743- Regression and Time Series

Mamikon S. Ginovyan

Statistics Review

A. Background. Point Estimation

- **Point Estimation**. Let $\underline{X} = (X_1, X_2, ..., X_n)$ be a sample form $X : F(x) = F(x, q)$ with unknown $q = (q_1, \ldots, q_p)$.

- **A point estimator** of q = any statistic $\hat{q}_n = \hat{q}_n(\underline{X})$, which is a **random variable**, and the distribution of this RV is called **sampling distribution** of $\hat{q}_n$.

- **A point estimate** of $q : \hat{q}_n = \hat{q}_n(\underline{x}), \underline{x} = (x_1, x_2, \ldots, x_n)$ is a **single number.**

Common point estimators

1. If $q = m =$ **Mean**, then

$$\hat{q} = \hat{q}_n = \overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$ is a point **estimator** for $q = m$;

$$\hat{q} = \hat{q}_n = \overline{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$ is a point **estimate** for $q = m$.

2. If $q = s^2 =$ **Variance**, then

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2$$ is a point **estimator** for $q = s^2$;

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2$$ is a point **estimate** for $q = s^2$.

Common point estimators

3. If $\theta = \sigma_{X,Y} = Cov(X,Y) =$ **Covariance**, then

$$\hat{\sigma}_{X,Y} = S_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \text{point } \textbf{estimator},$$

and $s_{x,y}$ = point **estimate** for $\sigma_{X,Y}$.

$$\boxed{S_{XX} = S_X^2.}$$

4. If $\theta = \rho_{X,Y} = \rho(X,Y) =$ **Correlation**, then

$$\hat{\rho}_{X,Y} = \frac{S_{X,Y}}{S_X S_Y} = \text{ point } \textbf{estimator}$$

and $\hat{\rho}_{x,y} = \dfrac{s_{x,y}}{s_x s_y} =$ point **estimate** for $\rho_{X,Y}$.

Properties of Point Estimators: Unbiasedness

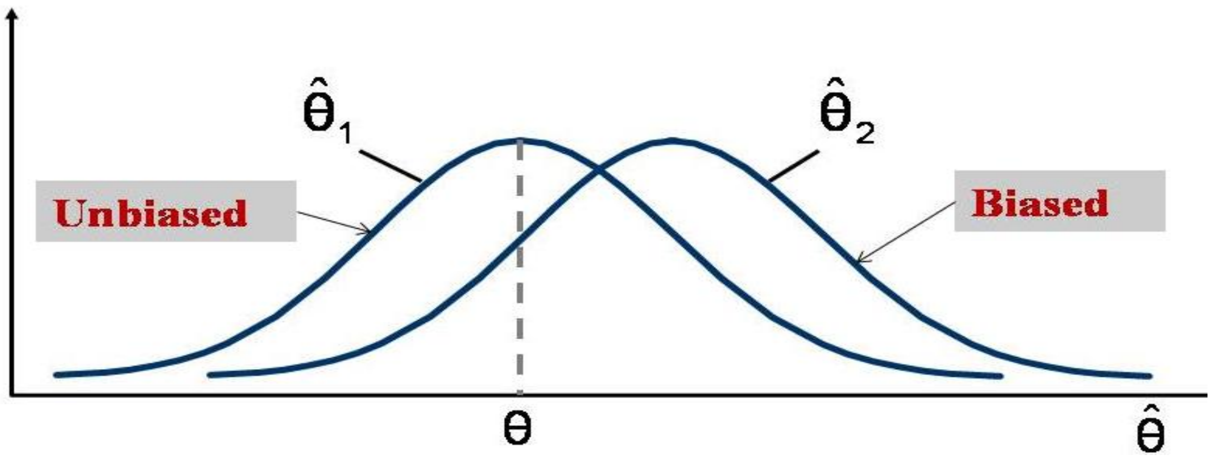
- **Def 1**. Let $\hat{q} = \hat{q}_n$ be a point estimator of q. The difference

$$B(\hat{q}) = E[\hat{q}] - q$$

is called the **bias** of $\hat{q}$.

- **Def 2**. A point estimator $\hat{q}$ is called an **unbiased** estimator of q if

$$B(\hat{q}) = 0 \Leftrightarrow E[\hat{q}] = q .$$



Properties of Point Estimators: UMVUE

Let $\boldsymbol{U}$ denote the set of all unbiased estimators $\hat{q}_n = \hat{q}_n(\underline{X})$ of unknown parameter q, and let $\hat{q}_1, \hat{q}_2 \in U$.
Then, $\hat{q}_1 \in U$ is said to be more efficient than $\hat{q}_2 \in U$ if

$$Var(\hat{q}_1) \leq Var(\hat{q}_2)$$

The relative efficiency of $\hat{q}_1$ with respect to $\hat{q}_2$ is defined by

$$\text{Relative Efficiency} = \frac{\text{Var}(\hat{q}_2)}{\text{Var}(\hat{q}_1)}$$

An estimator $\hat{q}_1 \in U$ is called **UMVUE (or best)** estimator of q, if

$$Var(\hat{q}_1) \leq Var(\hat{q}) \quad \text{for all} \quad \hat{q} \in U.$$

The Mean Squared Error (MSE)

The **MSE** $(\hat{q})$ of a point estimator $\hat{q}$ is defined to be

$$MSE(\hat{q}) = E(\hat{q} - q)^2.$$

If $B(\hat{q}) = E[\hat{q}] - q$ is the bias of $\hat{q}$, then

$$MSE(\hat{q}) = Var(\hat{q}) + \left[B(\hat{q}) \right]^2.$$

If $\hat{q}$ is unbiased estimator, then

$$MSE(\hat{q}) = Var(\hat{q}).$$

Properties of Point Estimators: Consistency

- **Def.** A point estimator $\hat{q}_n$ is called **consistent** estimator of q if

$$\hat{q}_n \xrightarrow{P} q \quad as \quad n \to \infty.$$

A Criteria for consistency:

- If $\hat{q}_n$ is an **unbiased** estimator for q, then $\hat{q}_n$ will be **consistent** if
$$\lim_{n \to \infty} Var[\hat{q}_n] = 0.$$

- If $\hat{q}_n$ is a **biased** estimator for q, then $\hat{q}_n$ will be **consistent** if
$$\lim_{n \to \infty} MSE[\hat{q}_n] = 0.$$

Asymptotically Normal Estimators

A point estimator $\hat{q}_n$ is called **asymptotically normal** estimator of q if the asymptotic distribution of

$$\sqrt{n}(\hat{q}_n - q) \quad as \quad n \to \infty$$

is **normal**, that is,

$$\sqrt{n}(\hat{q}_n - q) \xrightarrow{D} N\left(0, s_q^2\right).$$

Asymptotically Efficient Estimators

A point estimator $\hat{q}_n$ is called **asymptotically efficient** estimator of q if it is **asymptotically normal**, that is,

$$\sqrt{n}(\hat{q}_n - q) \xrightarrow{D} N\left(0, s_q^2\right), \quad s_q^2 = \frac{1}{I(q)},$$

where

$$I(q) = I_n(q) = E\left[\left(\frac{\partial}{\partial q} \ln f(\underline{X}, q)\right)^2\right]$$

is the Fisher information for sample $\underline{X} = (X_1, \ldots, X_n)$.

