B.SC.(H) STATISTICS
SEM. – IV
PAPER- STATISTICAL INFERENCE
UPC - 32371401
TOPIC- CONFIDENCE INTERVALS

REMEMBER

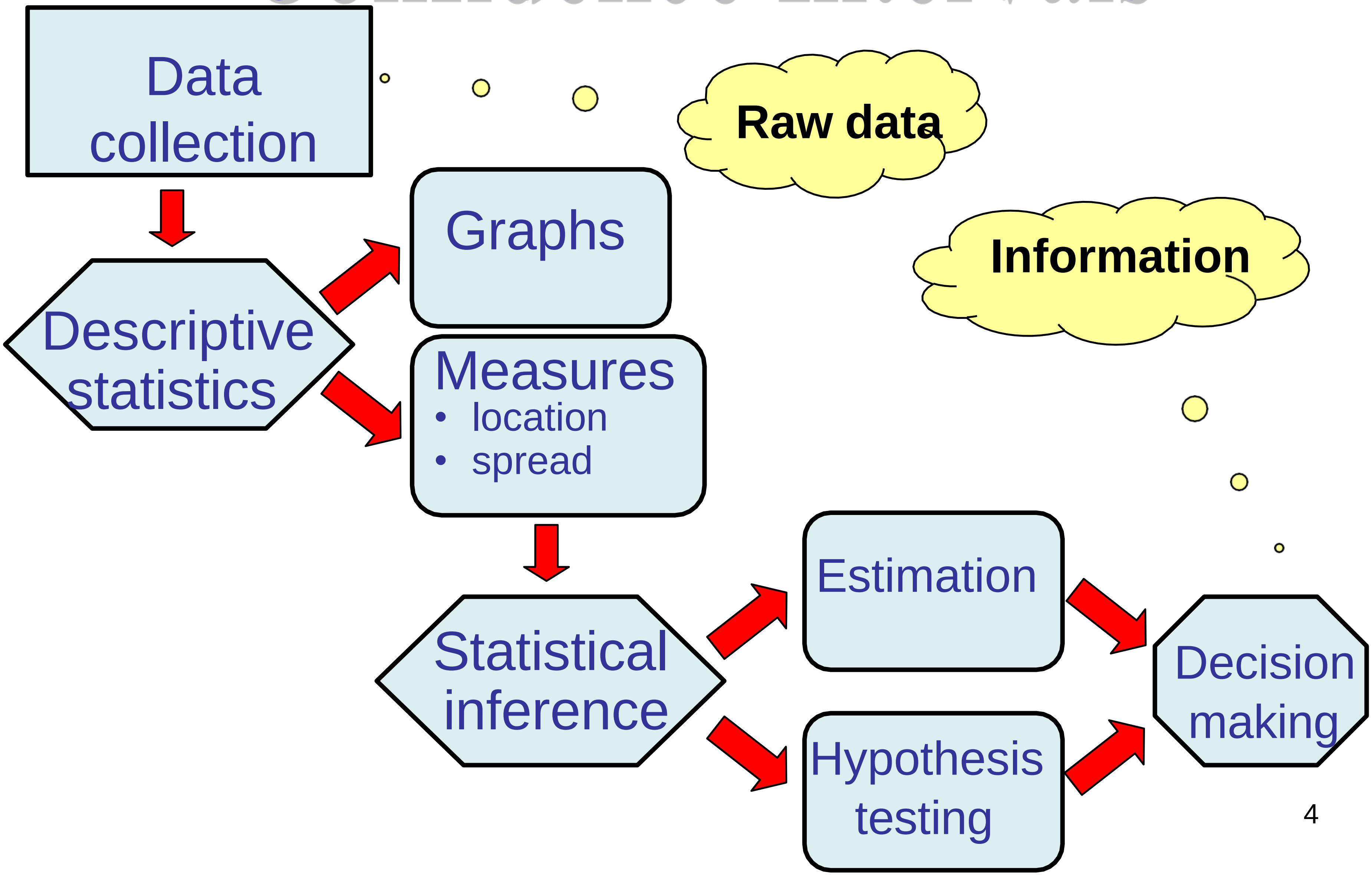
- A ***PARAMETER*** is a number that describes a population statistic

- A ***STATISTIC*** is a number that describes a characteristic in the sample data

Confidence Intervals

- Inferential statistics
  - Draw conclusion from data
  - Sample
    - Describe data
  - Use ***sample statistic*** to infer ***population parameter***
    - Estimation
    - Hypothesis testing

