Estimation

-Dr. Umesh R A

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Estimation related Topics

- What is Statistical Inference?
- Point Estimation
- Sample Mean and Sample Variance
- Interval Estimation
- Confidence interval: one parameter
- Confidence interval: two parameter

Statistical Inference

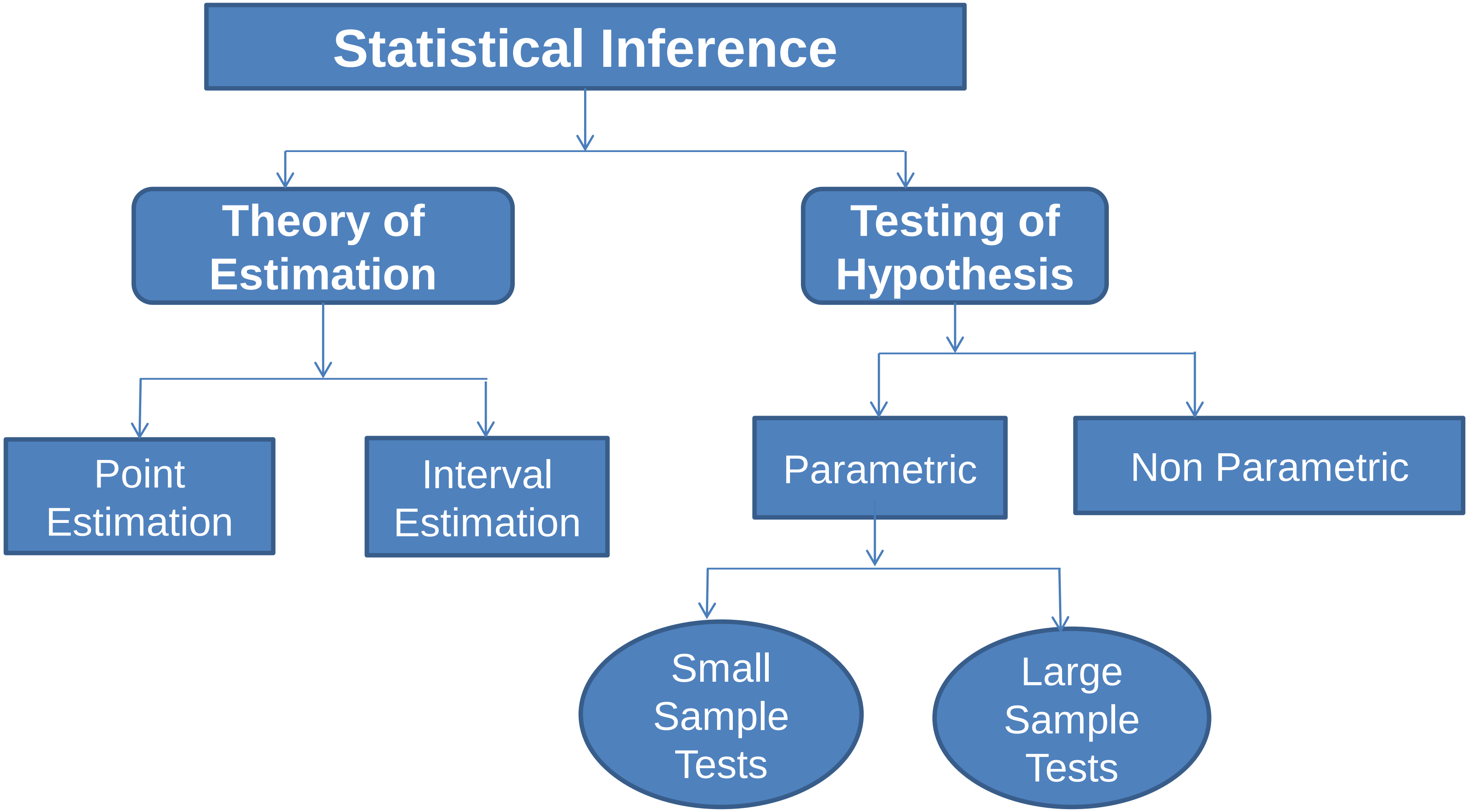
- **Statistical inference** is the process of drawing conclusions about populations or scientific truths from data.

