Module 4: Probability Distribution and Sampling Theory

Sr. No.	Topics	Hours
4.1	Probability Distribution: Poisson and Normal distribution	7 hr
4.2	Sampling distribution, Test of Hypothesis, Level of Significance, Critical region, One-tailed, and two-tailed test, Degree of freedom.	
4.3	Students' t-distribution (Small sample). Test the significance of mean and Difference between the means of two samples. Chi-Square Test: Test of goodness of fit and independence of attributes, Contingency table.	
	Self-learning Topics: Test significance for Large samples, Estimate parameters of a population, Yate's Correction.	

Basic terminology and logic

Population

The populations we wish to study are almost always so large that we are unable to gather information from every case.

Eg. : If we are interested in the weights of students enrolled in the Engineering college

Population Size

The number of elements in the population is called the population size and is denoted by N.

Sample

A sample is a part of a population. - From the population, we select various elements on which we collect our data. This part of the population on which we collect data is called the sample.

Eg: Suppose that we are interested in studying the characteristics of the weights of the students enrolled in the college of engineering. If we randomly select 50 students among the students of the college of engineering and measure their weights, then the weights of these 50 students form our sample.

Sample Size

The number of elements in the sample is called the sample size and is denoted by n.

Sampling and Statistical Inference

There are several types of sampling techniques, some of which are:

1. Simple random sampling

If a sample of size ‘n’ is selected from a population ‘N’ in such a way that each element in the population has the same chance to be selected, the sample is called a simple random sample.

Eg. Suppose we want to know what percent of students at a large university work during the semester •then draw a sample of 500 from a list of all students (N=20,000) and write the name of each students in single paper and folded, mixed and now draw 500 of them. But its lengthy method so another way

Each student has a unique, 6 digit ID number that ranges from 000001 to 999999. Use a table of random numbers or a computer program to select 500 ID numbers with 6 digits each. Each time a randomly selected 6 digit number matches the ID of a student, that student is selected for the sample

2. Stratified Random Sampling:

In this type of sampling, the elements of the population are classified into several homogenous groups (strata). From each group, an independent simple random sample is drawn. The sample resulting from combining these samples is called a stratified random Sample.

Eg. Suppose we want to conduct a sample survey relating to lung cancer among smokers. Suppose the population of smokers is 1000 and that there are 300 pipe smokers, 500 cigarette smokers and 200 bidi smokers. Suppose we have to select sample of 250. i.e. $1/4^{th}$ of the size of population. We have therefore to select $1/4^{th}$ from each group/strata i.e. 75,125,and 50 from three strata respectively.

Sampling Distribution

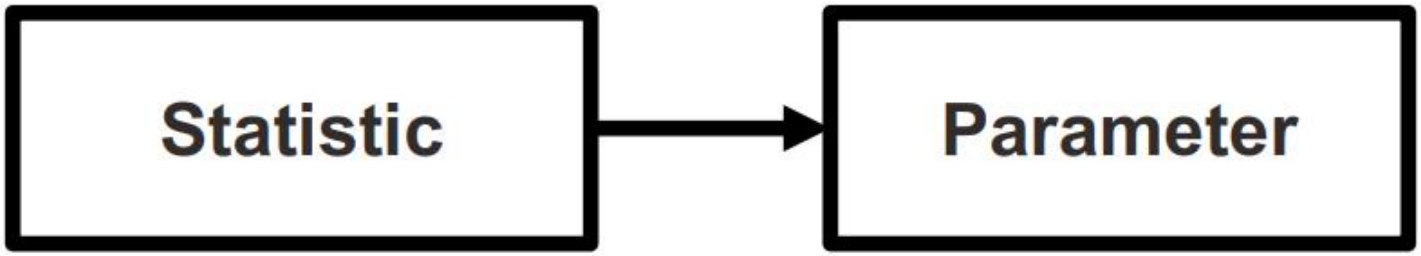
Sampling distribution is the link between sample and population. The probability distribution of a statistic is called the sampling distribution of that statistic. The sampling distribution of the statistic is used to make statistical inference about the unknown parameter.

How ???

- Statistics are used to estimate parameters

➢ Statistics are mathematical characteristics of samples

➢ Parameters are mathematical characteristics of populations



Distinctions Between Parameters and Statistics

	Parameters	**Statistics**
Source	Population	Sample
Notation	Greek (e.g.)	Roman (e.g. $\bar{}$)
Vary	No	Yes
Calculated	No	yes

Sampling Distribution

A sampling distribution acts as a frame of reference for statistical decision making.

- It is a *theoretical probability distribution* of the possible values of some sample statistic that would occur if we were to draw all possible samples **of a fixed size** from a given population.
- The sampling distribution allows us to determine whether, given the variability among all possible sample means, the one we observed is a common out come or a rare outcome.
- Imagine that each one of you asks a random sample of 10 people in this class what their height is.
- You each calculate the average height of your sample to get the sample mean.
- When you report back, would you expect all of your sample means to be the same?
- How much would you expect them to differ?

Sampling Distribution of the Mean

- Random samples rarely exactly represent the underlying population. We rely on sampling distributions to give us a better idea whether the sample we've observed represents a common or rare outcome.
- **Sampling distribution of the mean:** *probability distribution of means* for ALL possible random samples OF A GIVEN SIZE from some population, It describes the behavior of a sampling mean
- ALL possible samples is a lot!
- Example: All possible samples of size 5 from a class of 90 = **43949268**

Suppose that we select several random samples of size n=5.

	1st sample	2nd sample	3rd sample	…	Last sample
Sample values	28	31	14	.	17
	30	20	31	.	32
	34	31	25	.	29
	34	40	27	.	31
	17	28	32		30
Sample mean $\overline{x}$	28.4	29.9	25.8	…	27.8

Sampling Distribution of the Mean

The value of the sample mean . varies from random sample to another.

- The value of . is random and it depends on the random sample.
- The sample mean . is a random variable.
- The probability distribution of . is called the sampling distribution of the sample mean . .
- Questions:

➢ What is the sampling distribution of the sample mean . ?

➢ What is the mean of the sample mean . ?

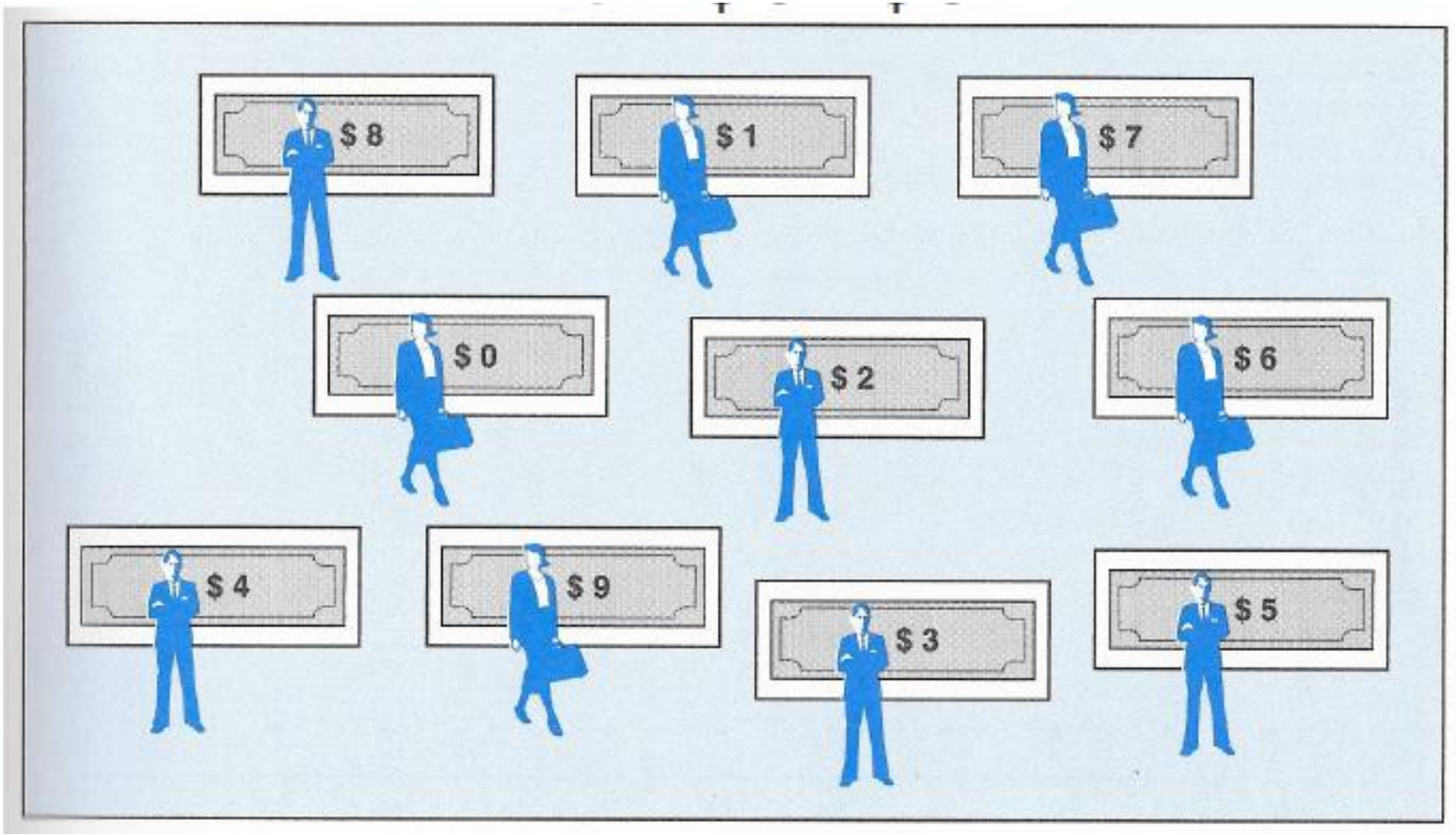
➢ What is the variance of the sample mean . ?

