BITS Pilani
K K Birla Goa Campus

Probability & Statistics

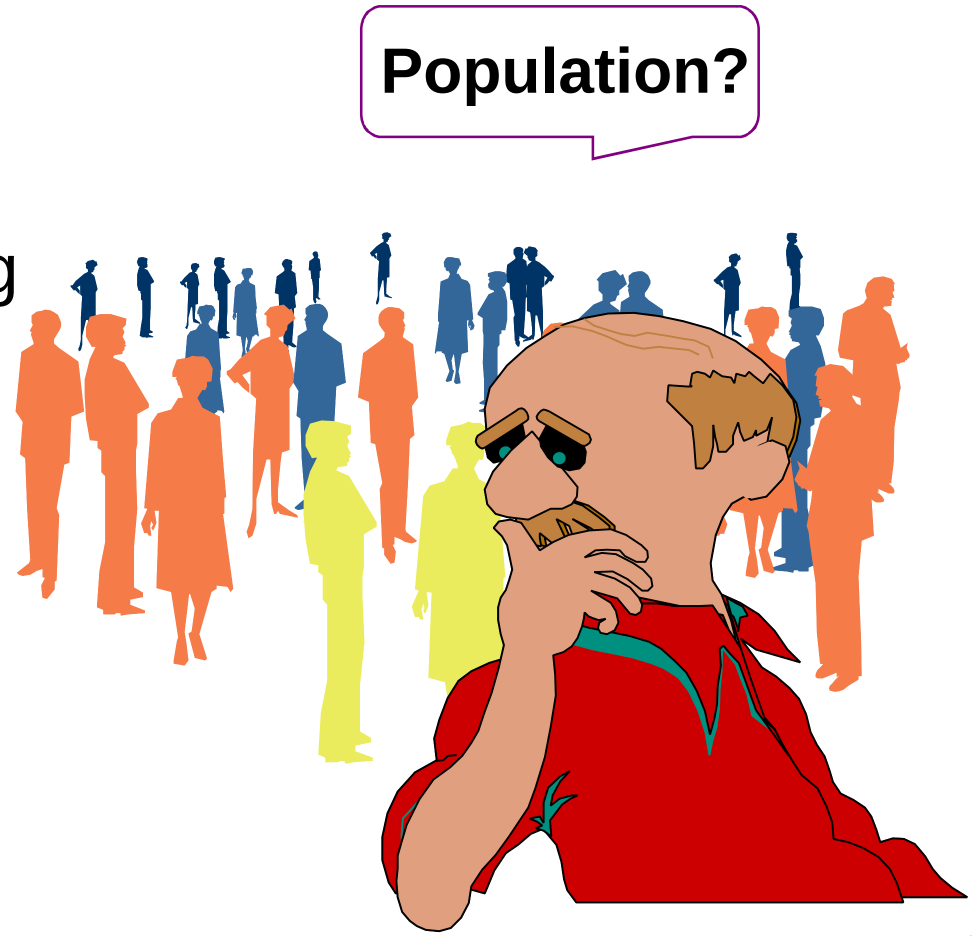
Dr. Jajati Keshari Sahoo
Department of Mathematics

