Estimation and Test of Hypothesis

Bijay Lal Pradhan

Sampling

- The process of selecting a number of individuals for a study in such a way that the individuals represent the larger group from which they were selected

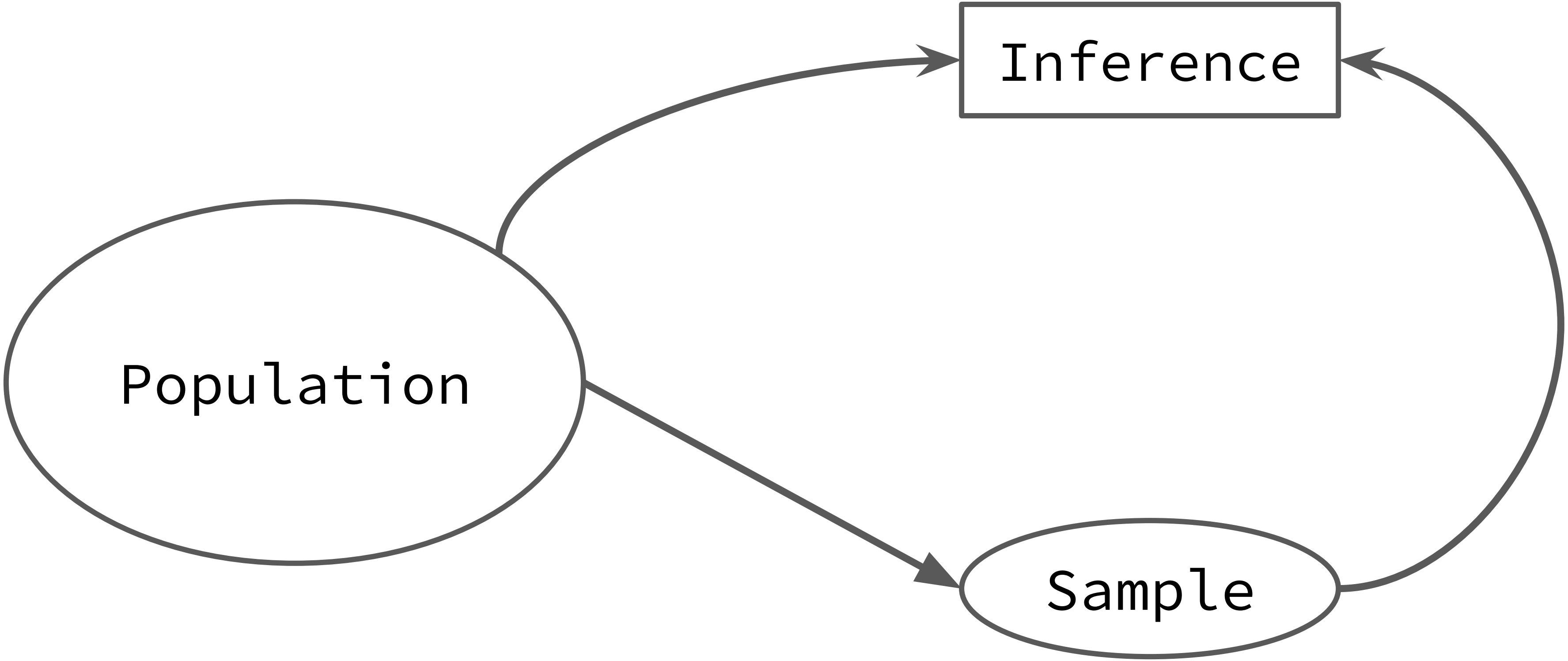
What is sample

the representatives selected for a study whose characteristics exemplify the larger group from which they were selected

Purpose

- To gather data about the population in order to make an inference that can be generalized to the population

The sampling process…



Advantage of sampling over census

- Reduce Cost
- Greater Speed
- Greater Scope
- Greater Accuracy
- Suitable for Uncountable population
- Suitable for destructive/ vanishable tests
- Hypothetical Population (coin toss)

Steps in Sampling

- State Objective of sampling
- Define population to be sampled
- Define data to be collected
- Define degree of precision required
- State Method of measurement
- Define Sampling frame
- Selection of proper sampling design
- Organization of field work

Types of Sampling

- Random sampling

1. Simple random sampling
   a. Lottery method
   b. Use of random number
   c. Use of software
2. Stratified sampling
3. Cluster sampling
4. Systematic sampling
5. Multistage sampling

- **Non Random sampling**

1. **Judgmental sampling**
2. **Convenient sampling**
3. **Quota sampling**
4. **Snow ball sampling**

Simple Random Sampling

the process of selecting a sample that allows individual in the defined population to have an equal and independent chance of being selected for the sample.

- **Simple random Sampling with Replacement (SRSWR)**
- **Simple random sampling without replacement (SRSWOR)**

Random Number Table

Draw a random sample without replacement of size 15 from a population of size 15 from population of size 500.

The following are 40 four digits' number from Tippet's random number table

2952	6641	3992	9792	7967	5911	3170	5624
4167	9524	1545	1396	7203	5356	1300	2963
2370	7483	3408	2762	3563	1089	6913	7691
0560	5246	0112	6107	6008	8126	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446

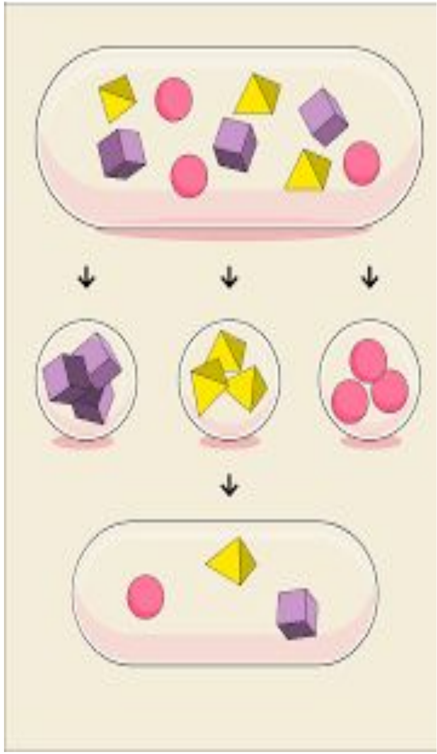
Now starting with first number and moving column wise the sample units are

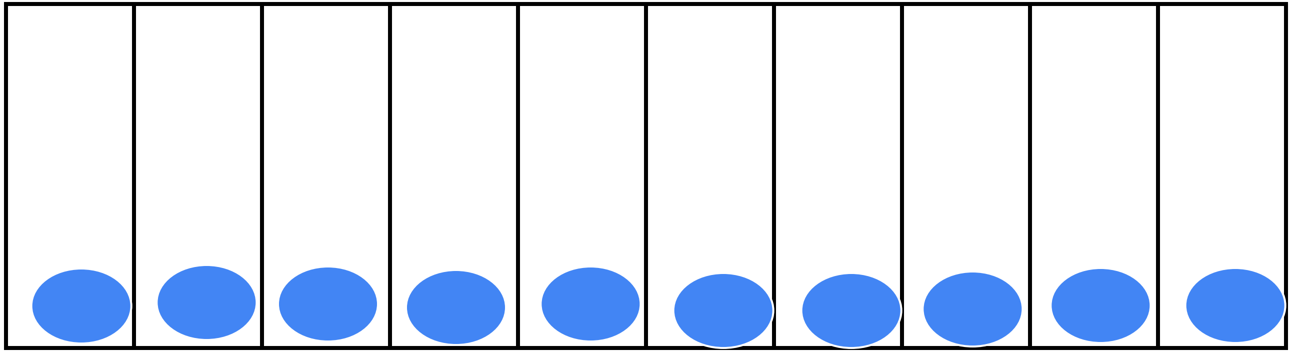
295, 416, 237, 056, 275, 266, 074, 052, 491, 413, 241, 460, 431, 408, 112.

If the population is heterogeneous then SRS may not give representative data

Stratified Random sampling

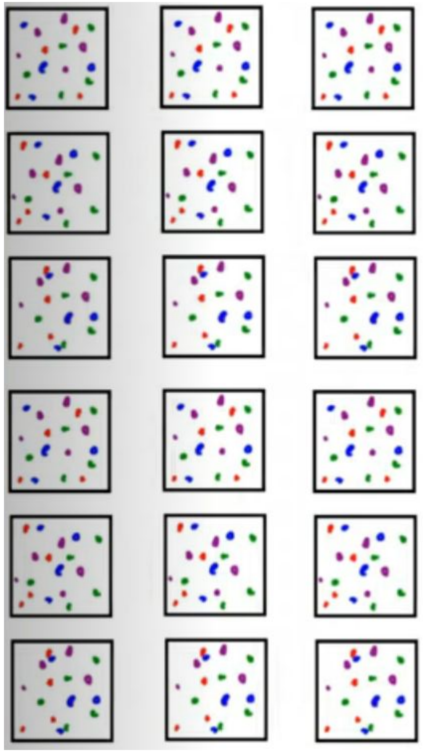
Each class is said to be strata





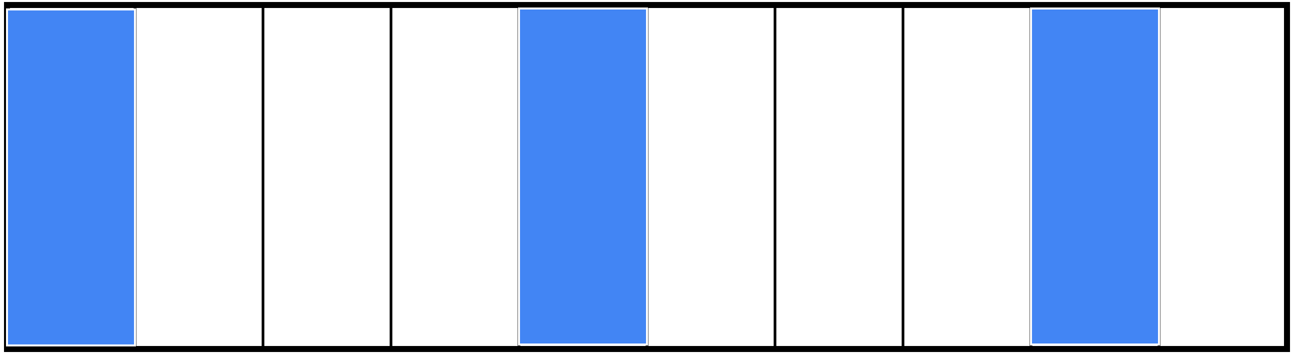
within class homogeneous;
between class heterogeneous

If the population is heterogeneous then SRS may not give representative data

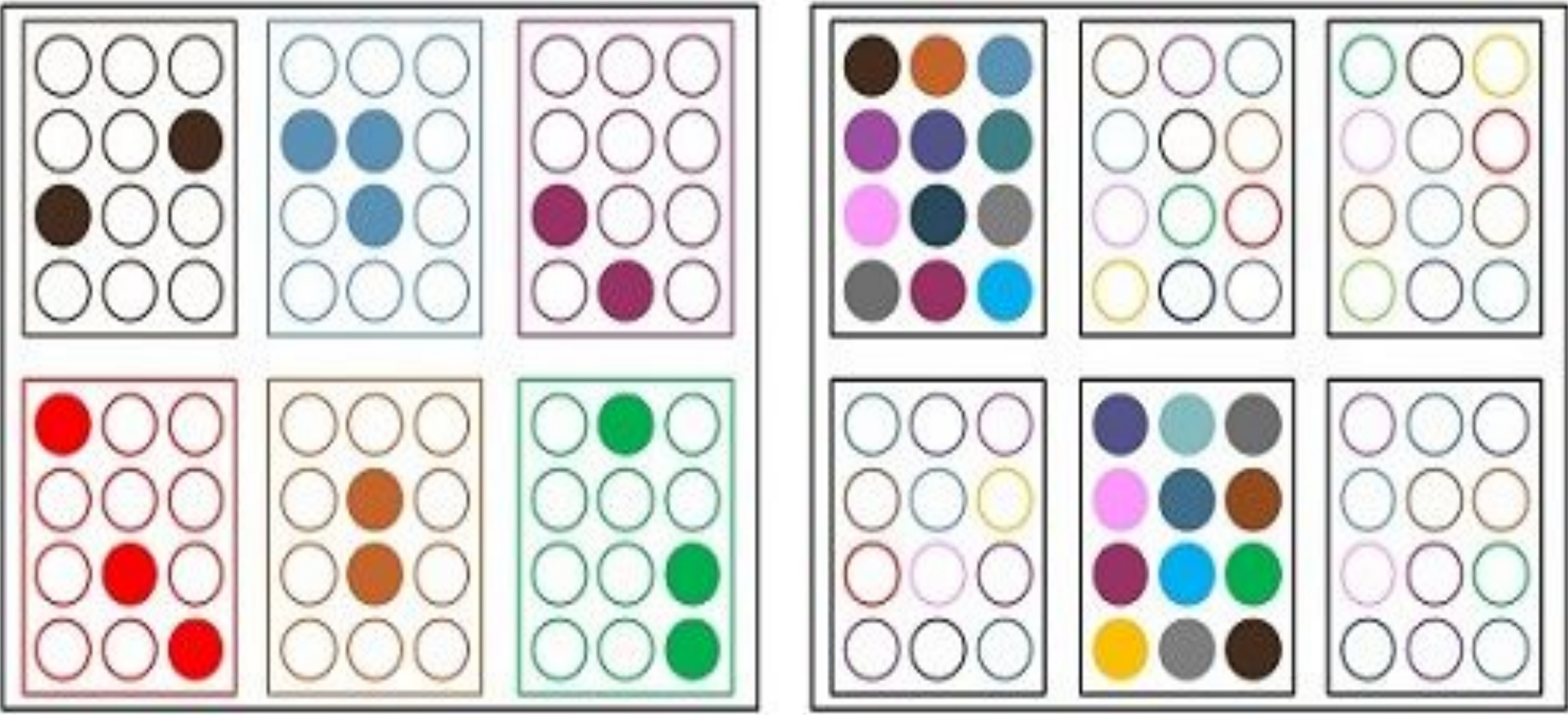


Cluster sampling

Each class is said to be cluster



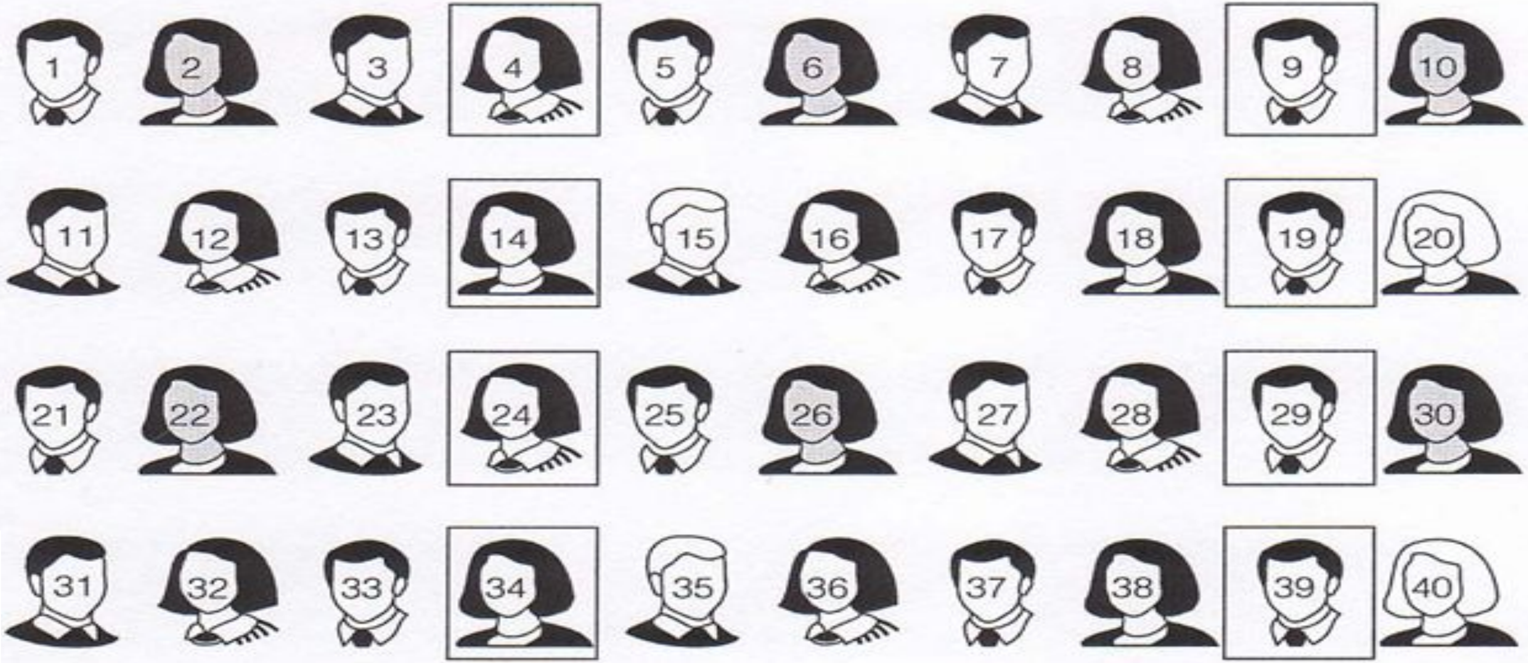
within class heterogeneous;
between class homogeneous

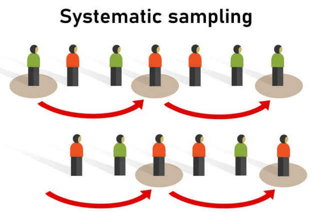


Stratified Sampling Vs Cluster Sampling

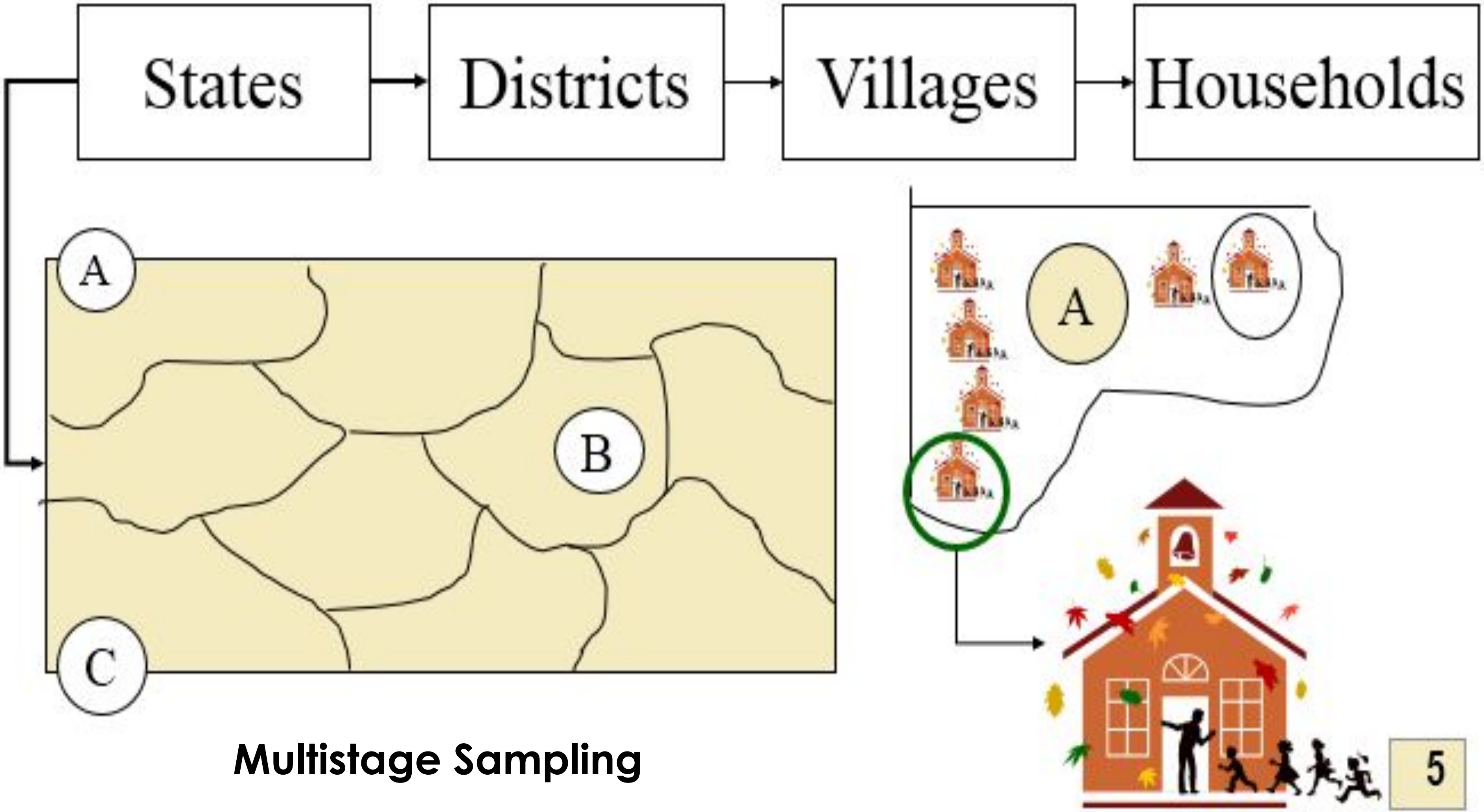
Figure 11.2 **Systematic Random Sampling**

From a population of 40 students, let's select a systematic random sample of 8 students. Our skip interval will be 5 (40 ÷ 8 = 5). Using a random number table, we choose a number between 1 and 5. Let's say we choose 4. We then start with student 4 and pick every 5th student:





Our trip to the random number table could have just as easily given us a 1 or a 5, so all the students do have a chance to end up in our sample.



States → Districts → Villages → Households

Multistage Sampling

Nonrandom sampling methods...

1. **Judgmental sampling**
2. **Convenient sampling**
3. **Quota sampling**
4. **Snow ball sampling**

Judgmental sampling: choice of sample items depends exclusively on the judgment of the investigator.

Convenience sampling: A sample obtained from readily available lists. the process of including whoever happens to be available at the time

Quota sampling: In quotas are set up according to some specific characteristics and sample will be taken according to specified quota. Sampling will be depend upon the field representative

Snowball sampling: Assumption of this method is that " if small ball is let roll from the top of snow-peak, it gathers substantial amount of snow and looks like a big ball when it arrives at the bottom of snow hill.