"Statistical inference is that branch of statistics which deals with the theory and techniques of decision making decisions regarding the statistical nature of the population using sample drawn from the population"

It has two branches:

1. **Theory of estimation**
2. **Testing of hypothesis**

- Statistical inference:



Theory of Estimation

Estimation

- Estimation
  - First step of Inferential Statistics (Second step is Hypothesis Testing)

*“**Estimation** is the process of finding an **estimate**, or approximation, which is a value that is usable for some purpose even if input data may be incomplete, uncertain, or unstable. The value is nonetheless usable because it is derived from the best information available”*

*Also, it would be great if you could **quickly guess** how many people are in a room, how many cars in the street, how many boxes on the shelf, etc.*

- Estimation is finding a number that is **close enough** to the right answer.
  - You are **not** trying to get the **exact** right answer
  - What you want is something that is **good enough** (usually in a hurry!)

In everyday life a few cents here or there are not going to make much difference ... you should focus on the dollars!

- **Estimator:** Estimator of an unknown parameter is a statistics which specifies the likely value of that parameter.
- **Estimate:** Estimate is a specific value of the estimator for a specified sample.

Or simply

Estimator is statistic

Estimate is a numerical value.

Samples, Parameters & Statistics

- **Sampling**

  –Allows us to make inferences about a population based on a sample of that population

- **Parameters**

  –Numerical characteristics about the population that are of interest

- **Statistics**

  –Parameters cannot be exactly determined. They can only be estimated from samples

  –These estimates or summaries, based on the sample, are known as Statistics

- **Major aspects of samples and statistics:**

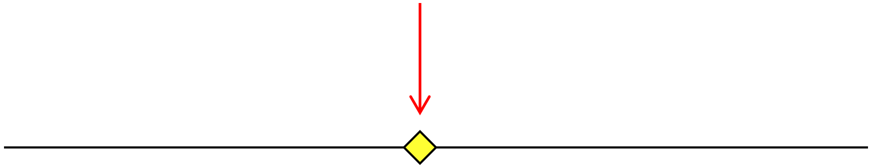
  –How accurate are the estimators (statistics)?

  –Is the sample truly representative of the population?

- Estimation is of two type:
  - Point estimation
  - Interval estimation

Point Estimation

- While estimating an unknown parameter if a single-value is proposed as the estimate, such estimation is called point estimation.



What is a Point Estimate?

- In simple terms, any statistic can be a point estimate. A statistic is an estimator of some parameter in a population. For example:
  - The sample standard deviation (s) is a point estimate of the population standard deviation (σ).
  - The sample mean (x-bar) is a point estimate of the population mean, μ
  - The sample variance (s^2 is a point estimate of the population variance (σ^2).
  - The sample proportion (p-hat) is a point estimate of the population Proportion , p