Statistics

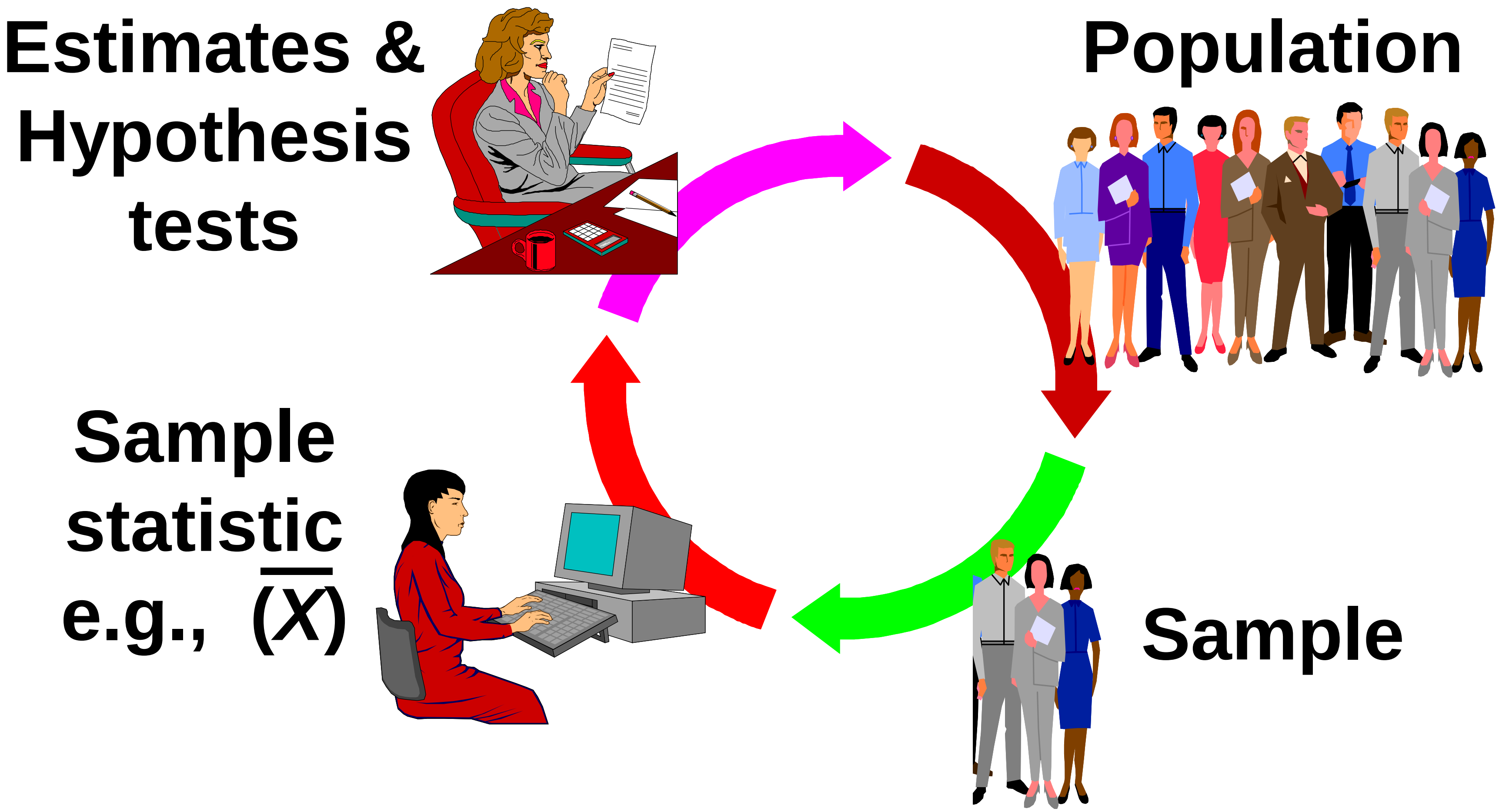
- Descriptive statistics
- Inferential statistics

Inferential Statistics

- 1. Involves:
  - Estimation
  - Hypothesis Testing
- 2. Purpose
  - Make Inferences about Population Characteristics



Inference Process



Key terms

- Population
  - All items of interest
- Sample or random sample
  - Portion of population
- Parameter
  - Summary Measure about Population
- Statistic
  - Summary Measure about sample

Population and Sample

- **Population**: refer to a population in term of its probability distribution or frequency distribution.
  - Population *X* or *f(x)* means a population described by a probability distribution *f(x).*
- Population might be infinite or it is impossible to observe all its values even finite, it may be impractical or uneconomical to observe it.