Sampling Distribution of the Mean

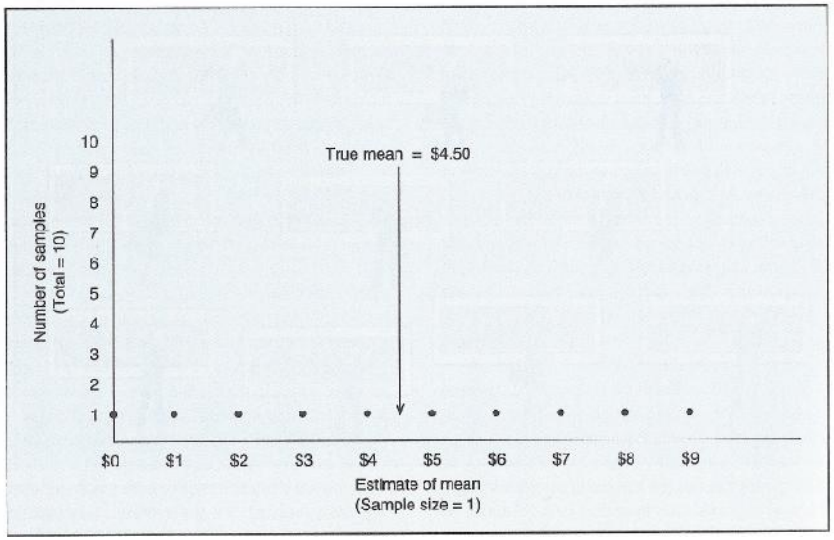
Type of Distribution	Mean	Standard Deviation
Sample	$\bar{x}$	s
Population	μ	σ
Sampling Distribution of the mean	$\mu_{\bar{x}}$ (mean of all sample means)	$\sigma_{\bar{x}}$ (standard error of the mean)

Sampling Distribution of the Mean

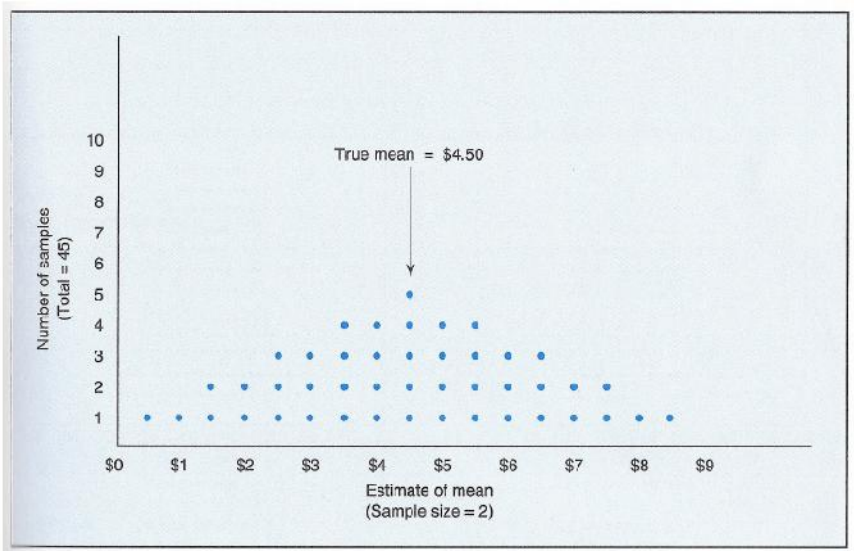
- A population of 10 people with $0–$9



Sampling Distribution of the Mean

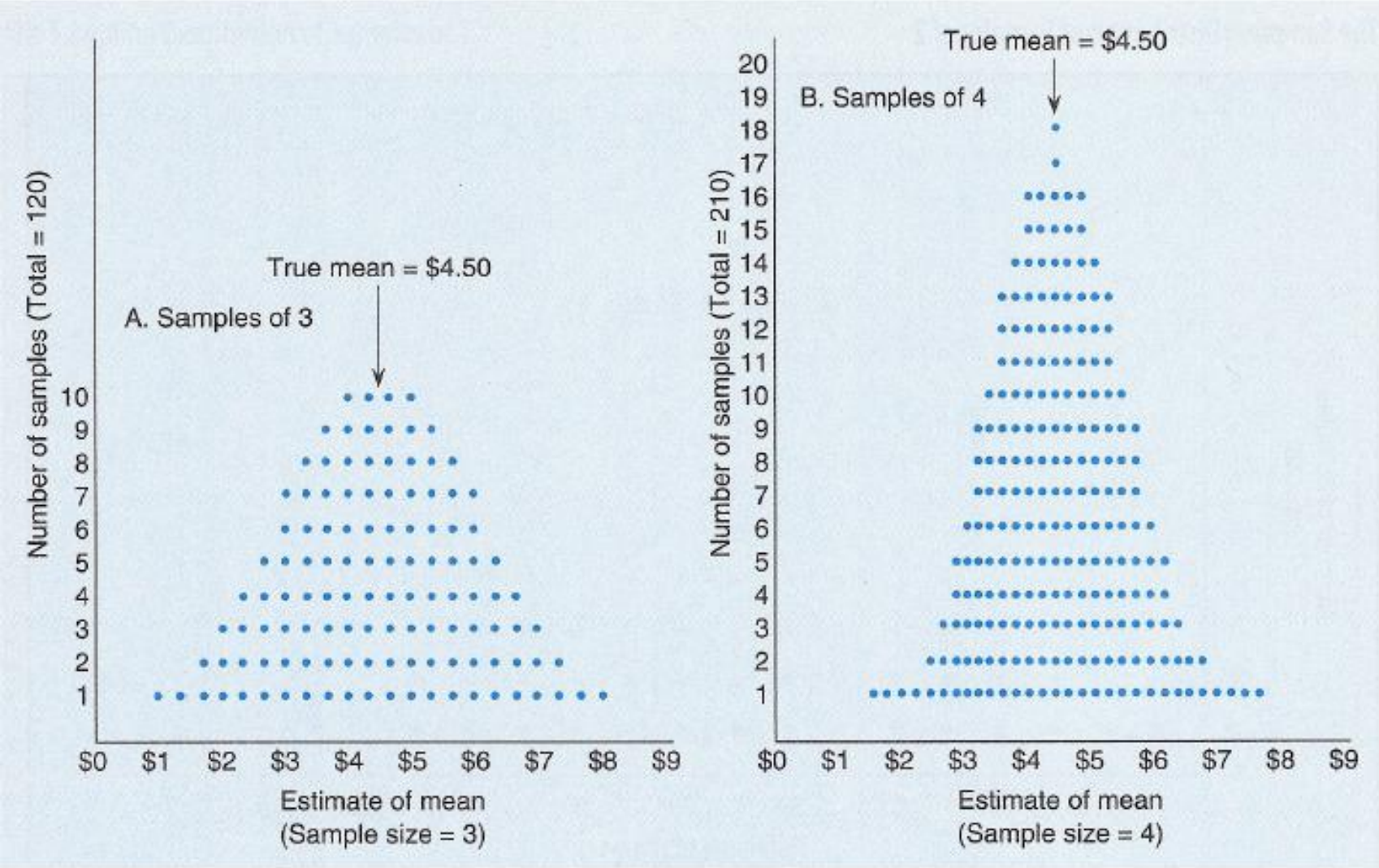


The sampling distribution (n=1)



