

# Statistical Methods for Decision Making

# Statistics Foundation Concepts

## Suggested Readings

Material	Details
Slides	Provided by Faculty
Course Notes & Exercises	Provided by faculty

Book	Author
Statistics for Business and Economics	Anderson, Sweeney, Williams
Statistics for Management	Richard I. Levin, David S. Rubin
Business Statistics - An applied Orientation	P K Viswanathan
Data Preparation for Data Mining	Dorian Pyle

# Myself

**C.K. Chandrasekhar**

B.E., PGDM(IIM-A), Ph.D. (Data Sciences)

3 decades experience in

- Training & Education
- Information Systems
- Business Analytics
- Database Systems
- Quality Management - 6 $\sigma$

Worked in

- ASCI, ECIL, WIPRO, RFO(ME), GE(Japan),  
IGATE(USA), IT KIDS, IBS, VIT, SSN, UOM, LIBA
- India – 16 years
- Overseas – 14 Years

# Types of Statistics

Descriptive Statistics: Transformation of raw data to information.

- Summarization and visualization, presentation, profiling.

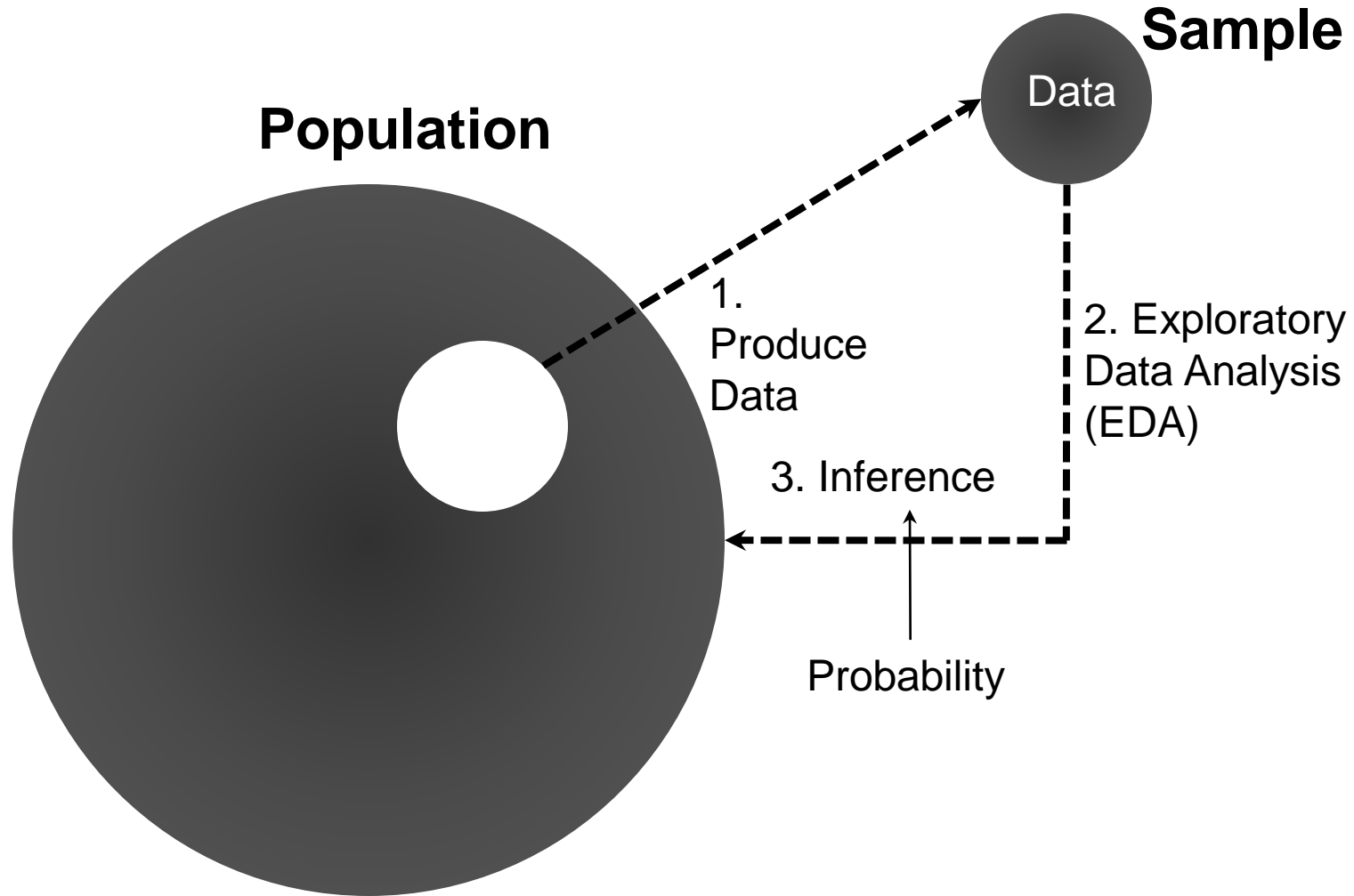
Example: comparison of mean performances, Noting magnitudes and volatility.

Inferential Statistics: Projection of data character from sample to population.