Rao-Cramer Lower Bound:

$$Var[\hat{q}(\underline{X})] \geq \frac{1}{I(q)}.$$

The Method of Maximum Likelihood

- **Likelihood function.** Let $\underline{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution having *pdf* (or *pmf*) $f(x, q), q \in \Theta$ (*for simplicity*, first we consider the one-dimensional parameter $q \in \Theta \subset R^1$).

The joint *pdf* of $\underline{X} = (X_1, \ldots, X_n)$ is

$$f(\underline{x}; q) = f(x_1, x_2, \ldots, x_n; q) = \prod_{k=1}^{n} f(x_k; q), \underline{x} \in R^n.$$

The Method of Maximum Likelihood

This function can be considered as a function of the parameter q, in this case, it is denoted by $L(q)$ and is called **likelihood function**. Thus,

$$L(q) = L(q;\underline{x}) = \prod_{k=1}^{n} f(x_k;q), q \in \Theta. \qquad (1)$$

The Method of Maximum Likelihood

- **Def.** The **maximum likelihood estimate (MLE)** $\hat{q} = \hat{q}_n$ of the unknown parameter q is the value of the parameter that **maximize the likelihood function**, that is,

$$L(\hat{q}) \geq L(q) \text{ for all } q \in \Theta.$$

We also use the notation

$$\hat{q}_{MLE} = \hat{q} = Arg \max_{q \in \Theta} L(q ; \underline{X}).$$

- The **maximum likelihood estimate (MLE)** is

$$\hat{q}_{MLE} = \hat{q}(\underline{x}) = \hat{q}(x_1, \dots, x_n).$$

The Method of Maximum Likelihood

- To find the MLE $\hat{q}$ we often apply calculus. Specifically we solve the equation

$$\frac{d}{dq} L(q) = 0 \qquad (2)$$

for q and check that the solution $\hat{q}$ is a maximum point for $L(q)$.

$$\left.\frac{d^2}{dq^2} L(q)\right|_{q=\hat{q}} < 0.$$

- Equation (2) is called **Estimation or Likelihood Equation.**

.

The Method of Maximum Likelihood

- In some cases, it is convenient, instead of likelihood function $L(q)$, to consider the **log-likelihood function**

$$l(q) = \log L(q).$$

Since $\log L(q)$ increases with $L(q)$, they will have **the same maximum point** $\hat{q}$.

So (2) is equivalent to

$$\frac{d}{dq} l(q) = 0 \qquad (3)$$

The Method of Maximum Likelihood

- **Remark.** If we have ***p***-dimensional parameter $q = (q_1,\dots,q_p)$, then the MLE $\hat{q} = (\hat{q}_1,\dots,\hat{q}_p)$, can be found by solving the system of equations

$$\begin{cases} \dfrac{\partial^k}{\partial q_k} l(q_1,\dots,q_p) = 0 \\ k = 1, 2, \dots, p. \end{cases}$$

The Method of Maximum Likelihood

- **Proposition.** **(The Invariance Principle of MLE).**

Let $\hat{q} = (\hat{q}_1, \ldots, \hat{q}_p)$ be the MLE of the parameter $q = (q_1, \ldots, q_p)$.

Then the MLE of the parameter $h = g(q)$, where $\boldsymbol{g}$ is any function, is given by

$$\hat{h} = g(\hat{q}).$$

The Method of Maximum Likelihood

- **Ex. 1.** An experimenter has reason to believe that the pdf describing the variability in a certain type of measurement is the continuous model :

(1) $f(x,q) = \dfrac{1}{q^{2}} x \cdot e^{-x/q}, 0 < x < \infty, 0 < q < \infty.$

Five data points have been collected:

(2) $x_1 = 9.2;\ \ x_2 = 5.6;\ \ x_3 = 18.4;\ \ x_4 = 12.1;\ \ x_5 = 10.7.$

(a) Find the maximum likelihood **estimate** for q.

(b) Find the maximum likelihood **estimate** for $h = g(q) = q^{3}$.

Example 1-Solution

Solution.

(a) **First** we find a **general formula** for MLE $\hat{q}$.

The **likelihood function** is

$$L(q) = L(q, \underline{x}) = \prod_{i=1}^{n} f(x_i, q) = \prod_{i=1}^{n} \frac{1}{q^2} x_i \cdot e^{-x_i/q}$$

$$= \frac{1}{q^{2n}} \left(\prod_{i=1}^{n} x_i\right) \cdot \left(e^{-\frac{1}{q}\sum_{i=1}^{n} x_i}\right)$$

The **log-likelihood** function is

$$l(q) = \ln L(q) = -2n \ln q + \ln \prod_{i=1}^{n} x_i - \frac{1}{q} \sum_{i=1}^{n} x_i .$$

Example 1-Solution

The **Estimation Equation** is

$$\frac{d}{dq} l(q) = -\frac{2n}{q} + \frac{1}{q^2} \sum_{i=1}^{n} x_i = 0.$$

Solving for q, we find the following general formula for maximum likelihood estimator and estimate:

$$\hat{q} = \frac{1}{2n} \sum_{i=1}^{n} X_i$$ = maximum likelihood **estimator**

$$\hat{q} = \frac{1}{2n} \sum_{i=1}^{n} x_i.$$ = maximum likelihood **estimate.**

Example 1-Solution

Finally, the MLE $\hat{q}$ for the specified sample (2) is ***n=5***

$$\hat{q} = \frac{1}{2n}\sum_{i=1}^{n} x_i = \frac{1}{2\cdot 5}\sum_{i=1}^{5} x_i = \frac{56}{10} = 5.6. \qquad \text{So, } \hat{q}_{MLE} = 5.6.$$

(b) Find the maximum likelihood estimate for $h = g(q) = q^3$.

Use the Invariance Principle of MLE

$$\hat{h} = g(\hat{q}) = (\hat{q})^3 = \left(\frac{1}{2n}\sum_{i=1}^{n} X_i\right)^3,$$

$$\hat{h} = g(\hat{q}) = (\hat{q})^3 = 5.6^3 = 175.61.$$

The Method of Maximum Likelihood

- **Ex.2.** Let $X_1, ..., X_n$ be a Random Sample of size ***n*** from the normal distribution $N(m, s^2)$ with unknown parameters $q_1 = m \quad and \quad q_2 = s^2$:

$$f(x_1, q_1, q_2) = f(x, m, s^2) = \frac{1}{\sqrt{2p\,s}} \cdot e^{-\frac{(x-m)^2}{2s^2}},$$

$$-\infty < x < \infty, -\infty < m < \infty, 0 < s^2 < \infty.$$

Find the ML Estimators and Estimates for $m \quad and \quad s^2$.

Example 2-Solution

Solution.