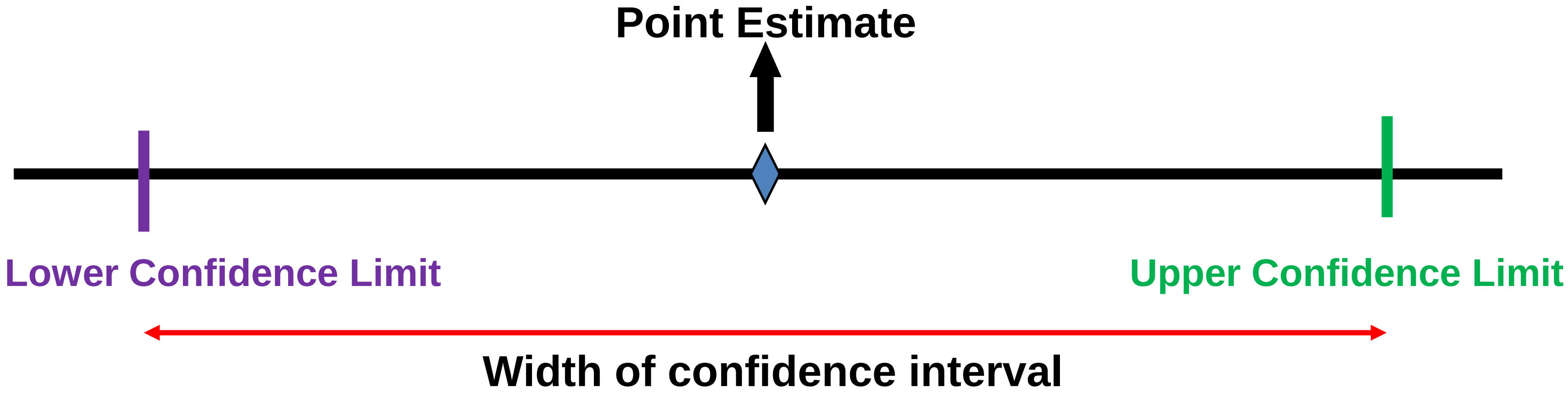
The following table displays some population characteristic researchers might try to estimate;

Characteristic	Point Estimate	Symbol	Parameter	Symbol
Mean IQ of 3rd graders	sample mean	x-bar	population mean	μ
Proportion of college students who ride their bicycle to school	sample proportion	p-hat	population proportion	p
What is the variation in weight among Olympic female gymnasts	sample standard error	s	population standard deviation	σ

Estimating Parameters

- Here are a few examples of point estimates and when you might use each one:
  - Sample means are used to find the center of continuous data.
  - Sample proportions are used to find the mean part or share per whole.
  - Sample standard errors describe the spread of data for means and proportions.

- A point estimate is a single number,
  - How much uncertainty is associated with a point estimate of a population parameter?
- An interval estimate provides a confidence level for the estimate.

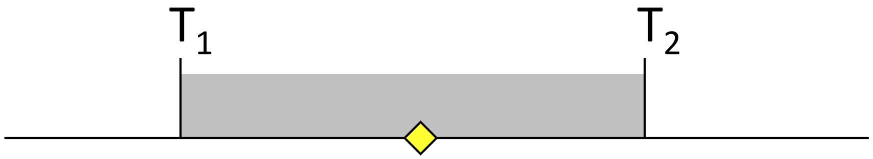


Interval Estimation

- Instead of proposing single value as estimate of population parameter, If we propose a small interval around the point estimate as the **likely interval to contain the parameter**, our proposition would be stronger.

- This interval which is likely to contain the parameter is called **Interval Estimation**.

- It is denoted such as (T_1, T_2) -------- (say)



Important Terms

- **Confidence Interval:** The interval (T_1,T_2) is called Confidence Interval.

- **Confidence Coefficient:** The probability that the confidence interval contains the parameter is called confidence coefficient. Ex. 90% or 95% etc.

- **Confidence Limits:** the limits T_1 and T_2 of the confidence interval is called confidence limits.

- **The general formula for all confidence intervals is equal to:**

Point Estimate ± (Critical Value)(Standard Error)

For Ex:

The general form of an interval estimate of a population mean is

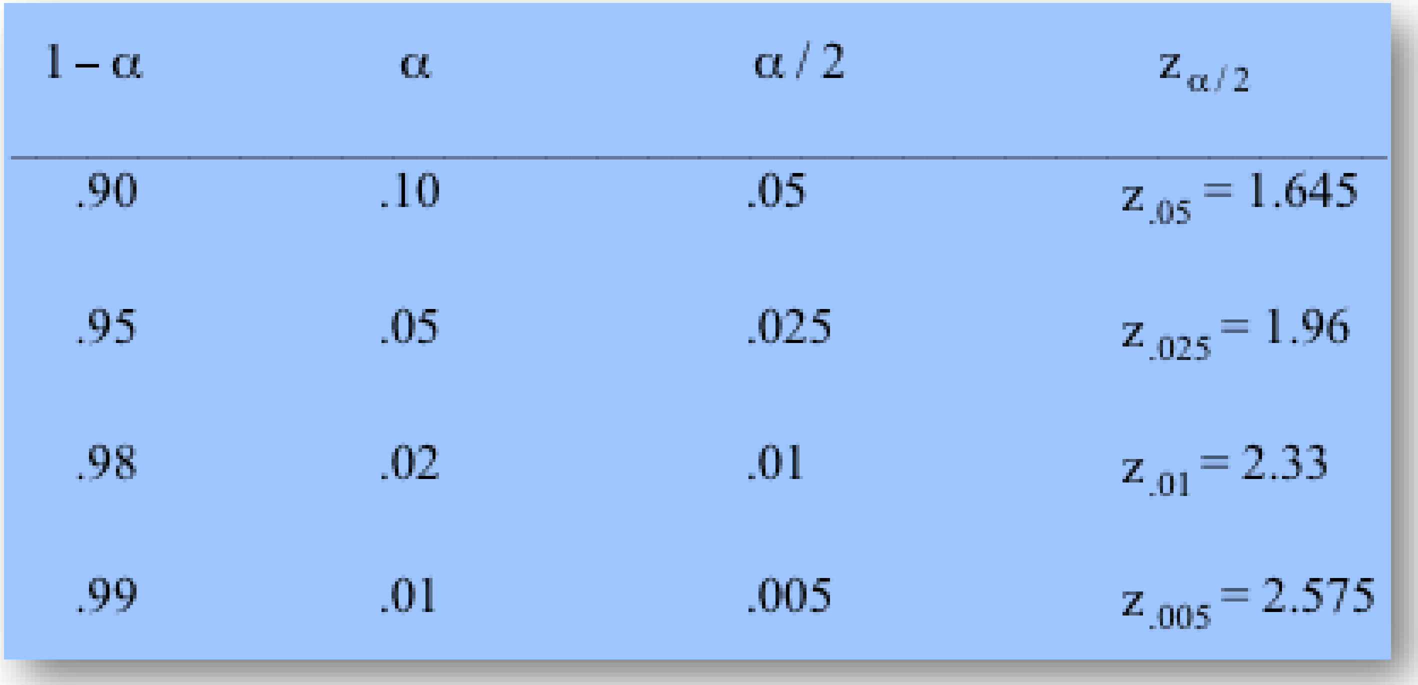
$$\bar{x} \pm \text{ Margin of Error}$$

- Suppose confidence level = 95%
- Also written $(1 - \alpha) = .95$
- α is the proportion of the distribution in the two tails areas outside the confidence interval

- A relative frequency interpretation:
  - If all possible samples of size n are taken and their means and intervals are estimated, 95% of all the intervals will include the **true value of that the unknown parameter**
- A specific interval either will contain or will not contain the true parameter (due to the 5% risk)

- Table values:

cut & keep handy!



$1-\alpha$	α	$\alpha/2$	$z_{\alpha/2}$
.90	.10	.05	$z_{.05} = 1.645$
.95	.05	.025	$z_{.025} = 1.96$
.98	.02	.01	$z_{.01} = 2.33$
.99	.01	.005	$z_{.005} = 2.575$

Confidence Interval Estimation

One Parameter

Confidence Interval Estimation of Population Mean (μ)

Confidence Interval Estimation of Population Mean, μ, when σ is known

- Assumptions
  - Population standard deviation σ is known
  - Population is normally distributed
  - If population is not normal, use large sample

- Confidence interval estimate:

$$\left\{\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} < \boldsymbol{\mu} < \left\{\bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\}$$