Sample

- Sample: a part of population.
- Random samples (Why we need?): such results can be useful only if the sample is in some way “representative”.
- Negative example: performance of a tire if it is tested only on a smooth roads; family incomes based on the data of home owner only.

Sampling

- Representative sample
  - Same characteristics as the population

- Random sample
  - Every subset of the population has an equal chance of being selected

Random sample

❖ ***Definition (Random sample)***:

A set of observations

$$X_1, X_2, \ldots, X_n$$

constitutes a random sample of size $\boldsymbol{n}$ from a population or distribution X if

(a) X_i's are independent

(b) X_i's have same distribution as of X.

Statistic

❖ ***Definition (Statistic)***:

A statistic is a random variable whose numerical value can be determined from a random sample i.e., a function of random sample.

Examples:

$$\text{(a) } \max\{X_i\} \quad \text{(b) } \min\{X_i\} \quad \text{(c) } \sum_{i=1}^{n} X_i$$

$$\text{(d) } \sum_{i=1}^{n} X_i / n = \overline{X}.$$

Sample mean

❖ ***Definition***:

Let $X_1, X_2, \cdots, X_n$ be a random sample from the distribution of X. The statistic

$$\frac{\sum_{i=1}^{n} X_i}{n}$$ is called the sample mean and is denoted by $\overline{X}$.

Sample variance and sample SD

❖ ***Definition***:

Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from the distribution of X. The statistic

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$ is called the sample variance

and the statistic $S = \sqrt{S^2}$ is called sample standard deviation.

Sample variance

❖ ***Computation formula for sample variance***:

$$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n-1}$$

$$\left(\text{or } S^2 = \frac{n.\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}{n(n-1)} \right)$$

Sampling Distribution

Mean and variance of sample mean:

If a random sample of size n is taken from a population having the mean μ and the variance σ^2, then $\overline{X}$ is a random variable whose distribution has the mean μ and variance σ^2/n.

Standard deviation of $\overline{X}$ is $\frac{\sigma}{\sqrt{n}}$ and called standard error of the mean.

Sampling Distribution

Mean of sample variance:

Theorem:

If a random sample of size n is taken from a population having the mean μ and the variance σ^2, then $E(S^2) = \sigma^2$.

Central limit theorem

Theorem:

If $\overline{X}$ is the mean of a sample of size n taken from a population having the mean μ and variance σ^2. Then for large n, $\overline{X}$ is approximately normal with mean μ and variance $\frac{\sigma^2}{\text{n}}$.