1) The **Likelihood function** is

$$L(q_1, q_2) = L(m, s^2) == \prod_{i=1}^{n} f(x_i, m, s^2)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2p} \cdot s} \cdot e^{-\frac{(x_i - m)^2}{2s^2}} = (2ps^2)^{-n/2} e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - m}{s}\right)^2}.$$

2) The **Log-Likelihood** function is

$$l(m, s^2) = \log L(m, s^2) = -\frac{n}{2}\ln(2ps^2) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - m}{s}\right)^2.$$

Example 2-Solution

3) Compute the partial derivatives of $l(m, s^2)$ w.r.t. m and s^2.

$$\frac{\partial}{\partial m} l(m, s^2) = -\sum_{i=1}^{n} \left(\frac{x_i - m}{s} \right) \left(-\frac{1}{s} \right)$$

$$\frac{\partial}{\partial s^2} l(m, s^2) = -\frac{n}{2} \cdot \frac{1}{2p s^2} 2p - \frac{1}{2} \sum_{i=1}^{n} (x_i - m)^2 \left(-\frac{1}{s^4}\right)$$

Example 2-Solution

4) The system of Estimation Equation is

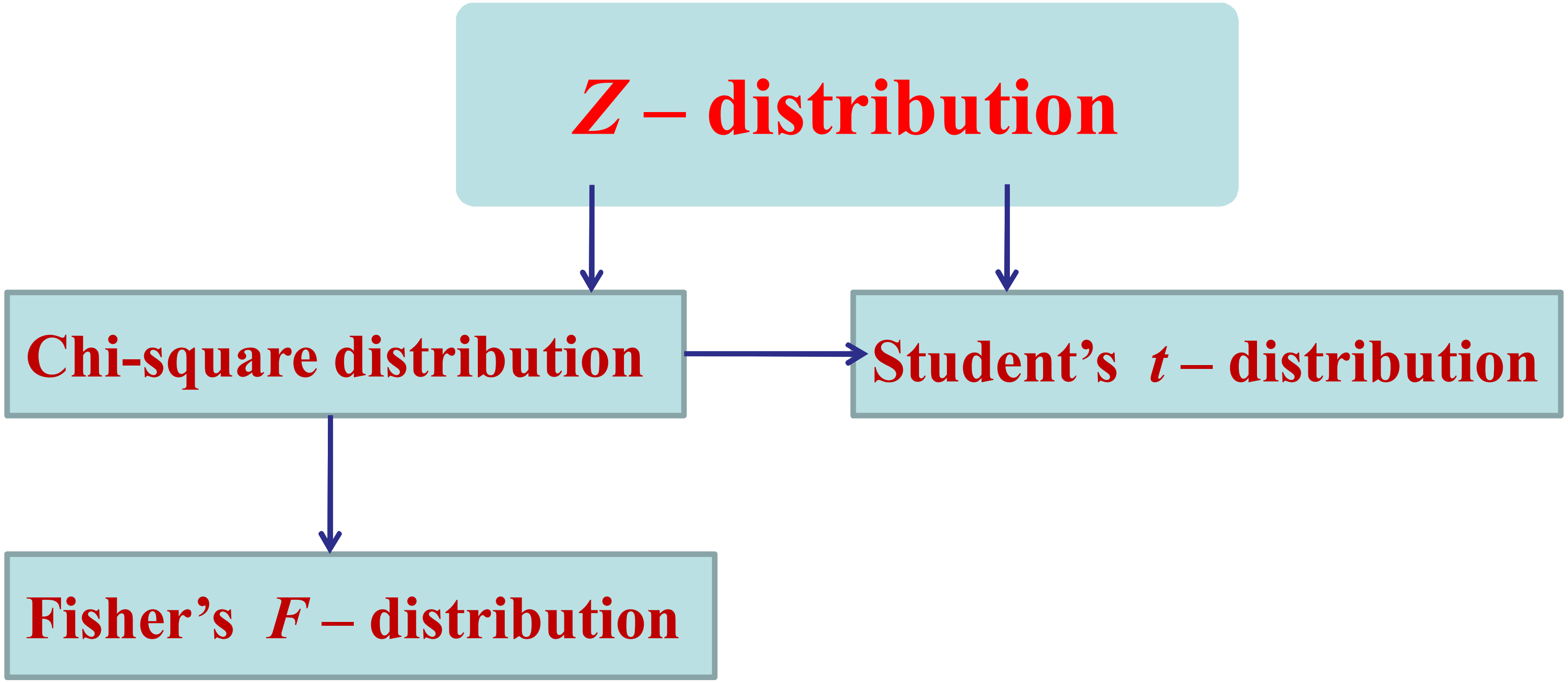
$$\begin{cases} \dfrac{\partial}{\partial m} l(m, s^2) = 0 \\ \dfrac{\partial}{\partial s^2} l(m, s^2) = 0 \end{cases} \Leftrightarrow \begin{cases} \displaystyle\sum_{i=1}^{n} (x_i - m) = 0 \\ -ns^2 + \displaystyle\sum_{i=1}^{n} (x_i - m)^2 = 0 \end{cases}$$

5) Solving for m and s^2 we obtain

$$\begin{aligned} \hat{m} &= \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x} \\ \hat{s}^2 &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 \end{aligned} \rightarrow$$

MLEstimates for m and s^2.

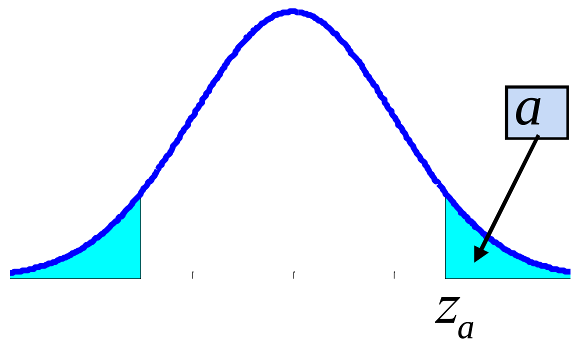
B. Sampling Distributions



Standard Normal Distribution

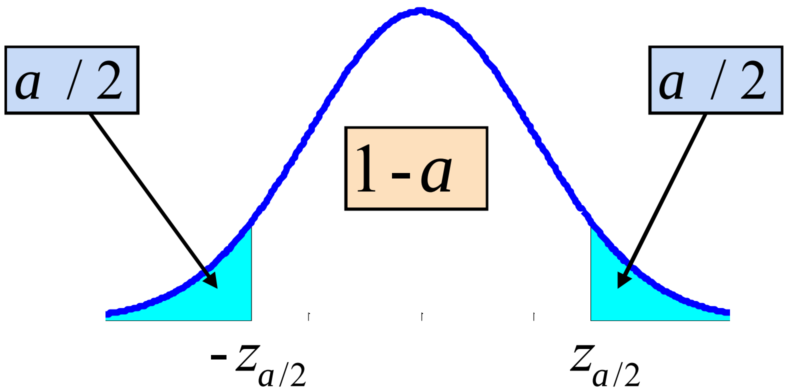
Definition: Let α be given number such that $0<\alpha<1$, the value z_α of $Z \sim N(0,1)$ satisfying

$$\boxed{P(Z > z_\alpha) = \alpha}$$



is called α - **upper percentile**, or **Critical Value** of ***Z.***

$$\boxed{P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha}$$



z_α - values can be found from **Z-Table**

Example:

For $\alpha = .025$: $z_\alpha = z_{.025} = 1.96$

The Chi-square Distribution

Definition: Let $Z_1, \ldots, Z_n$ be a sequence of **independent standard normal** random variables (a random sample from standard normal distribution) , then the distribution of the RV defined by

$$Y_n = c^2 = \sum_{i=1}^{n} Z_i^2 = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