Parameter Estimation

We use statistics to estimate parameters,

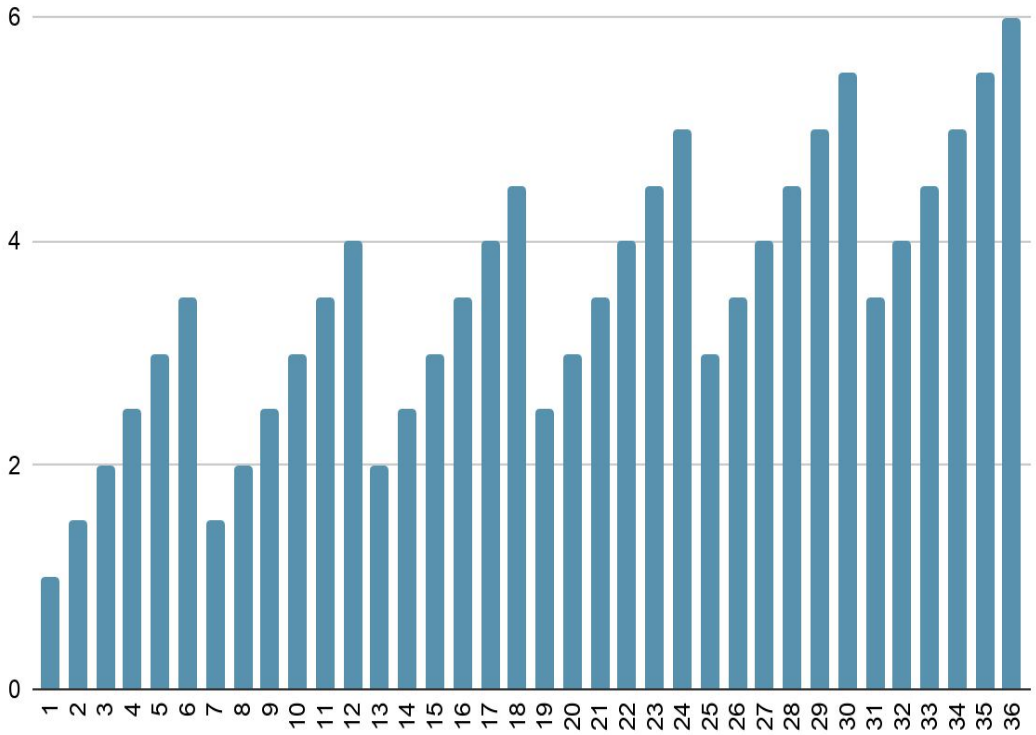
$$\bar{X} \rightarrow \mu \qquad SD \rightarrow \sigma$$

Sampling Distribution

- A sampling distribution is a distribution of a statistic over all possible samples.
- Suppose
- Population has 6 elements: 1, 2, 3, 4, 5, 6 (like numbers on dice)
- We want to find the sampling distribution of the mean for n=2
- If we sample with replacement, what can happen?

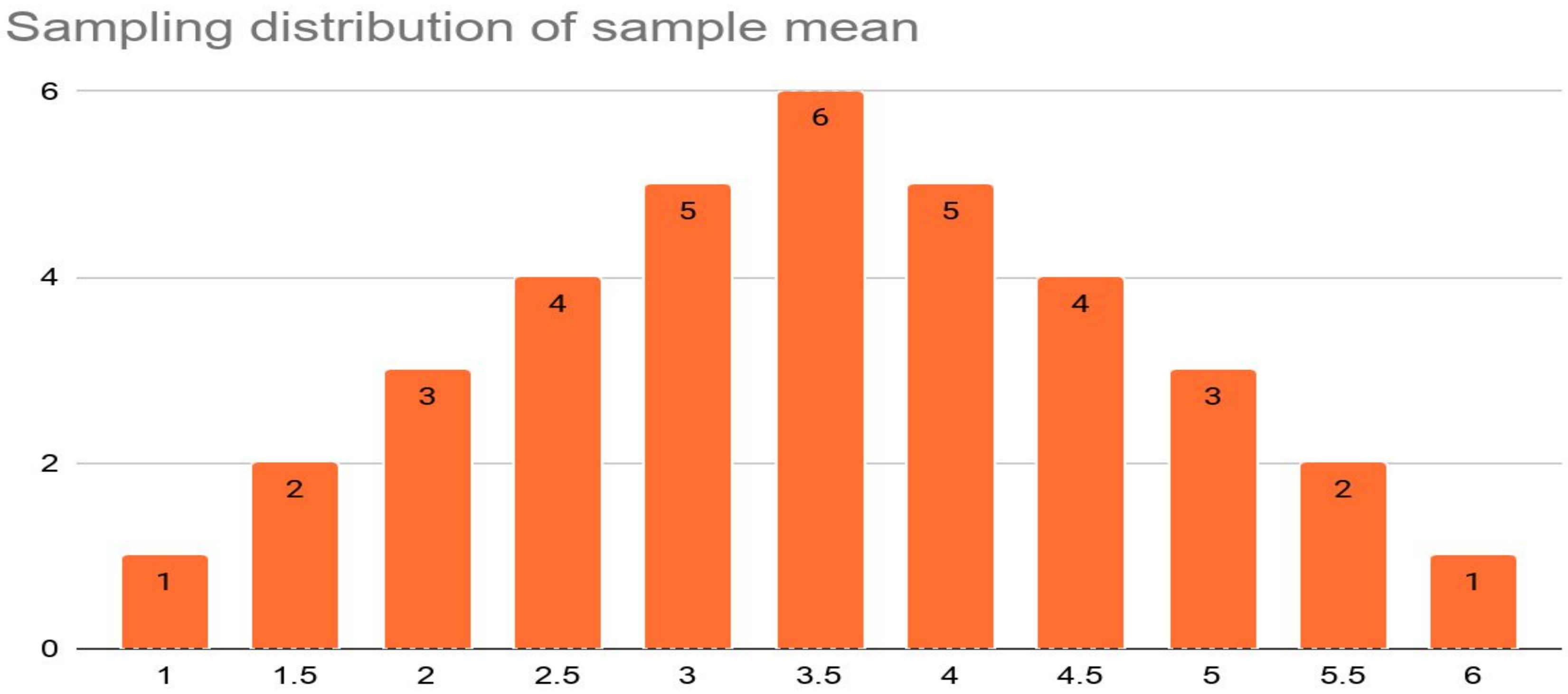
Sampling Distribution

1st	2nd	M	1st	2nd	M	1st	2nd	M
1	1	1	3	1	2	5	1	3
1	2	1.5	3	2	2.5	5	2	3.5
1	3	2	3	3	3	5	3	4
1	4	2.5	3	4	3.5	5	4	4.5
1	5	3	3	5	4	5	5	5
1	6	3.5	3	6	4.5	5	6	5.5
2	1	1.5	4	1	2.5	6	1	3.5
2	2	2	4	2	3	6	2	4
2	3	2.5	4	3	3.5	6	3	4.5
2	4	3	4	4	4	6	4	5
2	5	3.5	4	5	4.5	6	5	5.5
2	6	4	4	6	5	6	6	6

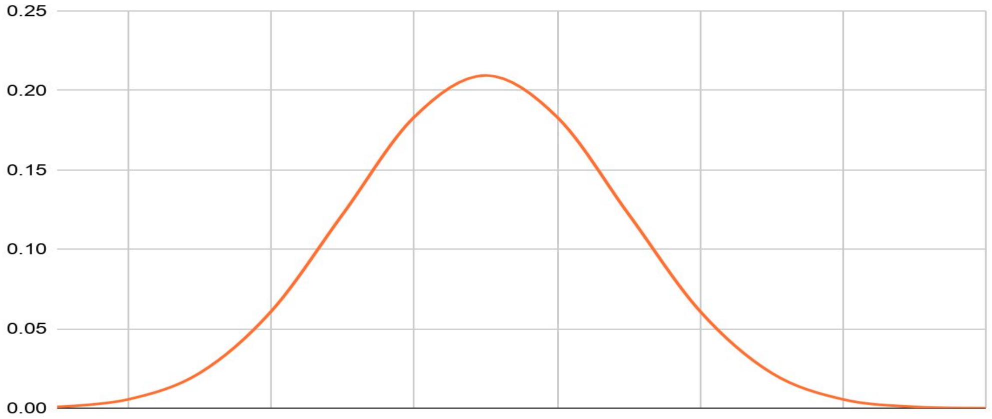


Possible Outcomes

Sampling Distribution



Sampling Distribution

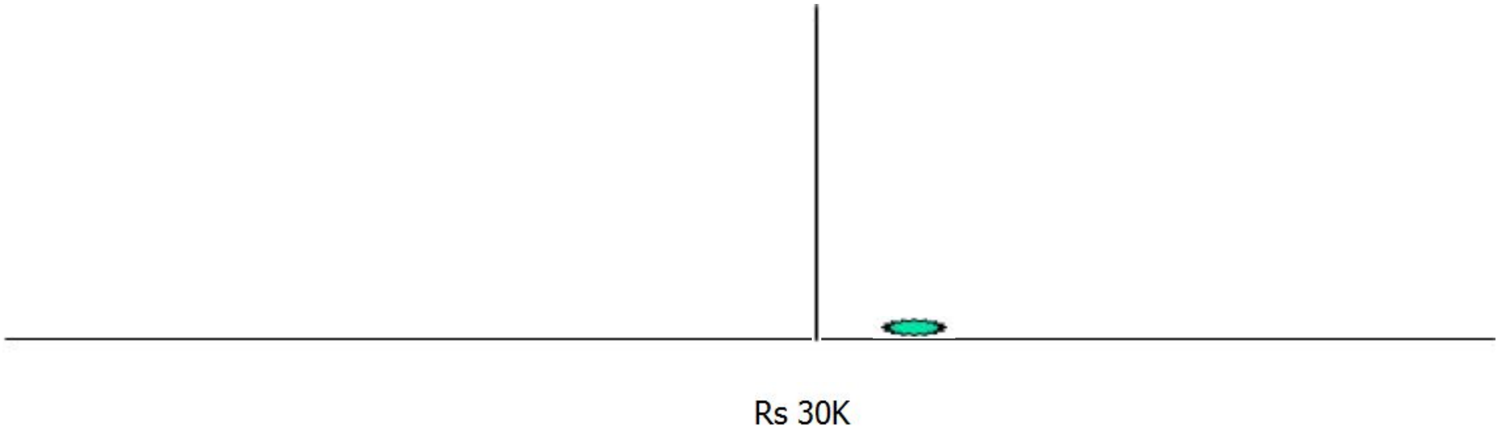


- The sampling distribution shows the relation between the probability of a statistic and the statistic value for all possible samples of size n drawn from a population.

Sampling Distribution

Let’s create a sampling distribution of means…

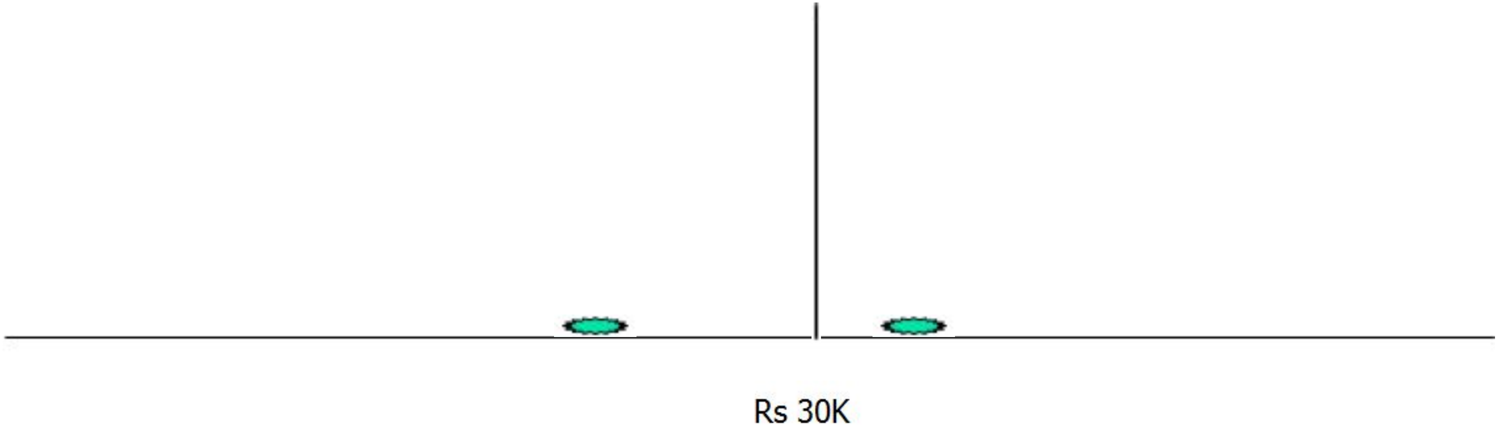
Take a sample of size 1,500 people. Record the mean income. Our census said the mean is Rs 30K.



Sampling Distribution

Let’s create a sampling distribution of means…

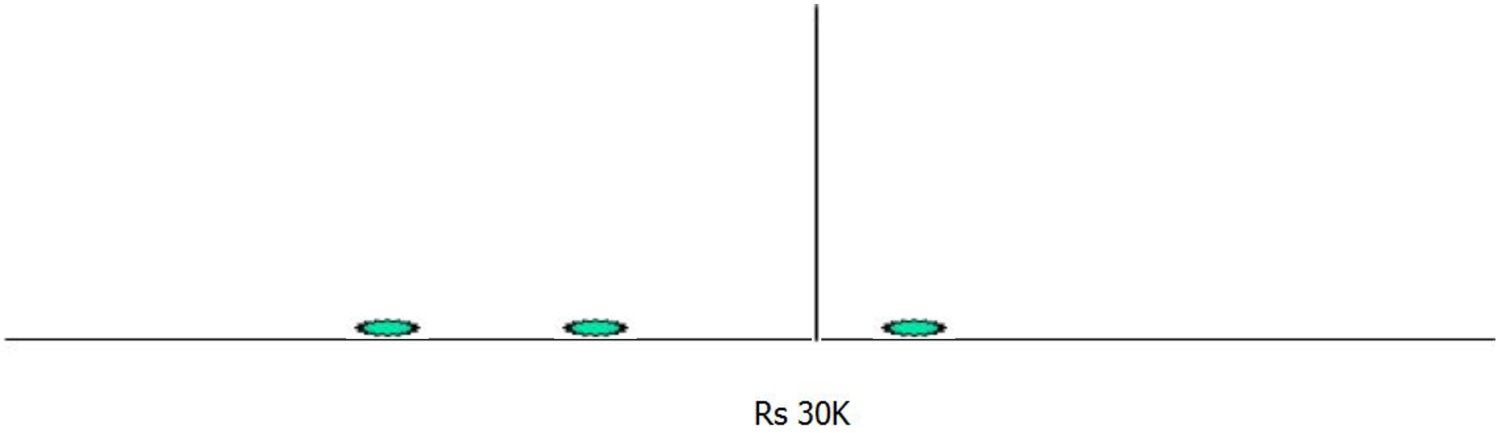
Take a sample of size 1,500 people. Record the mean income. Our census said the mean is Rs 30K.



Sampling Distribution

Let’s create a sampling distribution of means…

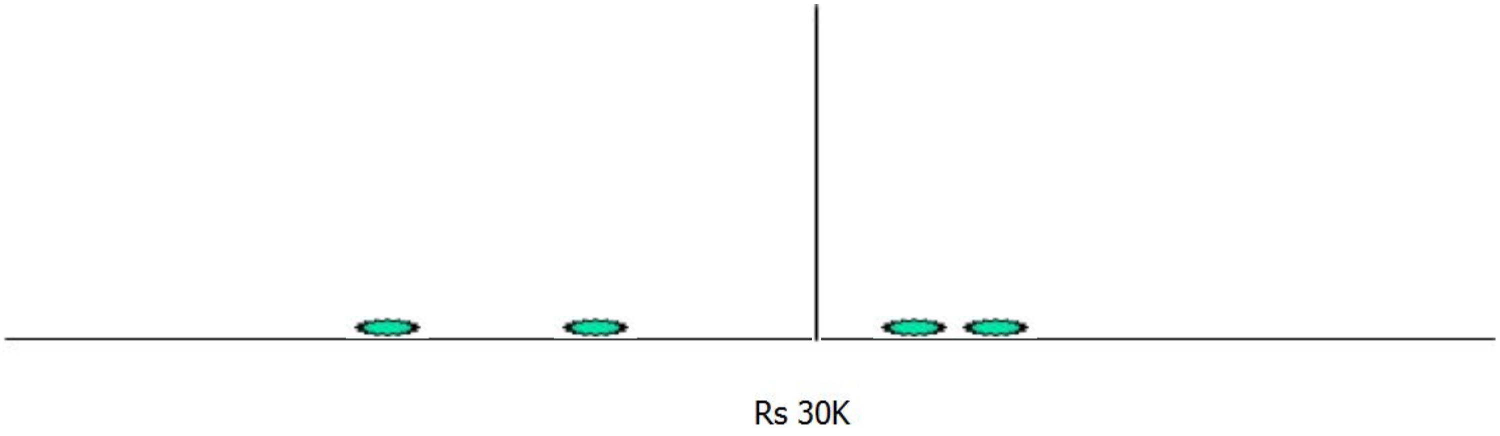
Take a sample of size 1,500 people. Record the mean income. Our census said the mean is Rs 30K.



Sampling Distribution

Let’s create a sampling distribution of means…

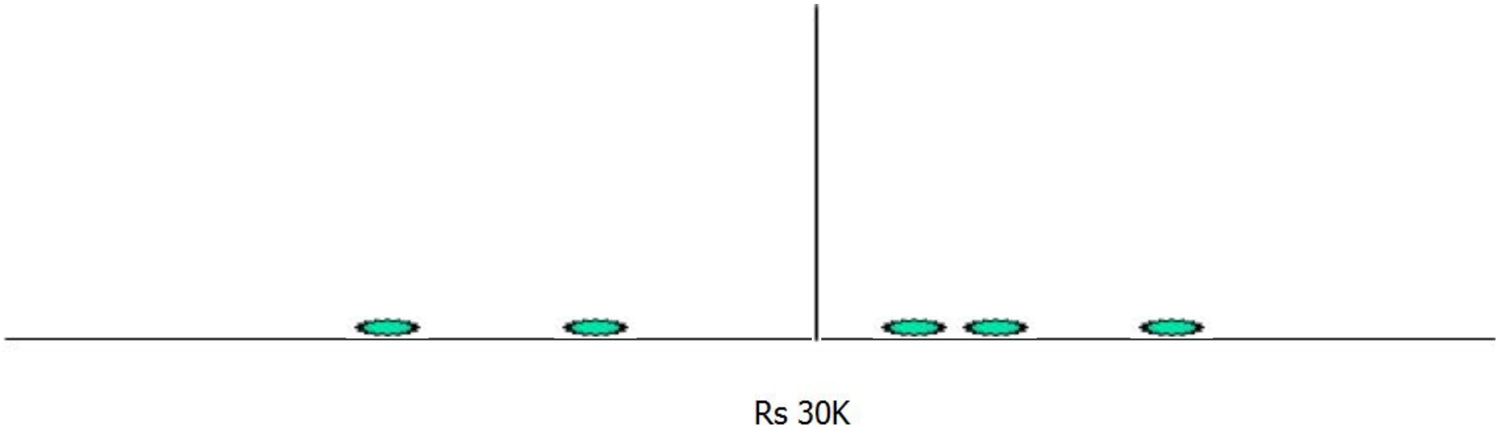
Take a sample of size 1,500 people. Record the mean income. Our census said the mean is Rs 30K.



Sampling Distribution

Let’s create a sampling distribution of means…

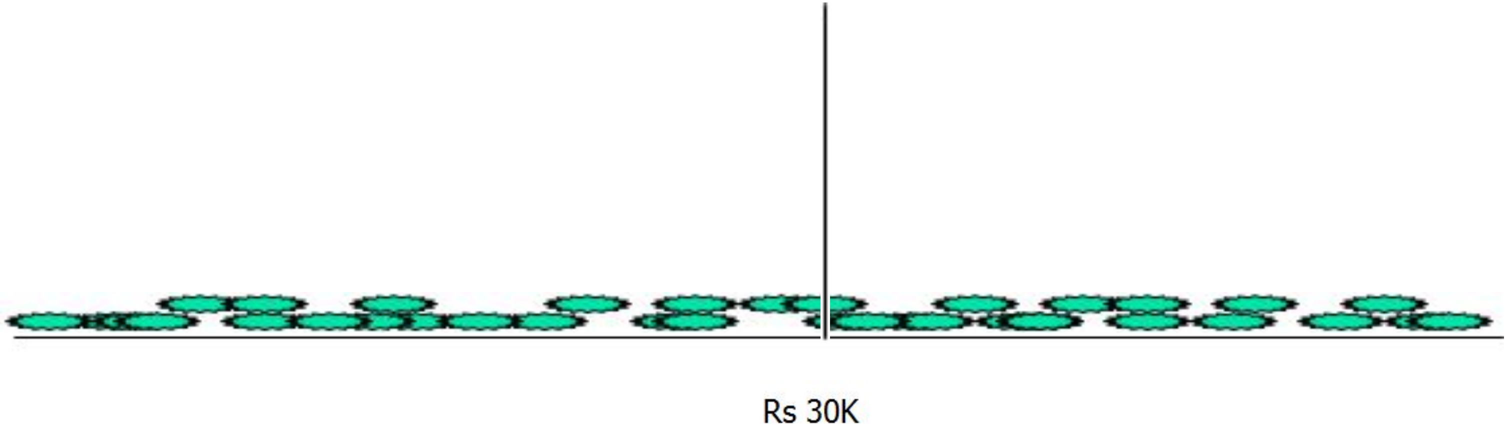
Take a sample of size 1,500 people. Record the mean income. Our census said the mean is Rs 30K.



Sampling Distribution

Let’s create a sampling distribution of means…

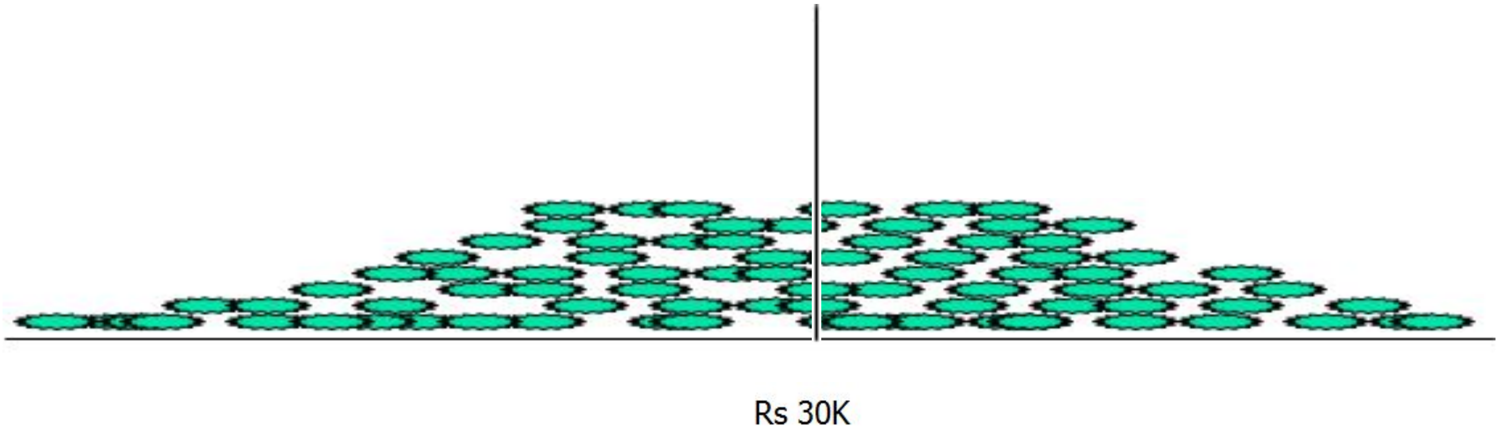
Take a sample of size 1,500 people. Record the mean income. Our census said the mean is Rs 30K.