The sampling distribution (*n*=2)

Sampling Distribution of the Mean



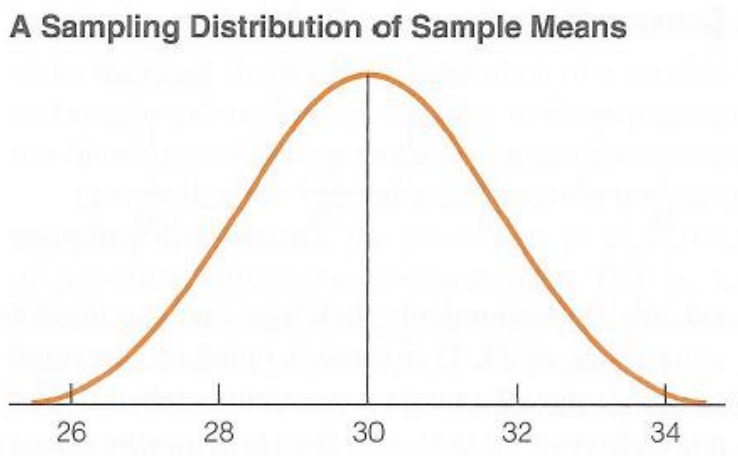
Properties of sampling distribution

- It has a mean ( ) equal to the population mean (μ) $\mu_{\bar{x}} = \mu$

- It has a standard deviation (standard error,)σ) equal to the population standard deviation (σ) divided by the square root of *n*. it is also called ***Standard Error*** of the mean measures the variability in the sampling distribution (roughly represents the average amount the sample means deviate from the mean of the sampling distribution)

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

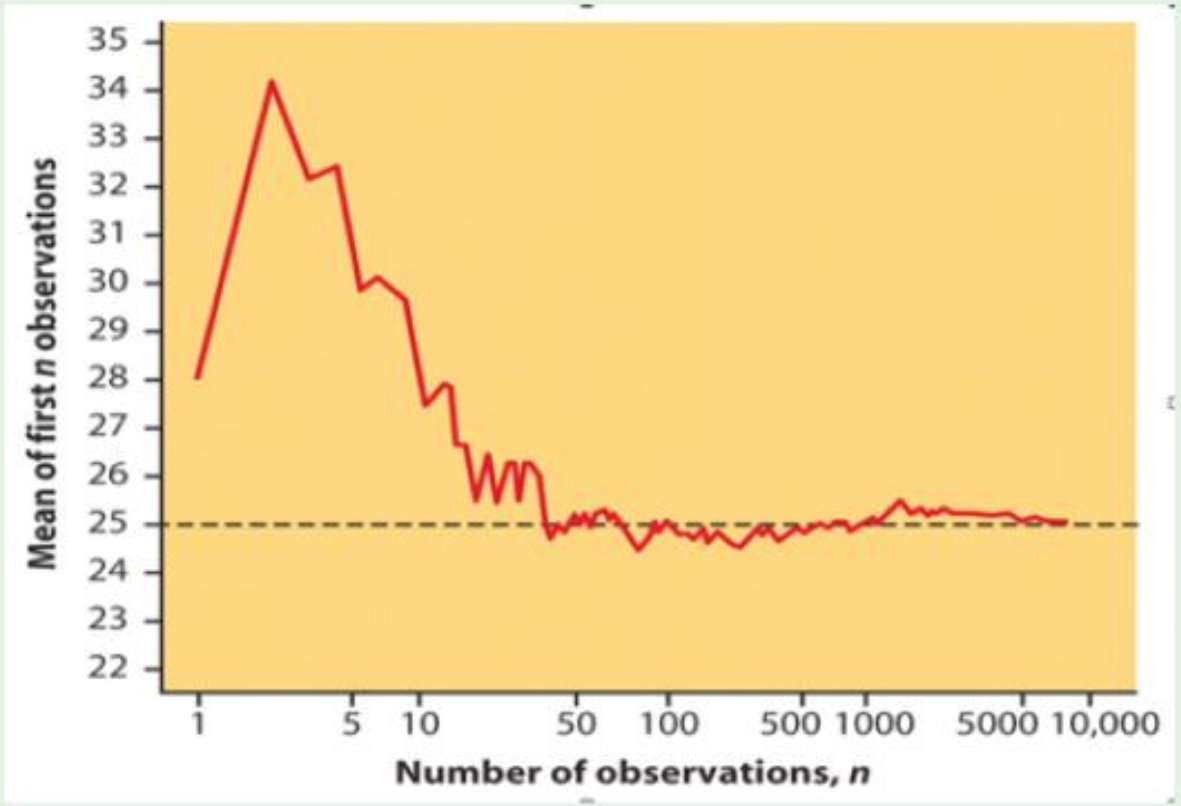
- It has a normal distribution.



Sampling Distribution of the Mean

Draw observations at random from any population with finite mean mu. As the number of observations drawn increases, the sample mean of the observed values $\bar{x}$ gets closer and closer to the mean μ of the population.

Example: How sample means approach the population mean ($\mu = 25$).



Central limit theorem

- If repeated random samples of size *N* are drawn from **any population** with mean μ and standard deviation σ
  - Then, as *N* becomes large, the sampling distribution of sample means will **approach normality** with...
  - A mean: $\mu_{\bar{X}} = \mu$
  - A standard error of the mean: $\sigma_{\bar{X}} = \sigma/\sqrt{N}$
- This is true for any variable, even those that are not normally distributed in the population
  - As sample size grows larger, the sampling distribution of sample means will become normal in shape

Central limit theorem

- The importance of the central limit theorem is that it removes the constraint of normality in the population
  – Applies to large samples (n≥30)

Parent Populations (can be of any shape, size, etc.)

Sampling Distribution of the Mean (approximates a normal curve, no matter the parent population)

- If the sample is small (N<30)
  – We must have information on the normality of the population before we can assume the sampling distribution is normal

Some Results of sampling distribution of .

Result (1): (mean & variance of $\overline{X}$)

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from any distribution with mean μ and variance σ^2; then:

1. The mean of $\overline{X}$ is: $\mu_{\overline{X}} = \mu$.
2. The variance of $\overline{X}$ is: $\sigma^2_{\overline{X}} = \dfrac{\sigma^2}{n}$.
3. The Standard deviation of $\overline{X}$ is call the standard error and is defined by: $\sigma_{\overline{X}} = \sqrt{\sigma^2_{\overline{X}}} = \dfrac{\sigma}{\sqrt{n}}$.

Result (2): (Sampling from normal population)

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from a normal population with mean μ and variance σ^2; that is Normal(μ, σ^2), then the sample mean has a normal distribution with mean μ and variance σ^2/n, that is:

1. $\overline{X} \sim$ Normal $\left(\mu, \dfrac{\sigma^2}{n}\right)$.
2. $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim$ Normal $(0,1)$.

We use this result when sampling from normal distribution with known variance

.

Some Results of sampling distribution of .

Result (3): (Central Limit Theorem: Sampling from Non-normal population)

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample of size n from non-normal population with mean μ and variance σ^2. If the sample size n is large $(n \geq 30)$, then the sample mean has approximately a normal distribution with mean μ and variance σ^2 / n, that is

1. $\overline{X} \approx \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$ (approximately)

2. $Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \approx \text{Normal}(0,1)$ (approximately)

Note: "≈ " means "approximately distributed.

We use this result when sampling from non-normal distribution with known variance and with large sample size.

Some Results of sampling distribution of .

Result (4): (used when σ^2 is unknown + normal distribution)

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from a normal distribution with mean μ and unknown variance σ^2; that is Normal(μ, σ^2), then the statistic:

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

has a t- distribution with $(n-1)$ degrees of freedom, where S is the sample standard deviation given by:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

We write:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

Notation: degrees of freedom = df = ν

Statistical Inferences:
(Estimation and Hypotheses Testing)

It is the procedure by which we reach a conclusion about a population on the basis of the information contained in a sample drawn from that population.

There are two main purposes of statistics;

- Descriptive Statistics: Organization & summarization of the data
- Statistical Inference: Answering research questions about some unknown population parameters.

(1) Estimation: Approximating (or estimating) the actual values of the unknown parameters:

- **Point Estimate:** A point estimate is single value used to estimate the corresponding population parameter.
- - **Interval Estimate (or Confidence Interval):** An interval estimate consists of two numerical values defining a range of values that most likely includes the parameter being estimated with a specified degree of confidence.

(2) Hypothesis Testing: Answering research questions about the unknown parameters of the population (confirming or denying some conjectures or statements about the unknown parameters).