- Point estimation, Interval estimation, Hypothesis testing

Example: Computing average household income

# Statistics – The Big Picture



# Types of Data

Qualitative: Can not be quantified numerically

Quantitative: Can be quantified numerically

## Measurement Scales

Nominal	Can be divided to categories. Categories are labeled.	Gender Part No
Ordinal	Can be ranked.	Low, Med, High
Interval	Difference can be measured.	Temperature
Ratio	All arithmetic is possible.	Sales, Height, Age

More  
Info



# Cross-Sectional and Time Series Data

Name	Gender	Sales	Time	Ad Exp	Share	Change	Accounts	Work	Rating	Poten Index
Adam	M	3669.88	43.1	4582.9	2.51	0.34	74.86	15.05	Excellent	90
Beatrice	F	3473.95	108.13	5539.8	5.51	0.15	107.32	19.97	Excellent	78
Charles	M	2295.1	13.82	2950.4	10.91	-0.72	96.75	17.34	Good	33
Dodsworth	M	4675.56	186.18	2243.1	8.27	0.17	195.12	13.4	Very Good	87
Elizebeth	F	6125.96	161.79	7747.1	9.15	0.5	180.44	17.64	Excellent	74
Frederik	M	2134.94	8.94	402.4	5.51	0.15	104.88	16.22	Excellent	55
Gertrude	F	5031.66	365.04	3140.6	8.54	0.55	256.1	18.8	Excellent	67
Harley	M	3367.45	220.32	2086.2	7.07	-0.49	126.83	19.86	Good	52
Ingrid	F	6519.45	127.64	8846.2	12.54	1.24	203.25	17.42	Excellent	63



# Cross-Sectional and Time Series Data

Sales Data				
Month/Yr	2005	2006	2007	2008
January	171422	170006	189863	195538
February	191003	196112	205745	212362
March	193991	200181	213248	221730
April	207399	211874	228480	236999
May	205494	219528	231781	245221
June	197367	215573	237120	256440
July	208396	232726	254384	277823
August	214413	236883	253862	274502
September	200260	221034	232810	250585
October	187900	216940	226728	249351
November	182203	206932	223668	245733
December	180789	198068	218328	236601

# Time Series Data



# Types of Studies

- **Observational: Variables are not controlled**
  - Surveys, Observations, Questionnaire, Mailer
- **Experimental: Variables are controlled**
  - Reference groups, Before and after, Experiments

# Inferential Statistics

Hours until Burnout – Sample of 200 bulbs (Norris)									
107	73	68	97	76	79	94	59	98	57
54	65	71	70	84	88	62	61	79	98
66	62	79	86	68	74	61	82	65	98
62	116	65	88	64	79	78	79	77	86
74	85	73	80	68	78	89	72	58	69
92	78	88	77	103	88	63	68	88	81
75	90	62	89	71	71	74	70	74	70
65	81	75	62	94	71	85	84	83	63
81	62	79	83	93	61	65	62	92	65
83	70	70	81	77	72	84	67	59	58
78	66	66	94	77	63	66	75	68	76
90	78	71	101	78	43	59	67	61	71
96	75	64	76	72	77	74	65	82	86
66	86	96	89	81	71	85	99	59	92
68	72	77	60	87	84	75	77	51	45
85	67	87	80	84	93	69	76	89	75
83	68	72	67	92	89	82	96	77	102
74	91	76	83	66	68	61	73	72	76
73	77	79	94	63	59	62	71	81	65
73	63	63	89	82	64	85	92	64	73

# Descriptive Statistics





Hours	
Mean	76
Standard Error	0.853
Median	75
Mode	77
Standard Deviation	12.065
Sample Variance	145.558
Kurtosis	0.036
Skewness	0.288
Range	73
Minimum	43
Maximum	116
Sum	15200
Count	200