Confidence Intervals



Confidence Intervals

- Estimation
  - Numerical values assigned to a population parameter using a sample statistic
- Sample mean $\overline{x}$ used to estimate population mean μ
- Sample variance $\boldsymbol{s^2}$ used to estimate population variance $\boldsymbol{\sigma^2}$
- Sample stand dev $\boldsymbol{s}$ used to estimate population stand dev $\boldsymbol{\sigma}$
- Sample proportion $\hat{p}$ used to estimate population proportion $\boldsymbol{p}$

Confidence Intervals

- Steps in estimation
  - Select sample
  - Get required information from the sample
  - Calculate sample statistic
  - Assign values to population parameter

ESTIMATORS

- Sample statistic used to estimate a population parameter is called an ***ESTIMATOR***

- An estimator is a rule that tells us how to calculate the estimate and it is generally expressed as a formula

ESTIMATORS

	POPULATION PAPARMETER	ESTIMATE (VALUE OF STATISTIC)	ESTIMATOR (Formula)
MEAN	μ	$\bar{x}$	$\bar{x} = \frac{1}{n\Sigma x}$
VARIANCE	σ^2	s^2	$\frac{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}{n-1}$
PROPORTION	p	$\hat{p}$	$\hat{p} = \frac{x}{n}$

TYPES OF ESTIMATE

- Two types of estimate:-
  - Point estimates
  - Interval estimates

POINT ESTIMATES

- A single number that is calculated from sample data

- Resulting number then used to estimate the true value of the corresponding population parameter

Disadvantages of Point Estimates

- Point estimates do not provide information about how close the point estimate is to the population parameter
- Point estimates do not consider the sample size or variability of the population from which the sample was taken

So what do we do?

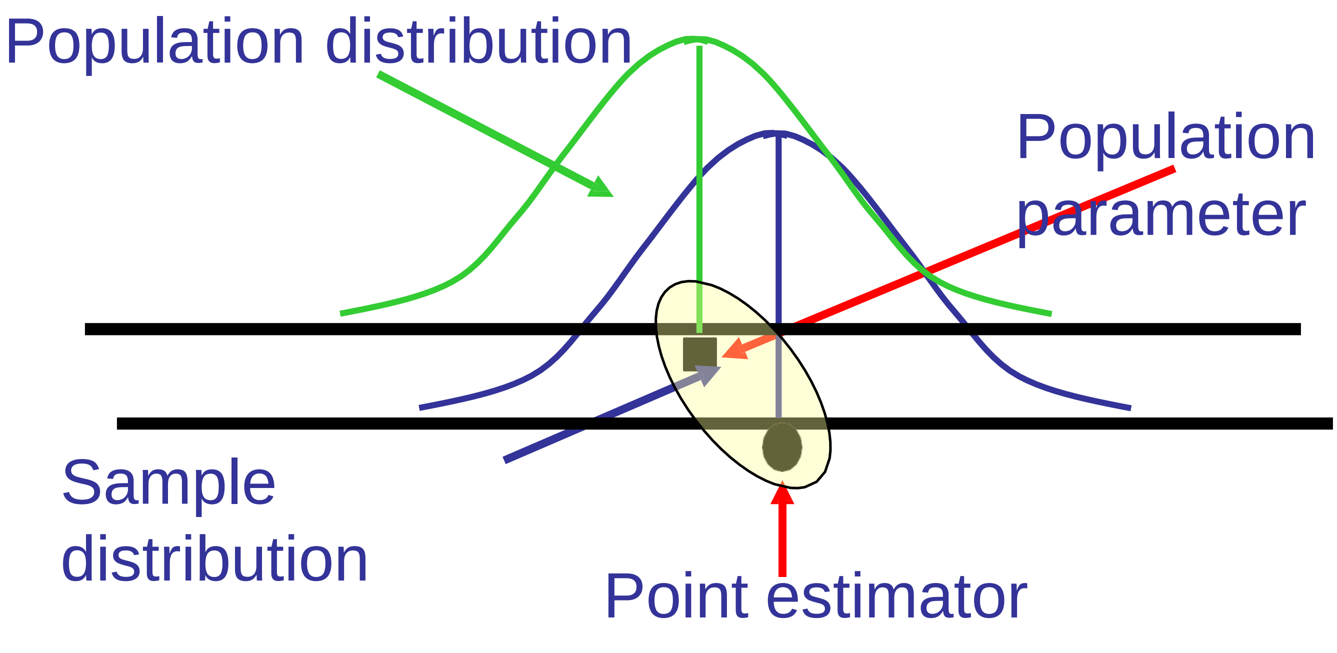
- Sample size and variability of population will affect the accuracy of the estimate so a point estimate is really not very useful