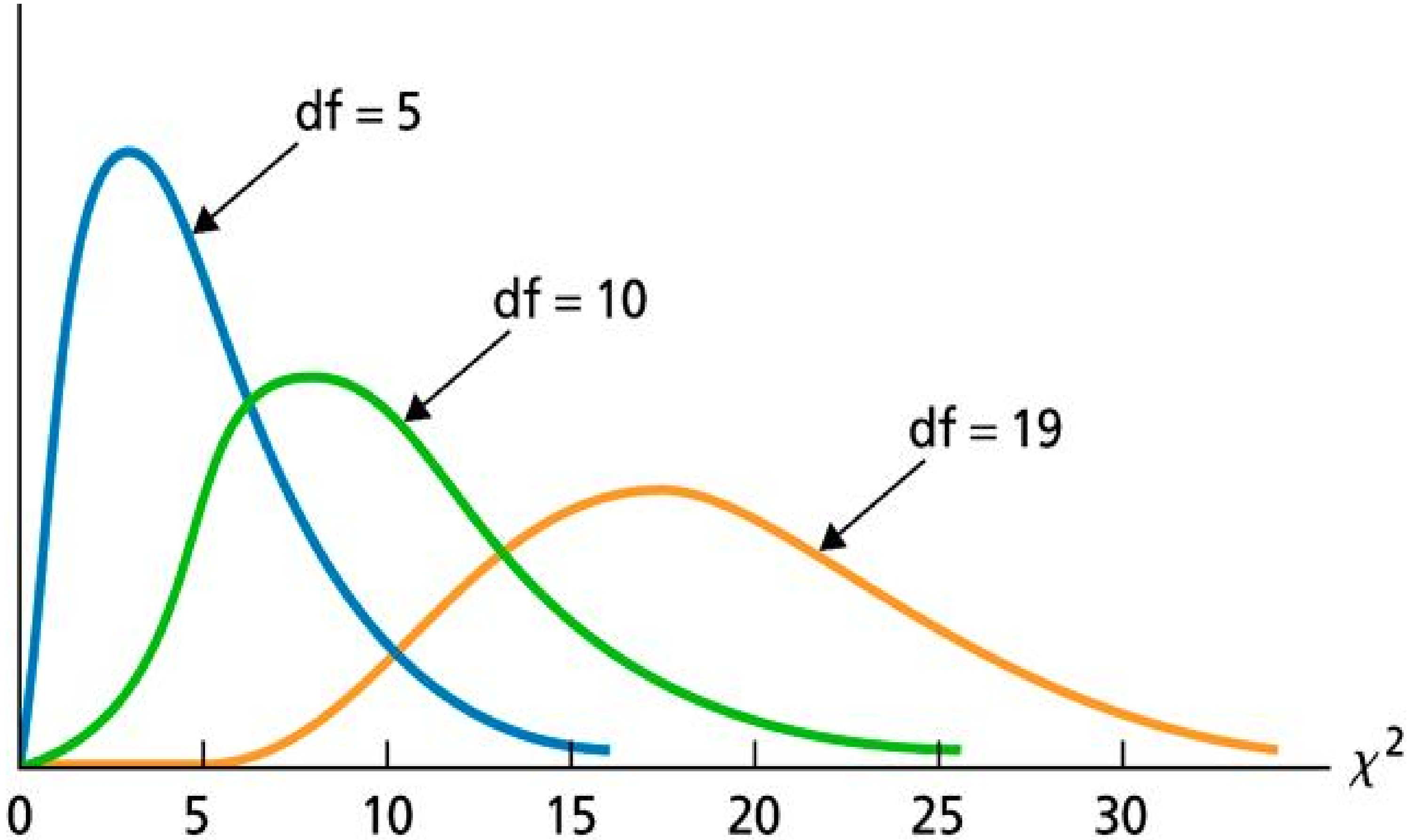
is called **chi-square distribution** with ***n*** degrees of freedom (df).

Notation: $Y \sim c_n^2 = c^2(n)$

Remark: Actually, there are **infinitely many** chi-square distributions, and we identify the chi-square distribution in question by its number of degrees of freedom (df).

The Chi-square Distribution

The following Figure shows three c^2 - curves and illustrates some basic properties of c^2 - distribution.



The Chi-square Distribution

Properties of the Chi-Square χ^2 - Distribution:

1. The χ^2-distribution is **not symmetric**. The χ^2--distribution is **skewed right**.

2. The shape of the chi-square distribution depends on the degrees of freedom.

3. As the number of degrees of freedom increases, the chi-square distribution becomes more nearly symmetric.

4. The values of χ^2 are nonnegative, that is, the values of χ^2 are greater than or equal to 0.

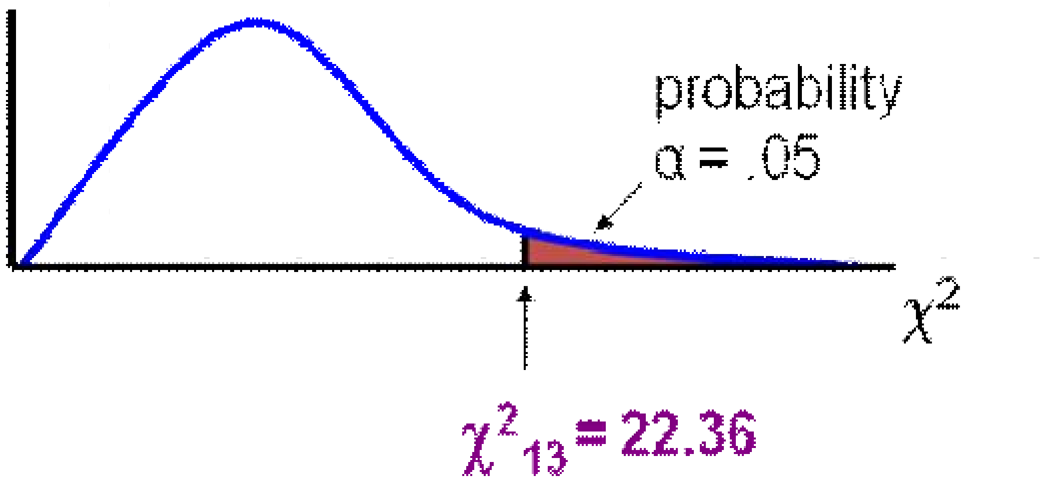
The Chi-square Distribution

Definition: Let a be given number such that $0<a<1$, the value

$c_a^2 = c_{a,n}^2$ of $c^2 \sim c_n^2$ satisfying

$$\boxed{P(c^2 > c_a^2) = a}$$

is called a - **upper percentile**, or **Critical Value** of χ^2–distribution. c_a^2 -values can be found from χ^2-**Table**



Example:

For $a = .05$; $n = 13$

$$c^2_{.05,13} = 22.36$$

Student's *t*- Distribution

Definition: Let ***Z*** and ***Y*** be two **independent** random variables such that $Z \sim N(0,1)$ and $X \sim c^2(n)$, then the distribution of the RV defined by