(where Z is the normal distribution's critical value for a probability of α/2 in each tail)

Where sample size is large, and population SD is not known

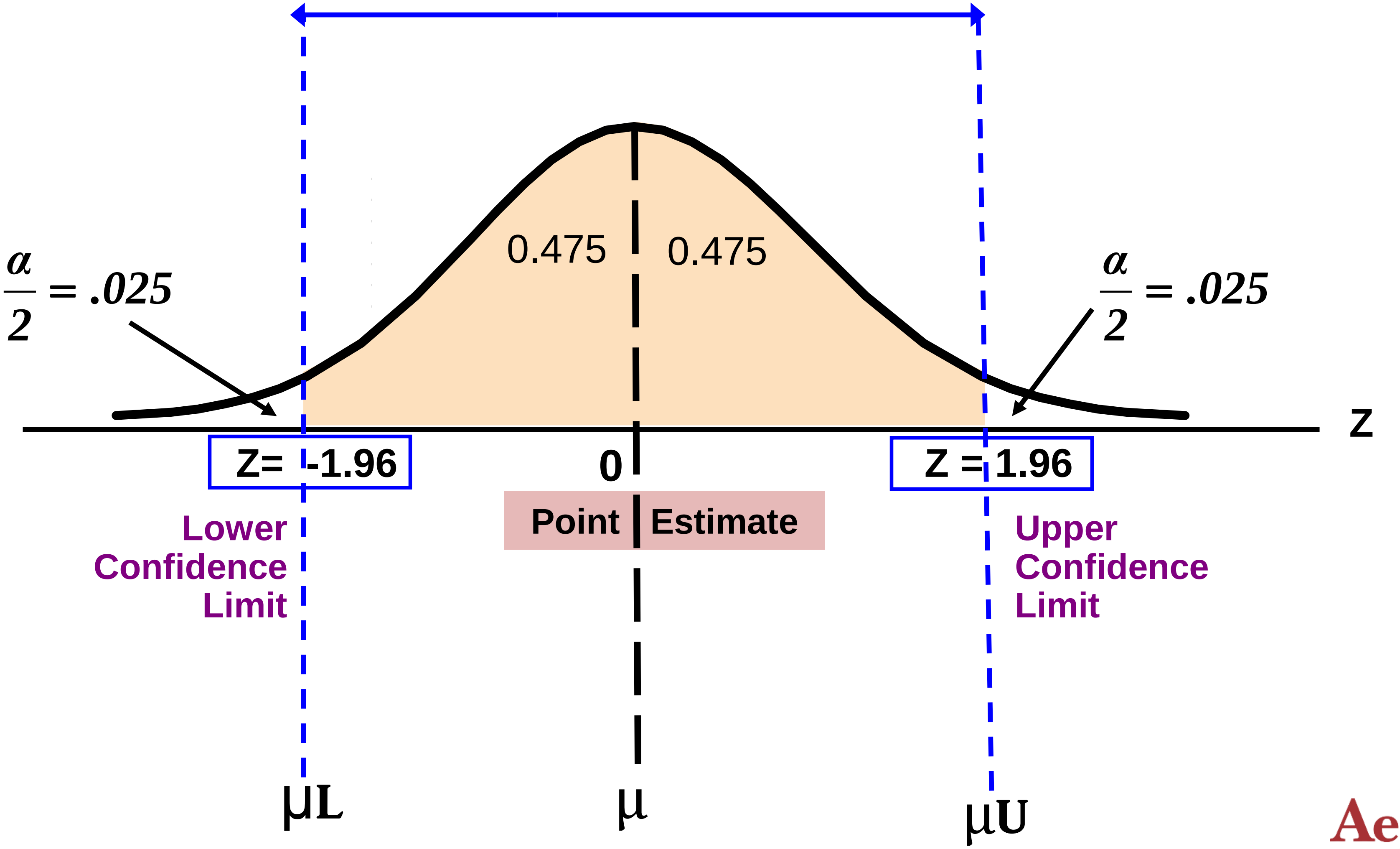
- Assumptions
  - –Population standard deviation σ is not known
  - –Population is normally distributed
  - –Sample Size is large (≥30)
  - –If population is not normal, use large sample
- Confidence interval estimate:

$$\left\{\bar{x} - Z_{\alpha/2}\frac{s}{\sqrt{n}}\right\} < \boldsymbol{\mu} < \left\{\bar{x} + Z_{\alpha/2}\frac{s}{\sqrt{n}}\right\}$$

(where Z is the normal distribution’s critical value for a probability of α/2 in each tail)

- Consider a 95% confidence interval:

$$1-\alpha = .95 \qquad \alpha = .05 \qquad \alpha / 2 = .025$$



Confidence Interval Estimation of Population Mean, when σ is Unknown and Sample Size is small

– If the population standard deviation σ is unknown, we can substitute the sample standard deviation, S

– Sample size is small (<30)

– This introduces extra uncertainty, since S varies from sample to sample

– So we use the student's *t* distribution instead of the normal Z distribution

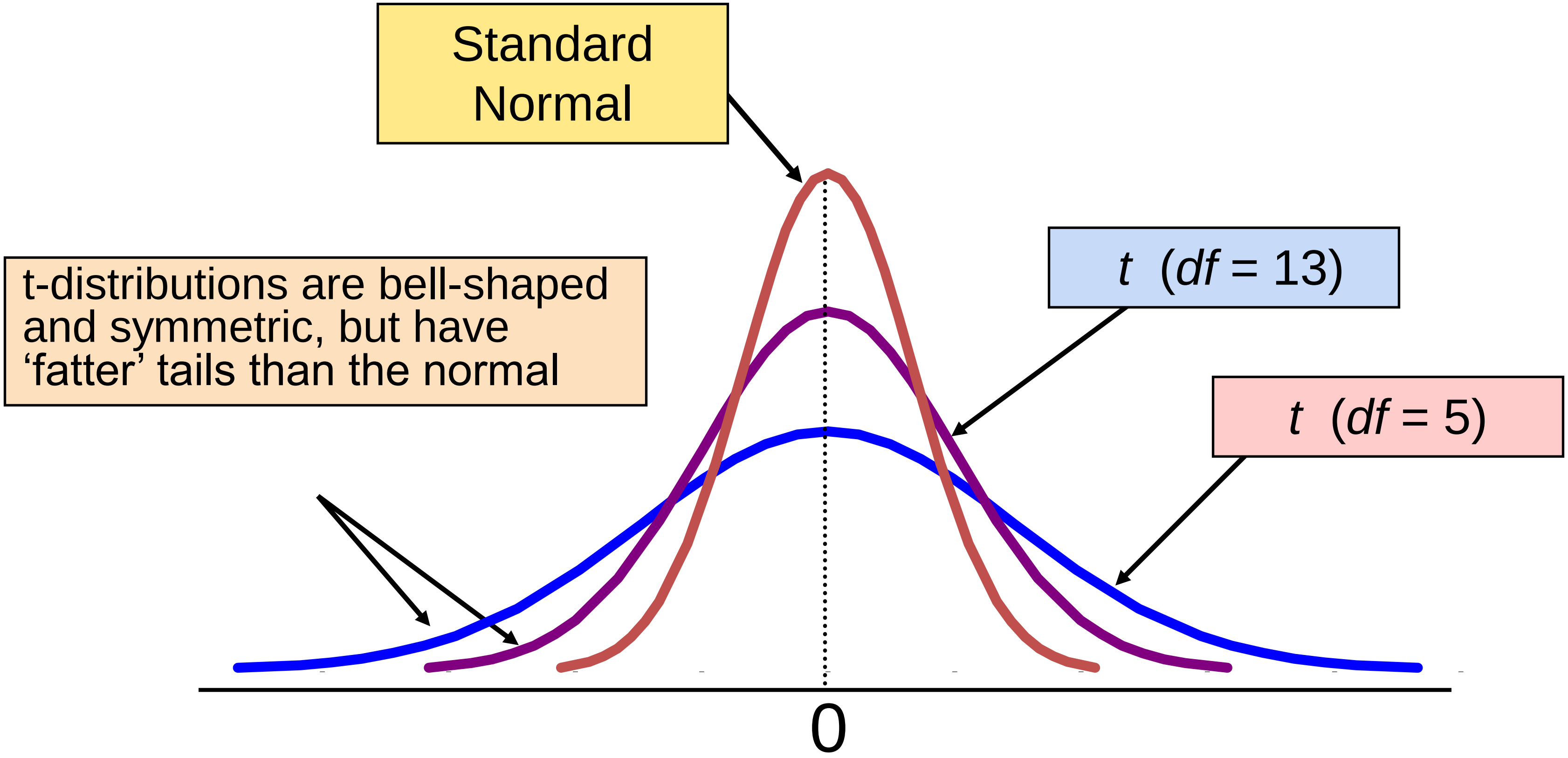
Confidence Interval Estimate Use Student's *t* Distribution :