- This problem can be overcome by using ***<u>INTERVAL ESTIMATES</u>***

Confidence Intervals

Point Estimates

– A single sample statistic used to estimate the population parameter

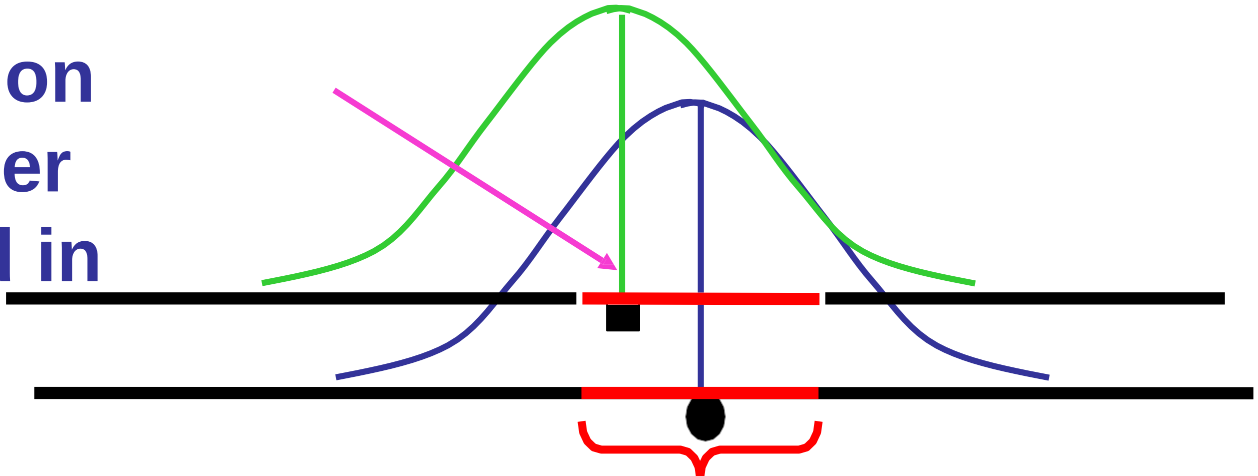


Confidence Intervals

Confidence interval

– An interval is calculated around the sample statistic

Population parameter included in interval



Confidence interval

Confidence Intervals

Confidence interval

- An upper and lower limit within in which the population parameter is expected to lie limits will vary from sample to sample.

- specify the probability that the interval will include the parameter
Typical used 90%, 95%, 99%

- Probability denoted by

• *(1 – α)* known as the level of confidence

• *α* is the significance level

<u>90% Confidence Interval Means:</u>

90% of all possible samples taken from population will produce an interval that will include the population parameter

Confidence Intervals

- An interval estimate consists of a range of values with an upper & lower limit
- The population parameter is expected to lie within this interval with a certain level of confidence
- Limits of an interval vary from sample to sample therefore we must also specify the probability that an interval will contain the parameter
- Ideally probability should be as high as possible

Level of Confidence

SO REMEMBER

- We can choose the probability
- Probability is denoted by $(1-\alpha)$
- Typical values are 0.9 (90%); 0.95 (95%) and 0.99 (99%)
- The probability is known as the ***LEVEL OF CONFIDENCE***
- α is known as the ***SIGNIFICANCE LEVEL***
- α corresponds to an area under a curve
- Since we take the confidence level into account when we estimate an interval, the interval is called ***CONFIDENCE INTERVAL***

REFERENCE BOOKS

1. GUN, A.M. ; GUPTA M.K. ; DASGUPTA, B. :An outline of STATISTICAL THEORY, volume two, Third Edition

2.John E. Freund's Mathematical Statistics with Applications Seventh Edition Irwin Miller and Marylees Miller, Pearson Education.