$$T = \frac{Z}{\sqrt{X/n}}$$

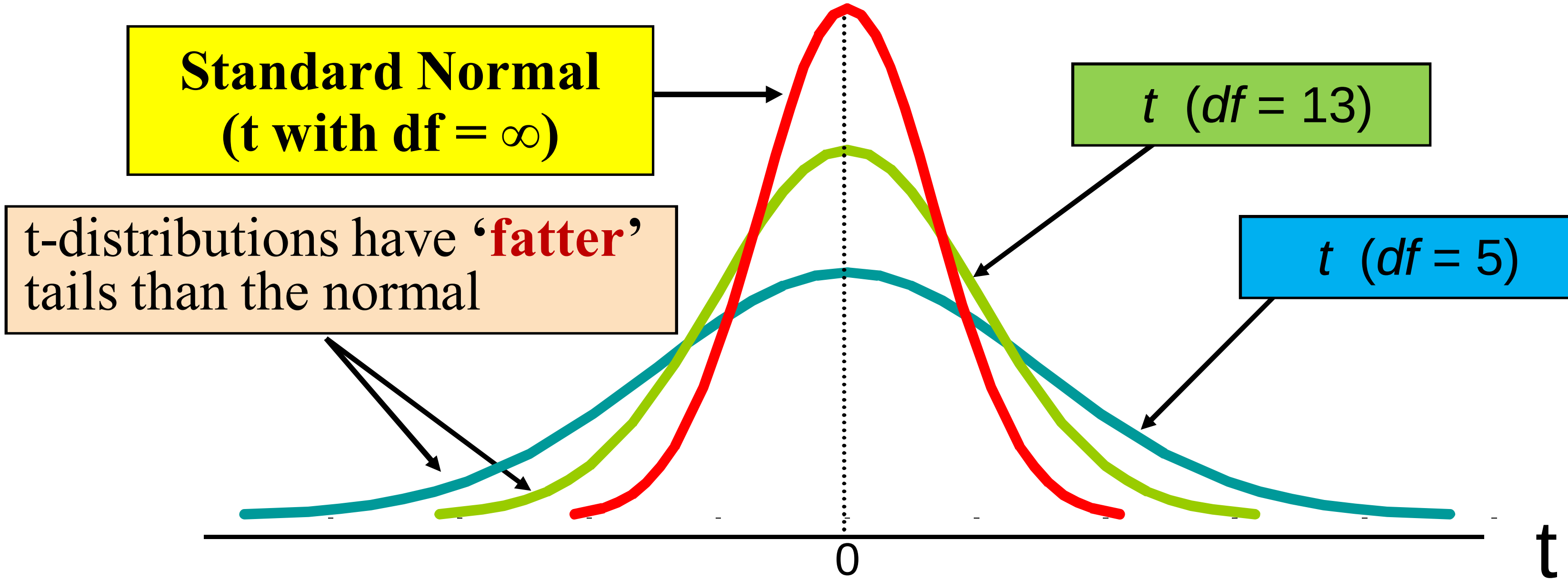
is called **Student's *t*- distribution** with ***n*** degrees of freedom (df).

Notation: $T \sim t(n) = t_n$.

Remark: Actually, there are **infinitely many *t***-distributions, and we identify the ***t***-distribution in question by its number of degrees of freedom (***n = df***).

Properties of Student's t - Distribution

1. **t- distributions** are **bell-shaped**, **centered** at **0**, and **symmetric** about 0: **P(T > 0) = P(T < 0) =1/2.**
2. **t- distributions depend** on $n = df$: $T \sim t_n$
3. For $n < 30$ **t- distribution** and **Z** ~ **N(0, 1)** are **quite different**.
4. For $n \geq 30$ **t- distribution** and **Z** ~ **N(0, 1)** are **close**: $T \longrightarrow Z$.

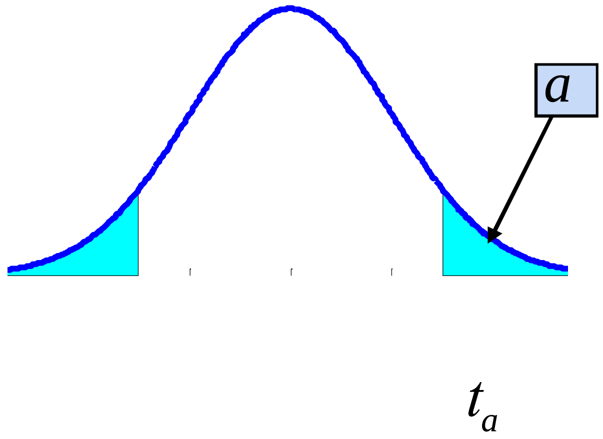


Critical Values of t – Distribution: t_a -values

Definition: Let a be given number such that $0 < a < 1$, the value

$t_a = t_{a,df} = t_{a,n}$ of $T \sim t_n$ satisfying

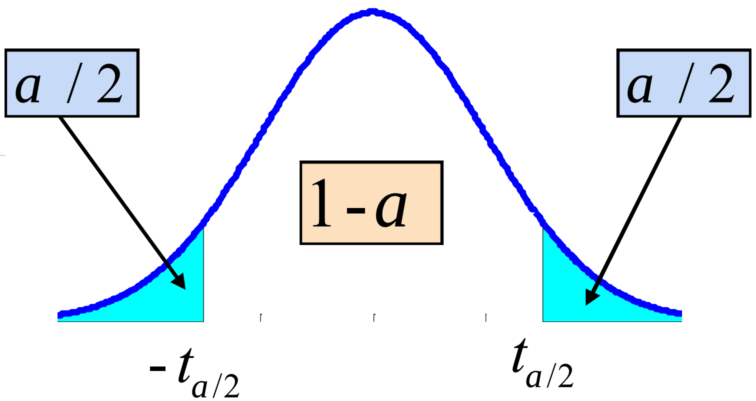
$$\boxed{P(T > t_a) = a}$$



is called a - **Critical Value (CV) of t – Distribution.**

We have $\boxed{P(-t_{a/2} < T < t_{a/2}) = 1 - a}$

t_a -values can be found from **T-Table (Table B3).**



Example: $t_{.025,7} = \boxed{2.365}$

Fisher's *F*- Distribution

Definition: Let X_1 and X_2 be two **independent** random variables such that $X_1 \sim c^2(n_1)$ and $X_2 \sim c^2(n_2)$, then the distribution of the RV defined by