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

(where t is the critical value of the *t* distribution with ***(n-1) degrees of freedom (d.f.)*** and an area of α/2 in each tail)

- *t* distribution is symmetrical around its mean of zero, like Z dist.
- Compare to Z dist., a larger portion of the probability areas are in the tails.
- As n increases, the *t* dist. approached the Z dist.
- t values depends on the degree of freedom.

- Student’s *t* distribution
- Note: t → Z as n increases



Confidence Interval Estimation
of
Population Proportion (p)

Confidence Interval: Proportions

- Another important “statistic” is the “proportion”
- We are often interested in:

Proportion of the population satisfying a certain criteria

  - Proportion of population above / below poverty line
  - Proportion of traveller’s reporting sick on arrival
  - Proportion of population using public transport
  - Proportion of kids dropping out of school by the age of 15

Confidence Interval: Proportions

- Let ‘P’ be the TRUE value of the proportion in the population
- Let ‘n’ be the size of the sample drawn from the population
- Let ‘X’ be the number of elements in the sample that exhibit the attribute under study
- The ‘estimated value’ of the TRUE proportion of the population is given by : $\hat{p}$= X/n

Confidence Interval: Proportions

- It can be shown that Z (it is normally distributed)
- Test Statistics is; $Z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$
- Confidence interval estimate

$$\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < \mathbf{p} < \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Thus, if number of elements ‘n’ in the sample and the proportion p in thê sample are known, one can estimate the interval of the TRUE proportion of the population at a given confidence level.

Confidence Interval Estimation of Population SD (σ)

Confidence Interval: Variance & SD

- The Chi-squared distribution : best represents the probability distribution of such "sum of squared items"

The test statistic is;

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

Degrees of freedom= (n-1)

- Therefore Chi-squared distribution is used to derive the confidence interval of sampling distribution of variance (and, hence, the standard deviation)

SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Confidence Interval: Variance & SD

- Confidence interval estimate for SD

$$\sqrt{\left\{\frac{(n-1)s^2}{\chi^2_{(n-1),\alpha/2}}\right\}} < \boldsymbol{\sigma} < \sqrt{\left\{\frac{(n-1)s^2}{\chi^2_{(n-1),1-(\alpha/2)}}\right\}}$$

Degrees of freedom= n-1

- Confidence interval estimate for Variance

$$\left\{\frac{(n-1)s^2}{\chi^2_{(n-1),\alpha/2}}\right\} < \boldsymbol{\sigma^2} < \left\{\frac{(n-1)s^2}{\chi^2_{(n-1),1-(\alpha/2)}}\right\}$$