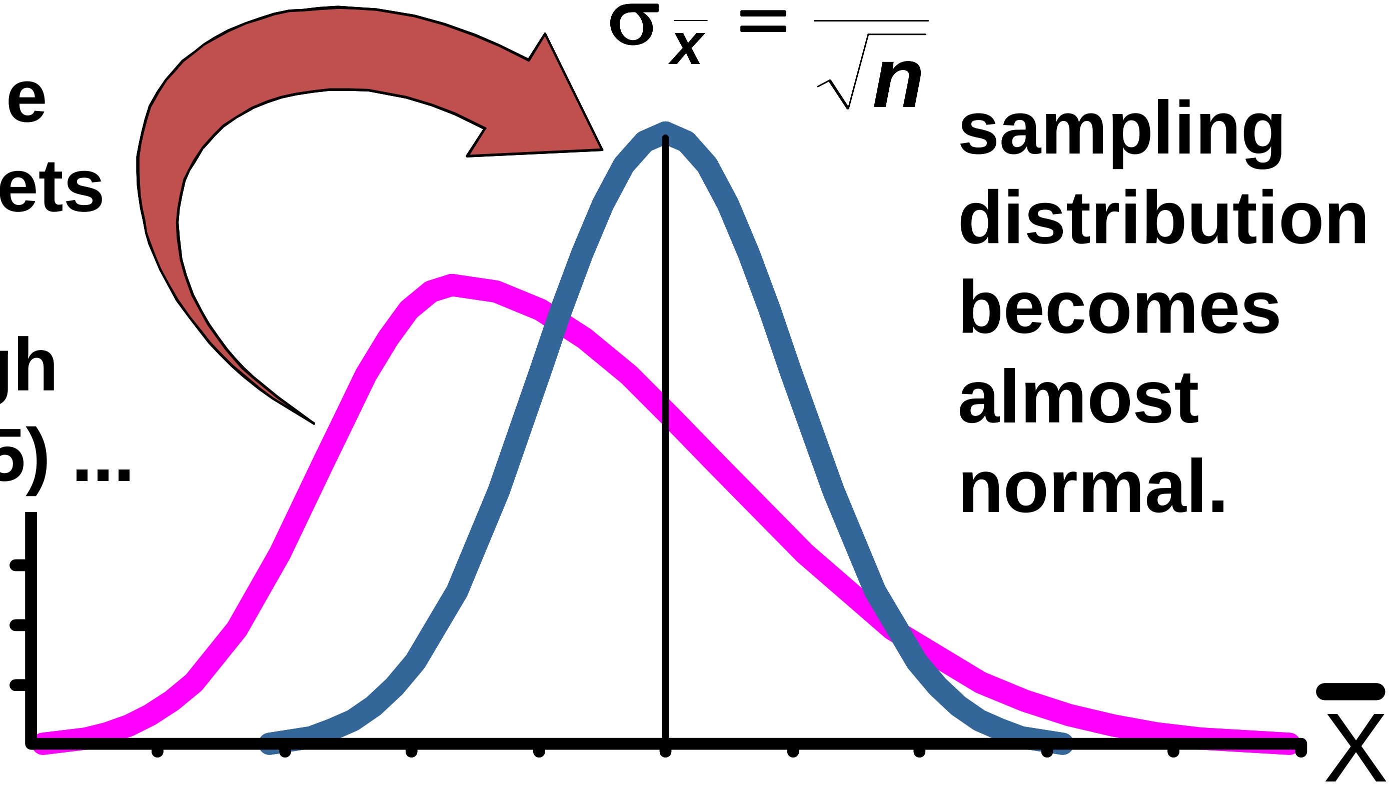
Hence $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is standard normal.

Central Limit Theorem

As sample size gets large enough (n ≥ 25) ...

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

sampling distribution becomes almost normal.



$$\mu_{\bar{x}} = \mu$$

Central Limit Theorem

Example:

The mean length of life of a certain cutting tool is 41.5 hours, with a standard deviation of 2.5 hours. What is the probability that a sample of size 50 drawn from this population will have a mean of between 40.5 hours and 42 hours ?

Ans: 0.9184

Central Limit Theorem

Example:

One has 100 light bulbs whose life times are independent exponential with mean 5. If the bulbs are used one at a time, with a failed bulb being replaced immediately by a new one, what is the probability that there is still a working bulb after 525 hours ?

Ans: 0.3085

Central Limit Theorem

Example:

If X is a Poisson random variable with mean μ then X is normal with mean μ and variance μ.

Central Limit Theorem

Example:

If the number of people entering a store per hour is Poisson distribution with parameter 100, how long should the shopkeeper wait in order to get a probability of 0.9 that more than 200 people have entered in the store ?

Central Limit Theorem

Example:

A certain component is critical to the operation of an electrical system and must be replaced immediately upon failure. If the mean life time of this type of component is 100 and its standard deviation is 30 hours, how many of these components must be in stock so that the probability that the system is in continual operation for the next 2000 hours is at least 0.95 ?

Ans: n=23.

Sampling Distribution

Distribution of sample variance:

Theorem:

If a random sample of size n is taken from a normal population having the mean μ and the variance σ^2, then the random variable $(n-1)S^2/\sigma^2$ has a chi-squared distribution with $(n-1)$ degrees of freedom.