3. GUPTA, S.C. ; KAPOOR V.K.: Fundamental of Mathematical Statistics, Sultan Chand.

A simple method of obtaining confidence limits:

Suppose there exists a statistic T and a function $\varphi(T,\theta)$ of T and θ, which is measurable for θ, such that the distribution of $\psi(T,\theta)$ is independent of θ (i.e. the same for each θ).One can then find two constants k_1 and k_2, depending on α_1 and α_2 but not on θ, such that

$$P_\theta\,[\psi(T,\theta) < k_1] = \alpha_1,\ P_\theta\,[\psi(T,\theta) < k_2] = \alpha_2$$

This implies

$P_\theta\,[k_1 \leq \psi(T,\theta) \leq k_2] = 1 - \alpha$ for all $\theta \in \Theta$

[for proof see Example 8.1, Page No.384 (Reference book .1)]

Suppose that further that it is possible to re-write the inequality

$$k_1 \leq \psi(T,\theta) \leq k_2$$

In the form

$$C_1(T) \leq \gamma(\theta) \leq C_2(T)$$

Where $C_1(T)$ and $C_2(T)$ are independent of θ.

Then

$$P_\theta\,[C_1(T) \leq \gamma(\theta) \leq C_2(T)]$$
$$= P_\theta\,[k_1 \leq \psi(T,\theta) \leq k_2]$$
$$= 1 - \alpha \text{ for all } \theta \,\varepsilon\, \Theta$$

Hence $C_1(T)$ *and* $C_2(T)$ may be taken as the statistics T_1 and T_2.

In other words, for any given set of observation X, the values $C_1(t)$ *and* $C_2(t)$ of $C_1(T)$ *and* $C_2(T)$ respectively, are a pair of confidence limits to $\gamma(\theta)$ with confidence coefficient 1- α.

Confidence Intervals – LARGE SAMPLES

Confidence interval for Population Mean, $n \geq 30$

- **population need not be normally distributed**
- **sample will be approximately normal**

$$CI(\mu)_{1-\alpha} = [\bar{x} \pm Z_{\frac{\alpha}{2}} . \frac{\sigma}{\sqrt{n}}], \text{ if } \sigma \text{ is known}$$

$$CI(\mu)_{1\text{-}\alpha} = [\bar{x} \pm Z_{\frac{\alpha}{2}} . \frac{S}{\sqrt{n}}], \text{ if } \sigma \text{ is unknown}$$

[see Example 8.2, Page No.386 (Reference book .1)]

Confidence Intervals

$$CI(\mu)_{1-\alpha} = \left[\bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right], \text{ if } \sigma \text{ is known}$$

$$CI(\mu)_{1-\alpha} = \left[\bar{x} \pm Z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right], \text{ if } \sigma \text{ is not known}$$

Example:

90% confidence interval

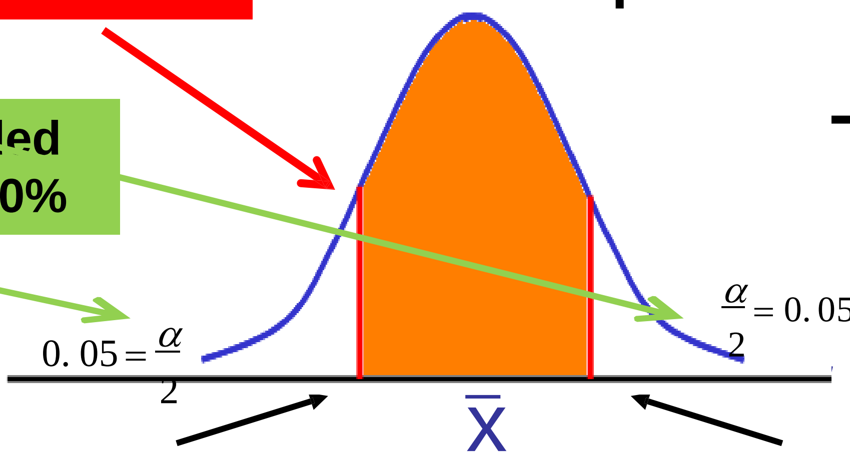
$$1 - \alpha = 0.90$$

$$\alpha = 0.10$$

$$\frac{\alpha}{2} = 0.05$$

90% of all sample means fall in this area

These 2 areas added together = α i.e. 10%



Lower conf limit

Upper conf limit

Confidence level = 1 - α

Example

A random sample of repair costs for 150 hotel rooms gave a mean repair cost of R84.30 and a standard deviation of R37.20. Construct a 95% confidence interval for the mean repair cost for a population of 2000 hotel rooms