Sampling Distribution

Let's create a sampling distribution of means…

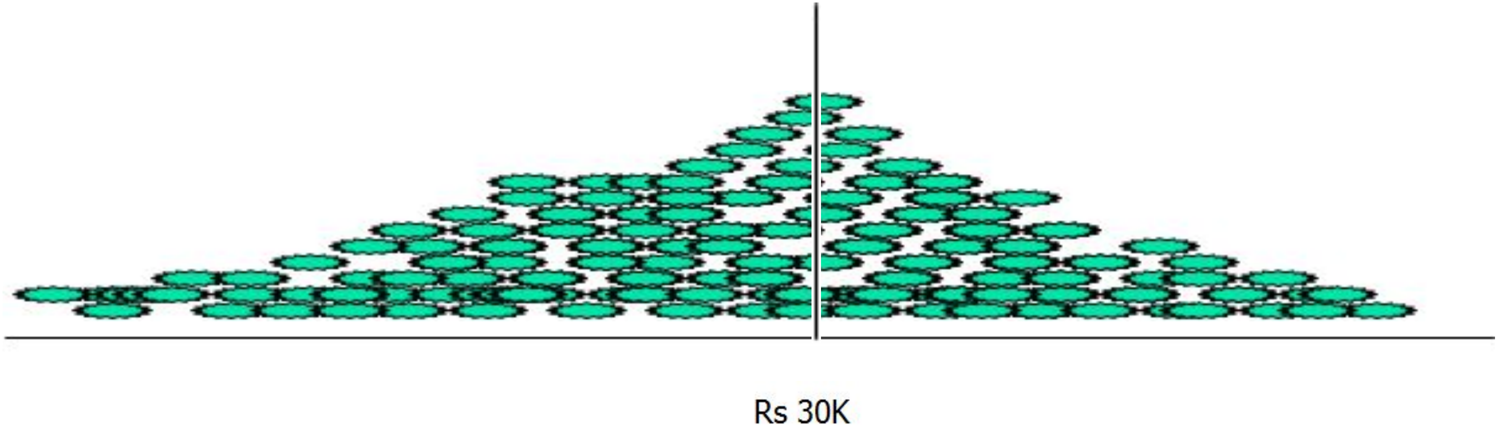
Take a sample of size 1,500 people. Record the mean income. Our census said the mean is Rs 30K.



Sampling Distribution

Let’s create a sampling distribution of means…

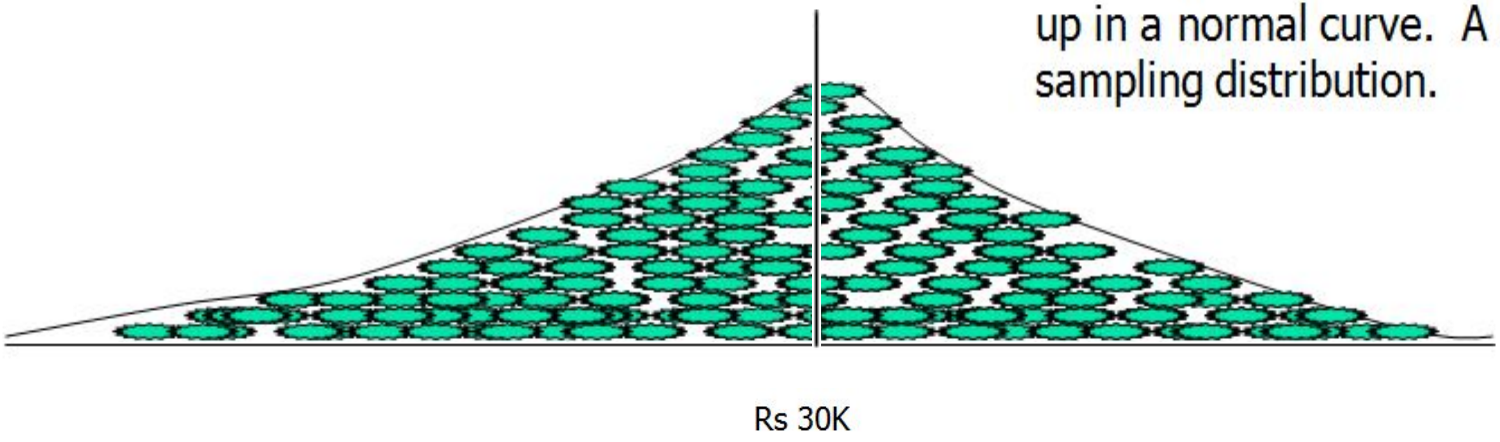
Take a sample of size 1,500 people. Record the mean income. Our census said the mean is Rs 30K.



Sampling Distribution

Let’s create a sampling distribution of means…

Take a sample of size 1,500 people. Record the mean income. Our census said the mean is Rs 30K.



The sample means would stack up in a normal curve. A normal sampling distribution.

Inference

- In real life calculating parameters of populations is usually impossible because populations are very large.
- Rather than investigating the whole population, we take a sample, calculate a **statistic** related to the **parameter** of interest, and make an inference.
- The **sampling distribution** of the **statistic** is the tool that tells us how close is the statistic to the parameter.

Sampling Distribution Mean and SD

- The Mean of the sampling distribution is defined the same way as any other distribution (expected value).
- The SD of the sampling distribution is the **Standard Error.** Important and useful.

Standard Error

The standard deviation of the sampling distribution.

Denoted by S.E.($\bar{y}$)

Has a great practical application, as it describes the amount of variation of sample values while estimating parameters using statistics

Example

Population consists of five numbers 1,3,5,7 and 9.

Enumerate all possible samples of size two which can be drawn from the population without replacement.

Possible samples

(1,3), (1,5), (1,7), (1,9), (3,5), (3,7), (3,9), (5,7), (5,9), (7,9)

Possible number of samples = ncr=5c2=10

Sampling Distribution

y	$y-\bar{y}$	$(y-\bar{y})^2$
1 3 5 7 9	-4 -2 0 2 4	16 4 0 4 16
25		40

Population Mean

$\mu = \bar{y} = 25/5=5$

Population variance

$\sigma^2=40/5 = 8$

Possible samples	Sample Means	$x-\bar{x}$	$(x-\bar{x})^2$
(1,3)	2	-3	9
(1,5)	3	-2	4
(1,7)	4	-1	1
(1,9)	5	0	0
(3,5)	4	-1	1
(3,7)	5	0	0
(3,9)	6	1	1
(5,7)	6	1	1
(5,9)	7	2	4
(7,9)	8	3	9
	$\sum\bar{x}=50$		$\sum(x-\bar{x})^2=$ 30

Mean of the sample means

$= \sum\bar{x}$ **/ncr**

$= 50/10 = 5 = \mu$

Variance of sample means

$= \sum(x-\bar{x})^2$ **/ncr**

=30/10

=3

Standard error (SE) = $\sqrt{3}$=1.732

Variance of SD also can be calculated as

$$\frac{\sigma^2}{n}\frac{(N-n)}{(N-1)}$$

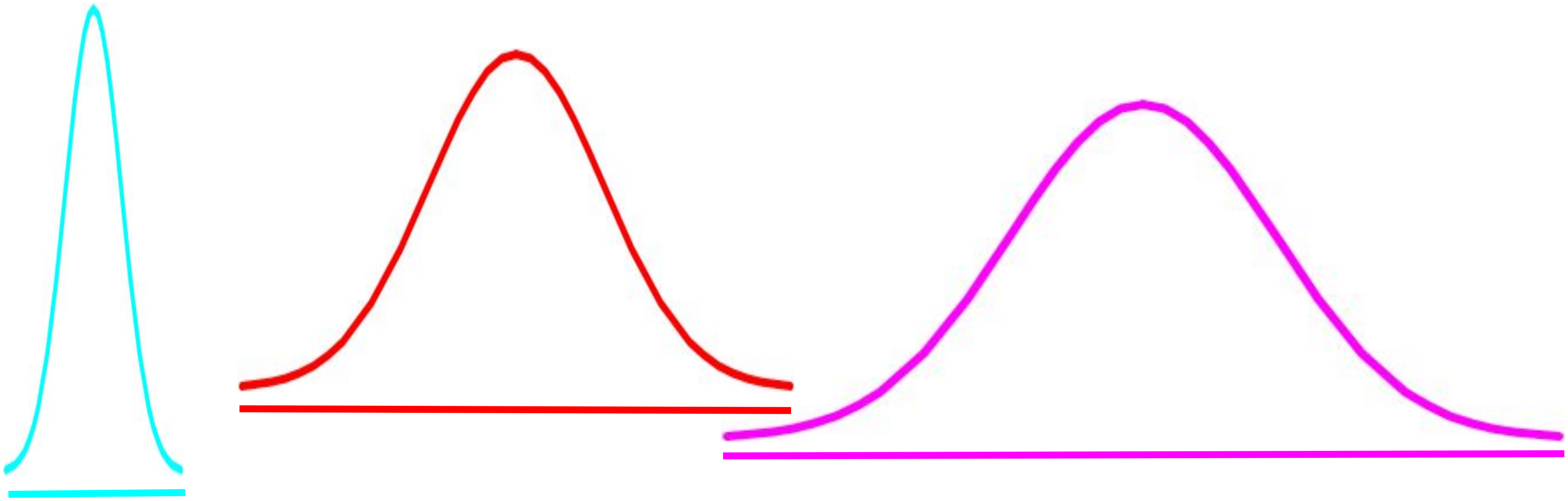
=8x3/(2x4)=3

Questions

1. **In a population the values of a characteristics Y_i , (i=1, 2, 3, 4, 5, 6.), are 6, 7, 3, 4, 8 and 5. Random samples of size two are drawn without replacement. Verify that E ($\bar{y}$) = $\bar{Y}$and E(s^2) = S^2. Also calculate V($\bar{y}$).The notations have their usual meaning.**
2. **In selecting 3 units with simple random sampling without replacement from a population having 6 units with values 1, 5, 8, 12, 15 and 19. Show that i) The sample mean is an unbiased estimator of the population mean ii) The sample mean square is an unbiased estimator of population mean square by enumerating all possible samples. Also find the variance of sample mean.**
3. **Consider a population of 6 units with values 1, 2, 3, 4, 5, 6. Write down all possible samples of 2 (without replacement) from this population and verify that sample mean is an unbiased estimate of population mean. Also calculate its sampling variance and verify that - (i) it agrees with the formula for the variance of the sample mean, and (ii) this variance is less than the variance obtained from sampling with replacement.**

Which sampling distribution has the lower variability

Precision of estimation = 1/variability = 1/ SE



Standard Error

The standard error indicates not only to observe the variability in the sample means but also the accuracy of the estimating population parameter.

A distribution of sample means having lower standard error is a better estimator of the population means than a distribution of sample means having larger standard error.

Standard Errors

Standard error of mean is whe $\sigma/\sqrt{n}$ sampling with replacement or if sampling is done with large population

When sampling is done from finite population and sampling is done with out replacement

$$\sqrt{\frac{\sigma 2\,(N-n)}{n\,(N-1)}}$$

Standard error of proportion = $\sqrt{PQ/n}$ $\sqrt{\frac{p(1-p)}{n}}$

For finite & SWOR $\sqrt{\frac{PQ}{n}}\cdot\sqrt{\frac{N-n}{N-1}}$

Standard Error

Standard error of sample Standard Deviation

$$\sqrt{\frac{\sigma^2}{2n}}$$

SE of difference of Means= **S.E.** ($\bar{\mathbf{x}}_{\mathbf{1}}$ **-** $\bar{\mathbf{x}}_{\mathbf{2}}$)=

$$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

SE of difference of proportion =

$$\text{S.E. }(p_1 - p_2) = \sqrt{PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

SE of difference of standard deviation =

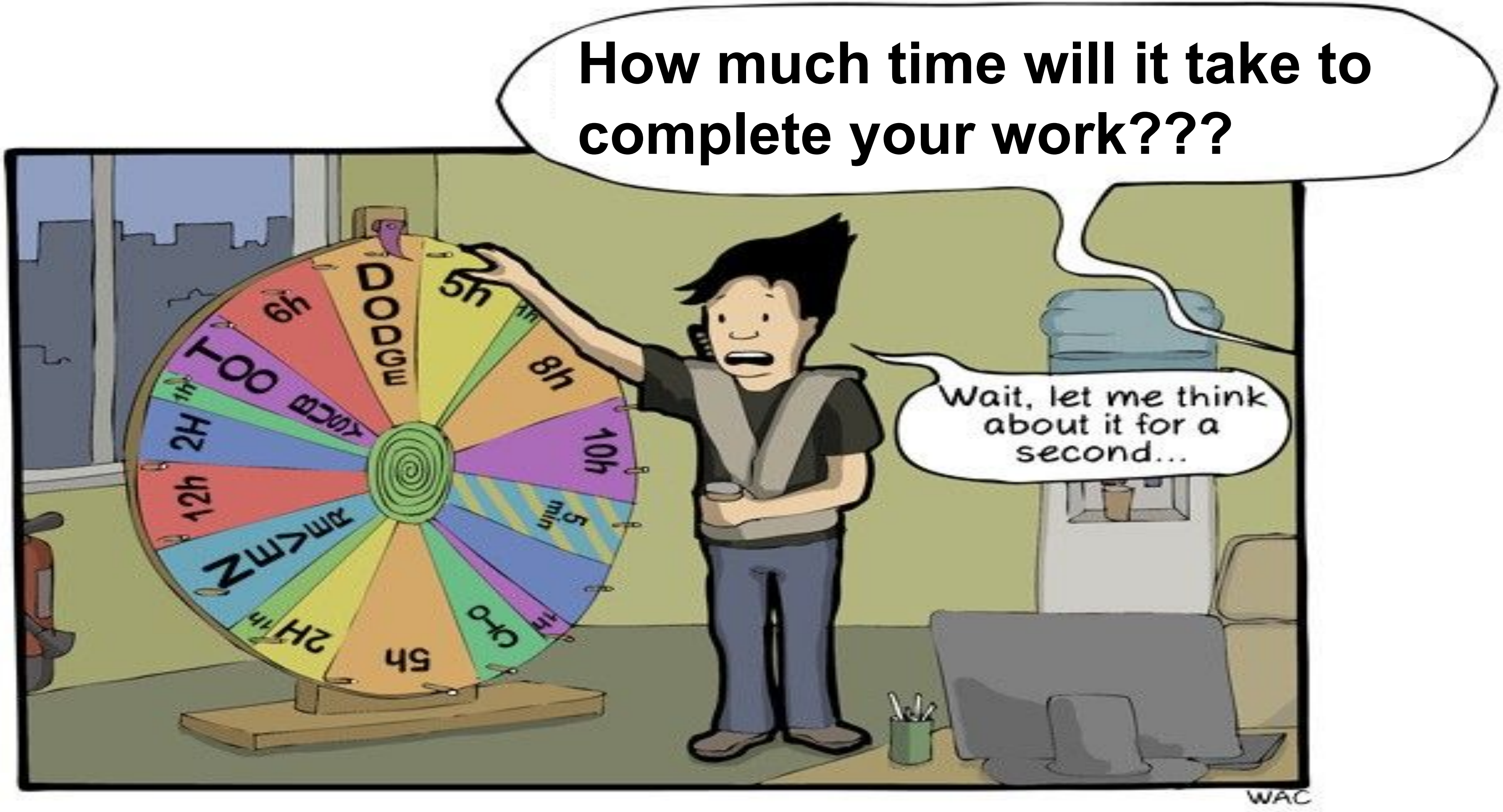
$$\text{S.E. }(s_1 - s_2) = \sqrt{\frac{\sigma_1^2}{2n_1}+\frac{\sigma_2^2}{2n_2}}$$

Statistics as Estimators

- We use sample data compute statistics.
- The statistics estimate population values, e.g.,

$$\overline{X} \to \mu$$

- An estimator is a method for producing a best guess about a population value.
- An estimate is a specific value provided by an estimator.
- We want good estimates. What is a good estimator? What properties should it have?

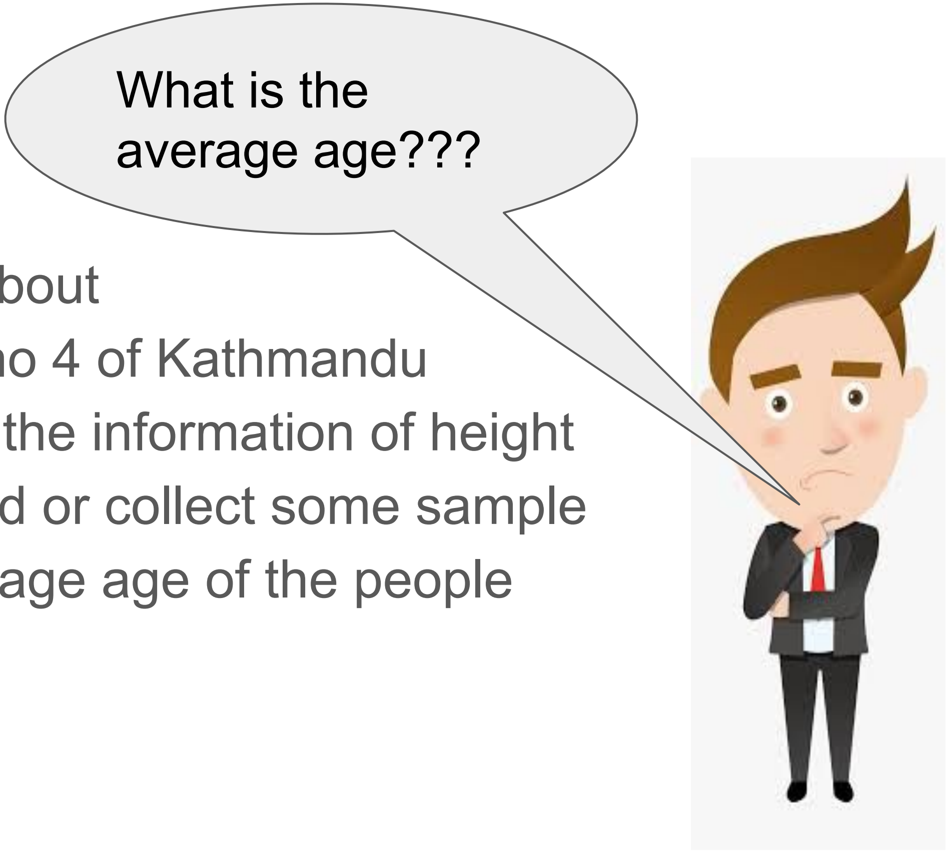

How much time will it take to complete your work???
Wait, let me think about it for a second...
DODGE
5h
1h
8h
10h
5 min
1h
0h
5h
2H
1h
NEVER
12h
2H
1h
TOO BUSY
6h
WAC

Estimation?

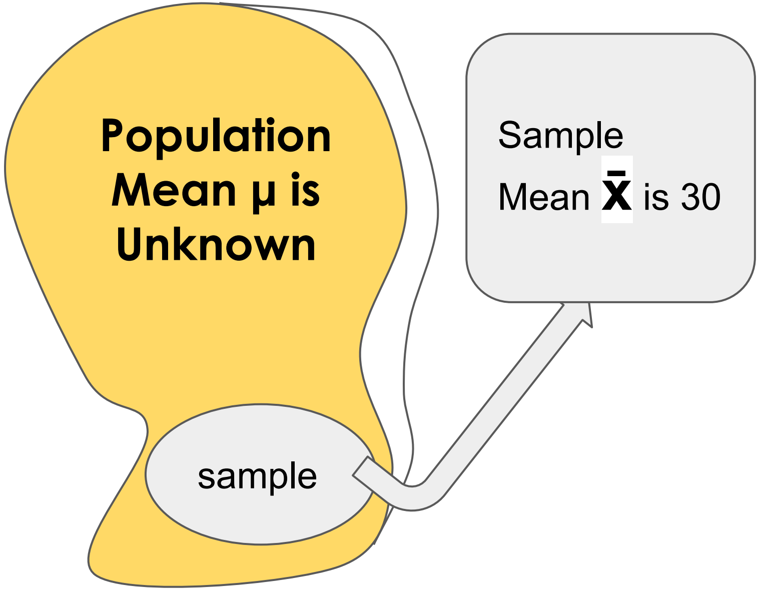
For example: the authority of District Administrative office of Kathmandu would like to know how pleased Kathmandu citizens are with their service. The best way to do this would be to ask every citizen of Kathmandu that how he or she feels about their service. Because the population of Kathmandu is about 17,50,000 people, this interviewing will take more cost and time. An alternative is to randomly select a subset of persons (say, 100) and ask them about the system. From this sample, we will infer what the people of Kathmandu think.

Estimation

Ward secretary would like to know about the average age of people of Ward no 4 of Kathmandu Municipality. Either he has to collect the information of height from all the people reside in this ward or collect some sample from the ward and estimate the average age of the people reside in this ward.