$$F = \frac{X_1 / n_1}{X_2 / n_2}$$

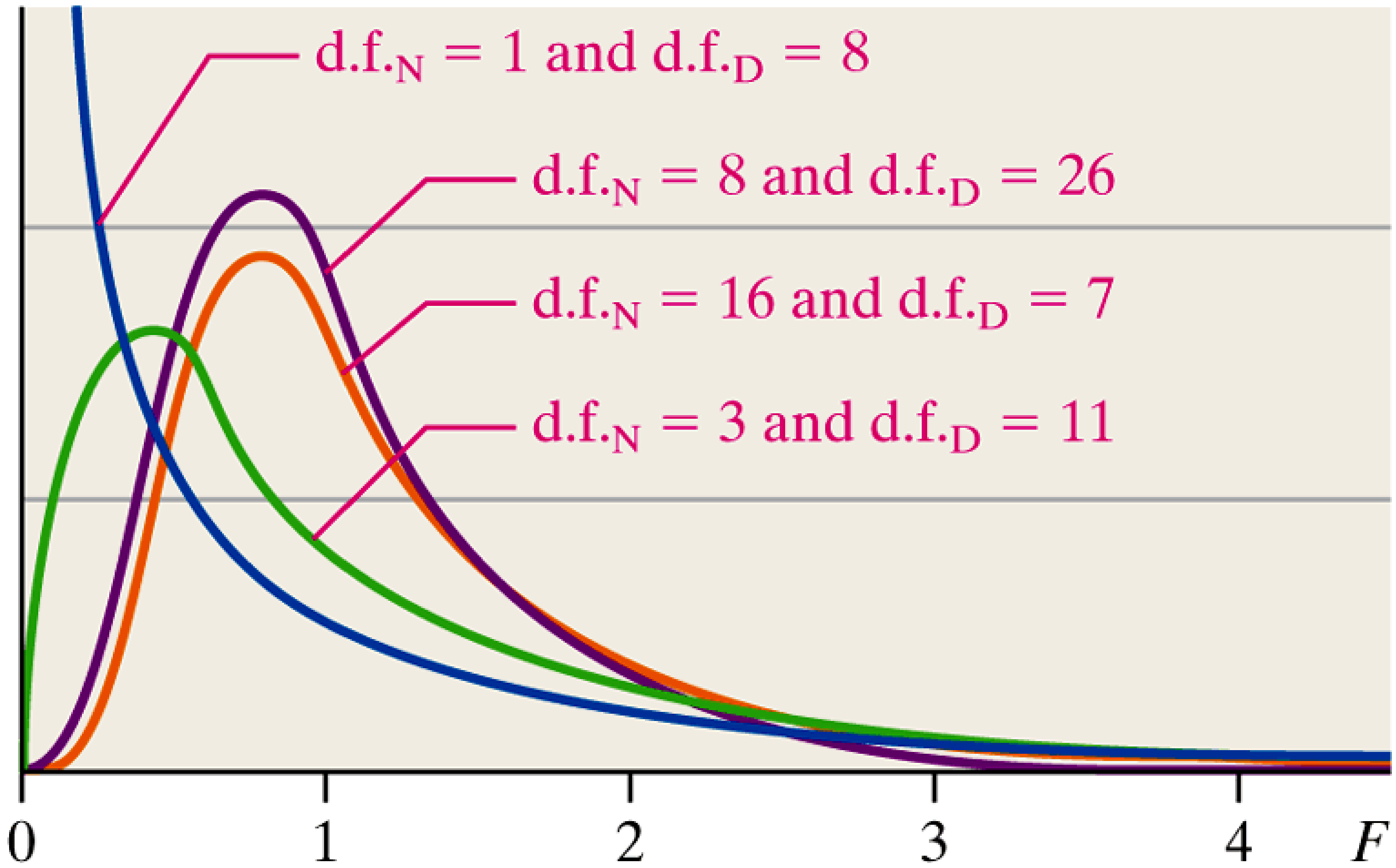
is called **Fisher's *F*- distribution** with n_1 ***numerator df*** and n_2 ***denominator df*** .

Notation: $F \sim F_{n_1, n_2}$

Remark: Actually, there are **infinitely many** ***F***-distributions, and we identify the ***F***-distribution in question by its number of degrees of freedom ($n_1 = df_1$; $n_2 = df_2$).

Fisher's *F*- Distribution

The following Figure shows three ***F*- *curves*** and illustrates some basic properties of ***F*- distribution**.



Fisher's *F*- Distribution

Properties of *F*- distribution:

1. The **F-distribution** is **not symmetric**.

2. The F-distribution is skewed right.

3. The shape of **F-distribution** depends on the degrees of freedom, just as with Student's *t*-distribution and χ^2-distribution.

4. The values of ***F*** are nonnegative.

Fisher's *F*- Distribution

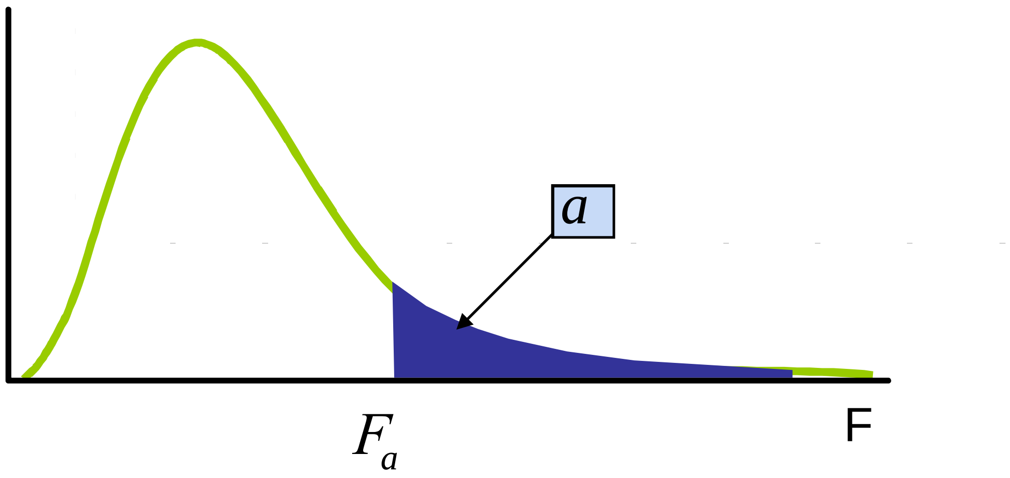
Definition: Let a be given number such that $0 < a < 1$, the value

$F_a = F_{a, n_1, n_2}$ of $F \sim F_{n_1, n_2}$ satisfying

$$\boxed{P(F > F_a) = a}$$

$$\boxed{F_{1-a, n_1, n_2} = \frac{1}{F_{a, n_2, n_1}}}$$

is called a - **upper percentile**, or **Critical Value** of ***F***–distribution. F_a -values can be found from **F-Table .**



Example:

For $a = .025$;

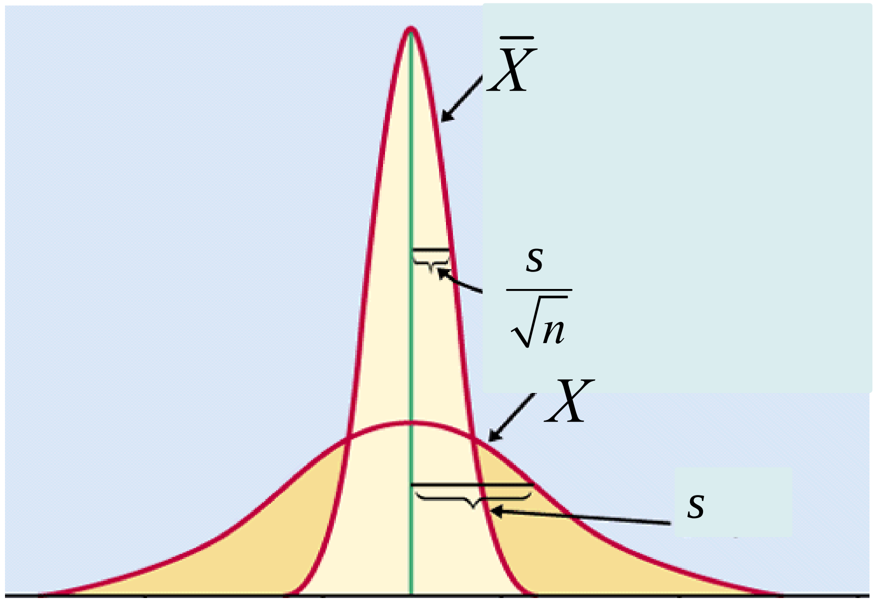
$\boldsymbol{n_1 = 20}$; $\boldsymbol{n_2 = 15}$

$F_{.025, 20, 15} = 2.76$

Sampling Distribution of Sample Mean

Theorem 1. (Sampling distribution of the sample mean for normally distributed population).

If we have a population as a possible values of a **normal** RV $\boldsymbol{X}$ with mean $\boldsymbol{\mu}$ and **known** variance s^2, then for **any sample of size** $\boldsymbol{n}$ the sampling distribution of $\bar{X}$ **is also normal** with the same mean $m_{\bar{X}} = m$ and variance $s^2_{\bar{X}} = \dfrac{s^2}{n}$.



Symbolically:

If $X \sim N(m, s^2)$,

then $\bar{X} \sim N(m, s^2 / n)$.