Confidence Intervals

- Confidence interval for Population Mean, $n \geq 30$
- Example:
  - Estimate the population mean with 90%, 95% and 99% confidence, if it is known that
  - $s = 9$ and $n = 100$
  - Solution: The confidence intervals are

$$90\% \quad \bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} = \bar{x} \pm 1.64 \frac{9}{\sqrt{100}} = \bar{x} \pm 1.48$$

$$95\% \quad \bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} = \bar{x} \pm 1.96 \frac{9}{\sqrt{100}} = \bar{x} \pm 1.76$$

$$99\% \quad \bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} = \bar{x} \pm 2.57 \frac{9}{\sqrt{100}} = \bar{x} \pm 2.31$$

Confidence Intervals

Confidence level influence width of interval

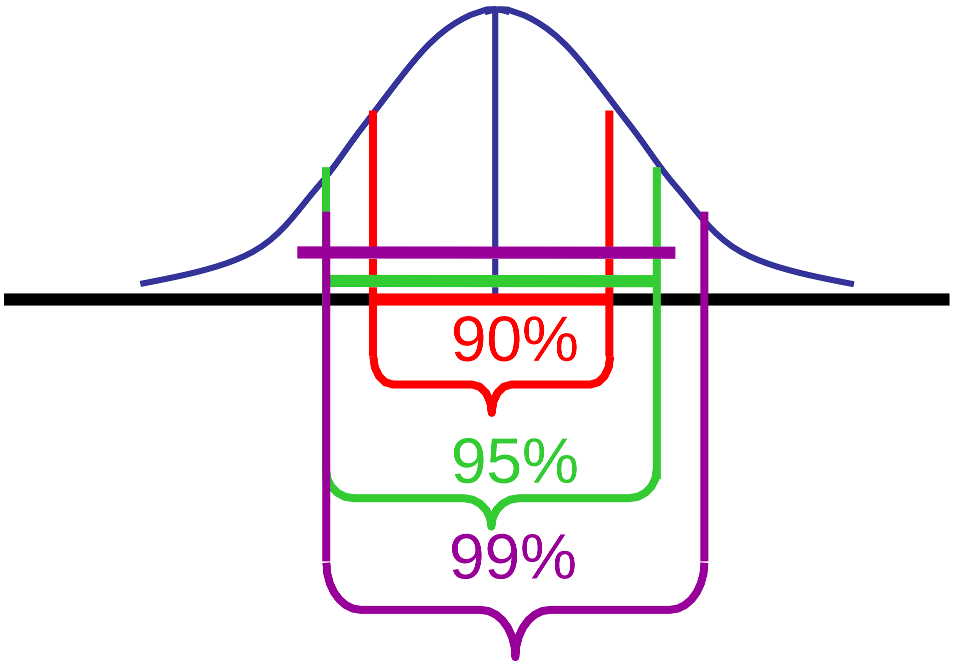
90% $\bar{x} \pm 1.48$ $\therefore$ Width of interval = 2 x 1.48 = 2.96

95% $\bar{x} \pm 1.76$ $\therefore$ Width of interval = 2 x 1.76 = 3.52

99% $\bar{x} \pm 2.31$ $\therefore$ Width of interval = 2 x 2.31 = 4.62

Margin of error becomes smaller if:

- *z*-value smaller
- *σ* smaller
- *n* larger



Confidence Intervals

- Example
  - A survey was conducted amongst 85 children to determine the number of hours they spend in front of the TV every week.
  - The results indicate that the mean for the sample was 24,5 hours with a standard deviation of 2,98 hours.
  - Estimate with 95% confidence the population mean hours that children spend watching TV.

$$\left[\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right] = \left[24.5 \pm 1.96 \frac{2,98}{\sqrt{85}}\right]$$