Note:

$$S^2 = \Gamma\left(\frac{n-1}{2}, \frac{2\sigma^2}{n-1}\right).$$

Sampling Distribution

Example:

The claim that the variance of normal population $\sigma^2 = 21.3$ is rejected if the variance of a random sample of size 15 excess 39.74. What will be the probability that the claim will be rejected even though $\sigma^2 = 21.3$.

Ans: 0.025

T- Distribution

Definition:

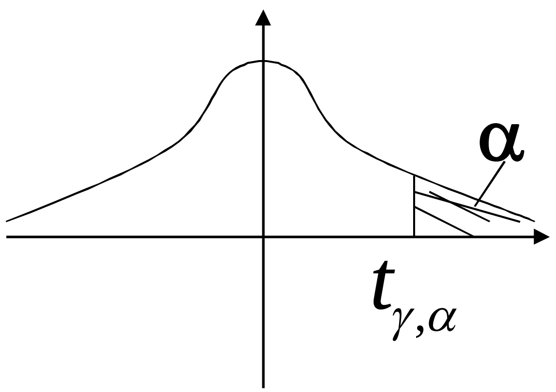
Let Z be a standard normal variable and let $\chi_\gamma^{\ 2}$ be an independent chi-squared random variable with γ degrees of freedom. The random variable $T_\gamma = \dfrac{Z}{\sqrt{\chi_\gamma^{\ 2}/\gamma}}$ is said to follow a T distribution with γ degrees of freedom.

T- Distribution

Theorem:

If a random sample of size n is taken from a normal population having the mean μ and the variance σ^2, then the random variable $\dfrac{\overline{X}-\mu}{S/\sqrt{n}}$ follows a T distribution with $(n\text{-}1)$ degrees of freedom.

T-distribution



$$P(T_\gamma \geq t_{\gamma,\alpha}) = \alpha$$

The sampling distribution of the Mean (unknown variance)

- If n is large, it doesn't matter whether σ is known or not, as it is reasonable in that case to substitute for it the sample standard deviation s.

Question: how about *n* is a small value?

We need to make the assumption that the sample comes from a normal population the use *T* distribution if σ is unknown otherwise use normal approximation.

ESTIMATION OF PARAMETERS

ESTIMATION:-

Procedure of estimating a Population (Parameter) by using sample information is referred as estimation.

i) POINT ESTIMATION

II) INTERVAL ESTIMATION

POINT ESTIMATION

POINT ESTIMATE :

An estimate of a population parameter given by a single number is called point estimate

POINT ESTIMATOR :

A point estimator is a statistic for Estimating the population Parameter θ and will be denoted by $\hat{\theta}$.

Point Estimator

Problem of point estimation of the population mean μ :

The statistic chosen will be called a point estimator for μ is $\overline{X}$

Logical estimator for μ is the Sample mean

Hence $\hat{\mu} = \overline{X}.$

Example

Market researcher use the number of sentences per advertisement as a measure of readability far magazine advertisement. The following represents a random sample of 54 advertisements. Find a point estimate of the population mean μ

9,20,18,16,9,16,16,9,11,13,22,16,5,18,6,6,5,12,25,17,2,3,7,10,9,
10,10,5,11,18,18,9,9,17,13,11,7,14,6,11,12,11,15,6,12,14,11,4,9,
18,12,12,17,11,20.

Solution:

The Sample mean of Data = 671/54 =12.4. So, Point Estimate for the mean length of all magazine advertisement is 12.4 sentences.

UNBIASED ESTIMATOR

Unbiased Estimator:

If the mean of sampling distribution of a Statistic equals the corresponding Population parameter, the Statistic is called an Unbiased Estimator of the Parameter

i.e if $E\left(\hat{\theta}\right) = \theta.$

Biased Estimator:

$$\text{if } E\left(\hat{\theta}\right) \neq \theta.$$

Unbiased Estimator

Example: The Sample mean $\overline{X}$ is an unbiased estimator for the population mean.

Example: S^2 is an unbiased estimator of population variance.