Point Estimation



Population Mean μ is Unknown

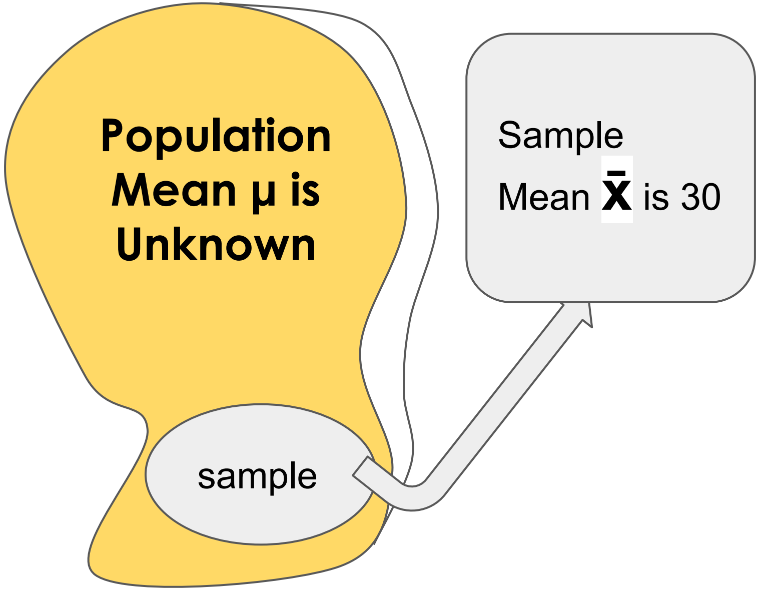
sample

Sample

Mean $\bar{\mathbf{x}}$ is 30

Population has mean 30.

Interval Estimation



Population Mean μ is Unknown

sample

Sample Mean $\bar{x}$ is 30

I am 95% confident that Population mean μ lies between 25 and 35.

Terminology

An estimator of a population parameter is

a random variable that depends on sample information . .

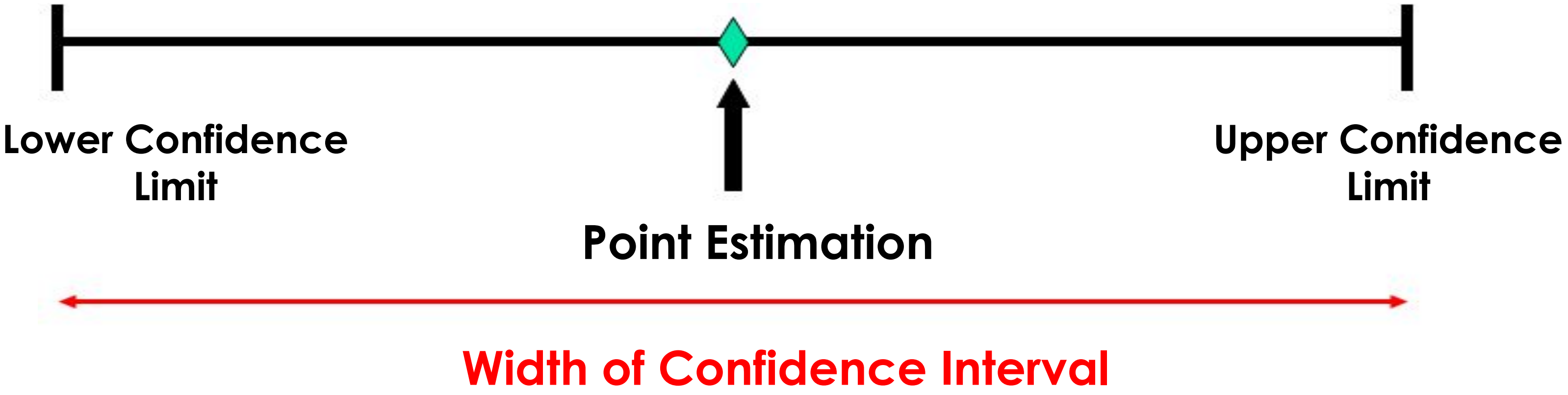
whose value provides an approximation to this unknown parameter

A specific value of that random variable is called an estimate

Estimation

A point estimate is a single number,

a confidence interval provides additional information about variability



Point Estimation

We can Estimate Population Parameter		With Sample Statistics
Mean	μ	$\bar{x}$
Proportion	P	p

Properties of good estimator (unbiasedness)

A point estimator **t** is said to be an unbiased estimator of the parameter Ɵ if the expected value, or mean, of the sampling distribution of **t** is Ɵ ,

E(t)=Ɵ

If E(statistic)=parameter, the estimator is unbiased.

Examples:

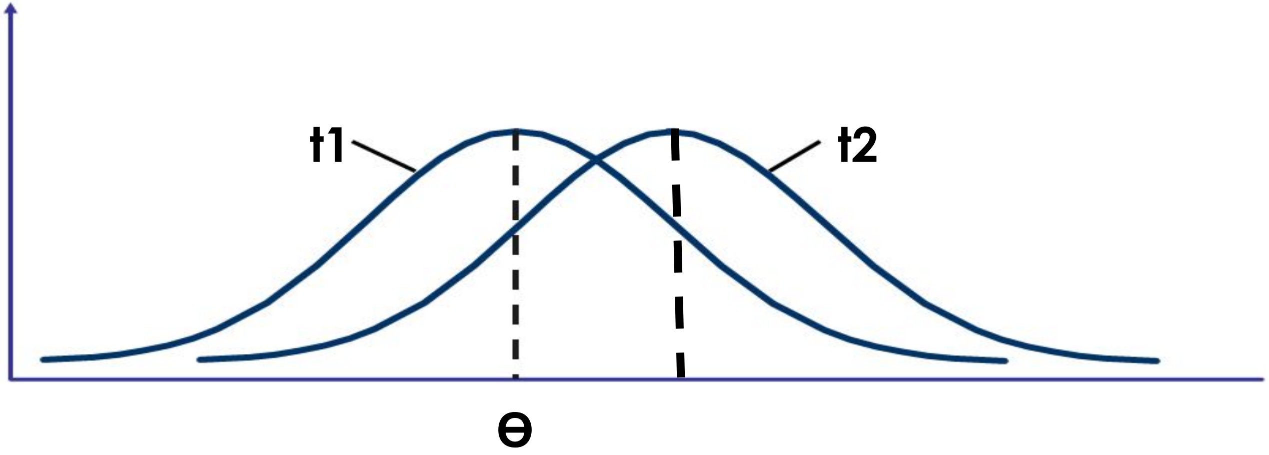
The sample mean is an unbiased estimator of μ

The sample mean square is an unbiased estimator of Pop S^2

$$E(\overline{X}) = \mu$$

The sample proportion is an unbiased estimator of P

Properties of good estimator (unbiasedness)



The bias is the difference between its mean and **ϴ**

Bias=E(t)-ϴ

The bias of an unbiased estimator is 0

The calculated value of estimator might not be equals to the population parametric value **θ**, but the average value of the estimates over all possible samples would be equal to the unknown population parameter **θ.**

Therefore the unbiasedness is a property of examining a good estimator through average.

Properties of good estimator (Consistency)

Let **t** be an estimator of **ϴ**

t is a consistent estimator of **ϴ** if the difference between the expected value of **t** and **ϴ** decreases as the sample size increases

$E(t) - ϴ$ = least when n is large enough

Consistency is desired when unbiased estimators cannot be obtained

Consistency

Consistency is a limiting property of estimators. So it is a procedure of checking quality of an estimator on the basis of large sample. This property implies that as the sample size increases, the variance of the estimator t tends to zero and consequently the estimate comes closer to the true value of the parameter θ.

Properties of good estimator (Efficiency)

Suppose there are several unbiased estimators of ϴ

The most efficient estimator or the minimum variance unbiased estimator of ϴ is the unbiased estimator with the smallest variance

Let $t_1$and t_2 be two unbiased estimators of ϴ, based on the same number of sample observations. Then,

t_1 is said to be more efficient than t_2 if $Var(t_1) < Var(t_2)$

Sample mean is more efficient estimator for μ than the sample median for large sample drawn from $N(\mu,\sigma^2)$

Confidence Interval

- How much uncertainty is associated with a point estimate of a population parameter?
- An interval estimate provides more information about a population characteristic than does a point estimate
- Such interval estimates are called confidence intervals

Confidence Interval Confidence Level

If $P(a < \theta < b) = 1 - \alpha$ then the interval from a to b is called a $100(1 - \alpha)\%$ confidence interval of θ.

The quantity $(1 - \alpha)$ is called the confidence level of the interval (between 0 and 1)

In repeated samples of the population, the true value of the parameter θ would be contained in $100(1 - \alpha)\%$ of intervals calculated this way.

The confidence interval calculated in this manner is written as $a < \theta < b$ with $100(1 - \alpha)\%$ confidence.

A Large-Sample Interval for μ

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population having a mean μ and standard deviation σ. Provided that n is large, the Central Limit Theorem (CLT) implies that $\bar{X}$ has approximately a normal distribution whatever the nature of the population distribution.

It then follows that $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ has approximately a standard normal distribution, so that

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) \approx 1 - \alpha$$

If σ is not known then it has to be replaced by S.

A Large-Sample Interval for μ

From areas under normal probability curve, we have

$P\ (-1.96 < Z < 1.96) = 0.95$

$P(-Z_{\alpha/2} < Z < +Z_{\alpha/2}) = 1- \alpha$

$$P\left(-Z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < +Z_{\alpha/2}\right) = 1- \alpha$$

$P(-Z_{\alpha/2}\ \sigma/\sqrt{n} < \bar{X} - \mu < +Z_{\alpha/2}\ \sigma/\sqrt{n}) = 1- \alpha$

$\bar{X} \pm Z_{\alpha/2}\ \sigma/\sqrt{n}$

$P(-\bar{X} - Z_{\alpha/2}\ \sigma/\sqrt{n} < -\mu < -\bar{X} + Z_{\alpha/2}\ \sigma/\sqrt{n}) = 1- \alpha$

$P(\bar{X} + Z_{\alpha/2}\ \sigma/\sqrt{n} > \mu > \bar{X} - Z_{\alpha/2}\ \sigma/\sqrt{n}) = 1- \alpha$

$P(\bar{X} - Z_{\alpha/2}\ \sigma/\sqrt{n} < \mu < \bar{X} + Z_{\alpha/2}\ \sigma/\sqrt{n}) = 1- \alpha$

$\hat{\mu} = \bar{X} \pm Z_{\alpha/2}\ \sigma/\sqrt{n}$

If σ is not known then it has to be replaced by S.

General Formula

The value of the reliability factor depends on the desired level of confidence.

Point Estimate 土 Reliability Factor x Standard Error

Confidence interval for mean

Assumptions

Population variance σ^2 is known

Population is normally distributed

If population is not normal, use large sample

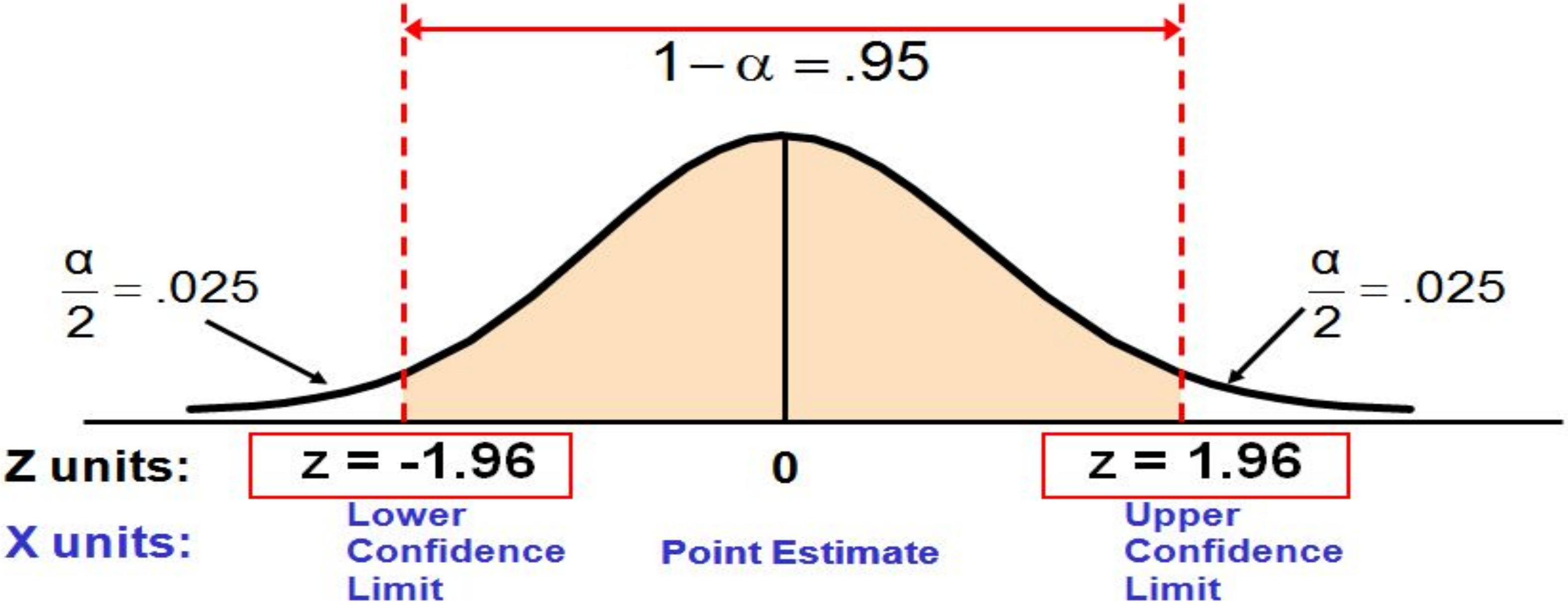
Confidence interval estimate:

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

(where $z_{\alpha/2}$ is the normal distribution value for a probability of $^{\alpha}/2$ in each tail)

Confidence Interval

- Consider a 95% confidence interval:



$1-\alpha = .95$

$\frac{\alpha}{2} = .025$

$\frac{\alpha}{2} = .025$

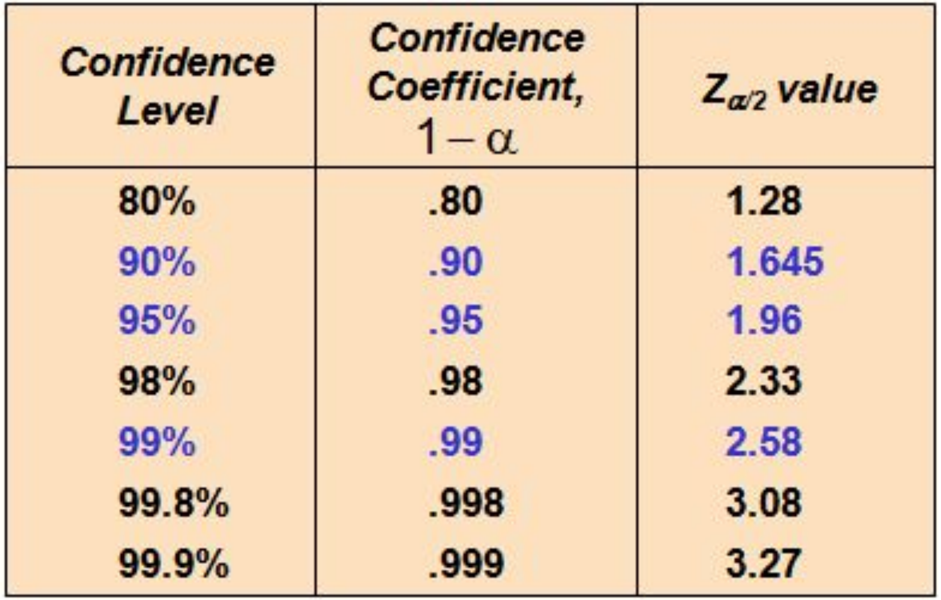
Z units: **z = -1.96** **0** **z = 1.96**

X units: **Lower Confidence Limit** **Point Estimate** **Upper Confidence Limit**

- Find $z_{.025} = \pm 1.96$ from the standard normal distribution table

Level of Confidence

Commonly used confidence level are 90%, 95% and 99%.



Confidence Level	Confidence Coefficient, $1-\alpha$	$Z_{\alpha/2}$ value
80%	.80	1.28
90%	.90	1.645
95%	.95	1.96
98%	.98	2.33
99%	.99	2.58
99.8%	.998	3.08
99.9%	.999	3.27

Margin of Error

The confidence interval,

$$\boxed{\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

Can also be written as

$$\boxed{\bar{x} \pm ME}$$

where ME is called the margin of error

$$\boxed{ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

The interval width, w, is equal to twice the margin of error

Reducing Margin of Error

$$ME = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

The margin of error can be reduced if

- the population standard deviation can be reduced (σ↓)
- The sample size is increased (n↑)
- The confidence level is decreased, (1 – α) ↓

Example

The Ministry of Federal Affairs and General Administration wishes to know the average income of general assistants. A sample of 60 assistants shows a sample mean of Rs17,400 with a standard deviation of Rs. 3,150. (a) Place a 90% and 95% of confidence limit around your best estimate of the average income of general assistance.

Confidence Interval for μ (σ^2 Unknown)

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

This introduces extra uncertainty, since s is variable from sample to sample

So we use the t distribution instead of the normal distribution

Confidence Interval for μ (σ^2 Unknown)

Assumptions

- Population standard deviation is unknown
- Population is normally distributed
- If population is not normal, use large sample

Use Student's t Distribution

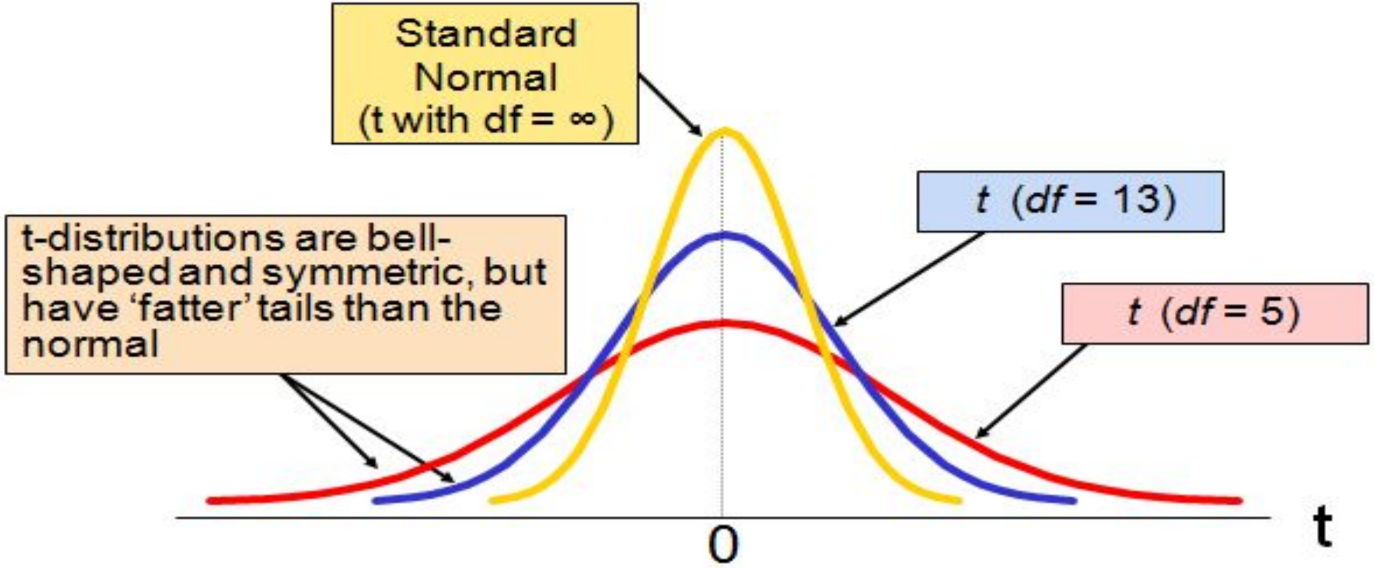
Confidence Interval Estimate:

$$\bar{x} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$$

where $t_{n-1,\alpha/2}$ is the critical value of the t distribution with n-1 d.f. and an area of α/2 in each tail:

Student’s t Distribution

As n increases t distribution tends to normal Distribution



Standard Normal (t with df = ∞)

t (*df* = 13)

t (*df* = 5)

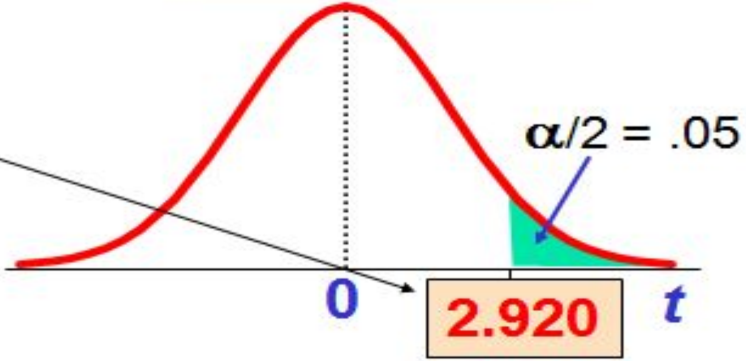
t-distributions are bell-shaped and symmetric, but have ‘fatter’ tails than the normal

t -table values

	Upper Tail Area		
df	.10	**.05**	.025
1	3.078	6.314	12.706
2	1.886	**2.920**	4.303
3	1.638	2.353	3.182

The body of the table contains t values, not probabilities

Let: n = 3
df = *n* - 1 = 2
α = .10
$\alpha/2$ =.05



Example

A random sample of n = 25 has $\bar{x}$ = 50 and s = 8. Form a 95% confidence interval for μ

- d.f. = n – 1 = 24, so $t_{n-1,\alpha/2} = t_{24,.025} = 2.0639$

The confidence interval is

$$\bar{x} - t_{n\text{-}1,\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{n\text{-}1,\alpha/2}\frac{S}{\sqrt{n}}$$

$$50 - (2.0639)\frac{8}{\sqrt{25}} < \mu < 50 + (2.0639)\frac{8}{\sqrt{25}}$$

$$\boxed{46.698 < \mu < 53.302}$$

Example II

Last year, there were 512 thiefs in Kathmandu. The police chief wants to know the average economic loss associated with thief in Kathmandu and wants to know it this afternoon. There isn’t time to analyze all 512 thief, so the department’s research analyst selects 10 thief at random, which show the following losses:

Rs.1,550 Rs.1,874 Rs.1,675 Rs.2,595 Rs.2,246

Rs.1,324 Rs.1,835 Rs.1,487 Rs.1,910 Rs.1,612

What is the best estimate of the average loss on a Thief? Place 80% confidence limits around this estimate.

Example II

x	(x-x̄)^2
1550	68016.64
1874	3994.24
1675	18441.64
2595	614969.6
2246	189399
1324	236974.2
1835	585.64
1487	104846.4
1910	9840.64
1612	39521.44
18108	1286590
1810.8	

Confidence Interval for difference of means $\mu_1 - \mu_2$

From areas under normal probability curve, we have

$P(-1.96 < Z < 1.96) = 0.95$

$P(-Z\alpha_{/2} < Z < +Z\alpha_{/2}) = 1 - \alpha$

$$Z = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}}$$

$$\hat{\mu}_1 - \hat{\mu}_2 = (\bar{x} - \bar{y}) \pm Z\alpha_{/2} \sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}$$

Question

A manager evaluates effectiveness of a major hardware upgrade by running a certain process 50 times before the upgrade and 50 times after it. Based on these data, the average running time is 8.5 minutes before the upgrade, 7.2 minutes after it. Historically, the standard deviation has been 1.8 minutes, and presumably it has not changed. Construct a 90% confidence interval showing how much the mean running time reduced due to the hardware upgrade.