Hypothesis Testing

Hypothesis

A statement or assumption about parameter is called hypothesis.

For eg. Average life time of table light is 1500 hrs.

Statistical Hypothesis

An assumption or statements about population parameters in numerical form is called statistical hypothesis

For eg. Height of indian soldiers is 6 feet. $\mu = 1500$ hrs

Types of Statistical Hypothesis

---Null Hypothesis

An assumption which is to be tested for possible rejection is called null hypothesis and it is denoted by .

---Alternative Hypothesis

An assumption which is opposite to null hypothesis is called altrnative hypothesis and it is denoted by .

Hypothesis Testing

Null hypothesis always involves equality i.e.

$$: =$$

$$: \geq$$

$$: \leq$$

Alternative hypothesis always involves inequality i.e.

$$: \neq$$

$$: <$$

$$: >$$

A procedure for deciding whether to accept or to reject a null hypothesis (or to reject or accept alternative hypothesis) is called Test of Hypothesis.

Type of Errors

- There are 4 possible situations in testing a statistical hypothesis:

		Condition of Null Hypothesis H_o (Nature/reality)	
		H_o is true	H_o is false
Possible Action (Decision)	Accepting H_o	Correct Decision	Type II error (β)
	Rejecting H_o	Type I error (α)	Correct Decision

- There are two types of Errors:
  - Type I error = Rejecting H_o when H_o is true
    P(Type I error) = P(Rejecting Ho | Ho is true) = α
  - Type II error = Accepting Ho when Ho is false
    P(Type II error) = P(Accepting Ho | Ho is false) = β

Level of Significance

- Probability of Type-I error is called level of significance. It is denoted by

  = ( )

  = (h)

- Usually expressed in %. Los usual values are = 5% = 1%
- = 5% means the probability of rejecting a true hypothesis is 0.05 or only 5 chances out of 100 that we reject hypothesis when it should be true.

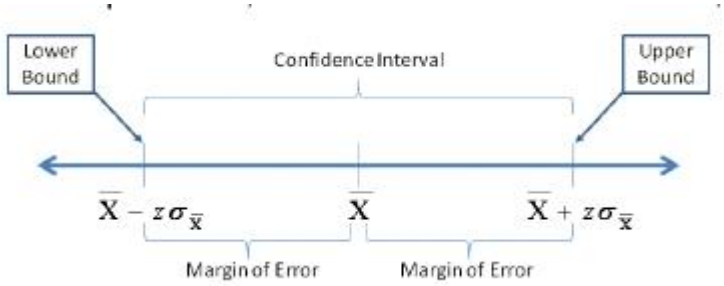
- **Note:**
- When a hypothesis is rejected it does not mean that the hypothesis is disproved. It only means that the sample value does not support the hypothesis. Same is true when we accept hypothesis.

- **Confidence limit**

- The limits with in which the hypothesis should lie with specified probability are called confidence limit
- Generally confidence limits set up with **5% or 1% LOS** or **95%,90% or 99%.**
- If sample values lies between the confidence limits the hypothesis is accepted otherwise rejected.

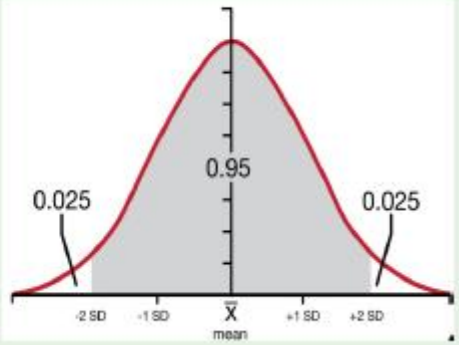
Confidence Interval

- Draw a sample it gives us a mean . bar that is our best guess at (for most samples . will be close to);
- . is a `point' estimate for the mean of the population.
- However, we can also give a range or interval estimate that takes into account the uncertainty involved in that point estimate.
- Condence interval equation is Limits = . ±
- where . bar is sample mean, z is value from normal curve, and is standard error of the mean



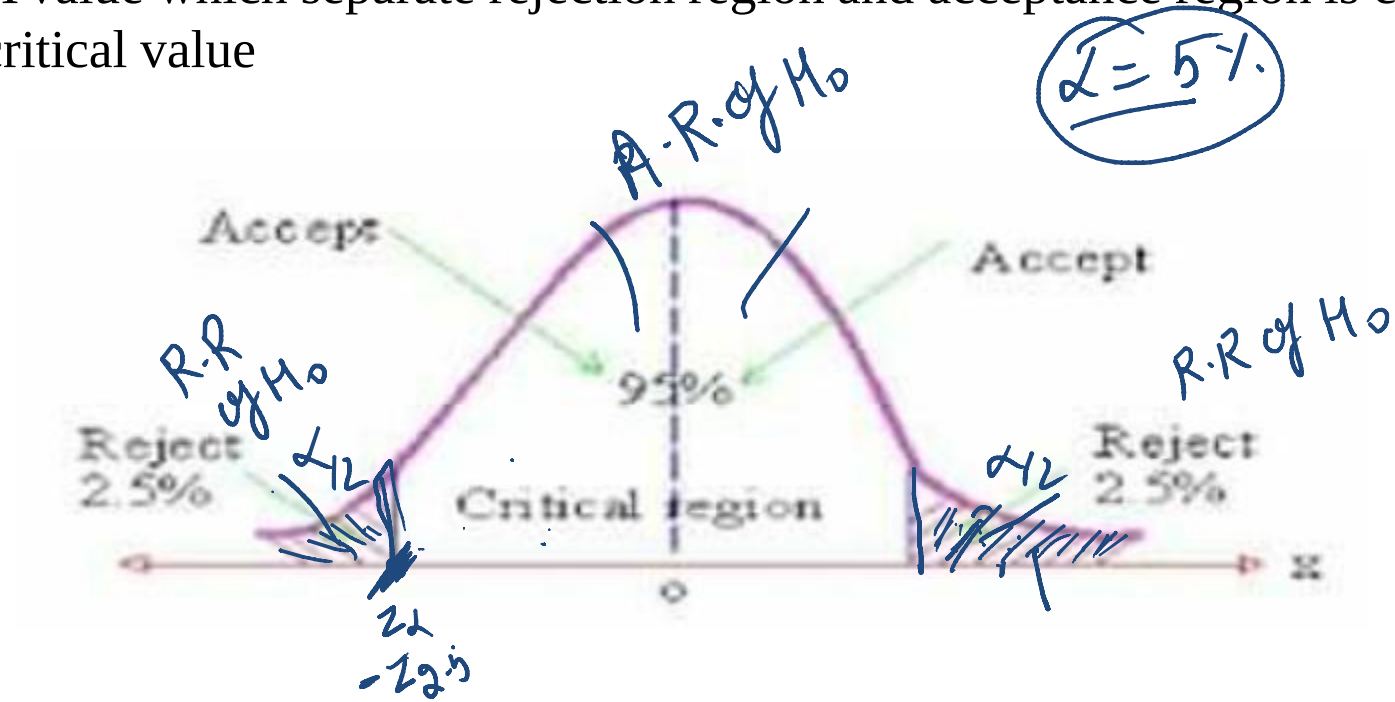
95% confidence interval

- Let's say we want a 95% confidence interval.
- Obtain the 'critical' z-score for $p = 0.025$
- If $p = 0.025$, then $z = 1.96$
- When the population standard deviation is not known, we use the t critical value (we will discuss it later on) instead $Limits = \bar{x} \pm t(s_{\bar{x}})$



Critical region

- **Critical or Rejection Region:**
- A region which leads us to reject . is called critical or rejection region.
- **Critical or Acceptance Region:**
- A region which leads us to accept . is called critical or acceptance region.
- **Critical Value:**
- A value which separate rejection region and acceptance region is called critical value



