
$\sum X^2$ = 25725.1 = 32.11^2 + ….+ 31.98^2

Mean is given by,

$$\overline{X} = \frac{\sum X}{n} = \frac{801.95}{25} = 32.078$$

S^2 = $\frac{1}{n-1}\left\{\sum X^2 - n\overline{X}^2\right\} = \frac{1}{25-1}\{25725.1 - 25*32.078\wedge 2\}$ = 0.00596

Now, the 95 % confidence interval for population variance σ^2 is given by,

$$\frac{(25-1)0.00596}{39.364} \le \sigma^2 \le \frac{(25-1)0.00596}{12.401}$$

Or, 0.003634 $\le \sigma^2 \le$0.01153

Conclusion: we are 95 % sure that the population variance of weight of packets of a guaranteed food is between 0.003634 ounce to 0.011536 ounce. Hence, packets produced by machine do not meet quality.

Interval estimation of mean when population standard deviation is known

1. The distribution of systolic and diastolic blood pressures for female diabetics between the ages of 30 and 34 have unknown means. However, their standard deviations are σ_S= 11.8 mmHg and σ_D= 9.1 mmHg respectively. A random sample of ten women is selected from this population. The mean systolic blood pressure for the sample is $\overline{x}_S$ = 130 mmHg, and the mean diastolic blood pressure for the sample is $\overline{x}_D$= 84 mmHg.
   (a) Calculate a two-sided 95 % confidence interval for μ_S , the true mean systolic blood pressure of the population and interpret this confidence interval.
   (b) Calculate a two-sided 99 % confidence interval for μ_D, the true mean diastolic blood pressure of the population.

Interval estimation of mean when population standard deviation is unknown

1. The weight at birth for 15 babies born in Hospital are given below, each figure is correct to the nearest tenth of a pound.

   6.2 5.7 8.1 6.7 4.8 5.0 7.1 6.8 5.8 6.9
   7.6 7.9 7.5 7.8 8.5

   From above data, estimate the mean weight of the babies born in the hospital. Also determine the percentage error in the estimated value

3. Height of 50 girls of certain campus gave following values

Mean	**Standard Deviation**
147.4 cm	**6.6 cm**

(a) Find the Standard Error of the mean

(b) Find the 99 % confidence interval for population mean height of girls.

4. Twenty air samples taken at the same site over a period of six months showed the following amounts of suspended particulate matter (micrograms per cubic meter of air):

   68 22 36 32 42 24 28 38 30 44
   28 27 28 43 45 50 79 74 57 21

   Consider these measurements to be a random sample from a population of normally distributed measurements, construct 98 % confidence interval for the population mean.

Interval estimation of population total

1. A simple random sample 30 households were drawn from a city area containing 14,848 households. The numbers of households in the sample were:

5	6	3	3	2	3	3	3	4	4
3	2	7	4	3	5	4	4	3	3
4	3	3	1	2	4	3	4	2	4

Estimate the total no. of people in the area and compute the probability that this estimate is within ± 10 % of the true value.

2.

Interval estimation of population proportion

1. In a locality of 18,000 families a sample of 840 families were selected. Of these 840 families, 206 were found to have a monthly income of Rs 30,000 and less.
   (a) Estimate the limits of percentage of families having such income.
   (b) Find the 95 % confidence limits for proportion of families having monthly income 30,000 or less.

2. In a simple random sample of 125 unemployed male high school dropouts between the ages of 16 and 21, inclusive, 88 stated that they were regular consumers of alcoholic beverages. Construct 95 percent confidence interval for the population proportion.

3. A student polls his school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation. Compute a 90% confidence interval for the true percent of students who are against the new legislation, and interpret the confidence interval.