The sampling distribution of the sample mean

Theorem 1 is a special case of the following **general result.**

- **Theorem 2** **(Additivity property of Normal distribution).**

Let $X_1,\ldots,X_n$ be **independent** RV's such that

$$X_i \sim N(m_i, s_i^2), i = \overline{1,n}.$$

Then for any constants $a_i \in R$ the linear combination

$$X = \sum_{i=1}^{n} a_i X_i : N(m, s^2),$$

where $m = \sum_{i=1}^{n} a_i m_i$ and $s^2 = \sum_{i=1}^{n} a_i^2 s_i^2.$

- **Remark.** Theorem 1 now follows from Theorem 2 with

$$a_i = \frac{1}{n}, m_i = m, s_i^2 = s^2.$$

Sampling Distribution of Sample Mean

Theorem 3. (CLT: Sampling distribution of the sample mean for arbitrary population).

Assume that we have a population as possible values of an **arbitrary** RV $\boldsymbol{X}$ with mean $\boldsymbol{\mu}$ and variance s^2 $(0 < s^2 < \infty)$, and that a random sample of size $\boldsymbol{n}$, is taken from this population.

Then the sampling distribution of $\bar{X}$ becomes **approximately normal** with mean $m_{\bar{X}} = m$ and variance $s_{\bar{X}}^2 = \dfrac{s^2}{n}$, as the sample size $\boldsymbol{n}$ becomes **large.**

Symbolically: If $X \sim ANY(m, s^2)$, then $\bar{X} \sim AN(m, s^2 / n)$.
A = Approximately, for large $\boldsymbol{n}$.

Note: $\boldsymbol{n}$ is **large** if $\boldsymbol{n \geq 30}$.

Sampling Distribution of Sample Mean

Z –transform for sampling mean:

$$\bar{X} \sim AN\left(m_{\bar{X}} = m, s^{2}_{\bar{X}} = \frac{s^{2}}{n} \right)$$

Z –transform:

$$\boxed{Z = \frac{\bar{X} - m_{\bar{X}}}{s_{\bar{X}}} = \frac{\bar{X} - m}{s / \sqrt{n}} \sim N(0,1)}$$

Thus, the **Sampling Distribution of** $\bar{X}$ **is Z –distribution.**

Note. In order to find probabilities for $\bar{X}$ it is enough to know the standard error $s_{\bar{X}} = \dfrac{s}{\sqrt{n}}$.

Sampling Distribution of Sample Mean

Theorem 4. **(Extended CLT).**

Assume that we have a population as possible values of an **arbitrary** RV $\boldsymbol{X}$ with mean $\boldsymbol{\mu}$ and **unknown variance** s^2, and that a random sample of size $\boldsymbol{n}$ is taken from this population, and the sample variance $\mathbf{S^2}$ is used as a **point estimator** for **unknown variance** s^2, then as the sample size $\boldsymbol{n}$ becomes **large** **($\boldsymbol{n} \geq 30$) ,**

$$Z = \frac{\bar{X} - m}{S/\sqrt{n}} \sim N(0,1)$$

Thus, the **Sampling Distribution of** $\bar{X}$ **is Z –distribution.**

Sampling Distribution of Sample Mean

Theorem 5. (Student's Theorem).

Assume that we have a population as possible values of a **normally** distributed RV X with mean μ and **unknown variance** $s^2: \; X \sim N(m, s^2)$,
and that a random sample of size n is taken from this population, and the sample variance $\mathbf{S^2}$ is used as a **point estimator** for **unknown variance** s^2, then the statistic

$$\boxed{T = \frac{\bar{X} - m}{S / \sqrt{n}} \sim t_{(n-1)}}$$

has **Student's *t*-distribution** with $df = (n - 1)$.

Thus, the **Sampling Distribution of** $\bar{X}$ **is *t* –distribution.**

Properties of the Sample Variance

Theorem 6: Sampling Distribution of Sample Variance s^{2}.

If a simple random sample of size ***n*** is obtained from a **normally distributed** population with mean m and standard deviation s, then the statistic

$$c^{2} = \frac{(n-1)S^{2}}{s^{2}} \sim c^{2}_{n-1}$$

where $\mathbf{S^2}$ is a sample variance, has a **chi-square distribution** with **(*n*-1)** degrees of freedom (***df***).

Thus, the **Sampling Distribution of** s^{2} **is** χ^2–**distribution.**

C. Interval Estimation.

Summary. Confidence, Prediction and Tolerance Intervals

Let $X_1, X_2, ..., X_n$ be a RS from the distribution of a RV $X : F(x,q)$ with unknown parameter q, and let $a\ (0 \le a \le 1)$ be a given number.

1.The interval $[\hat{q}_L, \hat{q}_U]$ such that $P_q(\hat{q}_L \le q \le \hat{q}_U) = 1 - a$ is called a $100(1-a)\%$ **Confidence Interval** for q.

So **CI covers unknown parameter** q.

Interval Estimation.

2. The interval $[\hat{q}_L, \hat{q}_U]$ such that $P_q(\hat{q}_L \le X_{n+1} \le \hat{q}_U) = 1 - a$ is called a $100(1-a)\%$ <u>**Prediction Interval**</u> for X_{n+1}.

So <u>**PI covers a new RV**</u> X_{n+1}.

3. Given a number $p (0 \le p \le 1)$. The interval $[\hat{q}_L, \hat{q}_U]$ such that

$$P_q[F(\hat{q}_U(\underline{X}), q) - F(\hat{q}_L(\underline{X}), q) \ge p] = 1 - a$$

is called a $100(1-a)\%$ <u>**Tolerance Interval**</u> for ***100p%*** of the population. So <u>**TI covers a proportion of population**</u>.

Confidence Intervals about population mean μ

1. Normal population or Large-sample, SD s is known.

The $100(1-a)\%$ **Confidence Interval (CI)** for m is the interval $(\hat{m}_L, \hat{m}_U)$,
where

$$\hat{m}_L = \bar{X} - z_{a/2} \cdot \frac{s}{\sqrt{n}} \quad \text{and} \quad \hat{m}_U = \bar{X} + z_{a/2} \cdot \frac{s}{\sqrt{n}}.$$

where $z_{a/2}$ is the value of ***Z***-statistic satisfying:

$$P(Z > z_{a/2}) = a / 2.$$

Confidence Intervals about population mean μ

2. Large-sample, SD s is known.

The $100(1-a)\%$ **Confidence Interval (CI)** for m is the interval $(\hat{m}_L, \hat{m}_U)$,

where

$$\hat{m}_L = \bar{X} - z_{a/2} \cdot \frac{s}{\sqrt{n}} \quad \text{and} \quad \hat{m}_U = \bar{X} + z_{a/2} \cdot \frac{s}{\sqrt{n}}.$$

where $z_{a/2}$ is the value of ***Z***-statistic satisfying:

$$P(Z > z_{a/2}) = a / 2.$$

Confidence Intervals about population mean μ

3. Normal population or Small-sample ($n < 30$), SD s is unknown.

The $100(1-a)\%$ **Confidence Interval (CI)** for m is the interval $(\hat{m}_L, \hat{m}_U)$,

where

$$\hat{m}_L = \bar{X} - t_{a/2} \cdot \frac{s}{\sqrt{n}} \quad \text{and} \quad \hat{m}_U = \bar{X} + t_{a/2} \cdot \frac{s}{\sqrt{n}}.$$

where $t_{a/2} = t_{a/2,(n-1)}$ is the value of ***t*-statistic** satisfying:

$$P(t > t_{a/2}) = a / 2.$$

D. Test of Hypotheses

- **Specification of the Test Statistic and its Distribution under** H_0

- This step depends on the problem of interest.