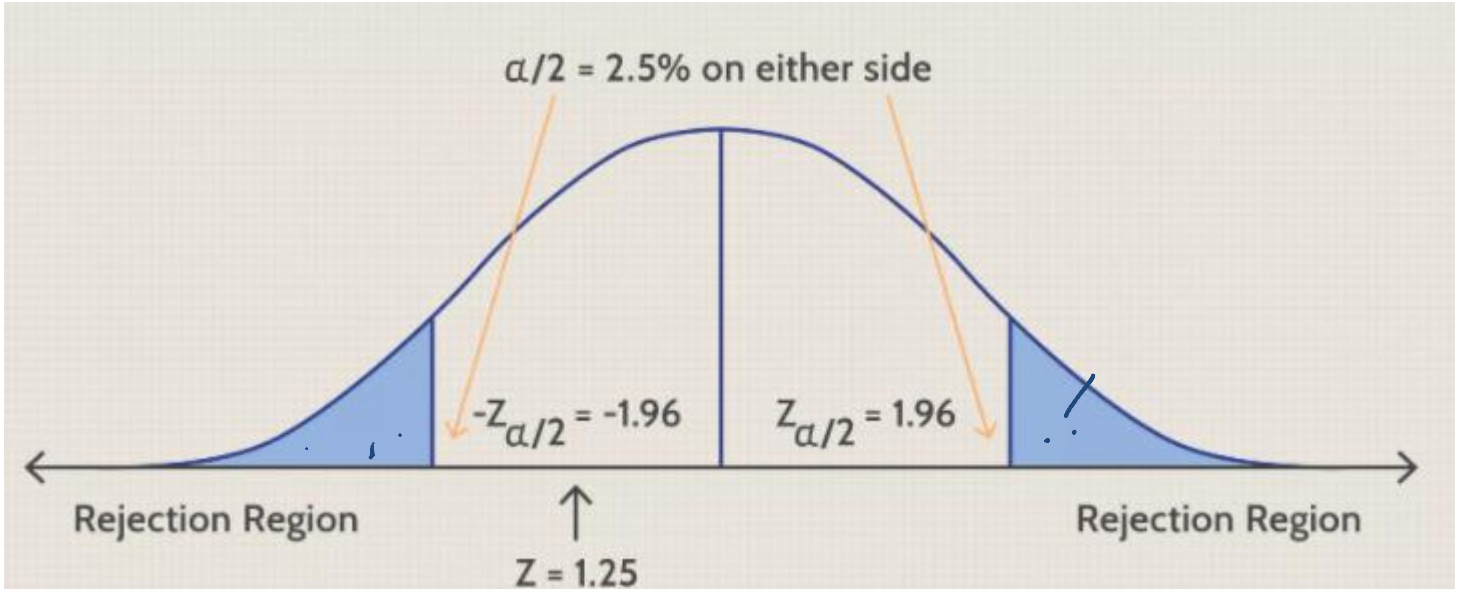
Type of Test

- Two type of Test

- **Two Tailed Test:**

When rejection region is on both sides of the distribution of population parameters. Such test is called two-tailed test.

: $=$ $\mu_1 = \mu_2$: $\neq$ $\mu_1 \neq \mu_2$ $\alpha = 5\%$.



Type of Test

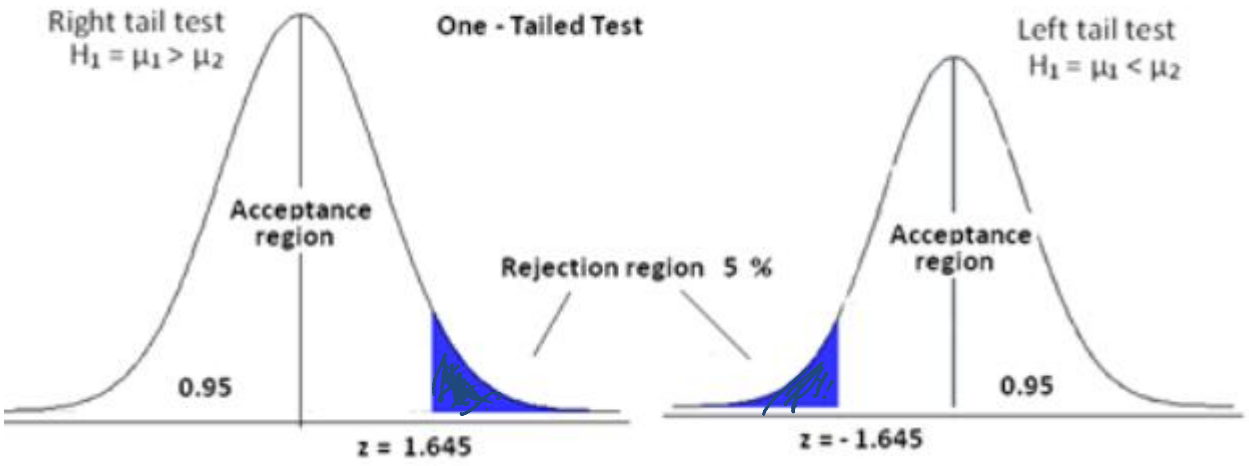
- Two type of Test

- **One Tail Test:**

When rejection region is on one sides of the distribution of population parameters. Such test is called one-tail test.

: ≥ : <

: ≤ : >



When to apply one tailed or two tailed test

Suppose that there are two population brands of bulbs, one manufactured by standard process (with mean _) and another manufactured by process (with mean _)

(a) If we want to test if the bulbs differ significantly.

(b) if the bulbs produced by process 2 have higher avg life than process 1.

(c) If the bulbs of new (2nd) process is inferior to that of std (1st) process.

(1) $H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

Two tailed

(2) $H_0: \mu_1 \geq \mu_2$

$H_1: \mu_1 < \mu_2$

one tail

(3) $H_0: \mu_1 \leq \mu_2$

$H_1: \mu_1 > \mu_2$

one tailed.

Small Sample Test

Student t-Distribution

- T-Distribution can be used when the variance of the population is unknown and the distribution is not normal
- Student-t, whose tail longer. That means the fact that sample mean with unknown population variance is inclined to be an extreme value. If you use normal distribution for hypothesis testing instead of t distribution, probability of error becomes bigger.
- Formula :

Suppose we have a simple random sample of size n drawn from a Normal population with mean and standard deviation . Let denote the sample mean and s, the sample standard deviation. Then the quantity

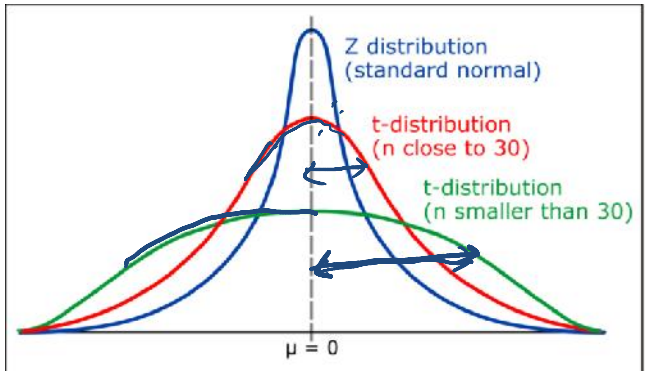
$$= \frac{}{\sqrt{}}$$ where $$= \frac{\sum_{i=1}(\quad\quad)}{}$$

or $$= \frac{}{\sqrt{}}$$ where s is given

- has a t distribution with n-1 **degrees of freedom**.
- Note that there is a different t distribution for each sample size,

Properties of Student t-Distribution

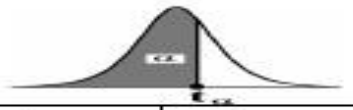
- The graph for the Student's t-distribution is similar to the standard normal curve.
- The mean for the Student's t-distribution is zero and the distribution is symmetric about zero. The variance is greater than one, but approaches one from above as the sample size increases.
- The Student's t-distribution has more probability in its tails than the standard normal distribution because the spread of the t-distribution is greater than the spread of the standard normal. So the graph of the Student's t-distribution will be thicker in the tails and shorter in the center than the graph of the standard normal distribution.
- The exact shape of the Student's t-distribution depends on the degrees of freedom. As the degrees of freedom increases, the graph of Student's t-distribution becomes more like the graph of the standard normal distribution. **For n > 30, the differences are negligible.**



T-Distribution

$t_{0.05} \rightarrow t_{0.95}$

Critical Values of the t-distribution (t_α)



v=df	$t_{0.90}$	$t_{0.95}$	$t_{0.975}$	$t_{0.99}$	$t_{0.995}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
35	1.3062	1.6896	2.0301	2.4377	2.7238
40	1.3030	1.6840	2.0210	2.4230	2.7040
45	1.3006	1.6794	2.0141	2.4121	2.6896
50	1.2987	1.6759	2.0086	2.4033	2.6778
60	1.2958	1.6706	2.0003	2.3901	2.6603
70	1.2938	1.6669	1.9944	2.3808	2.6479
80	1.2922	1.6641	1.9901	2.3739	2.6387
90	1.2910	1.6620	1.9867	2.3685	2.6316