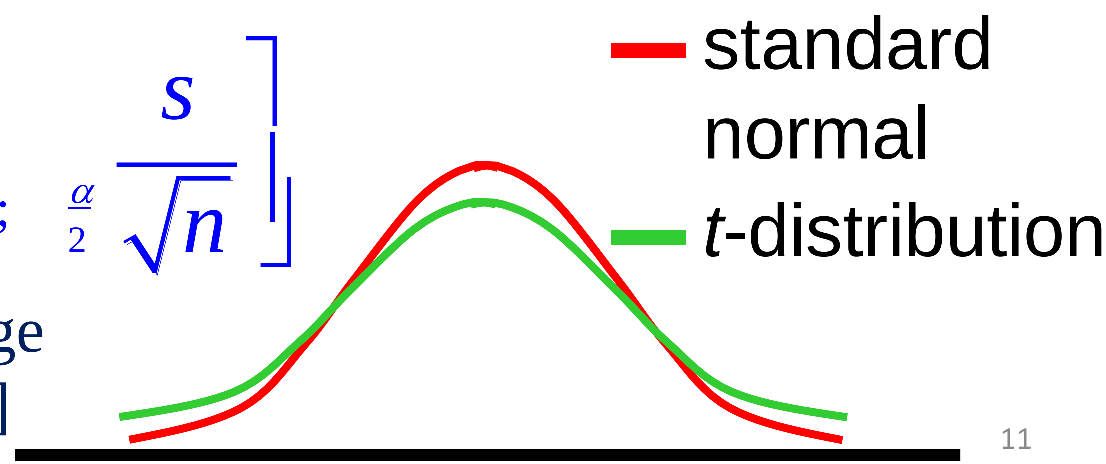
$$= \left[24.5 \pm 0.634\right]$$

$$= \left[23.866 \,;\, 25.134\right]$$

Confidence Intervals

- **<u>Confidence interval for Population Mean, *n* < 30</u>**
  - For a small sample from a normal population and *σ* is known, the normal distribution can be used.
  - If *σ* is unknown we use *s* to estimate *σ*
  - We need to replace the normal distribution with the t-distribution

$$CI(\mu)_{1-\alpha} = \left[\bar{x} \pm t_{n-1;\ \frac{\alpha}{2}} \frac{s}{\sqrt{n}} \right]$$



[see Example 8.3, Page No.386 (Reference book .1)]

Confidence Intervals

- Example
  - The manager of a small departmental store is concerned about the decline of his weekly sales.
  - He calculated the average and standard deviation of his sales for the past 12 weeks, $x = \text{R}12400$ and $s = \text{R}1346$
  - Estimate with 99% confidence the population mean sales of the departmental store.

$t_{11;0.995}$

$$\left[\bar{x} \pm t_{n-1;1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right] = \left[12400 \pm 3{,}106 \frac{1346}{\sqrt{12}}\right]$$

$$= [12400 \pm 1206{,}86]$$

$$= [11193{,}14 \,;\, 13606{,}86]$$

EXAMPLE 2

- A study of absenteeism among workers at a local mine during the previous year was carried out. A random sample of 25 miners revealed a mean absenteeism of 9.7days with a variance of 16 days. Construct a confidence interval for the average number of days of absence for miners for last year. Assume the population is normally distributed.

Confidence Intervals

- **Confidence interval for Population proportion**
  - Each element in the population can be classified as a success or failure

$$\text{Sample proportion } \hat{p} = \frac{\text{number of successes}}{\text{sample size}} = \frac{x}{n}$$

  - Proportion always between 0 and 1
  - For large samples the sample proportion $\hat{p}$ is approximately normal

$$CI(p)_{1-\alpha} = \left[\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

[see Article 11.4, Page No.362 (Reference book .2)]

Confidence Intervals

- Example
  - A sales manager needs to determine the proportion of defective radio returns that is made on a monthly basis.
  - In December 65 new radios were sold and in January 13 were returned for rework.
  - Estimate with 95% confidence the population proportion of returns for December.

$$\hat{p} = \frac{13}{65} = 0.2$$

$$\left[\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = \left[0.2 \pm 1.96 \sqrt{\frac{0.2(1-0.2)}{65}}\right]$$