- For example, in testing of hypothesis concerning mean of population, that is,

$$H_0 : m = m_0 \quad \text{vs} \quad H_1 : m \neq m_0 .$$

Test of Hypotheses

A sample mean $\bar{X}$ is used to specify the test statistic, and we have two cases:

a) ***Z***-statistic :

$$Z = \frac{\bar{X} - m}{s \ / \sqrt{n}};$$

$$Z = \frac{\bar{X} - m_0}{s \ / \sqrt{n}} \sim N(0,1) \qquad \textbf{\textit{(Z- procedure);}}$$

b) ***T***-statistic :

$$T = \frac{\bar{X} - m}{S / \sqrt{n}};$$

$$T = \frac{\bar{X} - m_0}{S / \sqrt{n}} \sim t_{(n-1)} \qquad \textbf{\textit{(T- procedure).}}$$

Methods and Steps of Hypotheses Testing

1. Critical - Value Method.

Step 1. Compute the **Critical Value** and specify the **Critical Region** **or** **Rej. Region** for the test.

Step 2. Take a random sample $X_1, \ldots, X_n$ from the population and compute the value of the test statistic, called the **Observed Value,**

$$z_0 = Z(obz) = \frac{\bar{X} - m_0}{s / \sqrt{n}} \qquad \text{for } (\ \boldsymbol{Z}\text{- procedure})$$

$$t_0 = T(obz) = \frac{\bar{X} - m_0}{s / \sqrt{n}} \qquad \text{for } (\ \boldsymbol{T}\text{- procedure})$$

Critical - Value Method

Step 3. Test the hypothesis using the following Decision Rule.

Decision Rule based on Critical-values:

- **Reject** H_0 if the observed value of the test statistic

$$z_0 = Z(obz) = \frac{\bar{X} - m_0}{s \ / \sqrt{n}}$$ falls the RR.

- **Do not reject** H_0 if $z_0 = Z(obz)$ does not fall the RR.

2. *P* - Value Method

Definition: The ***P***-value, (or observed significance level) corresponding to an observed value of a test statistic, is the smallest significance level at which the null hypothesis H_0 should be rejected.

Step 1. Take a random sample $X_1, \dots, X_n$ and compute the **Observed Value**,

$$z_0 = Z(obz) = \frac{\bar{X} - m_0}{s / \sqrt{n}} \quad \text{for} \quad (\textbf{\textit{Z}}\text{- procedure})$$

$$t_0 = T(obz) = \frac{\bar{X} - m_0}{s / \sqrt{n}} \quad \text{for} \quad (\textbf{\textit{T}}\text{- procedure})$$

P - Value Method

Step 2. Calculate the ***P***-value corresponding to $z_0 = Z(obz)$, depending on the alternative hypothesis.

Step 3. **Decision Rule** **based on *P*-values:**

- **Reject** H_0 if $P - value \leq a$.

- **Do not reject** H_0 if $P - value > a$.

3. Confidence Intervals Method

Step 1. For a given significance level a construct a CI for m :

$$\bar{X} \pm z_{a/2} \cdot \frac{s}{\sqrt{n}} \quad (\text{or} \quad \bar{X} \pm z_{a/2} \cdot \frac{S}{\sqrt{n}} \ ) \quad \text{for } (\ \boldsymbol{Z}\text{- procedure}).$$

$$\bar{X} \pm z_{a/2} \cdot \frac{S}{\sqrt{n}} \qquad \text{for } (\ \boldsymbol{T}\text{- procedure}).$$

Step 2. Test the hypothesis using the following Decision Rule.

Decision Rule based on Confidence Intervals:

- Ø**Reject** H_0 if CI does not contain m_0.
- Ø**Do not reject** H_0 if CI contains m_0.

Method and Steps of Hypotheses Testing

- **Example 1.** In order to test the null hypothesis

$$H_0 : m = 45 \quad \text{versus} \quad H_1 : m \neq 45.$$

A random sample of size ***n***=40 is obtained from a population with **known** standard deviation $s = 8$.
The sample mean is $\bar{X} = 48.3$.

Test the hypothesis at the $a = .05$ level significance using

a) Critical Value method

b) *P*-value method.

c) CI Method.

Example1.-Solution

Solution.

(a) Critical Value method

Step 1.

Specify test statistic (TS) and its distribution under H_0.

Since ***n***=40 > 30 we can apply CLT. Hence as a TS we can consider ***Z***-statistic:

$$Z = \frac{\bar{X} - m}{s \ / \sqrt{n}} : N(0,1), \quad \text{if } H_0 \text{ is true, that is,}$$

$$Z = \frac{\bar{X} - 45}{s \ / \sqrt{n}} : N(0,1).$$

Solution-(a) Critical Value method

Step 2. Compute the observed value of TS: .

We have

$$\bar{X} = 48.3, n = 40, s = 8, m_0 = 45,$$

hence $z_0 = Z(obz) = \dfrac{\bar{X} - m_0}{s / \sqrt{n}} = \dfrac{48.3 - 45}{8 / \sqrt{n_0}} = 2.61.$

So, $z_0 = Z(obz) = 2.61.$

Step 3. Compute the Critical Values and set up RR.

Since we are performing a **two-tailed test**,

Critical Values and RR are:

$\pm z_{a/2}$ and $|Z| > z_{a/2}$ respectively for $a = .05, \pm z_{a/2} = \pm 1.96.$

Solution-(a) Critical Value method

Step 4. Decision.

Since $z_0 = Z(obz) = 2.61$ falls RR (2.61>1.96), we **reject** H_0.

Step 5. Conclusion.

There is sufficient evidence to reject the null hypothesis

$$H_0 : m = 45$$

at the $a = .05$ level of significance.

Example 1.-Solution (b)

(b) *P*-value method.

Step 1. Test Statistic:

$$Z = \frac{\bar{X} - m}{s / \sqrt{n}} : N(0,1), \quad \text{if} \quad H_0 \quad \text{is true.}$$

Step 2. The observed value of TS is:

$$z_0 = Z(obz) = 2.61.$$

Step 3. Compute the P-value corresponding to $z_0 = 2.61.$

Since we are performing a **two-tailed test**, for ***P*** -value (from Z-Table we have

$$p - value = 2 \cdot P(Z > | z_0 |) = 2 \cdot P(Z > 2.61)$$

$$= 2 \cdot P(Z < -2.61) = 2 \cdot [.0045] = .009.$$

Solution-(b) *P*-value method

Thus, *P*-value = .009.

Step 4. Decision.

Since given significance level is $\alpha = .05$, and the observed significance level is

P-value $= .009$, and $.009 < .05$ the decision is:

Reject the hypothesis H_0.

Example 1.-Solution (c)

(c) CI Method.

Step 1. Construct a 95% CI for μ :

$$\bar{X} \pm z_{a/2} \cdot \frac{s}{\sqrt{n}} = 48.3 \pm 1.96 \cdot \frac{8}{\sqrt{40}} = 48.3 \pm 2.48.$$

That is, the **95% CI** is: **(45.82, 50.78).**

Step 2. Decision.

Since the **95% CI** = **(45.82, 50.78)**, **does not contain** the value $m_0 = 45$, we **reject** the null hypothesis $H_0 : m = 45$.