Confidence Interval for difference of means $\mu_1 - \mu_2$

From areas under normal probability curve, we have

$$Z = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}}$$

$P(-1.96 < Z < 1.96) = 0.95$

$$\hat{\mu}_1 - \hat{\mu}_2 = (\bar{x} - \bar{y}) \pm Z_{\alpha/2} \sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}$$

$P(-Z_{\alpha/2} < Z < +Z_{\alpha/2}) = 1 - \alpha$

Satterthwaite approximation for DF

When variances are unknown and unequal:
Then the σ_1^2 and σ_2^2 will be replaced by S_1^2 and S_2^2.

$$\text{DF จ} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{(n_1-1)} + \frac{(s_2^2/n_2)^2}{(n_2-1)}}$$

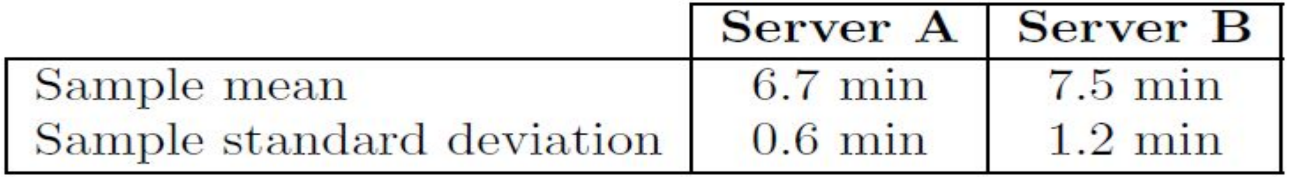
But when variances are unknown but equal then $\sigma_1^2 = \sigma_2^2 = \sigma^2$
Combined sample variance will be calculated by

$$Sp^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_2} \text{ (for large sample case)}$$

Question

An account on server A is more expensive than an account on server B. However, server A is faster. To see if it’s optimal to go with the faster but more expensive server, a manager needs to know how much faster it is.

A certain computer algorithm is executed 30 times on server A and 20 times on server B with the following results,



	Server A	Server B
Sample mean	6.7 min	7.5 min
Sample standard deviation	0.6 min	1.2 min

Construct a 95% confidence interval for the difference μ1 −μ2 between the mean execution times on server A and server B.

CI for Population Proportion

if the sample size is large, with standard error of sample proportion is equals to

$$\sigma_P = \sqrt{\frac{P(1-P)}{n}}$$

We will estimate this with sample data:

$$\sqrt{\frac{p(1-p)}{n}}$$

CI for Population Proportion

Upper and lower confidence limits for the population proportion are calculated with the formula

$$p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < P < p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

where

$z_{\alpha/2}$ is the standard normal value for the level of confidence desired

p is the sample proportion   and n is the sample size

example

A random sample of 100 people shows that 25 are left-handed.

Form a 95% confidence interval for the true proportion of left-handers

example

A random sample of 100 people shows that 25 are left-handed. Form a 95% confidence interval for the true proportion of left-handers

Solution:

$$p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < P < p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

$$\frac{25}{100} - 1.96\sqrt{\frac{.25(.75)}{100}} < P < \frac{25}{100} + 1.96\sqrt{\frac{.25(.75)}{100}}$$

$$\boxed{0.1651 < P < 0.3349}$$

Example

A survey of 500 people shopping at a shopping mall, selected at random showed that 350 of them used cash and 150 of them used credit cards, construct a 95% confidence interval estimate of the proportion of all the persons at the mall, who use cash for shopping.

Sample Size Estimation

To determine the sample size for the mean, researcher must know three factors:

1. The desired confidence level, which determines the value of Z, the critical value from the standardized normal distribution.

2. The acceptable sampling error ‘e’. (here e = x- m)

3. The estimated value of standard deviation.

The Size of sample will $n = \left(\frac{Z_{\alpha}\,\sigma}{E}\right)^2$

Where, E = error i.e. difference between sample mean and population mean = | x̄– μ|

Example

A researcher wants to estimate universe mean by using sampling technique. What should be the sample size when the permissible error between parameter value and sample statistic in 95% of chance will not be more than 1.5 and population standard deviation is 15.

Solution:

Error (E) = 1.5,

s.d. (s) = 15

Confidence level (1 – α) = 95%

Significant level (α) = 5% $z_{\alpha/2}$ = 1.96 [Two tailed]

We know, $$n = \left(\frac{z_\alpha \cdot \sigma}{E}\right)^2 = \left(\frac{1.96 \times 15}{1.5}\right)^2 = (19.6)^2 = 384.16 \approx 384$$

Sample Size Estimation

To determine the sample size for proportion, researcher must know three factors:

1. The desired level of confidence that determines the value of Z.
2. The acceptance sampling error ‘E’. (Where E = p-P)
3. The estimated value of proportion p.

The Size of sample will be:

Where, E = error i.e. difference between sample Proportion and population Proportion= | p– P|

$$n = \left(\frac{Z_\alpha}{E}\right)^2 PQ$$

Example

The Department of MTech wishes to estimate the percentage of students securing 60% marks or below. Department of MTech is 95% confident that the estimation will be within ± 3% of the true population proportion.

a. What sample size should be taken if the previous survey showed that 25% of students secured 60% marks or below?

b. What should be the minimum sample size for the same degree of confidence and same maximum allowable error, if no previous survey had been taken?

Central Limit Theorem

1. Sampling distribution of means becomes normal as N increases, regardless of shape of original distribution.
2. Binomial becomes normal as N increases.
3. Applies to other statistics as well (e.g., variance)

Properties of the Normal

- If a distribution is normal, the sampling distribution of the mean is normal regardless of N.
- If a distribution is normal, the sampling distributions of the mean and variance are independent.

What is Hypothesis

- Administrator often faced with decisions about program effectiveness, personnel productivity, and procedural changes.
- Decisions on such matters are based on the information relevant to them.
  - Is Ram Krishna an effective supervisor?
  - Is average time taken by a teller in ABC bank is 10 minutes?
  - Is work performance of Department A is better than that of B?
- A question that solicits information about managerial problems is called a **hypothesis.**
- The phrase will be use as a statement rather than as a question, a hypothesis is nothing more than the statement about the world that may be tested to determine whether it is true or false.

Hypothesis

Two words

Hypo: Under Thesis: reasoned theory

Theory which is not fully reasoned

Tentative answer of the research question

Imaginative idea or guess depending upon previous accumulated knowledge which can be put to test to determine its validity.

Generally specify relationship between variables or with specific value

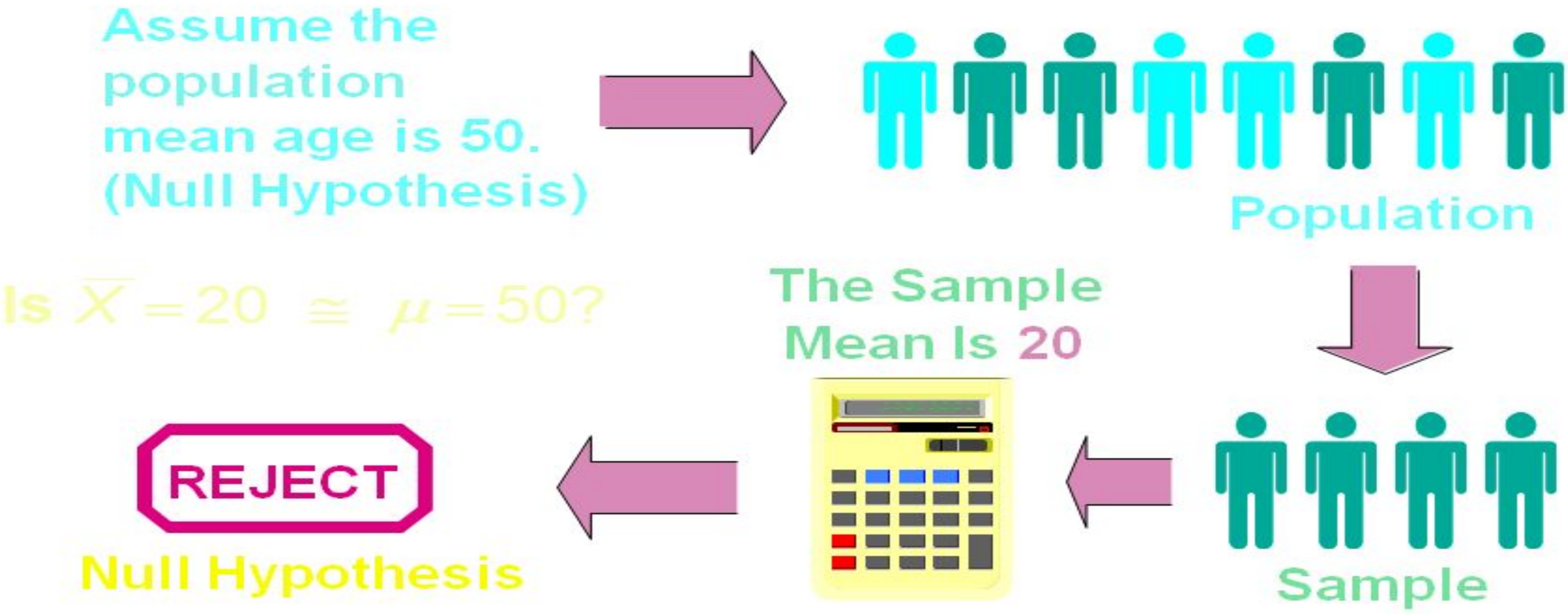
Hypothesis

- As an administrator or researcher you do not know about the phenomenon, situation, but you do have a hunch to form the basis of certain assumption or guesses. You test these by collecting information that will enable you to conclude if your hunch was right.
- It is tentative preposition
- Its validity is unknown
- In most cases, it specifies a relationship between two or more variables.
- It is the assumption about the population paramter.

Source of Hypothesis

- Culture of the society as culture has great influence upon the thinking process of people.
    - Caste is related to voting behavior among Nepalese
- Scientific study (Past Research)
- Personal experience
    - Very often researchers/ admin see the evidence of some behavior pattern in their daily lives.

Test of Hypothesis



Assume the population mean age is 50. (Null Hypothesis)

Population

The Sample Mean Is 20

Sample

Is $\overline{X} = 20 \cong \mu = 50$?

REJECT

Null Hypothesis

Type of Hypothesis

- Null Hypothesis
- A **null hypothesis** is a statement of no difference or no effect. If the null hypothesis is not rejected, no changes will be made, i.e. there is no differences. Null Hypothesis is denoted by Ho.
- Alternative Hypothesis
- An **alternative hypothesis** is one in which some difference or effect is expected. It is alternate to null hypothesis and is denoted by H_1.

Hypothesis Testing…

- **A criminal trial is an example of hypothesis testing without the statistics.**
- **In a trial a jury must decide between two hypotheses. The null hypothesis is**
- **H_0: The defendant is innocent**

- **The alternative hypothesis or research hypothesis is**
- **H_1: The defendant is guilty**

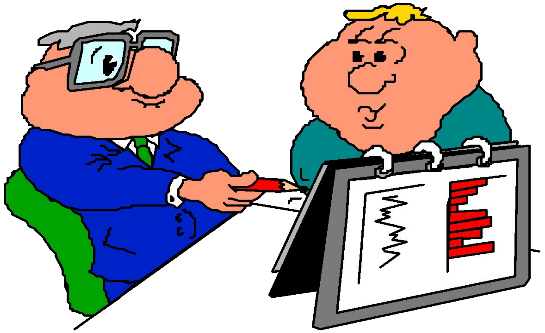
- **The jury does not know which hypothesis is true. They must make a decision on the basis of evidence presented.**

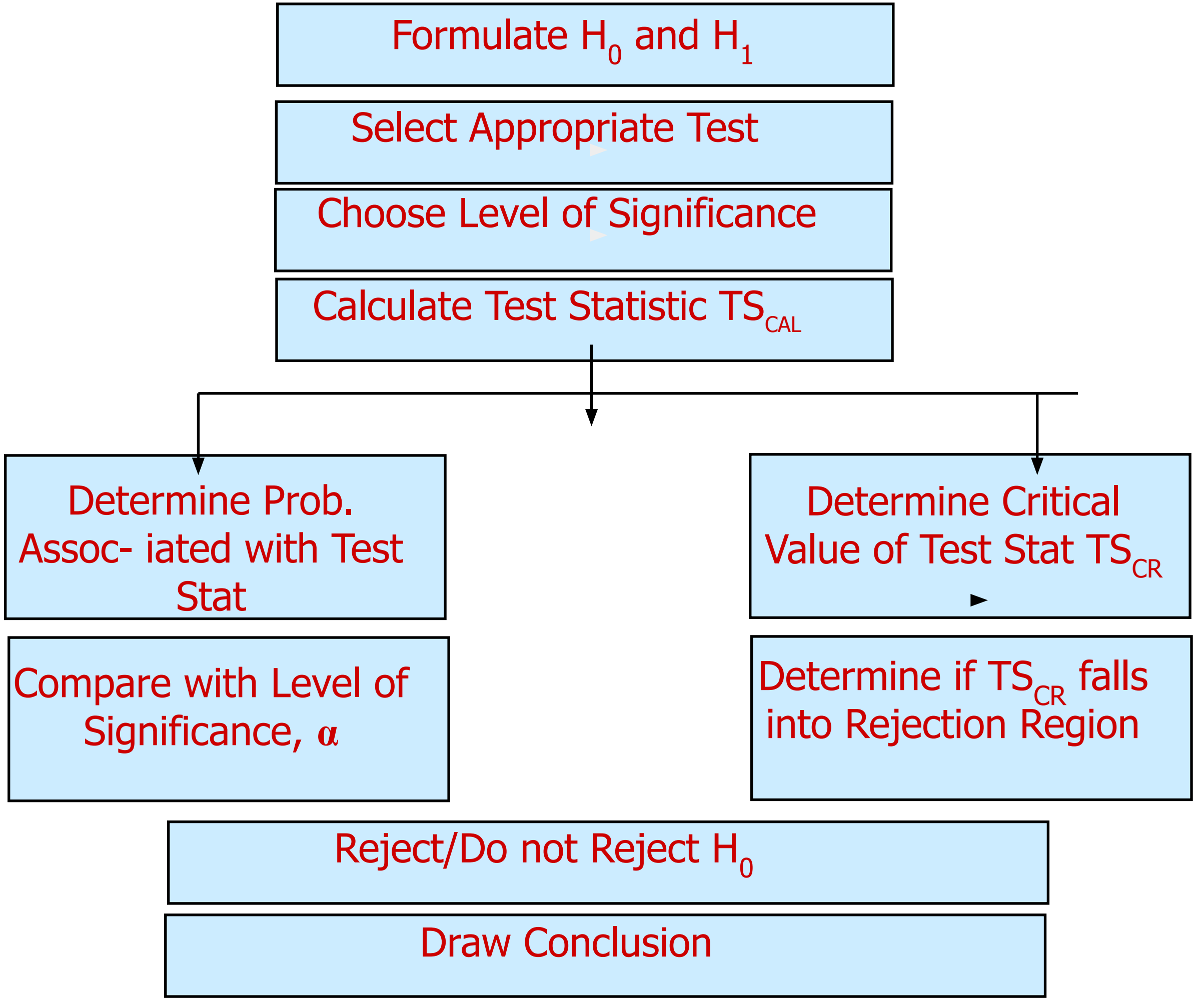
Hypothesis Testing…

- **In the language of statistics accusing the defendant is called *rejecting the null hypothesis in favor of the alternative hypothesis*. That is, the jury is saying that there is enough evidence to conclude that the defendant is guilty (i.e., there is enough evidence to support the alternative hypothesis).**

- **If the jury releases it is stating that *there is not enough evidence to support the alternative hypothesis*. Notice that the jury is not saying that the defendant is innocent, only that there is not enough evidence to support the alternative hypothesis.**

Steps in Test of Hypothesis





Formulate H_0 and H_1

Select Appropriate Test

Choose Level of Significance

Calculate Test Statistic TS_{CAL}

Determine Prob. Assoc- iated with Test Stat

Compare with Level of Significance, α

Determine Critical Value of Test Stat TS_{CR}

Determine if TS_{CR} falls into Rejection Region

Reject/Do not Reject H_0

Draw Conclusion

Concepts of Hypothesis Testing

- The **two** possible decisions that can be made:

□ Conclude that there ***is enough evidence*** to support the alternative hypothesis

(also stated as: **reject null hypothesis in favor of the alternative**)

□ Conclude that there ***is not enough evidence*** to support the alternative hypothesis

(also stated as: failing to reject the null hypothesis in favor of the alternative)

NOTE: Generally we **do not** say that we **accept** the null hypothesis……

Example of a Hypothesis Test

- The coordinator of CDPA about the cost of textbooks during a semester. A sample of 100 students enrolled in the Department indicates, sample average cost of Rs. 3050 with a sample S.D. of Rs. 150. Using 5% level of significance, is there evidence that the population average is significantly differ than Rs.3000?

Step 1: Setup Hypothesis

- The coordinator of CDPA about the cost of textbooks during a semester. A sample of 100 students enrolled in the Department indicates, sample average cost of Rs. 3050 with a sample S.D. of Rs. 150. Using 5% level of significance, is there evidence that the population average is significantly differ than Rs.3000?

The hypotheses may be formulated as:

$$\textit{Null hypothesis: } H_0: \quad \mu = 3000$$

$$\textit{Alternative hypothesis: } H_1: \quad \mu \neq 3000$$

Step 2: Chose Test Statistics

- The coordinator of CDPA about the cost of textbooks during a semester. A sample of 100 students enrolled in the Department indicates, sample average cost of Rs. 3050 with a sample S.D. of Rs. 150. Using 5% level of significance, is there evidence that the population average is significantly differ than Rs.3000?

The hypotheses may be formulated as:

Null hypothesis: H_0: $\mu = 3000$

Alternative hypothesis: H_1: $\mu \neq 3000$

$$Z = \frac{\text{Statistics} - E(\text{Statistics})}{\text{Standard Error}}$$

The Test statistics value will be : $Z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \dfrac{3050 - 3000}{150 / \sqrt{100}} = 3.33$

- The **test statistic** measures how close the sample has come to the null hypothesis.
- The test statistic often follows a well-known distribution (eg, normal, *t*, or chi-square).
- In our example, the *z* statistic, which follows the standard normal distribution, would be appropriate.

Since we are testing whether the mean value is differ from 3000. The Test Statistics will be Z value.

Step 3: Level of Significance

- The coordinator of CDPA about the cost of textbooks during a semester. A sample of 100 students enrolled in the Department indicates, sample average cost of Rs. 3050 with a sample S.D. of Rs. 150. Using 5% level of significance, is there evidence that the population average is significantly differ than Rs.3000?

The hypotheses may be formulated as:

Null hypothesis: H_0: $\mu = 3000$

Alternative hypothesis: H_1: $\mu \neq 3000$

The Test statistics value will be : $Z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \dfrac{3050 - 3000}{150 / \sqrt{100}} = 3.33$

The Level of Significance is = α = 0.05

Step 4: Critical Value

- The coordinator of CDPA about the cost of textbooks during a semester. A sample of 100 students enrolled in the Department indicates, sample average cost of Rs. 3050 with a sample S.D. of Rs. 150. Using 5% level of significance, is there evidence that the population average is significantly differ than Rs.3000?

The hypotheses may be formulated as:

$$\textit{Null hypothesis: } H_0\text{:} \quad \mu = 3000$$

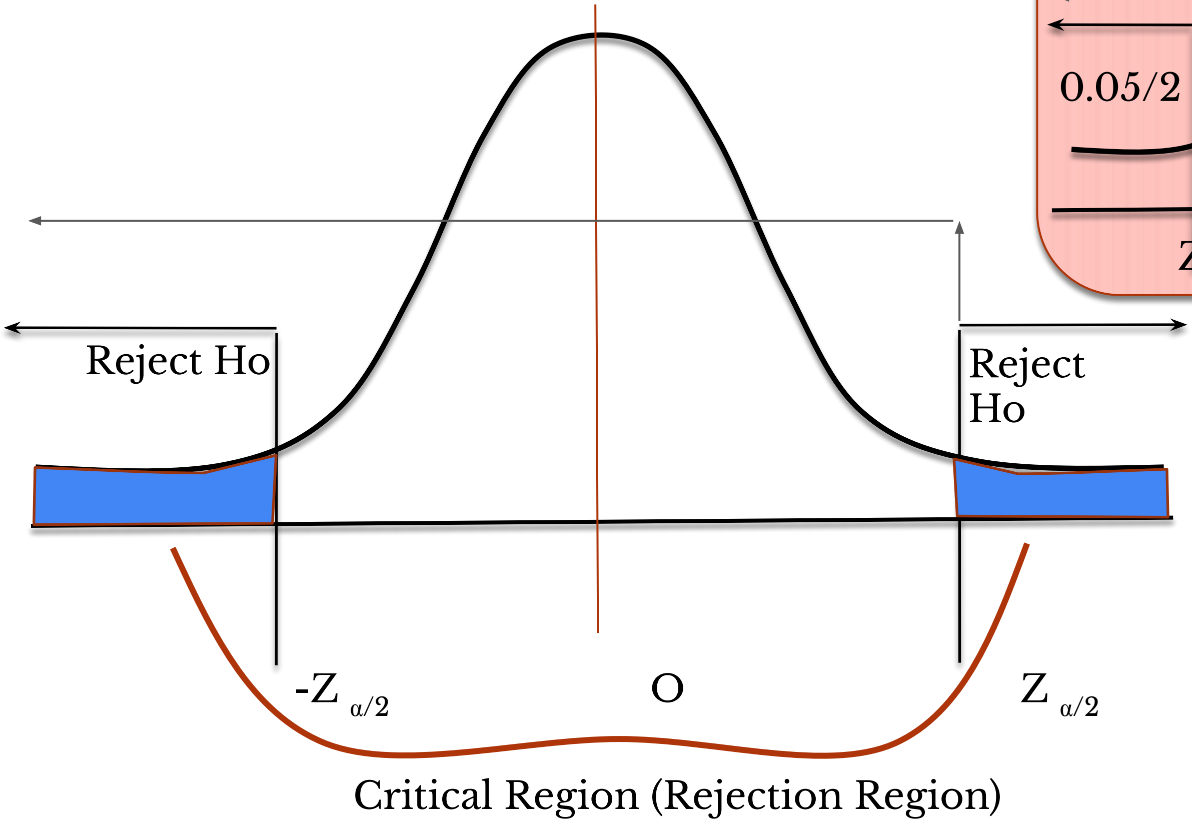
$$\textit{Alternative hypothesis: } H_1\text{:} \quad \mu \neq 3000$$

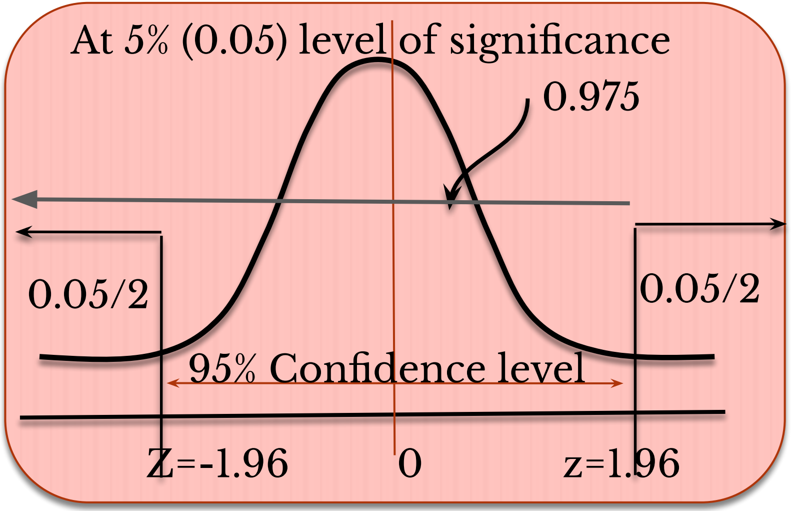
The Test statistics value will be : $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \dfrac{3050 - 3000}{150/\sqrt{100}} = 3.33$

The Level of Significance is = α = 0.05

The Critical value at 5% level of significance can be determined using Table values.