$\alpha = 1\%$

$n = 10$

$t_\alpha (n-1)$

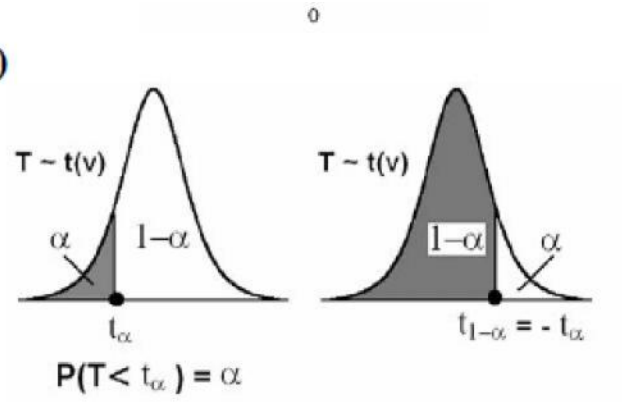
$t_{0.01}(9)$

$\downarrow$

$t_{0.99}$

T-Distribution

Notation: (t_α)



- t_α = The t-value under which we find an area equal to α
  = The t-value that leaves an area of α to the left.
- The value t_α satisfies: $P(T < t_\alpha) = \alpha$.
- Since the curve of the pdf of $T \sim t(\nu)$ is symmetric about 0, we have

$$t_{1-\alpha} = -t_\alpha$$

For example: $t_{0.35} = -t_{1-0.35} = -t_{0.65}$

$t_{0.82} = -t_{1-0.86} = -t_{0.14}$

- Values of t_α are tabulated in a special table for several values of α and several values of degrees of freedom.

Examples: t-Distribution

Example:

Find the t-value with $\nu=14$ (df) that leaves an area of:

(a) 0.95 to the left.

(b) 0.95 to the right.

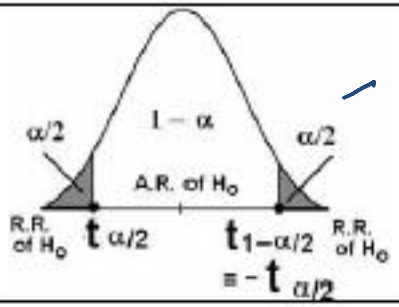
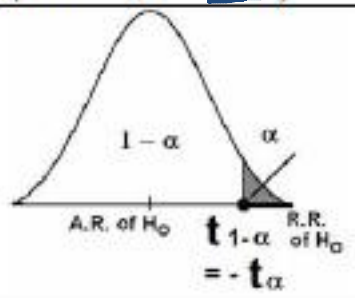
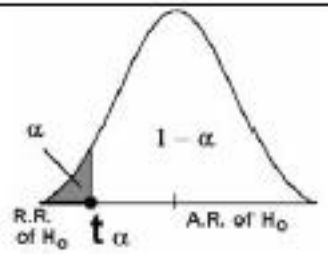
Example:

For $\nu = 10$ degrees of freedom (df), find $t_{0.93}$ and $t_{0.07}$.

The Procedure for hypotheses testing about the mean (μ)

Let be a given known value and Variance is unknown.

Test Procedures:

Hypotheses (1)	$H_o: \mu = \mu_o$ $H_A: \mu \neq \mu_o$	$H_o: \mu \leq \mu_o$ $H_A: \mu > \mu_o$	$H_o: \mu \geq \mu_o$ $H_A: \mu < \mu_o$
Test Statistic (T.S.) (2)	Calculate the value of: $t = \dfrac{\bar{X} - \mu_o}{S/\sqrt{n-1}} \sim t(n-1)$ $(df = \nu = n-1)$		
R.R. & A.R. of H_o (3)			
Critical value (s) (4)	$t_{\alpha/2}$ and $-t_{\alpha/2}$	$t_{1-\alpha} = -t_\alpha$	t_α
Decision (5)	We reject H_o (and accept H_A) at the significance level α if:		
	$t < t_{\alpha/2}$ or $t > t_{1-\alpha/2} = -t_{\alpha/2}$ Two-Sided Test	$t > t_{1-\alpha} = -t_\alpha$ One-Sided Test	$t < t_\alpha$ One-Sided Test

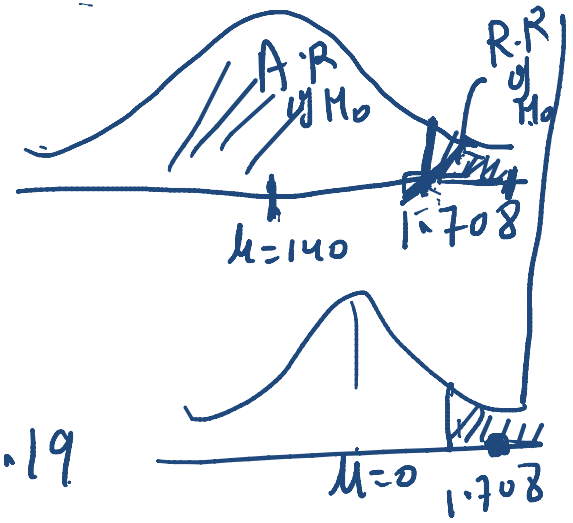
Test the significance of mean

$\alpha = 5\%$

Qu.1 A soap manufacturing company was distributing a particular brand of soap through a large number of retail shops. Before a heavy advertisement campaign the mean sale per week per shop was 140 dozens. After the campaign, a sample of 26 shops was taken and the mean sales was found to be 147 dozens with standard deviation 16. Can you consider the advertisement effective?

(1) $H_0 :- \mu \leq 140$

$H_1 :- \mu > 140$



$n = 26$

$\bar{x} = 147$ dozens

$S = 16$

Mean sale of Soap $= \mu$

(2) $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n-1}}$

$t = \dfrac{147 - 140}{16/\sqrt{25}} = \dfrac{35}{16} = 2.19$

|t|

(3) $t_\alpha = t_{0.05} \rightarrow$ dof $= n - 1 = 26 - 1 = 25 = 1.708$

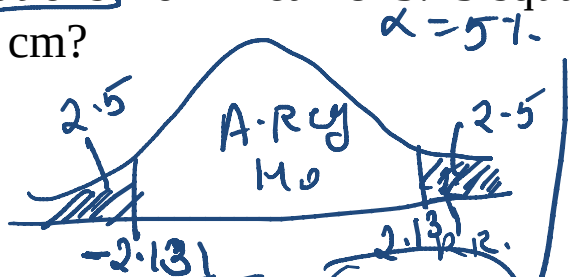
(4) Decision :- $2.19 > 1.708 \Rightarrow |t| > t_\alpha$ at $(n-1)$ H_0 is rejected

Avg sale quite than 140,

Test the significance of mean

Qu.2 A random sample of size 16 from normal population showed a mean of 103.75 cm. and sum of squares of deviations from mean 843.75 square cm. Can we say that population has a mean of 108.75 cm?

(1) $H_0: \mu = 108.75$ cm.
$H_1: \mu \neq 108.75$ cm.



(2) $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n-1}} = \dfrac{103.75 - 108.75}{52.73/\sqrt{15}} = -2.67$

$n = 16$
$\bar{x} = 103.75$
$s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$
$s^2 = \dfrac{843.75}{16} = 52.73$

(3) $t_\alpha \to t_{2.5} \to (15) \to 2.131$

$2.13 < 2.67$

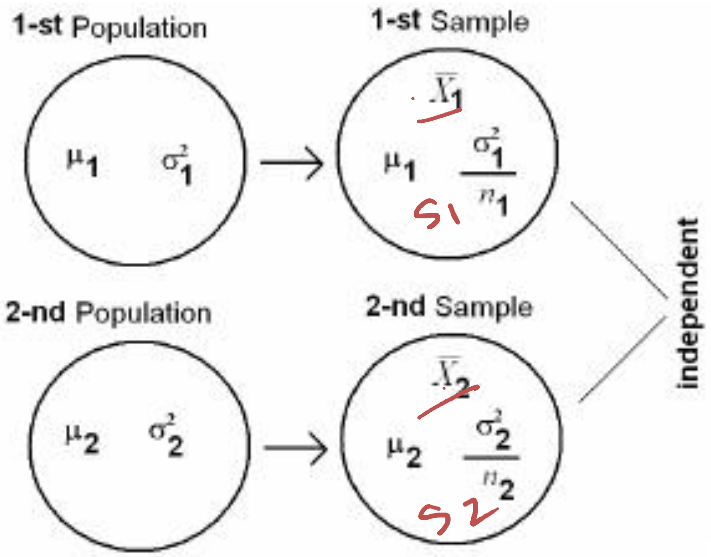
(5) H_0 is rejected, $\therefore$ we conclude that $\mu \neq 108.75$

Test the significance of mean

Qu.3 Problem 27.11

Qu.4 Problem 27.12

Test of Difference between the means of two samples.