Degrees of freedom= n-1

Confidence Interval Estimation

Two Parameter

Confidence Interval Estimation of difference of Two Population Proportion (p_1-p_2)

CI Estimation of difference of Two Population Proportion (p_1-p_2)

- Use Z- table

$$\boldsymbol{C.I.} = (\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Confidence Interval Estimation of difference of Two Population Mean (μ_2- μ_1)

Case 1) When $\boldsymbol{\sigma_1}$ and $\boldsymbol{\sigma_2}$ Known

- When σ_1 and σ_2 Known
- **We use Z-table**

$$C.I. = (\bar{x}_2 - \bar{x}_1) \pm Z_{\alpha/2}\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$$

Case 2) When $\boldsymbol{\sigma_1}$ and $\boldsymbol{\sigma_2}$ Unknown and Sample size is large

- When σ_1 and σ_2 Unknown and both $\mathbf{n_1}$ and $\mathbf{n_2}$ are greater than or equal to **30**
- **We use Z-table**

$$C.I. = (\bar{x}_2 - \bar{x}_1) \pm Z_{\alpha/2}\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

Case 3) When **σ_1** and **σ_2** Unknown and Sample size is small

- When σ_1 and σ_2 Unknown and either sample size is less than **30**
- **We use t-table**

$$C.I. = (\bar{x}_2 - \bar{x}_1) \pm t_{\alpha/2} \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

- With degrees of freedom,

$$d.f. = \frac{\left(\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{{s_1}^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{{s_2}^2}{n_2}\right)^2}$$

Case 4) Equal Population Variance

We use Pooled SD

- If we assume equal variances between groups, we can pool the information on variability (sample variances) to generate an estimate of the population variability. Therefore, the standard error (SE) of the difference in sample means is the pooled estimate of the common standard deviation **(Sp)**

$$S.E.(\bar{x}_1 - \bar{x}_2) = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and the pooled estimate of the common standard deviation is

Computing the CI for a Difference Between Two Means

- If the sample sizes are larger, that is both **n_1** and **n_2** are greater than or equal to **30**, then one uses the z-table.
- If either sample size is less than **30**, then the t-table is used.
- For both large and small samples **Sp** is (same formula as follow)

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

- If either $n_1 < 30$ or $n_2 < 30$, use the t-table:

$$C.I. = (\bar{x}_1 - \bar{x}_2) \pm t_{(n_1+n_2-2),\alpha/2}\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Use the t-table with degrees of freedom = ***$n_1 + n_2 - 2$***

Case 5) Paired Sample

- X_i – Before Y_i – After ________ (Paired Sample)

Here, $d_i = X_i - Y_i$

- Then the test Statistics is,

$$t = \frac{\bar{d} - (\mu_1 - \mu_2)}{s_d / \sqrt{n}} \qquad d.f. = n - 1$$

Where,

$$\bar{d} = \frac{\sum d_i}{n} \quad \text{and} \quad \text{Var(d)} = \frac{\sum (d_i - \bar{d})^2}{n-1}$$

- Confidence interval in case of Paired Sample

$$\mathrm{C.I.} = \bar{d} \pm t_{\alpha/2}\frac{s_d}{\sqrt{n}} \qquad d.f. = n - 1$$

Confidence Interval Estimation of ratio of Two Population variance $\frac{{\sigma_2}^2}{{\sigma_1}^2}$.

Confidence interval for the ratio of variances

- **Confidence interval for the ratio of variances of two normally distributed populations**
- The F distribution uses two values for degrees of freedom (n_2-1 , n_1-1)

Confidence interval for $\frac{{\sigma_2}^2}{{\sigma_1}^2}$ is

$$\frac{{s_2}^2}{{s_1}^2} \cdot \frac{1}{F_{1-\alpha/2}} \leq \frac{{\sigma_2}^2}{{\sigma_1}^2} \leq \frac{{s_2}^2}{{s_1}^2} \cdot \frac{1}{F\alpha_{/2}}$$

END