Step 4: Critical Value

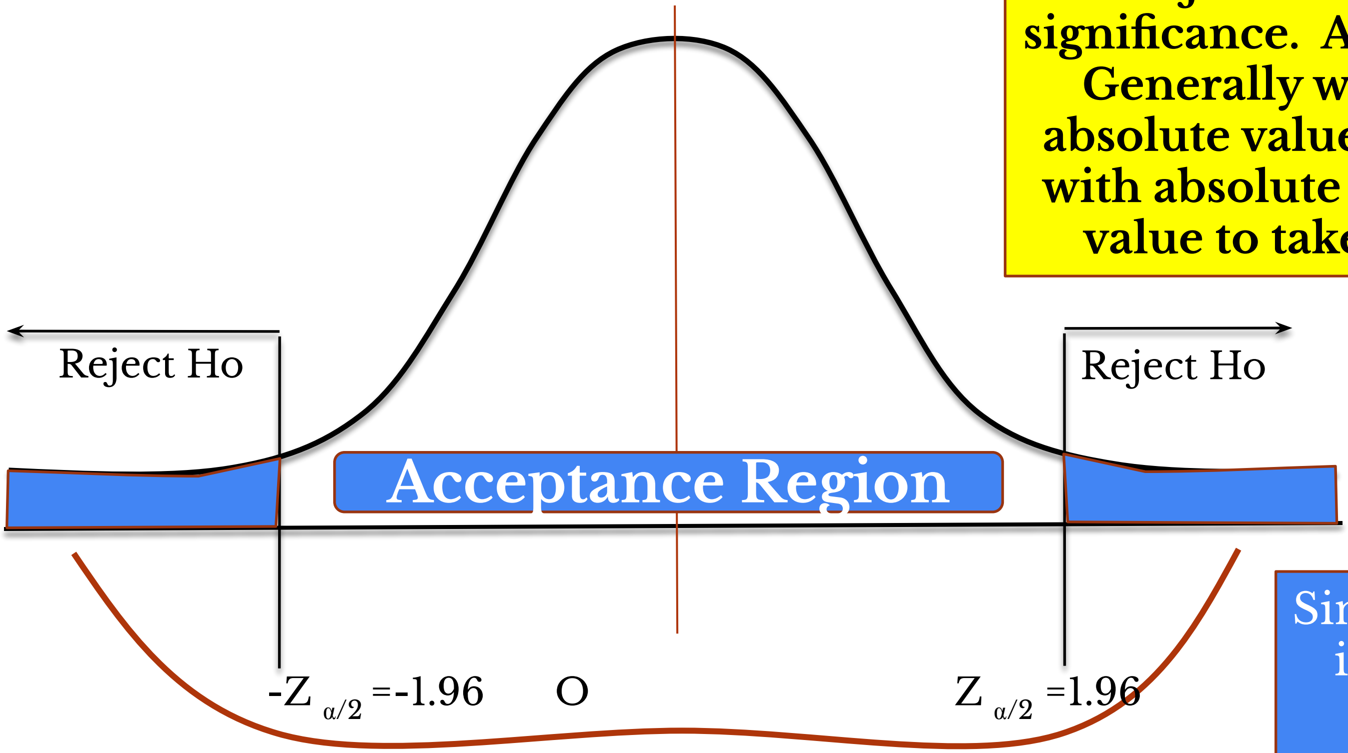




Note, in determining the critical value of the test statistic, the area to the right of the critical value is either α or α/2. It is α for a one-tail test and α/2 for a two-tail test.

Step 5: Decision Rule

If Critical Value Lies Between -1.96 to +1.96 , H_1 will be rejected at 5% level of significance. Accept Otherwise. Generally we compare the absolute value of test statistics with absolute value of Critical value to take this decision.



Critical Region (Rejection Region)

Since 3.33 > 1.96 Ho is Rejected at 5% level of significance.

Step 5: Decision Rule

- Since 3.33 > 1.96 Ho is Rejected at 5% level of significance.
- there is strong evidence that the population average is significantly differ than Rs.3000.

Directional Tests

- When a research study predicts a specific direction for the treatment effect (increase or decrease), it is possible to incorporate the directional prediction into the hypothesis test.
- The result is called a **directional test** or a **one-tailed test**. A directional test includes the directional prediction in the statement of the hypotheses and in the location of the critical region.

Directional Tests (cont.)

- For example, if the original population has a mean of μ = 80 and the treatment is predicted to increase the scores, then the null hypothesis would state that after treatment:

  H0: $\mu \leq 80$ (there is no increase)

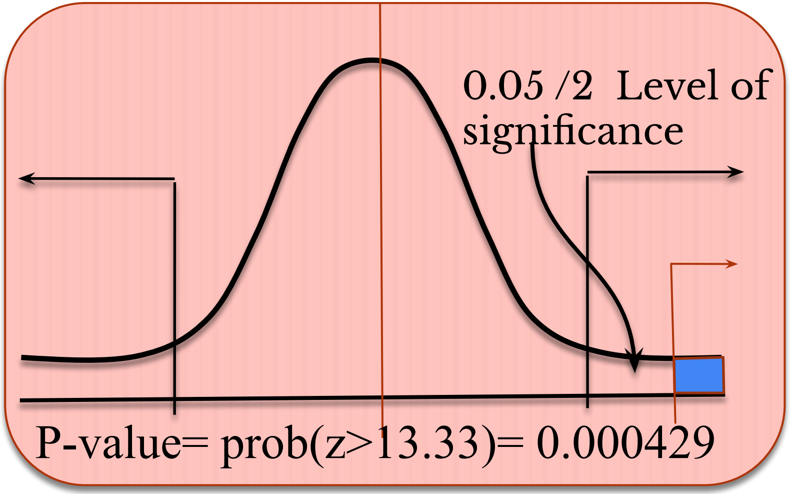
- In this case, the entire critical region would be located in the right-hand tail of the distribution because large values for μ would demonstrate that there is an increase and would tend to reject the null hypothesis.

Step 4,5: Determine Probability Value (P-value) approach

- Using standard normal tables the area to the right of z_{CAL} is .000429 (z_{CAL} =3.33)
- The shaded area between 0 and 3.33 is 0.499571. Therefore, the area to the right of 3.33 is

  0.5 - 0.499571 = .000429.

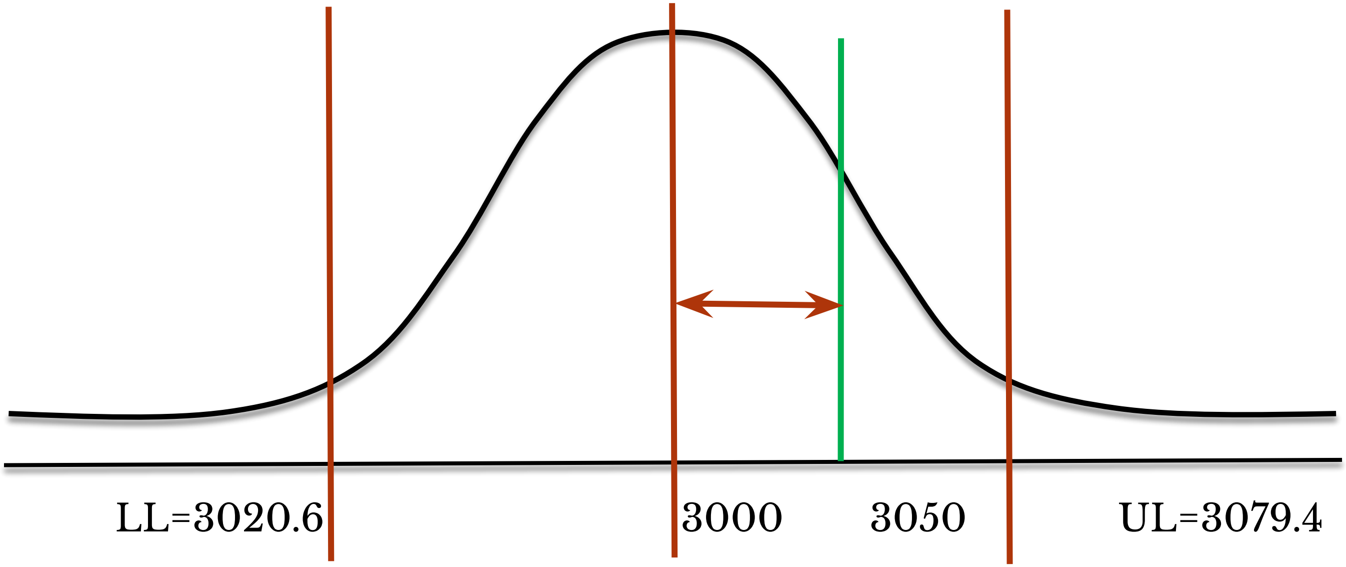
- Thus, the p-value is .000429
- While comparing p-value with level of significance, compare with α for one tail and compare with α /2 for two tail test.


0.05 /2 Level of significance
P-value= prob(z>13.33)= 0.000429

Note: Different software use 2x p-value to compare with level of significance for two tail test.

Use of Interval Estimation for Decision

- Since 3050 lies between the range of 3020.6 to 3079.4, we can accept that population mean can be 3050.



Decision Rule

- If the probability associated with the calculated value of the test statistic (P-value) is less than the level of significance ($\alpha/2$) for two tail and α for one tail, the null hypothesis is rejected.

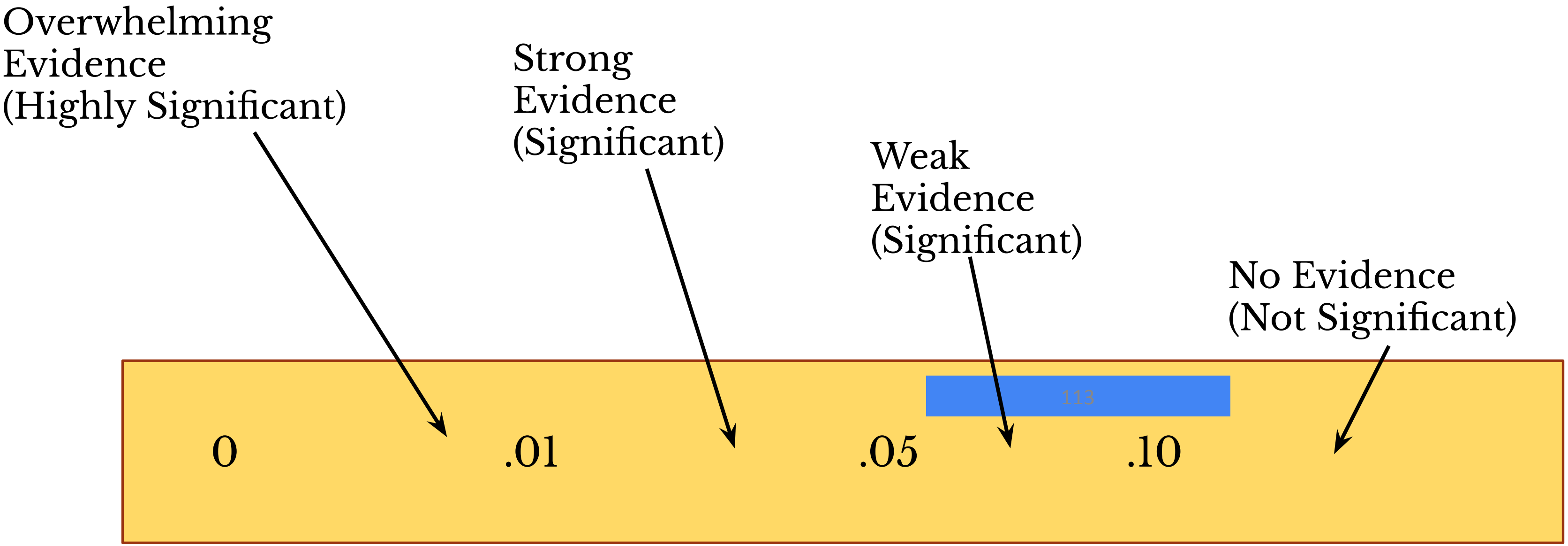
- Alternatively, if the calculated value of the test statistic is greater than the critical value of the test statistic (z_{α}), the null hypothesis is rejected.

- Or, if parametric value doesn’t lies between the Interval estimation, the null hypothesis is rejected.

Interpreting the p-value…

- The smaller the p-value, the more statistical evidence exists to support the alternative hypothesis.
- If the p-value is less than 1%, there is ***overwhelming evidence*** that supports the alternative hypothesis.
- If the p-value is between 1% and 5%, there is a ***strong evidence*** that supports the alternative hypothesis.
- If the p-value is between 5% and 10% there is a ***weak evidence*** that supports the alternative hypothesis.
- If the p-value exceeds 10%, there is ***no evidence*** that supports the alternative hypothesis.

Interpreting the p-value…



Overwhelming Evidence (Highly Significant)

Strong Evidence (Significant)

Weak Evidence (Significant)

No Evidence (Not Significant)

Type I and Type II Errors

Conclusion	Population Condition	
	H_0 True ($\mu \leq 12$)	**H_0 False** ($\mu > 12$)
Accept H_0 (Conclude $\mu \leq 12$)	Correct Decision	Type II Error
Reject H_0 (Conclude $\mu > 12$)	Type I Error	Correct Decision

- A Type I error occurs when we *reject* a *true* null hypothesis (i.e. Reject H_0 when it is TRUE)
- A Type II error occurs when we *don't reject* a *false* null hypothesis (i.e. Do NOT reject H_0 when it is FALSE)

Example: one tail test

- **A department store manager determines that a new billing system will be cost-effective only if the mean monthly account is *more than Rs*17000.**

- **A random sample of 400 monthly accounts is drawn, for which the sample mean is Rs17800. The accounts are approximately normally distributed with a standard deviation of Rs 6500.**

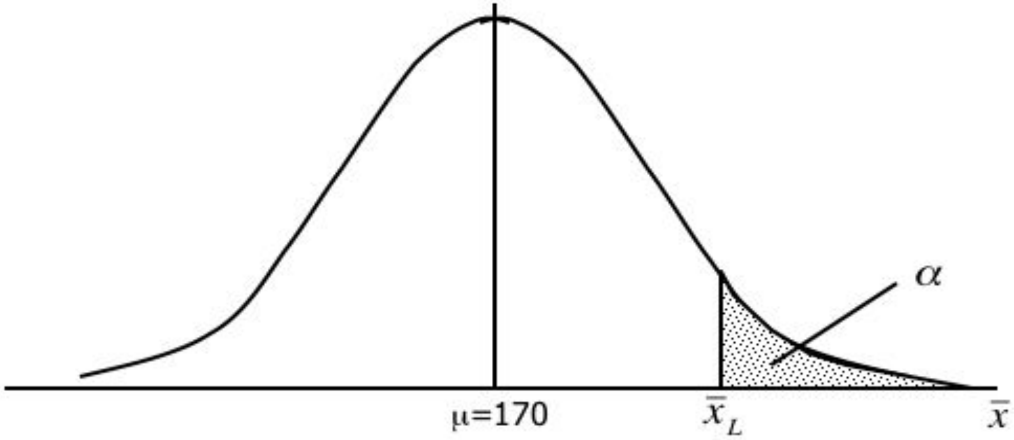
- ***Can we conclude that the new system is cost-effective?***

Example: one tail test

- The system will be cost effective if the mean account balance for all customers is greater than Rs17000.
- We express this belief as a our research hypothesis, that is:
- H_1: $\mu > 17000$ (this is what we want to determine)
- Thus, our null hypothesis becomes:
- H_0: $\mu = 17000$ (this specifies a single value for the parameter of interest)
  - Actually H_0: $\mu \leq 17000$

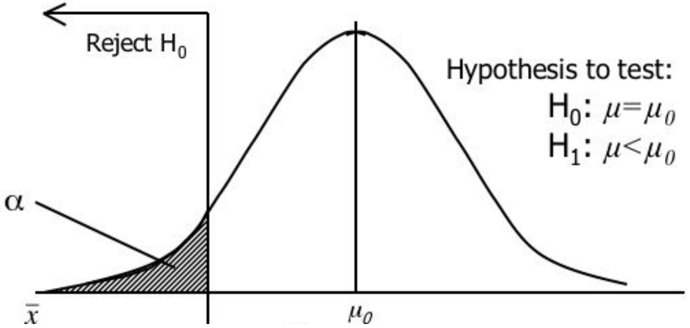
Example Rejection Region…(one tail)

- The ***rejection region*** is a range of values such that if the test statistic falls into that range, we decide to reject the null hypothesis in favor of the alternative hypothesis.



Rejection Region $\boxed{\bar{x} > \bar{x}_L}$

$\bar{x}_L$ is the critical value of $\bar{x}$ to reject H_0.



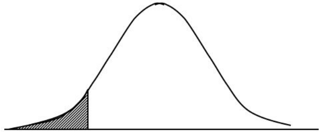
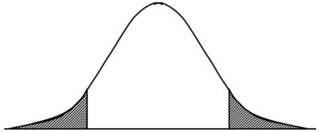
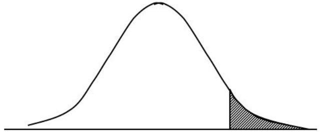
Hypothesis to test:
H_0: $\mu=\mu_0$
H_1: $\mu<\mu_0$

Conclusions of a Test of Hypothesis…

- If we reject the null hypothesis, we conclude that there is enough evidence to infer that the alternative hypothesis is true.

- If we fail to reject the null hypothesis, we conclude that there is not enough statistical evidence to infer that the alternative hypothesis is true. This does not mean that we have proven that the null hypothesis is true!

One- and Two-Tail Tests…

One-Tail Test (left tail)	Two-Tail Test	One-Tail Test (right tail)
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$
		

Hypothesis testing (Z,t F, Anova, Chi square)

- Z test
  - Test of significance of single mean
  - Test of significance of double mean
  - Test of significance of single proportion
  - Test of significance of double proportion

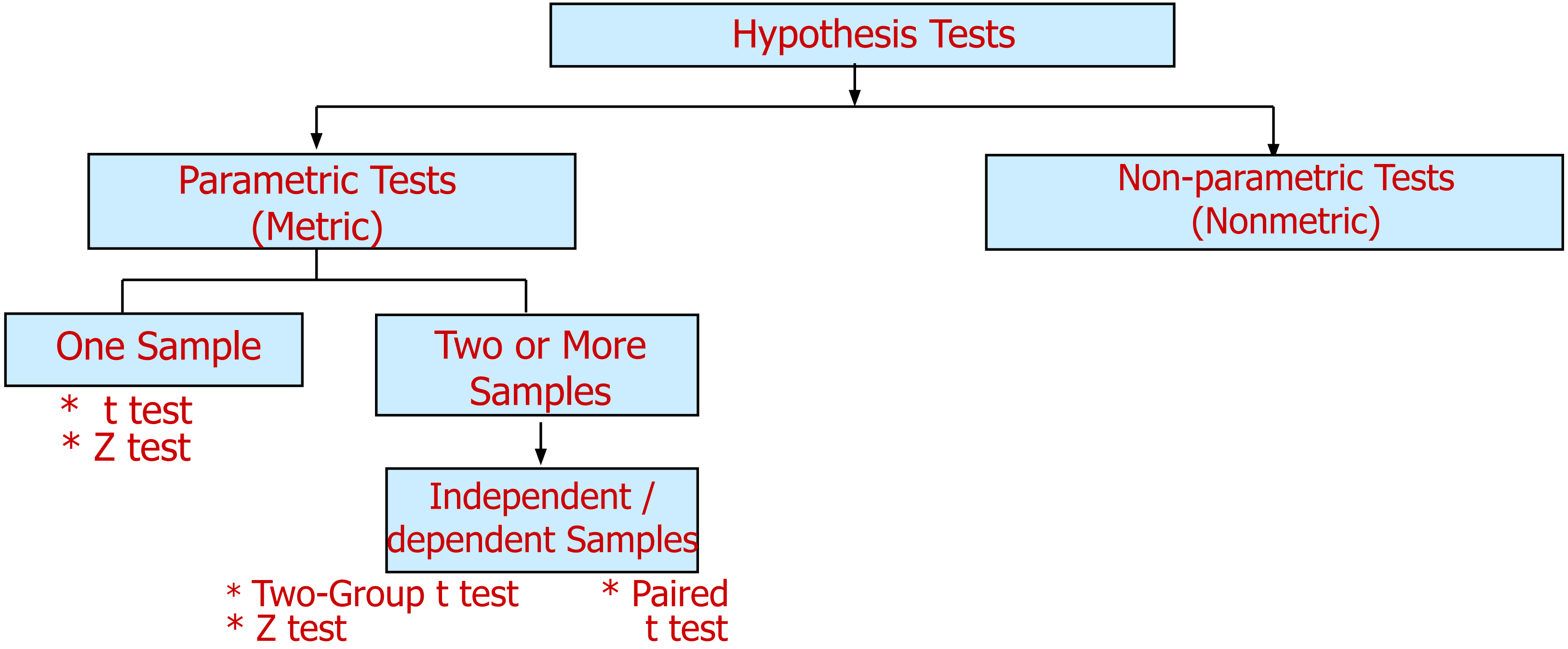
- T test
  - Test of significance of single mean
  - Test of significance of double mean (independent)
  - Test of significance of pair t-test (dependent two mean)

Hypothesis testing (Z,t F, Anova, Chi square)

- F test
  - Test of significance of more than two mean (ANOVA)
  - Test of significance of standard deviation

- Chi Square test
  - Test of significance of independence of attribute.
  - Test of goodness of fit.