Example: S is not an unbiased estimator of population standard deviation.

Example: $\sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n}$ is a biased estimator of population variance.

Unbiased Estimator

Example: If X is $b(n, p)$ then

(a) $\dfrac{X}{n}$ is an unbiased for p.

(b) $\dfrac{X+\sqrt{n}/2}{n+\sqrt{n}}$ is an unbiased for p iff $p = 0.5$.

(c) $\dfrac{X+p\sqrt{n}}{n+\sqrt{n}}$ is an unbiased for p.

Unbiased Estimator

Example: If X_i for $1 \le i \le n$ constitute a random sample from the population given by

$$f(x) = \begin{cases} e^{-(x-\theta)}, & x > \theta \\ 0, & \text{otherwise.} \end{cases}$$

Find an unbiased estimator for θ.

Methods For Finding Estimators

- Method of Moments: Using the k-th moment as $M_k = E(X^k) = \sum_{i=1}^{n} X_i^k / n,$ we estimate the parameters in terms of moments.
- Method of Maximum Likelihood:

We maximize the likelihood function with respect to the parameter θ and then the statistic at which the likelihood function gives maximum is called maximum likelihood estimator for θ.

Method of Moments

Example: If X is $\Gamma(\alpha,\beta)$. Using method of moments find estimators for α and β.

Example: If X is $N(\mu,\sigma^2)$. Using method of moments find estimators for μ and σ^2.

Example: If X is $U(-a,a)$. Using method of moments find an estimator for a.

Likelihood Function

Likelihood Function:

Let $x_1, x_2, \cdots, x_n$ be a random sample of size n from a population with density function $f(x)$ and parameter θ. Then the likelihood function of the sample value $x_1, x_2, \cdots, x_n$ is denoted by $L(\theta)$, and defined as

$$L(\theta) = \prod_{i=1}^{n} f(x_i)$$

Method of maximum likelihood

Example: If X is $B(n, p)$. Using method of maximum likelihood (ML) find an estimators for p.

Example: If X is $\mathrm{Exp}(\beta)$. Using method of ML find an estimator for β.

Example: If X is $U[0, \theta]$. Using method of ML find an estimator for θ.

Method of maximum likelihood

Example: If X is Poisson(λ). Using method of ML find estimator for λ.

Example: If X is $N(\mu, \sigma^2)$. Using method of ML find estimators for μ and σ.

Example: Find the ML estimator for a if the population is given by $f(x) = (a+1)x^a, 0 < x < 1$.

INTERVAL ESTIMATION

By using point estimation ,we may not get desired degree of accuracy in estimating a parameter. Therefore , it is better to replace point estimation by interval estimation.

INTERVAL ESTIMATION

Interval estimate:

An interval estimate of an unknown parameter is an interval of the form $L_1 \leq \theta \leq L_2$, where the end points L_1 and L_2 depend on the numerical value of the statistic and or parameter of population distribution.

100(1-α)% Confidence Interval:

A *100(1-α)%* confidence interval for a parameter θ is a random interval of the form $[L_1, L_2]$ such that

$$P(L_1 \leq \theta \leq L_2) = 1 - \alpha.$$

INTERVAL ESTIMATION

Theorem: [100(1-α)% confidence interval for μ when σ is known]:

Let $X_1, X_2, \cdots, X_n$ be a random sample of size n (large) from a population with mean μ and variance σ^2. A 100(1-α)% confidence interval for μ is $\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$.

Note: If n is small assume the sample from normal.

INTERVAL ESTIMATION

Example:

A random sample of size 100 is taken from a population with mean μ and σ=0.5. If the sample mean is 0.75, find a 95% confidence interval for μ.

Ans: (0.652, 0.848).

INTERVAL ESTIMATION

Example:

A random sample of size n is taken from a population with mean μ and $\sigma^2=20$. How large n should choose with 90% confidence that the random interval $\left(\overline{X}-2,\ \overline{X}+2\right)$ includes μ.