# Cross Tabulation





- Cross tabulation is a tabular summary of two variables

	Smoker	Non Smoker
Male		
Female		

# Cross Tabulation

<div><div>Boss Mood</div><div>Weather</div></div>		
	72%	28%
	72%	28%

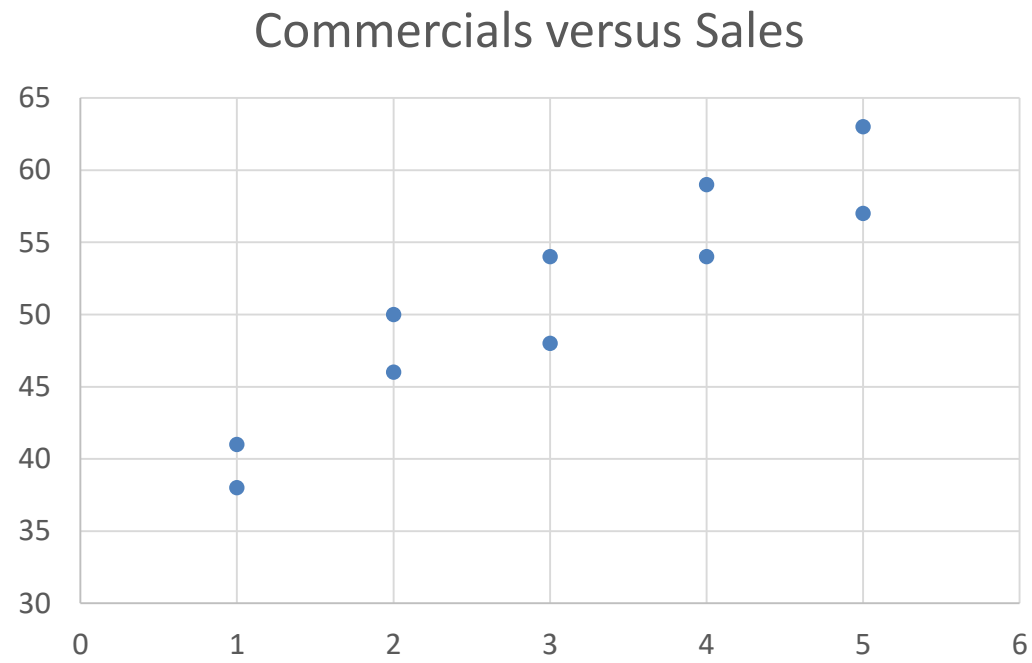
# Cross Tabulation

<div><div>Boss Mood</div><div>Weather</div></div>		
	82%	18%
	60%	40%

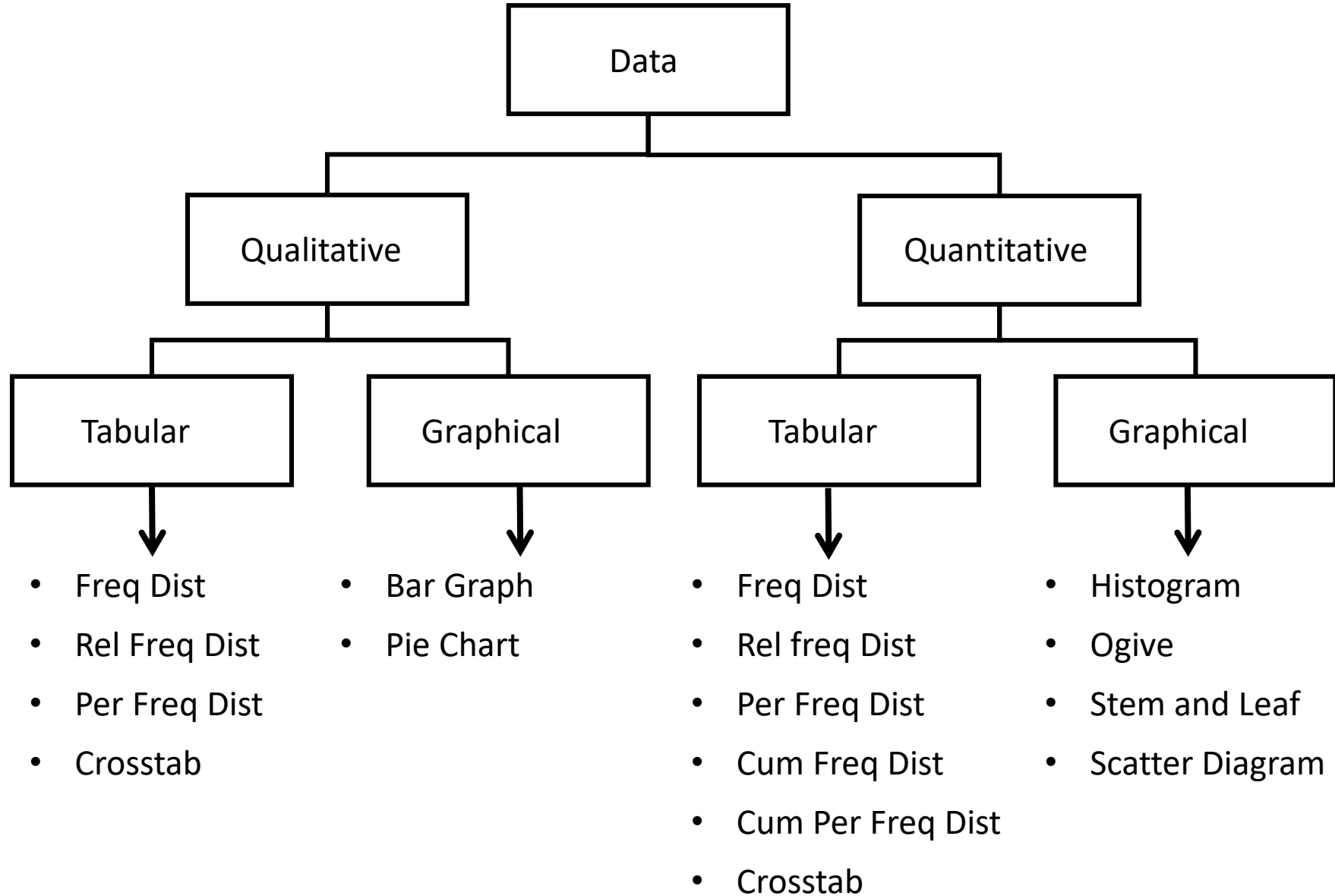


# Scatter Plots

Week	Commercials	Sales
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46



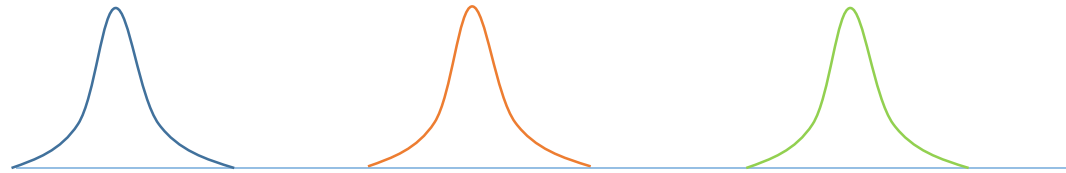
# Tools for Data Summarization



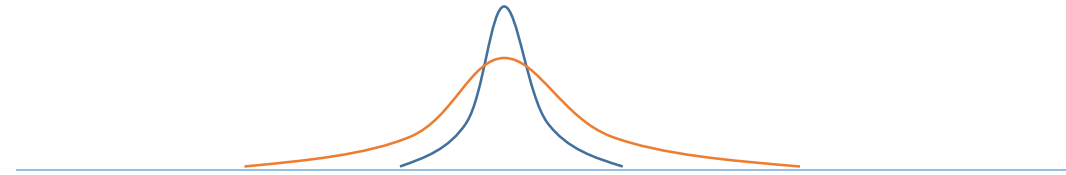
# Numerical Measures

Measures of

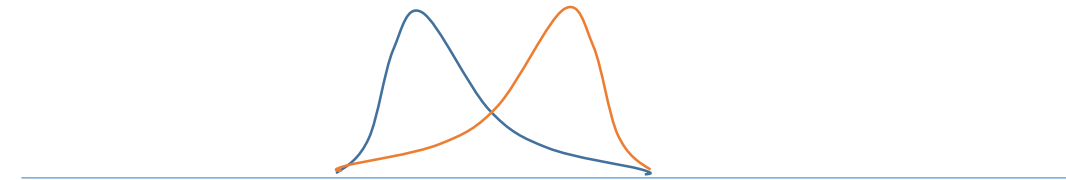
- Location



- Spread



- Shape



# Population & Sample

	Population	Sample
Mean	$\mu$	$\bar{x}$
Variance	$\sigma^2$	$s^2$
Standard Deviation	$\sigma$	$s$

# Descriptive Statistics

Descriptive Statistics		
Statistic	Expression	Measures
Mean	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$	Central Tendency
Median	Divides ordered set into half	Central Tendency
Mode	Most frequent observation	Central Tendency
Range	$X_{(\max)} - X_{(\min)}$	Spread
Variance	$S_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$	Spread
Std. Deviation	$S_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$	Spread
IQR	$Q_3 - Q_1$	Spread
Skewness	$Skewness = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{S_x^3} \frac{n}{(n-1)(n-2)}$	Asymmetry
Kurtosis	$Kurtosis = \frac{n(n+1) \sum_{i=1}^n (x_i - \bar{x})^4 - 3(n-1) S_x^4}{(n-1)(n-2)(n-3) S_x^4}$	Peakedness

# z-Scores

- Statisticians are often interested in relative location of items in data set.
- z-score of an observation is a measure for this purpose.
- The z-score uses both mean and SD to calculate this relative location.
- The z-score is defined as

$$Z_i = \frac{x_i - \bar{x}}{s}$$

- z-score is called standardized value. It is the number of standard deviations that an observation is away from the mean.
- A z-score of 1.2 means that the observation is 1.2 standard deviations away from mean on the right hand (+ve) side.
- Similarly z-score of -1.2 means that the observation is 1.2 standard deviations away from mean on the left hand (-ve) side.

# z-Scores

What are z scores for class size data?