$$= [0.2 \pm 0.097]$$

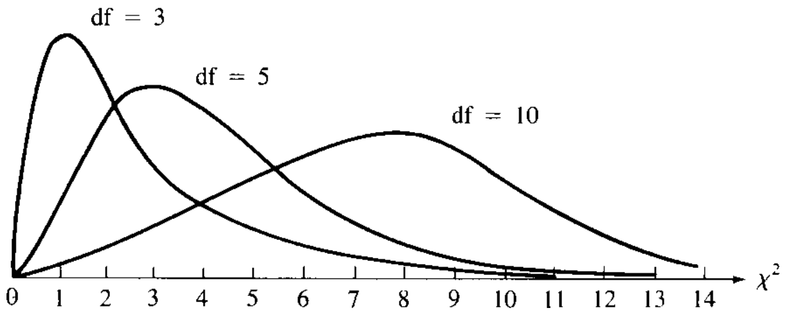
$$= [0.103\ ;\ 0.297]$$

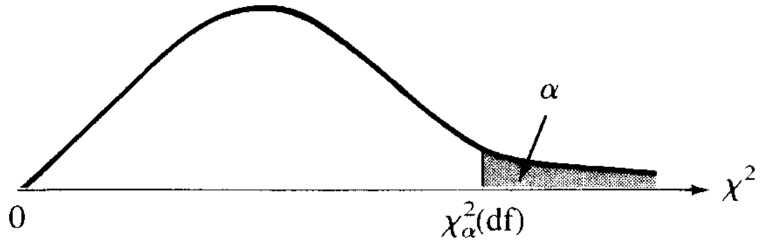
Confidence Intervals

- <u>**Confidence interval for Population Variance**</u>
  - Population variance very often important
  - Very often required for quality control
  - Sample drawn from a normal population
  - Sample variance is based on a random sample of size n
  - Distribution of s^2 resulted from repeated sampling is a χ^2 (chi-square) distribution

Confidence Intervals

- **<u>Confidence interval for Population Variance</u>**
  - χ^2 (chi-square) distribution
    - Skewed to the right distribution
    - Shape varies in relation to the degrees of freedom
    - Critical values from the χ^2-table A4(read same way as t distribution)
    - Critical value of $\chi^2_{1 - \alpha}$ specifies an area to the left
    - Critical value of χ^2_{α} specifies an area to the right





Confidence Intervals

- **Confidence interval for Population Variance**

$$CI(\sigma^2)_{1-\alpha} = \left[\frac{(n-1)s^2}{\chi^2_{n-1;(1-\alpha/2)}}; \frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}} \right]$$

[see Example 8.5, Page No.387 (Reference book .1)]

Confidence Intervals

- Example
  - For a binding machine to work on its optimum capacity the variation in the temperature of the room is vital.
  - The temperature for 30 consecutive hours were measured and sample standard deviation were found to be 0,68 degrees.
  - What will be a 90% confidence interval for σ^2?

$$CI(\sigma^2)_{1-\alpha} = \left[\frac{(n-1)s^2}{\chi^2_{n-1;1-\frac{\alpha}{2}}} ; \frac{(n-1)s^2}{\chi^2_{n-1;\frac{\alpha}{2}}} \right] = \left[\frac{29(0.68^2)}{\chi^2_{29;0.95}} ; \frac{29(0.68^2)}{\chi^2_{29;0.05}} \right]$$

n= 30; s = 0.68; α = 0.1

$$= \left[\frac{29(0.68^2)}{42.56} ; \frac{29(0.68^2)}{17.71} \right]$$

$$= [0.315; 0.757]$$

Where are we?

- So far we have looked at interval estimation procedures for μ, p and σ^2 for a **SINGLE POPULATION**
- We are now going to look at interval estimation procedures for:-
  - The difference between two population means
  - The difference between two population proportions
  - The ratio of two population variances

Confidence Intervals

- **Interval estimation for two populations**
  - There is different procedures for the differences in means, proportions and variances.

	Population 1	Sample 1	Population 2	Sample 2
Mean	μ_1	$\bar{x}_1$	μ_2	$\bar{x}_2$
Variance	σ^2_1	s^2_1	σ^2_2	s^2_2
Std dev	σ_1	s_1	σ_2	s_2
Size	N_1	n_1	N_2	n_2
Proportion	P_1	$\hat{p}_1$	P_2	$\hat{p}_2$

Confidence Intervals

- Confidence interval difference in means
  - Large independent samples

$$CI(\mu_1 - \mu_2)_{1-\alpha} = \left[(\overline{x_1} - \overline{x_2}) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$

if σ_1^2 and σ_2^2 is known

$$CI(\mu_1 - \mu_2)_{1-\alpha} = \left[(\overline{x_1} - \overline{x_2}) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

if σ_1^2 and σ_2^2 not is known

NOTE: If 0 is not included in the interval it means that 0 does not occur between the lower and upper boundaries of the interval

Example 1

Independent random samples of male and female employees selected from a large industrial plant yielded the following hourly wage results:-

MALE	FEMALE
$n_1 = 45$	$n_2 = 32$
$x = 6.00$	$x = 5.75$
$s_1 = 0.95$	$s_1 = 0.75$

Construct a 99% confidence interval for the difference between the hourly wages for all males and females and interpret the results