Hypotheses: We choose one of the following situations:

(i) Ho: μ1 = μ2 against H1: μ1 ≠ μ2

(ii) Ho: μ1 ≥ μ2 against H1: μ1 < μ2

(iii) Ho: μ1 ≤ μ2 against H1: μ1 > μ2 or equivalently,

(i) Ho: μ1-μ2 = 0 against H1: μ1 - μ2 ≠ 0

(ii) Ho: μ1-μ2 ≥ 0 against H1: μ1 - μ2 < 0

(iii) Ho: μ1-μ2 ≤ 0 against H1: μ1 - μ2 > 0

For normal populations, and if σ_1^2 and σ_2^2 are unknown but equal ($\sigma_1^2=\sigma_2^2=\sigma^2$), then the test statistic is:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} \sim t(n_1+n_2-2)$$

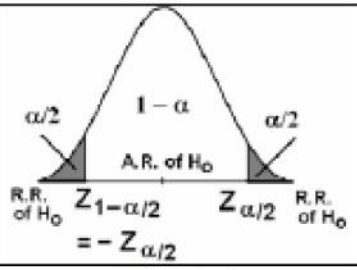
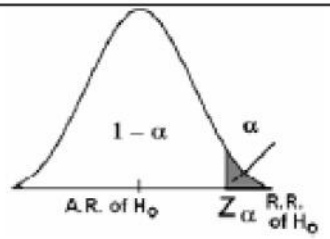
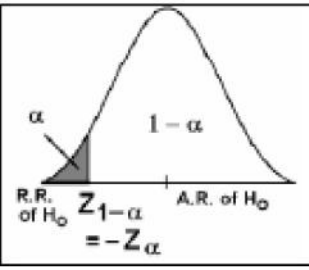
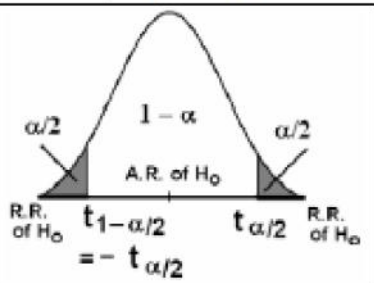
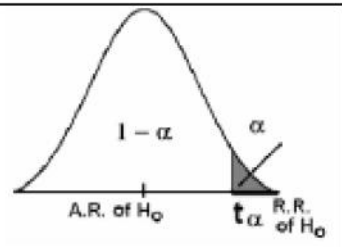
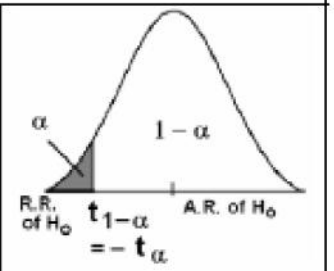
where the pooled estimate of σ^2 is

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

and the degrees of freedom of S_p^2 is df= $\nu=n_1+n_2-2$.

Test of Difference between the means of two samples.

Summary of Testing Procedure:

Hypotheses	$H_o: \mu_1 - \mu_2 = 0$ $H_A: \mu_1 - \mu_2 \neq 0$	$H_o: \mu_1 - \mu_2 \leq 0$ $H_A: \mu_1 - \mu_2 > 0$	$H_o: \mu_1 - \mu_2 \geq 0$ $H_A: \mu_1 - \mu_2 < 0$
R.R. and A.R. of H_o (For the First Case)			
Test Statistic For the Second Case:	$T = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{S_p^2}{n_1} + \dfrac{S_p^2}{n_2}}} \sim t(n_1+n_2-2)$ {if $\sigma_1^2=\sigma_2^2=\sigma^2$ is unknown}		
R.R. and A.R. of H_o (For the Second Case)			
Decision:	Reject H_o (and accept H_A) at the significance level α if:		
	T.S. ∈ R.R. Two-Sided Test	T.S. ∈ R.R. One-Sided Test	T.S. ∈ R.R. One-Sided Test

Test of Difference between the means of two samples.

Sample of two types of electric light bulbs were tested for length of life and following data obtained .

	Type-I	Type-II
Sample no.	$= 8$	$= 8$
Sample mean	$= 1234$	$= 1036$
Std deviation	$= 36$	$= 40$

μ_1 type-I, μ_2 type-II

μ_1, μ_2

μ_1 = mean length of bulbs of type I

Test at 5% LOS whether the difference in the sample means is significant.

(1) H_0 :- $\mu_1 = \mu_2$

H_1 :- $\mu_1 - \mu_2 \neq 0 \Rightarrow \mu_1 \neq \mu_2$

(2) $$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} = \frac{1234 - 1036}{\sqrt{\frac{1448}{8} + \frac{1448}{8}}} = 10.406$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

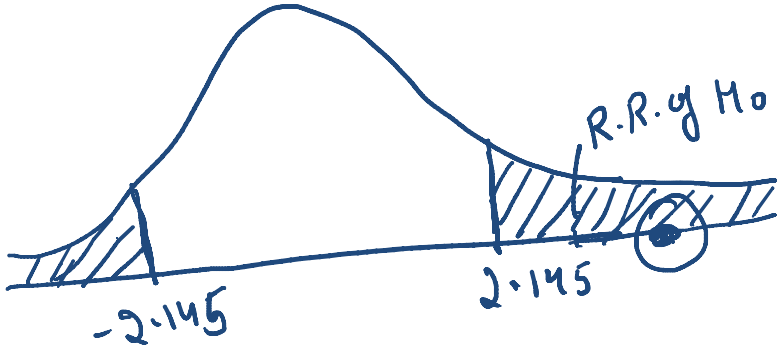
$$S_p^2 = \frac{7.36^2 + 7.40^2}{14} = 1448$$

$T_{cal} = 10.406$

$T_{tab} = 2.145$

$\alpha = 5\%$,

$dof = n_1 + n_2 - 2 = 8 + 8 - 2 = 14$



$T_{cal} \in$ RR of H_0

$T_{cal} > T_{tab}$

H_0 is rejected

$\therefore$ H_1 is accepted

There is a significant difference in the mean length of life of both type of bulbs.

Paired t-test for difference of mean.

- If n1=n2=n (sample size are same)
- And Two samples are not independent (or correlated).

⇒ Samples observations are paired together.

Examples of related populations are:

1. Height of the father and height of his son.
2. Mark of the student in MATH and his mark in STAT.
3. Pulse rate of the patient before and after the medical treatment.
4. Hemoglobin level of the patient before and after the medical treatment.

Example: (effectiveness of a diet program) Suppose that we are interested in studying the effectiveness of a certain diet program. Let the random variables X and Y are as follows:

X = the weight of the individual before the diet program

Y= the weight of the same individual after the diet program

Populations:

1-st population (X): weights before a diet program mean = μ_1

2-nd population (Y): weights after the diet program mean = μ_2

Hypotheses:

Ho: the diet program has no effect on weight

H1: the diet program has an effect on weight Equivalently,

Ho: $\mu 1 = \mu 2$ H1: $\mu 1 \neq \mu 2$

Equivalently, Ho: $\mu 1 - \mu 2 = 0$ H1: $\mu 1 - \mu 2 \neq 0$

Equivalently, Ho: $\mu D = 0$ H1: $\mu D \neq 0$ where: $\mu D = \mu 1 - \mu 2$

Test Statistic:

$$t = \frac{\overline{D}}{S_D / \sqrt{n}} \sim t(n-1)$$

Where

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} = \frac{D_1 + D_2 + \ldots + D_n}{n}$$

Di = Xi - Yi (i=1, 2, …, n)

$$S_D^2 = \frac{\sum_{i=1}^{n} (D_i - \overline{D})^2}{n-1} = \frac{(D_1 - \overline{D})^2 + (D_2 - \overline{D})^2 + \cdots + (D_n - \overline{D})^2}{n-1}$$

2.Q before training x_i : 110 120 123 132 125

2.Q after training y_i : 120 118 125 136 121

D_i : − − − − − $\Sigma D_i = ?$

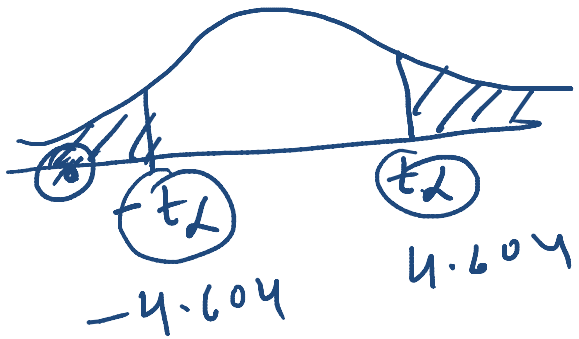
Test there is any Change in 2-Q after the training at 1%. Los