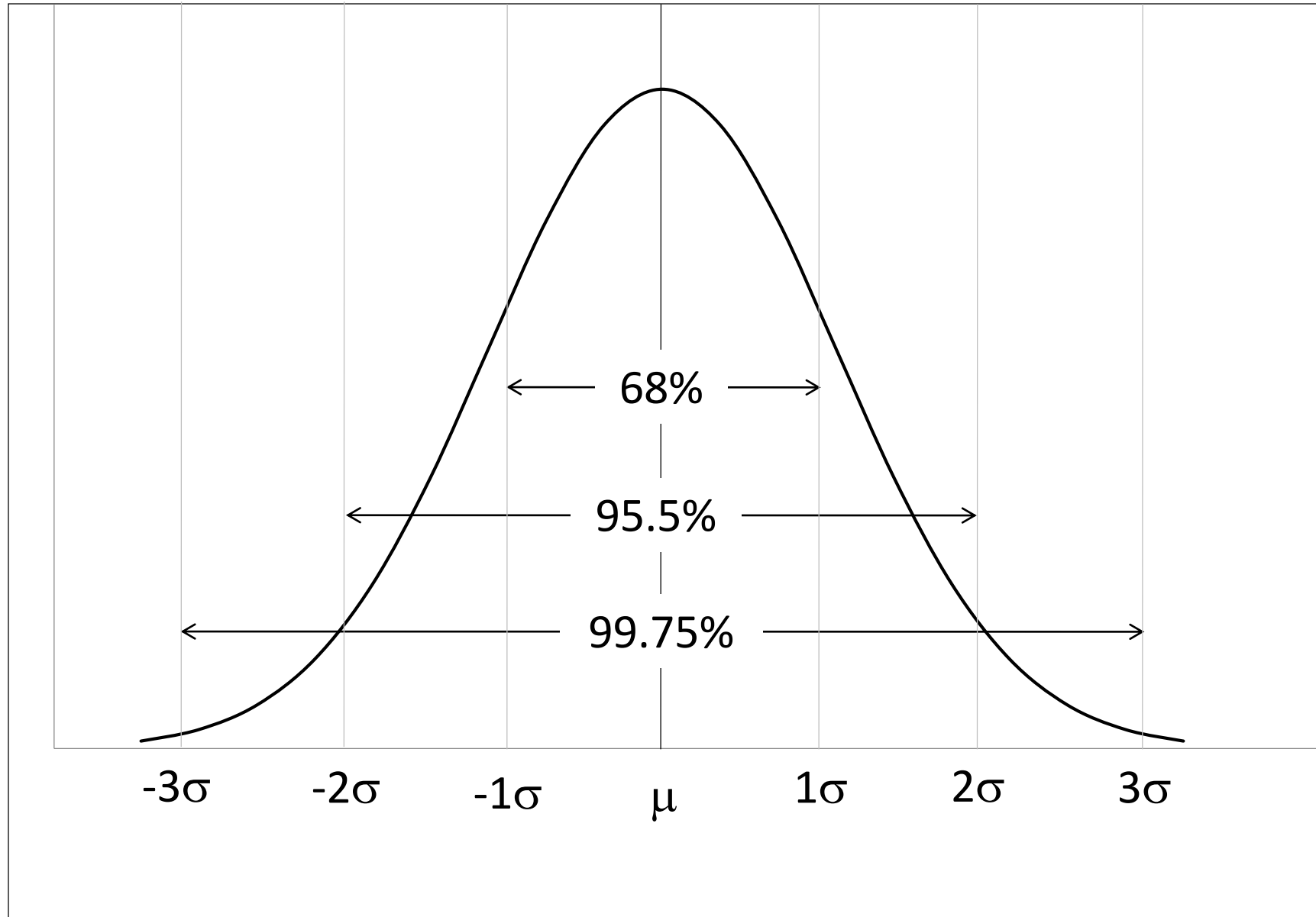
32, 42, 46, 46, 54

What are z scores for starting salary data?

2850 2950 3050 2880 2755 2710 2890 3130 2940 3325 2920 2880

z-scores for some sample data → (46 54 42 46 32)		
Frequency	Deviation from Mean	z-score
46	2	$2/8 = 0.25$
54	10	$10/8 = 1.25$
42	-2	$-2/8 = -0.25$
46	2	$2/8 = 0.25$
32	-12	$-12/8 = -1.50$
Mean = 44, St Dev = 8		

# The Normal Curve





# Probability Distributions

# Random Variable

- A random variable is a numerical description of the outcome of an experiment.
- An experiment's outcome may or may not be a numerical value
- Random variable is a method to transform the outcome of an experiment into a numerical value.
- The outcome of throwing a die is already in numerical form. The number so obtained  $\{0,1,2,3,4,5,6\}$  is a random variable.
- A successful or unsuccessful sales call resulting in an order or no order may not be readily expressed in numerical terms. In such case we transform  $x=0$  as failure to secure an order and  $x=1$  as successfully securing an order.
- A discrete random variable can assume only finite number of numerical values. Eg: Number of questions answered by a student in an examination.
- A continuous random variable can assume infinite number of numerical values. These experimental outcomes are based on interval or ratio scale. Eg: flight time to Delhi.

# Examples of Random Variable

Experiment	Random Variable (x)	Possible Values
<b>DISCRETE RANDOM VARIABLE</b>		
Contact Five Customers	No. of orders placed	0,1,2,3,4,5
Inspect a batch of 50 items	No. of defective items	0,1,2,...,49,50
Operate a bank counter	No. of customers	0,1,2,3,.....
Make a random call	Gender of respondent	1 if F, 0 if M
<b>CONTINUOUS RANDOM VARIABLE</b>		
Operate an ATM	Waiting time for customer	$X \geq 0$
Fill 150ml soft drink can	milliliters	$\leq 150$
Execute a s/w project	% Project completed in 6 months	$0 \leq x \leq 100$
Examine a patient	Body temperature	$96 \leq x \leq 106$

# Required Conditions for Discrete Probability Distribution

- $f(x) \geq 0$
- $\sum f(x) = 1$
- For a Discrete uniform probability distribution
  - $f(x) = 1/n$  for  $x = 1, 2, \dots, n$
- Discrete probability distribution may be described by a formula that satisfies conditions above.
- $f(x) = x/10$  for  $x = 1, 2, 3, 4$
- Examples of discrete probability distributions
  - Binomial distribution
  - Poisson distribution
- Mean or Expected value of a discrete random variable
  - $E(x) = \mu = \sum xf(x)$
- Variance of a discrete random variable
  - $\text{Var}(x) = \sigma^2 = \sum (x - \mu)^2 f(x)$

# Problems on $E(x)$ , $\text{Var}(x)$

1. A casino charges 10¢ per game. The player throws a die

- If he gets  $\{1,2,3\}$  he is paid 1¢
- If he gets  $\{4,5\}$  he is paid 5¢
- If he gets  $\{6\}$  he is paid 35¢

What is the long term expected profit for the casino?

2. What is  $E(x)$  and  $\text{Var}(x)$  for the following distribution?

x	1	2	3	4	5
#	72177	60133	16674	2230	623

# Binomial Probability Distribution

- Is a discrete probability distribution.
- It is a multistep experiment
- Business examples
  - How many of the customers visiting a supermarket make a purchase?
  - How many calls of an insurance agent will result in subscription?
  - How many purchases will be made using credit card?
- Properties (or assumptions) of binomial experiment
  - The experiment consists of a sequence on  $n$  trials.
  - Only two outcomes are possible in each trial. We refer to one of them as “success” and the other as “failure”.
  - The probability of success is denoted by  $p$ , does not change from trial to trial.
  - The probability of failure is given by  $(1-p)$  as a consequence of above.
  - The trials are independent.

# Binomial Probability Function

Binomial probability function gives probability of  $x$  successes in  $n$  trials.

$$f(x) = \binom{n}{x} p^x (1 - p)^{(n-x)}$$

Where


$n$  = number of trials

$x$  = number of successes

$$\binom{n}{x} = \frac{n!}{x! (n - x)!}$$

$p$  = the probability of success in any one trial

$(1-p)$  = probability of failure in any one trial



Combination for selecting  $x$  objects from  $n$  objects

# Bernoulli Process

- Each trial of a binomial experiment is a Bernoulli process.
- A Bernoulli process is a single trial resulting in either a success  $S = 1$  or a Failure  $F = 0$ .
- Probability of success  $s$  is  $p$  and that failure  $F$  is  $(1-p)$
- The mean of Bernoulli experiment is  $1 \cdot p + 0 \cdot (1-p) = p$
- The variance of Bernoulli trial is
  - $\text{Var}(x) = p(1-p)^2 + (1-p)(0-p)^2 = p - p^3 - p^2 + p^3 = p(1-p)$
- Please note
  - $\mu_{x \pm y} = \mu_x \pm \mu_y$
  - $\sigma^2_{x \pm y} = \sigma^2_x + \sigma^2_y$