Example 1- Answer

$$1-\alpha=0.99$$

$$\alpha=0.01$$

$$1-\frac{\alpha}{2}=1-\frac{0.01}{2}=0.995$$

$Z_{0.995}$= 2.57

$$CI(\mu_1-\mu_2)_{0.99}=\left[(\overline{x}_1-\overline{x}_2)\pm Z_{1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right]$$

$$=\left[(6-5.75)\pm 2.57\sqrt{\frac{(0.95)^2}{45}+\frac{(0.75)^2}{32}}\right]$$

$$=[-0.2486;0.7486]$$

Example 1- answer

Interpretation:-

At a 99% level of confidence, the difference between the hourly wages of males and females is between -0.2486 and 0.7486 rand. The value 0 is included in the interval which tells us that there is a possibility that there is no difference between the two population means. To make sure whether there is a difference or not, a hypothesis test (next chapter!!!!) has to be performed.

Confidence Intervals

- **Confidence interval difference in means**
  - **Small independent samples**
  - **When sample sizes are small, n_1 & $n_2 < 30$ we use the t distribution**

$$CI(\mu_1 - \mu_2)_{1-\alpha} = \left[(\bar{x}_1 - \bar{x}_2) \pm t_{n_1 + n_2 - 2; 1 - \frac{\alpha}{2}} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$$

$$\text{with } s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

[see Theorem 11.5, Page No.362 (Reference book .2)]

NOTE: If both the limits of the confidence interval are negative you should suspect that the mean of first population is smaller than mean of second population

Example 1

A plant that operates two shifts per week would like to consider the difference in productivity for the two shifts. The number of units that each shift produces on each of the 5 working days is recorded in the following table:-

	Monday	Tuesday	Wednesday	Thursday	Friday
Shift 1	263	288	290	275	255
Shift 2	265	278	277	268	244

Assuming that the number of units produced by each shift is normally distributed and that the population standard deviations for the two shifts are equal construct a 99% confidence interval for the difference in mean productivity for the two shifts and comment on the result.

Confidence Intervals

- **Confidence interval difference in proportions**
  - **Large independent samples**

$$CI\,(p_1 - p_2)_{1-\alpha} = \left[(\hat{p}_1 - \hat{p}_2) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right]$$

with $\hat{p}_1 = \dfrac{x_1}{n_1}$ and $\hat{p}_2 = \dfrac{x_2}{n_2}$

[see Theorem 11.8, Page No.362 (Reference book .2)]

Example 1

Two groups of males are polled concerning their interest in a new electric razor that has four cutting edges. A sample of 64 males under the age of 40 indicated that only 12 were interested while in a sample of 36 males over the age of 40, only 8 indicated an interest. Construct a 95% confidence interval for the difference between age froup populations

Confidence Intervals

- Confidence interval for the ratio of two population variances
- We use the f distribution.

[see Theorem 11.10, Page No.362 (Reference book .2)]

$$CI\left(\frac{\sigma_1^2}{\sigma_2^2}\right)_{1-\alpha} = \left[\frac{s_1^2}{s_2^2}\left(\frac{1}{F_{n_1-1;n_2-1;\frac{\alpha}{2}}}\right); \frac{s_1^2}{s_2^2}\left(F_{n_2-1;n_1-1;\frac{\alpha}{2}}\right)\right]$$

NOTE: If 1 does not lie in the confidence interval, there is some evidence that the population variances are not equal

EXAMPLE 1

A criminologist is interested in comparing the consistency of the lengths of sentences given to people convicted of robbery by two judges. A random sample of 17 people convicted of robbery by judge 1 showed a standard deviation of 2.53 years, while a random sample of 21 people convicted by judge 2 showed a standard deviation of 1.34 years. Construct a 95% confidence interval for the ratio of the two populations variances. Does the data suggest that the variances of the lengths of sentences by the two judges differ? Motivate your answer.

REFERENCE BOOKS

1. GUN, A.M. ; GUPTA M.K. ; DASGUPTA, B. :An outline of STATISTICAL THEORY, volume two, Third Edition.

2. John E. Freund's Mathematical Statistics with Applications Seventh Edition Irwin Miller and Marylees Miller, Pearson Education.

3. Hogg , R.V. and Tenis , E.A. (1998): Probability and Statistical Inference, Sixth Edition, Pearson Education.

4. GUPTA, S.C. ; KAPOOR V.K.: Fundamental of Mathematical Statistics, Sultan Chand.