Hypothesis Testing for Differences



Hypothesis Tests

Parametric Tests (Metric)

Non-parametric Tests (Nonmetric)

One Sample

* t test
* Z test

Two or More Samples

Independent / dependent Samples

* Two-Group t test
* Z test

* Paired t test

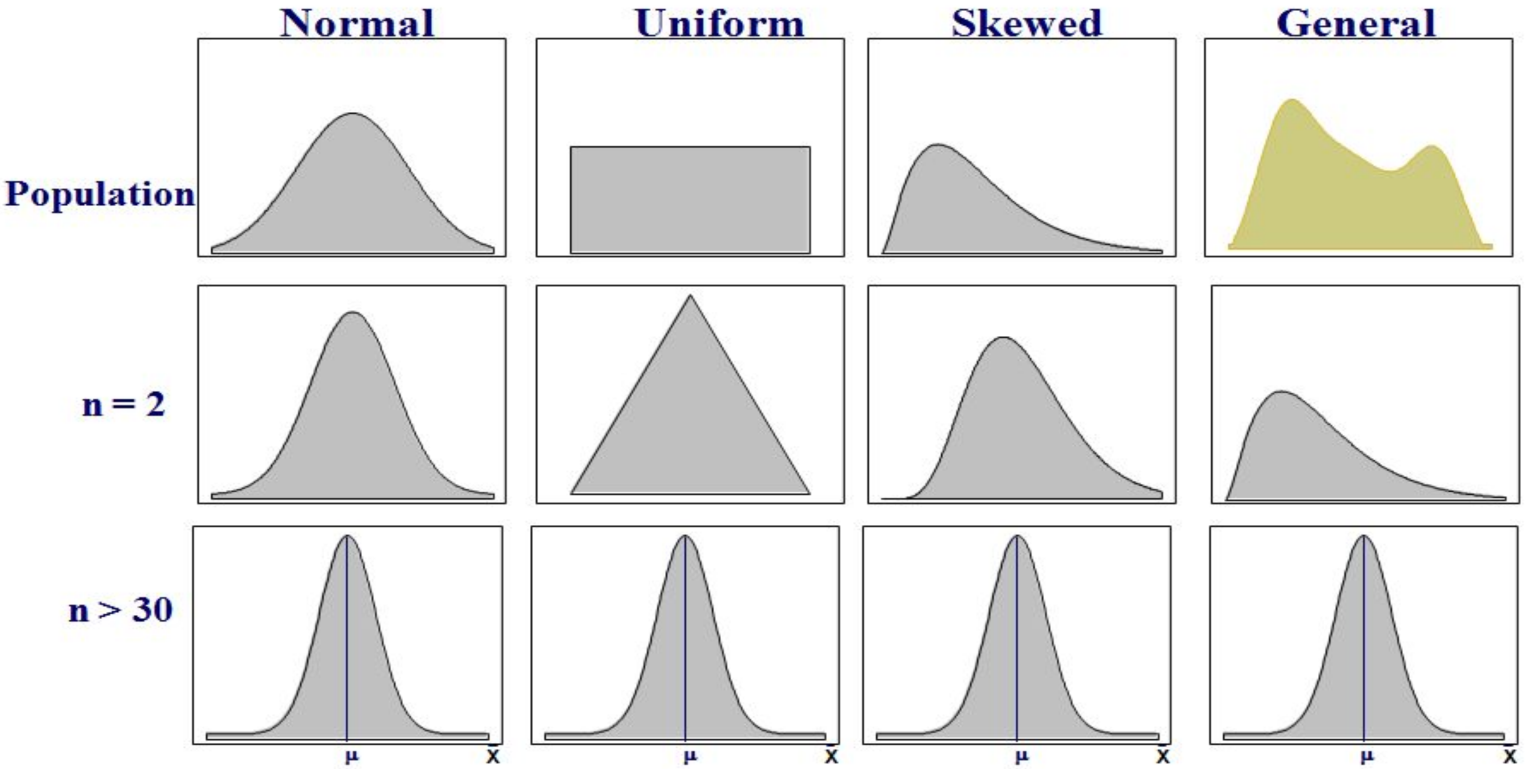
Question for practice

Kathmandu University and Kathmandu Metropolitan city jointly administer a student volunteerism program. Last year, students in the program volunteered an average of 7.3 hours of community service per month. Officials are concerned that students might not be putting in as many volunteer hours this year. A random sample of 75 student volunteers from the first 2 months of the year reveals an average of 6.8 hours of community service (s = 1.5 hours). Based on these data, what can program administrators conclude about their initial hypothesis (i.e., that student volunteering is decreasing)?

Question

The Ministry of Education wants to know the average days of absences for students of government schools in Nepal. The officials believe that the number of days of absence is 12. A sample of 150 students was observed and found that the average number of days a student was absent was found as 11 days with standard deviation of 3.2 days. What can you tell from this information about the official hypothesis?

Small Sample n<30

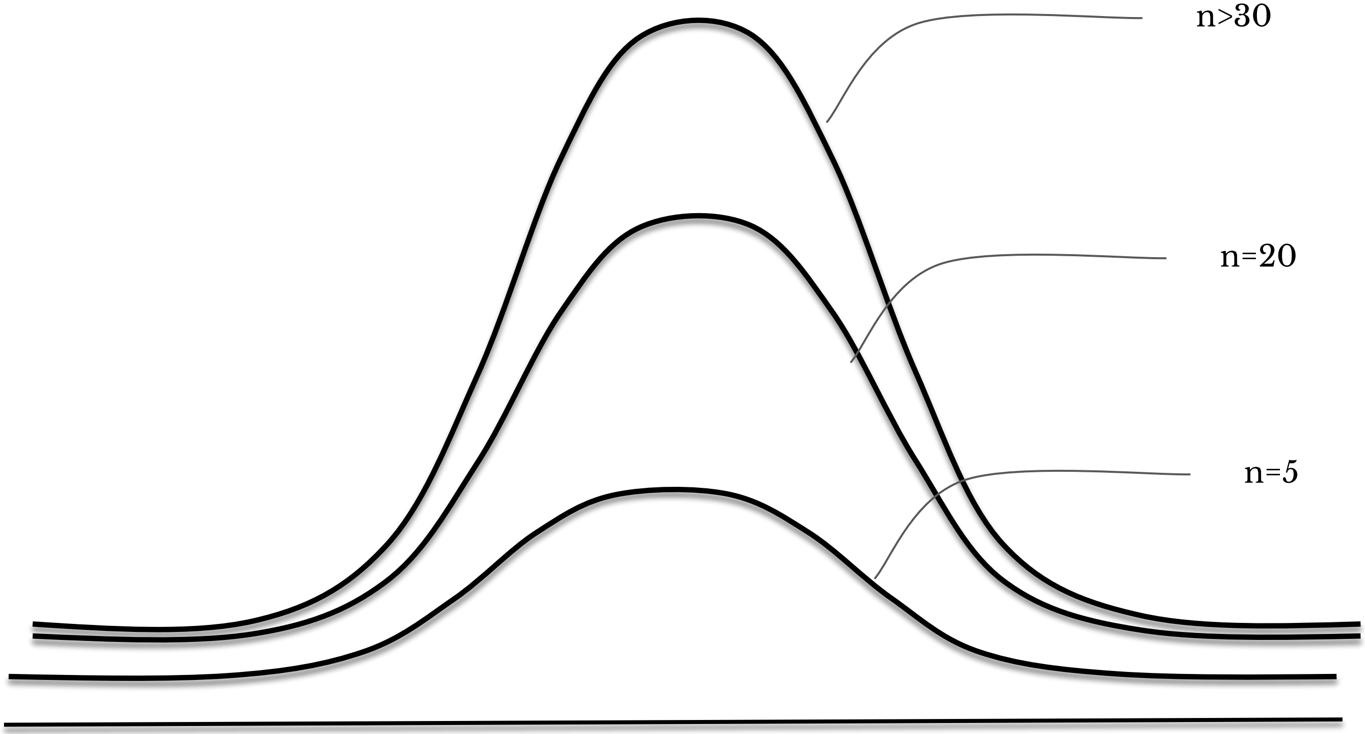


Central Limit Theorem

When sampling from a population with mean μ and finite standard deviation σ, the sampling distribution of the sample mean will tend to a normal distribution with mean μ and standard deviation σ/√n as the sample size becomes large (*n* >30).

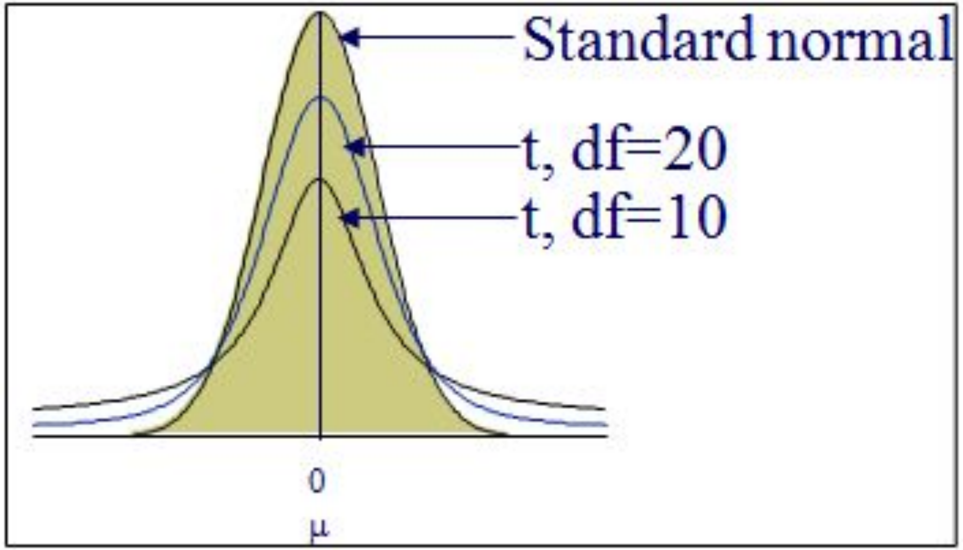
For “large enough” *n*: $\bar{X} \sim N(\mu, \sigma^2 / n)$

What happen when sample size decreases



Small sample test

- The *t* is a family of bell-shaped and symmetric distributions, one for each number of degree of freedom.
- The expected value of *t* is 0.
- The variance of *t* is greater than 1, but approaches 1 as the number of degrees of freedom increases. The *t* is flatter and has flatter tails than does the standard normal.
- The *t* distribution approaches a standard normal as the number of degrees of freedom increases.



If the population standard deviation, σ, is ***unknown***, replace σ with the sample standard deviation, *s*. If the population is normal, the resulting statistic:

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

has a ***t distribution* with *(n - 1) degrees of freedom***.

Example of a Hypothesis Test

- The coordinator of MTech about the cost of textbooks during a semester. A sample of 10 students enrolled in the Department indicates, sample average cost of Rs. 3050 with a sample S.D. of Rs. 150. Using 5% level of significance, is there evidence that the population average is significantly differ than Rs.3000?

Step 1: Setup Hypothesis

- The coordinator of MTech about the cost of textbooks during a semester. A sample of 10 students enrolled in the Department indicates, sample average cost of Rs. 3050 with a sample S.D. of Rs. 150. Using 5% level of significance, is there evidence that the population average is significantly differ than Rs.3000?

The hypotheses may be formulated as:

$$\textit{Null hypothesis: } H_0: \quad \mu = 3000$$

$$\textit{Alternative hypothesis: } H_1: \quad \mu \neq 3000$$

Step 2: Chose Test Statistics

- The coordinator of MTech about the cost of textbooks during a semester. A sample of 10 students enrolled in the Department indicates, sample average cost of Rs. 3050 with a sample S.D. of Rs. 150. Using 5% level of significance, is there evidence that the population average is significantly differ than Rs.3000?

$$t = \frac{\underline{\text{Statistics} - \text{E(Statistics)}}}{\text{Standard Error}}$$

The hypotheses may be formulated as:

Null hypothesis: H_0: $\mu = 3000$

Alternative hypothesis: H_1: $\mu \neq 3000$

The Test statistics value will be : $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}} = \dfrac{3050 - 3000}{150/\sqrt{10}} = 1.826$

- The **test statistic** measures how close the sample has come to the null hypothesis.
- The test statistic often follows a well-known distribution (eg, normal, *t*, or chi-square).
- In our example, the *t* statistic, which follows the t distribution, with degree of freedom 10-1=9

Since we are testing whether the mean value is differ from 3000. The Test Statistics will be t value.

Step 4: Critical Value

- The coordinator of MTech about the cost of textbooks during a semester. A sample of 100 students enrolled in the Department indicates, sample average cost of Rs. 3050 with a sample S.D. of Rs. 150. Using 5% level of significance, is there evidence that the population average is significantly differ than Rs.3000?

The hypotheses may be formulated as:

Null hypothesis: H_0: $\mu = 3000$

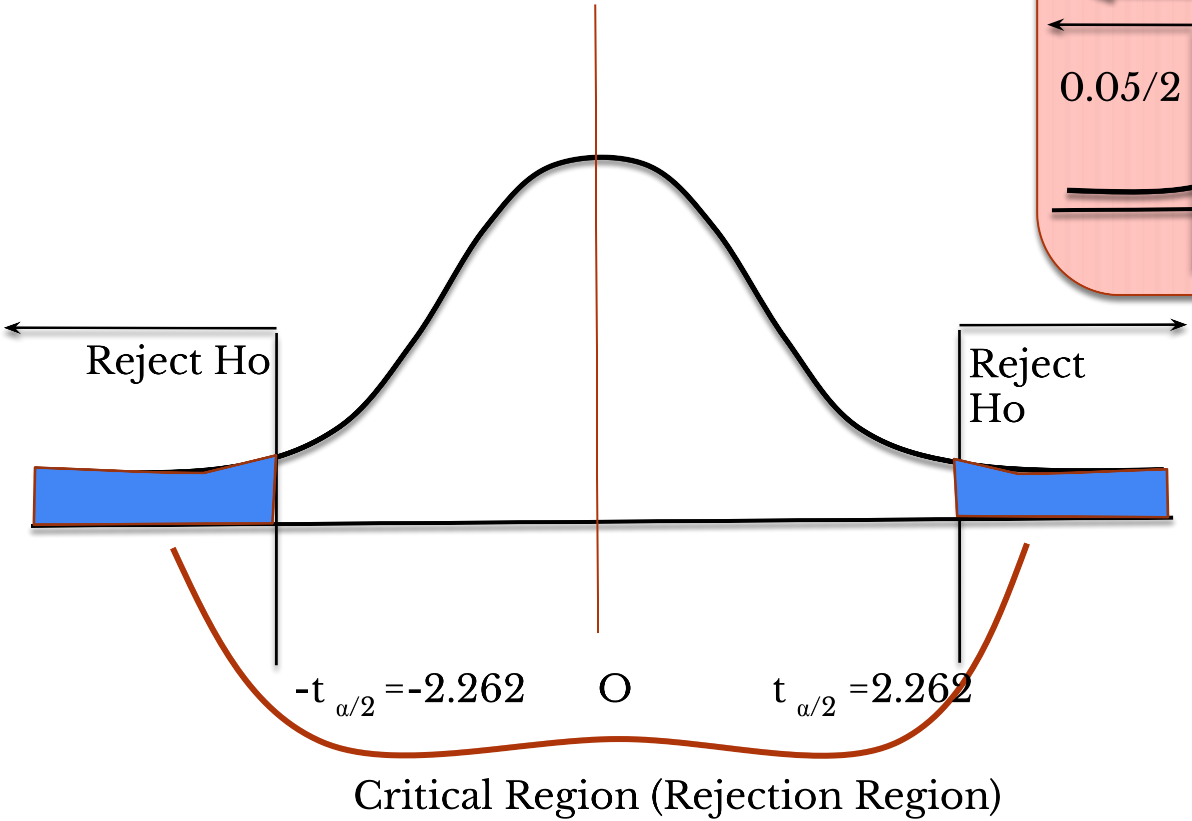
Alternative hypothesis: H_1: $\mu \neq 3000$

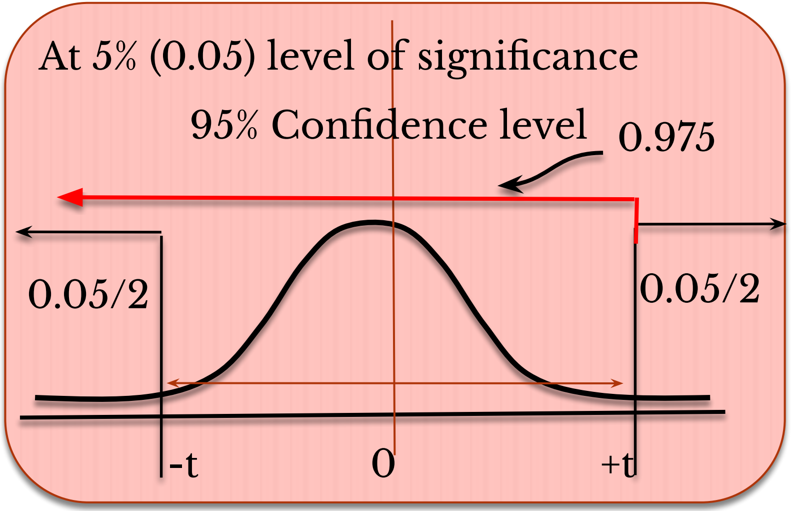
The Test statistics value will be : t = 1.826

The Level of Significance is = α = 0.05 with degree of freedom 9.

The Critical value at 5% level of significance can be determined using Table values.

Step 4: Critical Value

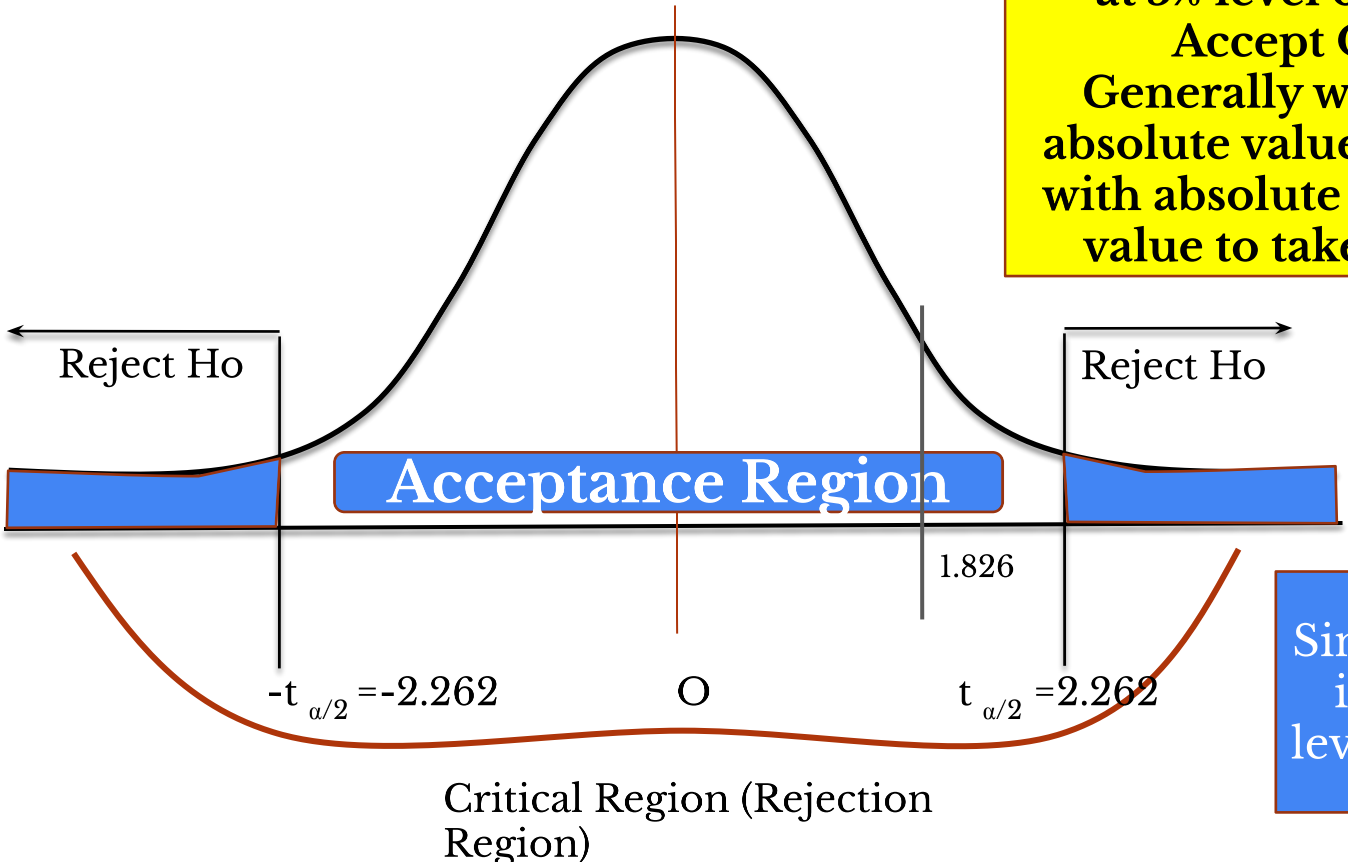




Note, in determining the critical value of the test statistic, the area to the right of the critical value is either α or α/2. It is α for a one-tail test and α/2 for a two-tail test.

Step 5: Decision Rule

If Critical Value Lies Between
$-t_{\alpha/2}$ to $+t_{\alpha/2}$, H_1 will be rejected
at 5% level of significance.
Accept Otherwise.
Generally we compare the
absolute value of test statistics
with absolute value of Critical
value to take this decision.



Since 3.33 > 1.96 Ho
is Rejected at 5%
level of significance.

Step 5: Decision Rule

- Since 1.826 < 2.262 H_1 is Rejected at 5% level of significance.
- there is no evidence that the population average is significantly differ than Rs.3000.

Degree of freedom

Consider a sample of size n=4 containing the following data points:

x_1=10 x_2=12 x_3=16 x_4=?

and for which the sample mean $\bar{x} = \frac{\sum x}{n} = 14$

Given the values of three data points and the sample mean, the value of the fourth data point can be determined:

In other words the three data points can be selected **freely** to get the mean value of 14 with the size of data points n=4. So degree of freedom is the number of data points can be selected freely to get the desired statistics.