(1) $H_0 :- \mu_1 = \mu_2$
$H_1 :- \mu_1 \neq \mu_2$

$T_{tab} = \alpha, \frac{n-1}{1\%, \; 5-1=4}$

(2) $\bar{D} = \frac{\Sigma D_i}{n} = -2$

$S_D^2 = \frac{\Sigma (D_i - \bar{D})^2}{n-1} = S_D = 5.47$



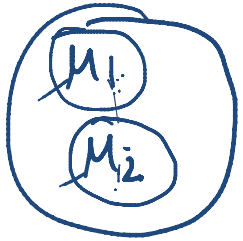
$T = \frac{\bar{D}}{S_D/\sqrt{n}} = \frac{-2}{5.47/\sqrt{5}} = -0.81$

H_0 rejected, & H_1 accepted

$T_{cal} = -0.81$

$T_{cal} \in$ R.R. of H_0

Individual (i)	1	2	3	4	5	6	7	8	9	10
Weight before (X_i)	86.6	80.2	91.5	80.6	82.3	81.9	88.4	85.3	83.1	82.1
Weight after (Y_i)	79.7	85.9	81.7	82.5	77.9	85.8	81.3	74.7	68.3	69.7



Does these data provide sufficient evidence to allow us to conclude that the diet program is effective? Use α=0.05 and assume that the populations are normal.

Solution:

μ1 = the mean of weights before the diet program

μ2 = the mean of weights after the diet program Hypotheses:

Ho: μ1 = μ2 (Ho: the diet program is not effective)

H1: μ1 ≠ μ2 (HA: the diet program is effective)

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} = \frac{54.5}{10} = 5.45$$

$$S_D^2 = \frac{\sum_{i=1}^{n} (D_i - \overline{D})^2}{n-1} = \frac{(6.9-5.45)^2 + \ldots + (12.4-5.45)^2}{10-1} = 50.3283$$

$$S_D = \sqrt{S_D^2} = \sqrt{50.3283} = 7.09$$

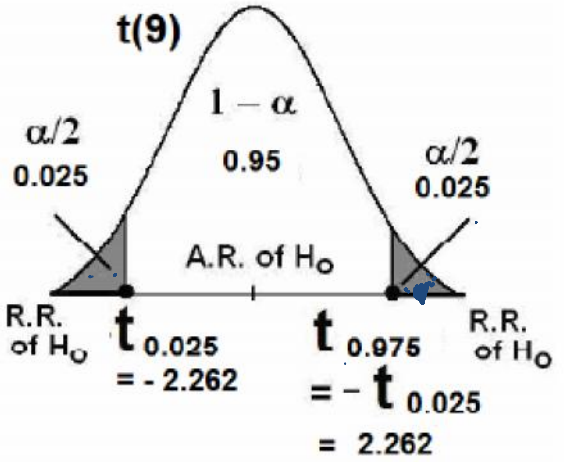
i	X_i	Y_i	$D_i = X_i - Y_i$
1	86.6	79.7	6.9
2	80.2	85.9	-5.7
3	91.5	81.7	9.8
4	80.6	82.5	-1.9
5	82.3	77.9	4.4
6	81.9	85.8	-3.9
7	88.4	81.3	7.1
8	85.3	74.7	10.6
9	83.1	68.3	14.8
10	82.1	69.7	12.4
sum	$\sum X$=842	$\sum Y$=787.5	$\sum D$=54.5

Test Statistic:

$$t = \frac{\bar{D}}{S_D / \sqrt{n}} = \frac{5.45}{7.09 / \sqrt{10}} = 2.431 = t_{cal.}$$

$t_{tab} = \alpha, (n)-1$ 5%. 10-1 = 9

Degrees of freedom: df= ν= n-1 = 10-1=9
Significance level: α=0.05
Rejection Region of Ho: Critical values: t 0.025 = - 2.262
2.262 Critical Region: t < - 2.262 or t > 2.262



Decision:
Since t= 2.43 ∈R.R., i.e., t=2.43 > 2.262,
we reject: Ho: μ1 = μ2 (the diet program is not effective) and
we accept: H1: μ1 ≠ μ2 (the diet program is effective)
Consequently, we conclude that the diet program is effective at α=0.05.

Chi-Square( )

Chi-Square( )

- Non-Parmetric Test
- A measure of the difference between the observed and expected frequencies of the outcomes of a set of events or variables.
- $\chi 2$ depends on the size of the difference between actual and observed values, the degrees of freedom, and the samples size.
- Can be used to test whether two variables are related or independent from one another or to test the goodness-of-fit between an observed distribution and a theoretical distribution of frequencies.
- Can be applied to only categorical data type e.g containing groups/categories of gender, marital status, inoculated, age group etc.
- Data to be presented in tabular form.

Chi-Square Distribution Table

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892

Hypothesis Testing Using Chi-Square( )

1. Set up Null Hypothesis (No significant difference between the observed and expected values/No association between the mentioned attributes) and Alternate Hypothesis.
2. Identify the degrees of freedom, n-1 OR(r-1)(c-1), where r = no. of rows, c = no.of columns.
3. Test statistc $= \sum \frac{(\quad)}{\quad} \rightarrow \chi^2_{cal}$ $<$ $\chi^2_{\alpha, dof}$ tab

- where:*c*=Degrees of freedom; =Observed value(s); =Expected value(s)

4. Determine the critical value of from the table.
5. Compare the calculated and tabulated results.
6. Make Decision as is rejected or is not rejected on the basis if the calculated test statistic value falls in rejection region or acceptance region respectively.

Problems-1Chi square test for goodness of fit

- A die is thrown 132 times with the following results:

Number turned up:	1	2	3	4	5	6
Frequency: O_i	16	20	25	14	29	28

Test the hypothesis that the die is unbiased.

$\frac{132}{6C_1} = \frac{132}{6} = 22$

(1) H_0 :- die is unbiased

H_1 :- die is biased

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
16	22	-6	36	36/22
20	22	-2	4	4/22
25	22	3	9	9/22
14	22	-8	64	64/22
29	22	7	49	49/22
28	22	6	36	36/22

$\sum \frac{(O_i - E_i)^2}{E_i} = 9$

$\chi^2_{cal} = 9$

$\chi^2_{tab} = (\alpha, n-1)$

5%, 5

$= 11.070$

$\chi^2_{cal} < \chi^2_{tab}$

H_0 is accepted

die is unbiased

Problems-2 Chi square test for goodness of fit

- The theory predicts the proportion of beans in the four groups A,B,C and D should be 9:3:3:1. In an experiment among 1600 beans, the number in the four groups were 882,313, 287,118. Does the experimental results support theory?

(1) H_0 :- there is no difference.

H_1 : there is a difference

(2)

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
882	900			
313	300			
287	300			
118	100			
1600				$\sum \frac{(O_i - E_i)^2}{E_i} = 4.726$

$A \rightarrow \frac{9}{16} \times 1600 = 900$

$B \rightarrow \frac{3}{16} \times 1600 = 300$

$\chi^2_{cal} = 4.726$

$\chi^2_{tab} = (\alpha, n-1)$

$5\%, 3$

$= 7.815$

$\chi^2_{cal} < \chi^2_{tab} \therefore$ H_0 accepted

H_1 rejected

Problems-3 Chi square test Independence of Attributes

- A certain drug was administered to 456 males out of total 720 in a certain locality to test its efficacy against typhoid. The incidence of typhoid is shown below. Find out the effectiveness of the drug against the disease.

$\chi^2_{cal} > \chi^2_{tab}$ H_0 is rejected H_1 accepted : drug is effected

	Infection	No Infection	Total
Administered the drug	$\frac{336 \times 456}{720}$ 144	312 $\frac{456 \times 384}{720}$	456
Without Administered the drug	192 $\frac{264 \times 336}{720}$	72 $\frac{264 \times 384}{720}$	264
Total	336	384	720

① H_0:- No association b/w attributes

H_1:- There is associb b/w attr.

②

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
144	212.8			
312	243.2			
192	123.2			
72	140.8			

$\sum \frac{(O_i - E_i)^2}{E_i}$

$\chi^2_{cal} = 113.75$

$\chi^2_{tab} = \alpha_{1}(r-1)(c-1)$ $(2-1)(2-1)$

5% 1

3.84

Problems-4 Chi square test Independence of Attributes

An app provides ratings to three categories of restaurants under three categories. Can we conclude that the ratings are related to the size of the restaurant?

	Small	Medium	Large	Total
Good:	20	10	17	47
Okay:	11	8	8	27
Not Recommended:	10	7	9	26
Total:	41	25	34	100