Ans: $n \simeq 9$.

INTERVAL ESTIMATION

Theorem: [100(1-α)% confidence interval for μ when σ is unknown]:

Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from a normal population with mean μ and variance σ^2. A 100(1-α)% confidence interval for μ is $\left[\overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}},\ \ \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right]$.

Note: t has (n-1) degrees of freedom.

Note: If n large(n ≥ 25), we can scrap normality assumption.

INTERVAL ESTIMATION

Example:

A signal is transmitted from location A to B. The value received at location B is nornally distributed with mean μ and variance σ^2. A particular value is transmitted 9 times. Find the 95% confidence interval for μ, when the values received are 5, 8.5, 12, 15, 7, 9, 7.5, 6.5, and 10.5.

Ans: (6.631, 11.369).

INTERVAL ESTIMATION

Example:

A random sample of size 80 is taken from a population with mean μ and $S^2 = 30.85.$ If the sample mean is 18.85, construct a 99% confidence interval for μ.

Ans: (17.0214, 20.6786).

INTERVAL ESTIMATION

Theorem: [100(1-α)% confidence interval for σ]:

Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from a normal population with mean μ and variance σ^2. A 100(1-α)% confidence interval for σ^2 is

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}}, \quad \frac{(n-1)S_2}{\chi^2_{1-\alpha/2}}\right].$$

Note: $\chi^2_{(n\text{-}1)}$ has (n-1) degrees of freedom.

INTERVAL ESTIMATION

Example:

An optical firm purchases glass for making lenses. Assume that the refractive index of 20 pieces of glass have a variance of 0.00012. Find a 95% confidence interval for σ^2.

Problem

To estimate the average time it takes to assemble a certain Computer component, the industrial engineer at an electronic firm timed 40 technicians in the performance of this task, getting a mean of 12.73 minutes and a standard deviation 2.06 minutes.

(a) What can we say with 99% confidence about the maximum error if is used as a point estimate of the actual average time required to do this job?

(b) Use the given data to construct a 98% confidence interval for the true average time it takes to assemble the computer component.

Interval Estimation (cont'd)

Solution: Given $\bar{x} = 12.73$, s = 2.06 and n = 40

(a) $\bar{x} = 12.73$ and (1 - α) = 0.99 $\Rightarrow \alpha$ = 0.01

Since sample is large (n = 40) The maximum error of estimation with 99% confidence is

$$E = z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} = z_{0.005} \cdot \frac{s}{\sqrt{n}} = 2.575 \cdot \frac{2.06}{\sqrt{40}} = 0.839$$

(b) 98% confidence interval (i.e. α = 0.02) is given by

$$\bar{x} - z_{0.01} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{0.01} \cdot \frac{s}{\sqrt{n}}$$

$$\Leftrightarrow 12.73 - 2.33 \cdot \frac{2.06}{\sqrt{40}} < \mu < 12.73 + 2.33 \cdot \frac{2.06}{\sqrt{40}}$$

$$\Leftrightarrow 11.971 < \mu < 13.489.$$

Problem

With reference to the previous problem with what confidence we can assert that the sample mean does not differ from the true mean by more than 30 seconds.

Solution: Given E = 30 seconds = 0.5 minute, s = 2.06, n = 40 and we have to get value of ($1 - \alpha$).

$$E = z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \Leftrightarrow z_{\alpha/2} = E \cdot \frac{\sqrt{n}}{\sigma} \Rightarrow z_{\alpha/2} = (0.5)\frac{\sqrt{40}}{2.06} = 1.54$$

$$F(z_{\alpha/2}) = F(1.54) = 0.9382 \quad \text{(from Table 3)}$$

$$\Rightarrow \frac{\alpha}{2} = 1 - 0.9382 \Rightarrow \alpha = 0.1236 \Rightarrow 1 - \alpha = 0.8764$$

Thus, we have 87.64% confidence that the sample mean does not duffer from the true mean by more than 30 seconds.