# Mean and Variance of Binomial Distribution

A binomial experiment is a series of  $n$  Bernoulli trials. Therefore

$$\mu_x = p + p + \dots n \text{ times} = np$$

Let  $(1-p) = q$  then

For a Bernoulli trial

$$\text{Var}(b) = pq$$

For  $n$  trials

$$\text{Var}(x) = pq + pq + \dots n \text{ times} = npq = np(1-p)$$

# Martin Cloth Store

1. On the basis of past experience martin cloth store manager estimates that the probability of purchase by any customer is 0.3 and purchase decision of one customer does not influence the other. What is the probability of 2 of the next three customers will make a purchase?
2. In the class test 1 last week, 15 MCQs with 5 choices for each question were given. One student who did not study for the test randomly ticked the answers. What is the probability of getting at least 3 correct answers?
3. A machine produces machine parts of which 3% are expected to be defective. Randomly a batch of 100 parts were inspected for defectives.
  - a. What is the probability of exactly 3 defectives?
  - b. What is the probability of at most 4 defectives?
  - c. What is the probability of at least 96 good parts
4. A bank issues credit cards in partnership with MASTER CARD. From history 40% of cardholders do not pay dues on-time. Construct a probability distribution of on-time payers for a sample of 7 accounts.

# Poisson Distribution

A probability Distribution that is used to estimating number of occurrences over a specified interval of time or space.

Business Examples:

- No. of persons arriving at a service station or ATM
- No. of calls coming to a call center in an given interval.
- No. of defects in a length of cloth from a loom
- No. of accidents in a given length of highway.

Properties of Poison Probability Function

- The probability of occurrence is same in two intervals of same length.
- The probability of occurrence or non-occurrence is independent of occurrence or non-occurrence in any other interval.

# Poisson Probability Function

The Poisson Probability Function

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}$$

Where

$f(x)$  = the probability of  $x$  occurrences in an interval

$\mu$  = expected value or mean number of occurrences in an interval

$e = 2.71828$

Mean of Poisson Probability Distribution =  $\mu$

Variance of Poisson Probability Distribution =  $\mu$

# Example

1. A bank manager is interested in number of persons arriving at a teller window on a weekday morning. It may be assumed that the probability of a person arriving is same for any two time periods of equal intervals and the arrival of non arrival of a person is independent of arrival or non arrival in any other time period.

The average number of arrival in any 15 minute interval is 10.

Calculate probability distribution function for the teller operation window.

What is the probability of exactly 5 arrivals in 15 minutes?

What is PDF function for 3 minute interval?

2. A loom produces on the average 2 defects per meter of cloth. What is likely hood that there is no defect in a 3 meter length?

# Questions

1. On the average 6 cars arrive every 2 minutes at a petrol station.
  - a. What is the probability that exactly four cars arriving in 1 minute?
  - b. What is the probability of 0 cars arriving in 3 minutes?
  - c. What is probability of at least 3 cars arriving in 2 minutes?
2. An airline gets on the average 4 missing baggage claims every day. Construct a probability distribution table for the missing baggage data (15 entries).
  - a. What is the probability of at least 50 claims a week?
  - b. What is the probability of no claim on a day?
  - c. What is the probability of at most 10 claims in a 3-day period?.

# Continuous Probability Functions

- Assume that to come to this class you travel for 45mins to 1 hour.
- If your travel time is a random variable it may take value between 45 and 60 including decimal values.
- So the travel time is a continuous random variable.
- All the random variables measured on interval and ratio scale are continuous random variables.
- For a continuous random variable
  - The probability at any point is 0. (Area under curve only makes sense)
  - Only probabilities exist for intervals. We do not talk about probability of RV assuming particular value. We talk about probability for a given interval.
  - The cumulative probability is equal to 1 i.e. sample space.
  - The probability density can never be negative.
  - The probability density can exceed 1.
  - **Area is a measure of probability.**

# Uniform Probability Distribution

- Assume that to come to this class you travel for 45mins to 1 hour.
- Let your travel time be a random variable that may assume a value between 45 and 60 with equal probability.
- Then the random variable is called uniform random variable.
- The probability density function for the random variable is