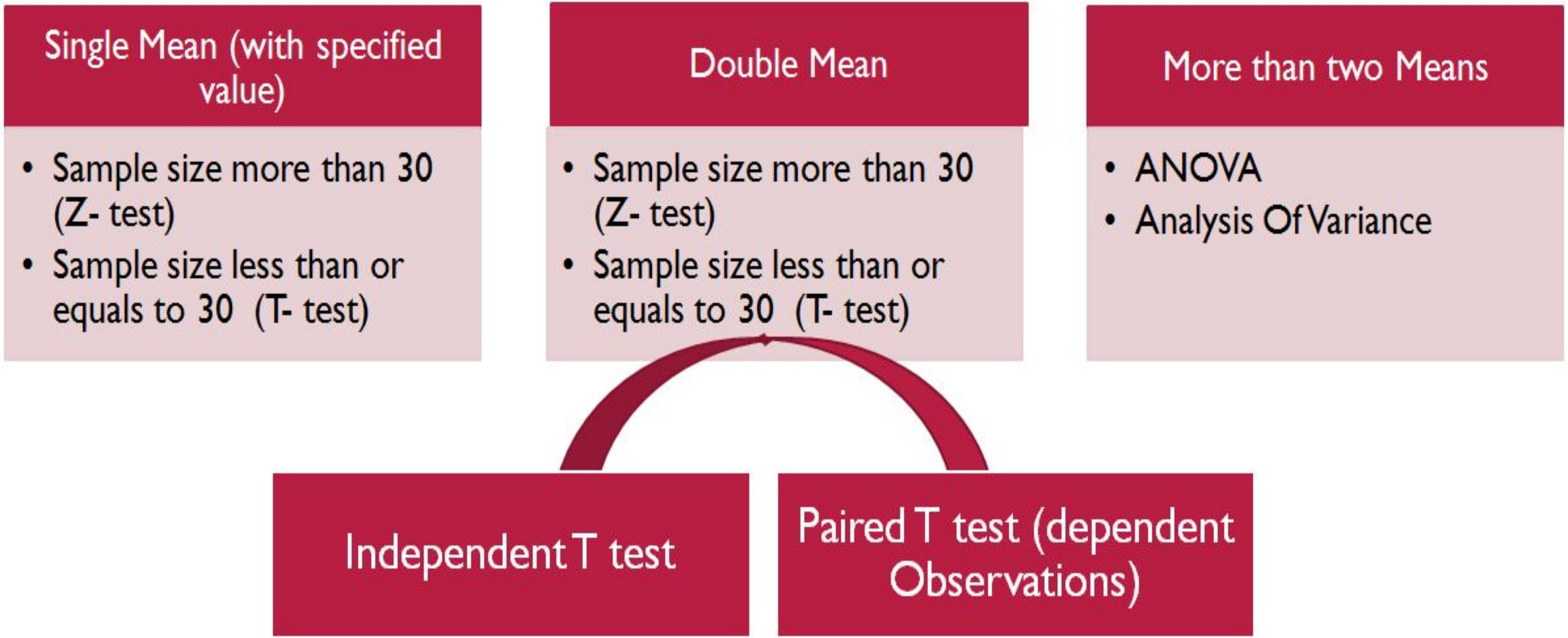
Degree of freedom

The number of ***degrees of freedom*** is equal to the total number of measurements (these are not always raw data points), less the total number of *restrictions* on the measurements. A restriction is a quantity computed from the measurements.

The sample mean is a restriction on the sample measurements, so after calculating the sample mean there are only ***(n-1)*** **degrees of freedom** remaining with which to calculate the sample variance. The sample variance is based on only *(n-1)* free data points:

$$s^2 = \frac{\sum (x - \bar{x})^2}{(n-1)}$$

Significance of mean / s



Single Mean (with specified value)

- Sample size more than 30 (Z- test)
- Sample size less than or equals to 30 (T- test)

Double Mean

- Sample size more than 30 (Z- test)
- Sample size less than or equals to 30 (T- test)

More than two Means

- ANOVA
- Analysis Of Variance

Independent T test

Paired T test (dependent Observations)

Significance of two means (n1, n2 >=30)

Example

An Intelligence tests on two groups of people gave the following results

	Group A	Group B
Nos of person	80	120
Average Intelligence score	78	75
SD of average Int. Score	12	15

Test the hypothesis that the average Intelligence score is significantly different.

Problem

	Group A	Group B
Nos of person	80	120
Average	78	75
SD	12	15

1. **Hypothesis setup**

Null Hypothesis: H0: $\mu_1 = \mu_2$ There is no significance difference between two average value (average Intelligence score)

Alternative Hypothesis: H0: $\mu_1 \neq \mu_2$ There is no significance difference between two average value.

2. Test Statistics

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}} = \frac{78 - 75}{\sqrt{(144/80 + 225/120)}} = 1.565$$

3. Level of significance: $\alpha = 0.05$

4. Critical value: at 5% risk the critical value is 1.96 (two tail)

5 Decision: Since z cal < z critical, So there is no significance difference between two average value

Significance of two means (n1 or n2 <=30)

Example

Kathmandu University Library would like to increase the subscription of online library. For it, Library has conducted training seminars. to test the effectiveness of the seminar a study has conducted and following results were obtained.

	experimental group	control group
Nos of person	8	10
Mean nos of downloads	50	20
SD of nos of downloads	12	6

Test the hypothesis that the average nos of downloads in experimental group is higher than control group.

Problem

	experimental group	control group
Mean nos of downloads	50	20
SD of nos of downloads	12	6

1. **Hypothesis setup**

Null Hypothesis: H0: $\mu_1=\mu_2$ There is no significance difference between two average value (average nos of download of two groups)

Alternative Hypothesis: H0: $\mu_1>\mu_2$ There is significantly higher nos of downloads in experimental group.

t-test, two means, equal variance assumed

2. Test Statistics

$t= \dfrac{\bar{x}_1 - \bar{x}_2}{S*\surd(1/n_1+ 1/n_2)}$ with df= n_1+n_2-2 where $S^2= \dfrac{n_1 s_1^2+n_2 s_2^2}{n_1+n_2-2}$

Now $S^2= \dfrac{n_1 s_1^2+n_2 s_2^2}{n_1+n_2-2} = \dfrac{8*12^2+10*6^2}{8+10-2} = 94.5$

$t = \dfrac{50 - 20}{\surd 94.5*(1/8+ 1/10)} = 6.506$

Level of significance degree of freedom

3. **Level of significance** = α = if not given take it as 0.05

   Degree of freedom = n_1+n_2-2 = 8+10-2=16

4. **Critical Value** at 5% level of significance at 16 degree of freedom from table $t_{tab}=t(\alpha,16)=2.120$

5. **Decision:** Since $|t\ cal|=6.5>2.12=t_{tab}$ Ho is rejected at 5% level of significance.

   Conclusion: There is significance difference between two average value (average nos of download of two groups). The training is effective.

t-test, two means, unequal variance assumed

$$t= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(S_1{}^2/n_1 + s_2{}^2/n_2)}} \quad \text{with} \qquad df = \frac{(s_1^2/N_1 + s_2^2/N_2)^2}{(s_1^2/N_1)^2/(N_1 - 1) + (s_2^2/N_2)^2/(N_2 - 1)}$$

$$t= \frac{50 - 20}{\sqrt{(144/8 + 36/10)}} = 6.46 \qquad \text{with df} = \frac{(144/8 + 36/10)^2}{(144/8)^2/7 + (36/10)^2/9} = 9.77$$

Level of significance = 0.05

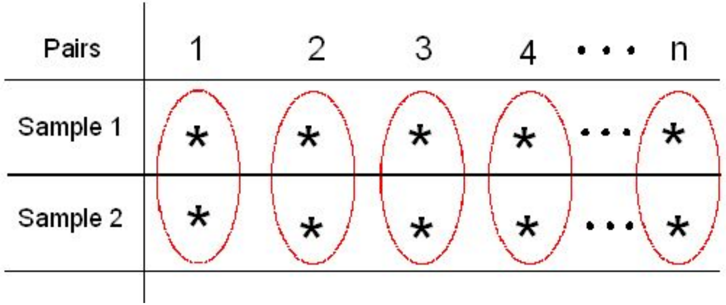
critical value = t(0.05/2,9.77)=2.262 (from Excel)

Decision: since t cal > t tab, Ho is rejected i.e H_1 is accepted

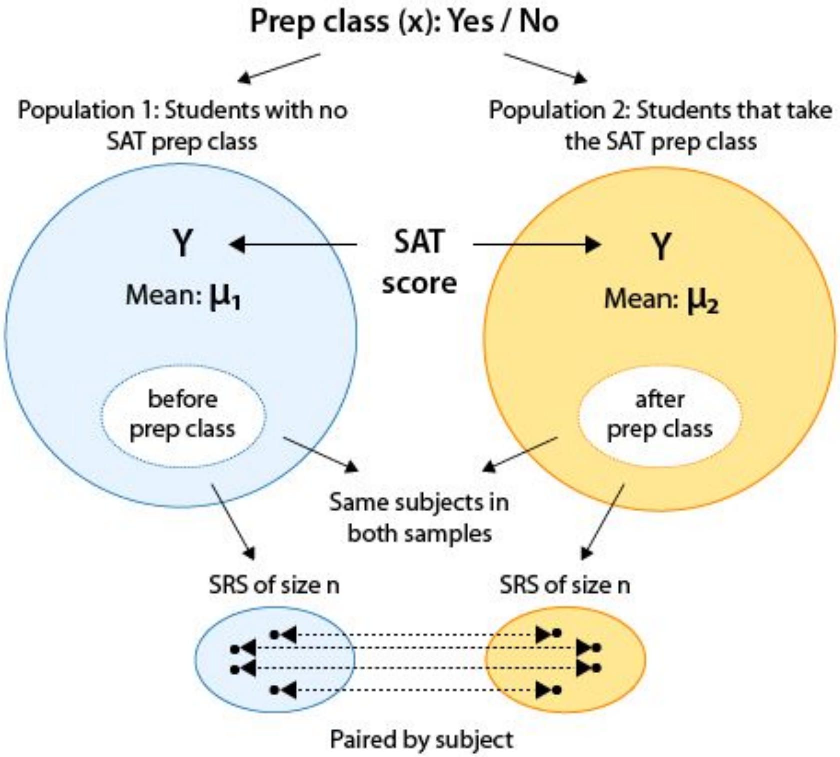
Significance difference of two means (dependent observations) Pair data

The idea behind the paired t-test is to **reduce** this **two-sample situation**, where we are comparing two means, **to a single sample situation** where we are doing inference on ingle mean, and **then use a simple t-test**.

we can easily reduce the raw data to a set of **differences** and conduct a **one-sample t-test**.



Pairs	1	2	3	4	• • •	n
Sample 1	*	*	*	*	• • •	*
Sample 2	*	*	*	*	• • •	*

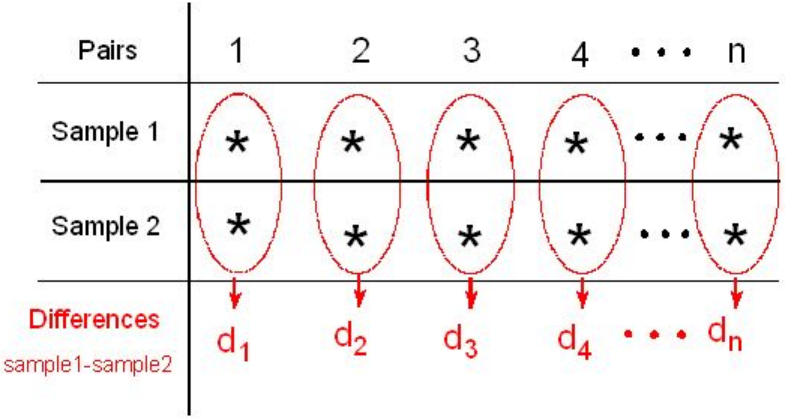


Pair - t -test

Ho: $\mu_d = 0$

There is no significance difference between samples (The difference is not significantly different than zero)

Ha: $\mu_d \neq 0$ or $\mu_d > 0$ or $\mu_d < 0$



A researcher is studying the influence of noise on one’s ability to solve statistics problems. The researcher randomly selects n = 10 students and exposes them to a noisy condition for 10 minutes and then a quiet condition for 10 minutes. In each condition, students are given a set of statistics problems to solve. The dependent variable is the number of mistakes made on the statistics problems during the ten minutes. Here, the researcher is testing a non-directional hypothesis, because she wants to know if there is any effect of noise on performance (errors); thus:

H0: µNoise = µQuiet

H1: µNoise ≠ µQuiet

	Condition	
	Noisy	**Quiet**
Stu	**(X_N)**	**(X_Q)**
A	9	6
B	9	7
C	6	7
D	7	5
E	6	4
F	7	4
G	9	6
H	11	9
I	7	5
J	9	7
	$\bar{x}_N = 8$	$\bar{x}_Q = 6$

Solution

Setup Hypothesis

H0: μNoise = μQuiet

H1: μNoise ≠ μQuiet

Test Statistics

$$t = \frac{\bar{d}}{S/\sqrt{n}} = \frac{2}{1.05/\sqrt{10}} = 5.477$$

Level of significance= 0.05 degree of freedom 9

Critical value: 2.262 p-value: 0.000

	Condition			
Stu	**Noisy (X_N)**	**Quiet (X_Q)**	**Diff $d=x_N-X_Q$**	**d^2**
A	9	6	3	9
B	9	7	2	4
C	6	7	-1	1
D	7	5	2	4
E	6	4	2	4
F	7	4	3	9
G	9	6	3	9
H	11	9	2	4
I	7	5	2	4
J	9	7	2	4
			$\Sigma d=20$	**$\Sigma d^2=50$**

$$\bar{d}=\Sigma d/n = 20/10 = 2$$

$$S = \sqrt{\frac{\sum d^2 - \frac{\left(\sum d\right)^2}{n}}{n-1}} = \sqrt{\frac{50 - \frac{(20)^2}{10}}{9}} = 1.05$$

Question

In a manufacturing company the new modern manager is in a belief that music enhances the productivity of the workers. He made observation on eight workers for a week and recorded the production before and after music was installed. From the data given below can one conclude that productivity has been changed due to music?

Employee	1	2	3	4	5	6	7	8
Without music	220	202	226	190	200	215	208	210
With music	236	190	240	200	220	205	212	215

Measurement Metric / Non-metric

Often a manager does not want to know the mean score of some population but rather the percentage of some population that does something.

A Department of traffic police, for example, might want to know the proportion of small vehicles that pass through Nagdhunga.

A criminal justice planner might want to know what percentage of persons released from prison will be arrested for another criminal act within 1 year.

A manager might want to know the percentage of volunteers who show up when they are scheduled. All these situations require the analyst to estimate a population proportion rather than a population mean.

z-Test for a Population Proportion

z-Test for a Population Proportion *P*

- A statistical test for a population proportion *P*.
- Can be used when a **binomial distribution** is given such that $np \geq 5$ and $nq \geq 5$.
- The **test statistic** is the sample proportion .
- The **standardized test statistic** is *z*.

$$Z= \frac{p - P}{\sqrt{(PQ/n)}}$$

Where p= sample proportion

P = Population proportion Q=1-P

Steps

1. State the claim mathematically and verbally. Identify the null and alternative hypotheses.
2. Find test statistics (z calculated value)
3. Specify the level of significance.
4. Determine the critical value.
5. Decision Making.

4. Find out P-value on z cal value
5. Decision Making.

Question

In a sample of 625 persons selected at random from a city, 48% were males. Calculate the standard error of sample proportion and test the hypothesis that males and females were in equal numbers in city at 1% level of significance.

Solution:

Here given,

Sample size $(n) = 625$

Sample proportion of males $(p) = 0.48$

Population proportions $= P = Q = 0.5$

SE of sample proportion, $SE(p)=\sqrt{(PQ/n)} = \sqrt{(0.5*0.5/625)}= 0.02$

Test the hypothesis that males and females were in equal numbers in city at 1% level of significance.

Hypothesis formulation

H_0: P = 0.5 The proportion of male in the city is 50% or the number of males and females are equal in the city.

H_1: P ‡0.5 (two tailed) The proportion of male in the city is not 50% or the number of males and females are equal in the city. Number of males and females are not equal in the city.

Test Statistic: Under H_0

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{0.48 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{625}}} = -1 \qquad \text{so, } |Z| = 1$$

Level of significance (α) = 1% = 0.01

Tab $Z_{0.01}$ (Two tailed) = 2.58

Decision: Since calculated Z is less than tabulated Z at 1% level of significance. So it is not significant and we accept null hypothesis and hence we conclude that males and females are equal in the city.

Question

A research center claims that more than 70% of adults have accessed the Internet over a wireless network with a laptop computer. In a random sample of 100 adults, 65% say they have accessed the Internet over a wireless network with a laptop computer. At $\alpha = 0.01$, is there enough evidence to support the researcher's claim?

Z-test for Double Sample Proportions

It is used to test the significance difference between two sample proportions. Following steps are involved under hypothesis testing procedure of Z-test for double sample proportions:

Step 1: Hypothesis formulation

Null Hypothesis (H_0): $P_1 = P_2$

That is there is no significant difference between two sample proportions or two population proportions are equal or two samples are drawn same normal population.

Alternative Hypothesis (H_1): $P_1 \neq P_2$ (Two tailed test)

That is there is significant difference between two sample proportions or two population proportions are not equal or two samples are not drawn from same normal population.

$P_1 < P_2$ (Left tailed test)

That is first sample proportions is less than second sample proportions.

$P_1 > P_2$ (Right tailed test)

That is first sample proportions is greater than second sample proportions.

Steps in Z-test for Double Sample Proportions

Step 1: Hypothesis formulation

Null Hypothesis (H_0): $P_1 = P_2$

Alternative Hypothesis (H_1): **$P_1 \neq P_2$** (Two tailed test)

Step 2: Test Statistics: Under H0

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ OR $\frac{X_1 + X_2}{n_1 + n_2}$

and $Q^\wedge = 1 - P^\wedge$

Step 3: Level of Significance (a) = either given or 5% as most commonly used

Step 4: Tabulated value (i.e. Critical value)

Step 5: Decision

Question

A large hotel chain is trying to decide whether to convert more of its rooms to non-smoking rooms. In a random sample of 400 guests last year, 166 had requested non-smoking rooms. This year, 175 guests in a sample of 380 preferred the smoking rooms. Would you recommended that the hotel chain convert more rooms to non-smoking? Support your recommendation by testing the appropriate hypothesis at 0.01 level of significance.

Would you recommended that the hotel chain convert more rooms to non-smoking?

Hypothesis formulation

H_0: $P_1 = P_2$ i.e., there is no significant in the difference in the sample proportion of choosing non smoking rooms of hotel in two years.

H_1: $P_1 < P_2$ (left tailed test) i.e., the proportion of choosing non smoking room is increased this year.

Last Year	**This Year**
$n_1 = 400$	$n_2 = 380$
$X_1 = 166$	$X_2 = 380 – 175 = 205$
$p_1 = 166/400 = 0.415$	$p_2 = 205/380 = 0.539$

Test Statistic: Under H0

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.415 - 0.539}{\sqrt{0.476 \times 0.524\left(\frac{1}{400} + \frac{1}{380}\right)}} = -3.47$$

$$\hat{P} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{166 + 205}{400 + 380} = 0.476$$

$$\hat{Q} = 1 - 0.476 = 0.524$$

Therefore $|Z| = 3.47$

Now **Level of significance (a) =** 1% = 0.01 Tab $Z_{0.01}$ (One tailed, n = 400 and 380) = 2.33

Decision: Since calculated Z is greater than tabulated Z at 1% level of significance. So it is significant and we reject null hypothesis and hence we conclude that the proportion of choosing non smoking room is increased this year.

Comparison of Two Population variances

We want to test the hypothesis that two population variances are equal, i.e.

$$\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2$$

$$\mathrm{H}_1 : \sigma_1^2 \neq \sigma_2^2$$

We need to rewrite the null and alternative hypotheses so that we can use a single value to represent the test statistic.

Ratio of variances

The null and alternative hypotheses are converted to the following form.

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

The Test Statistic

A estimated value for the test statistic for the ratio of two population variances is the ratio of the corresponding sample variances

$$\frac{s_1^2}{s_2^2}$$

The F-distribution

The ratio of two chi-square variables follows a new distribution known as the F-distribution.

If we have one χ^2 variable with n_1 -1 degrees of freedom, and another with $n_2 - 1$ degrees of freedom then the ratio has an F-distribution with $n_1 - 1$ degrees of freedom for the numerator and $n_2 - 1$ degrees of freedom for the denominator.

Therefore, for a specified cumulative area a

$$F(a; n_1 - 1; n_2 - 1) = \frac{\chi^2(a; n_1 - 1)}{\chi^2\left(a; n_2 - 1)\right)}$$

Extract of F-tables (1-α=.95) or α=.05

Denominator df	The F-distribution with 1 - α = .95 numerator df									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54

F-distribution examples

$F(.95;4,9) = 3.63$

$F(.95;8,3) = 8.85$

$F(.99;15,20) = 3.09$

$F(.99;40,30) = 2.30$

Test of Hypothesis for two variances

$$\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2$$

$$\mathrm{H}_1 : \sigma_1^2 \neq \sigma_2^2$$

Rewrite the hypotheses as:

$$\mathrm{H}_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$\mathrm{H}_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

$$\mathrm{TS} : F^* = \frac{s_1^2}{s_2^2}$$

EXAMPLE

The production manager of a textile company wants to test the hypothesis that the mean cost of producing a polyester fabric is the same for two different production processes. Assume that production costs are normally distributed for both processes.

Random samples of production costs for several production runs using the two different production processes are as follows:

Process I	**20**	**15**	**20**	**23**	**24**	**21**
Process II	**27**	**19**	**41**	**30**	**16**	

Test the hypothesis that the two population variances are equal with a 2% level of significance.

Sample Data

	Pop 1	Pop 2
Sample size	$n_1 = 6$	$n_2 = 5$
Mean	20.5	26.6
Variance	9.9	97.3

Testing the Hypothesis

Null Hypothesis: Ho: $\sigma_1^2=\sigma_2^2$
Alt Hypothesis: Ho: $\sigma_1^2 \neq \sigma_2^2$

Test Statistics F= S_1^2/S_2^2 = 9.9/97.3 = 0.1017

Level of significant= 2%

F critical =
F(0.02/2, 5,4)=1/F(0.99,4,5)=0.088
F(1-0.02/2,5,4)=F(0.99,,4)=15.52

Decision: Since 0.088<F<15.52
H0 is accepted