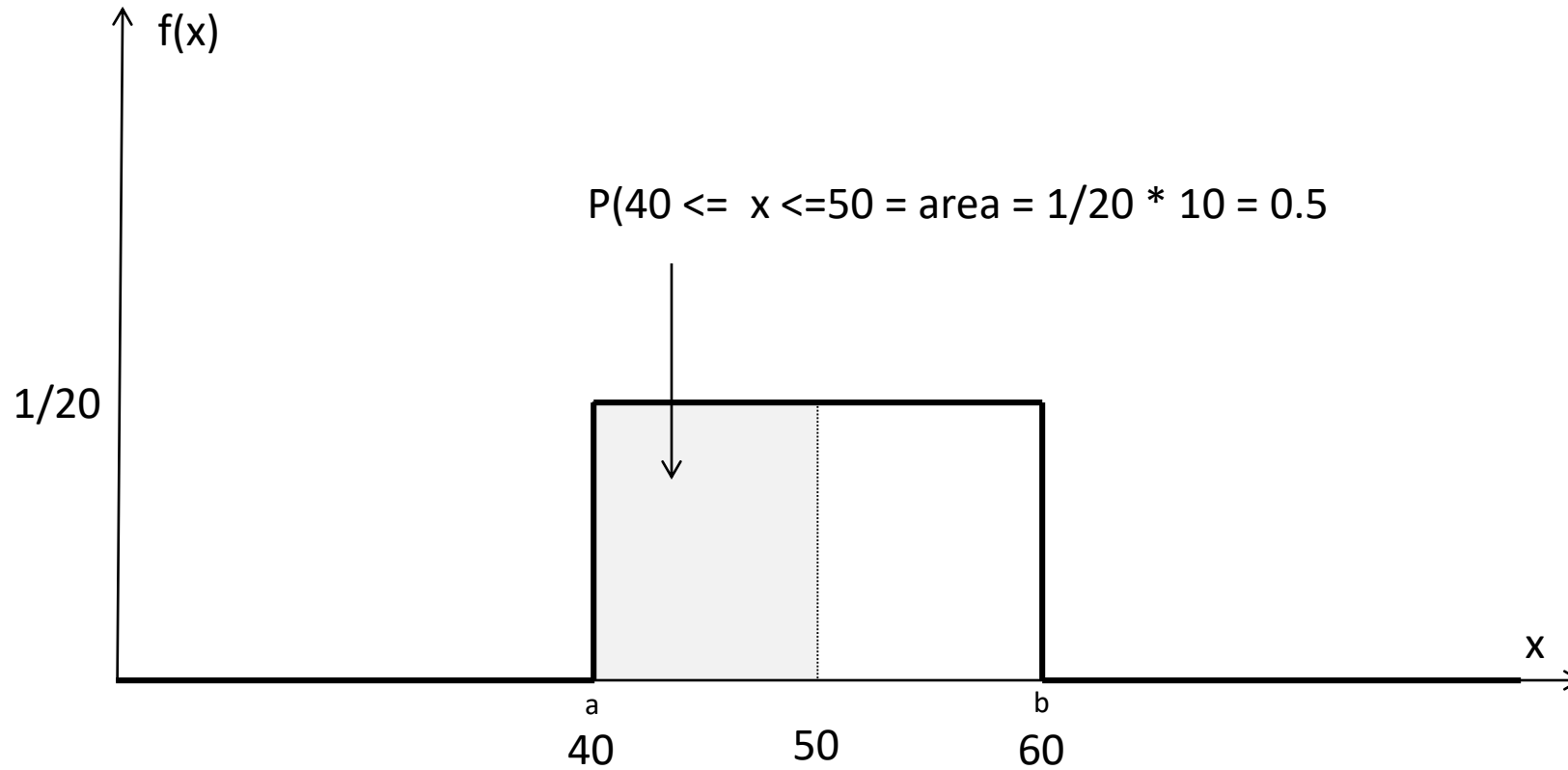
$$f(x) = \begin{cases} \frac{1}{15} & \text{for } 45 \leq x \leq 60 \\ 0 & \text{elsewhere} \end{cases}$$

- In general the probability density function for a uniform random variable is given by

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{elsewhere} \end{cases}$$



# Uniform Probability Distribution



Mean of Uniform Distribution Function =  $E(x) = (a + b)/2$

Variance of Uniform Distribution Function =  $(b - a)^2/12$

# Questions on Uniform Probability

1. You are participating in an auction for a piece of property. The value of the property is uniformly distributed between 7 lakhs and 10 lakhs.
  - a. You have bid 8.8 Lakhs. What is your probability of winning the bid?
  - b. You want to be 90% sure of winning the bid. How much should you bid?
2. A hall was booked for an important function. The function will have to be cancelled if it did not start within an hour. The chief guest has a habit of coming 30 minutes to 2 hours late for any function with uniform probability. What is the chance of cancelling the function?

# Normal Probability Distribution

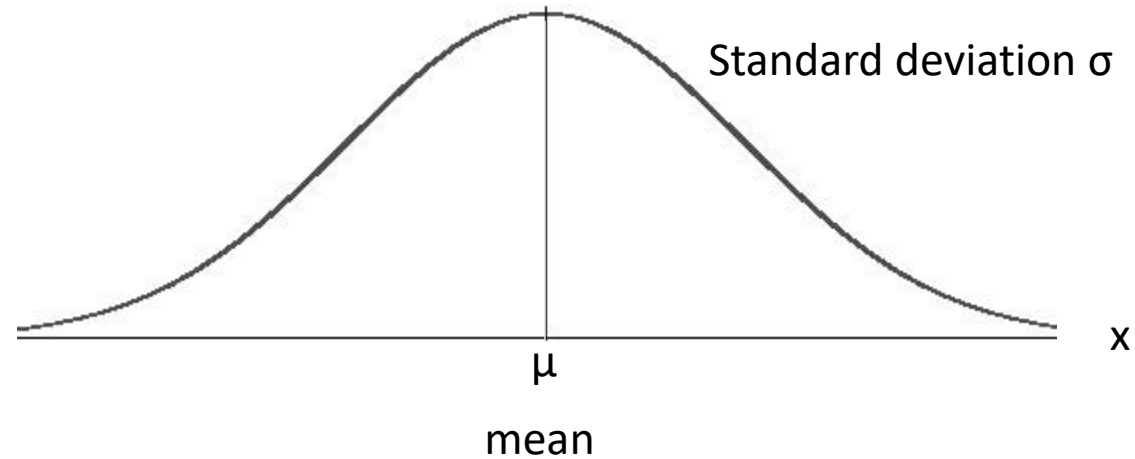
- The most important continuous probability distribution.
- Many natural and economic and demographic data follow normal distribution.
- Heights and weights of people, test scores, scientific and industrial measurements, rainfall data all follow normal distribution.
- Normal distribution is widely used in statistical inference and therefore it is important to know it.
- The normal probability density function

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$\mu$  = Mean

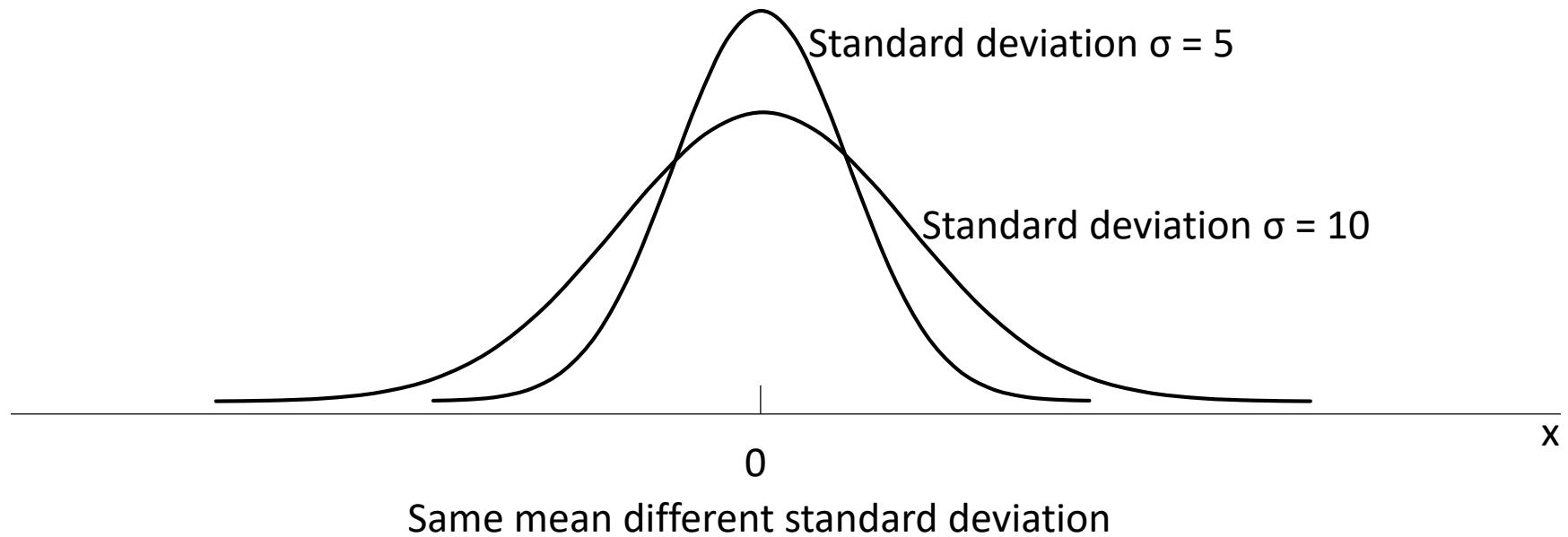
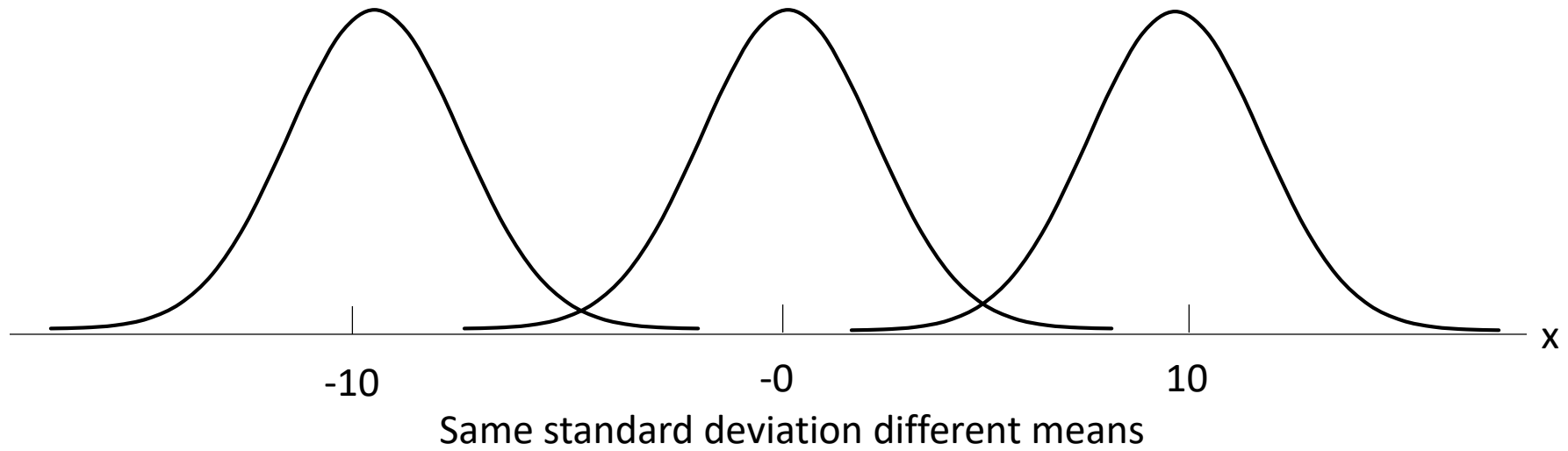
$\sigma$  = Standard deviation

# Normal Probability Distribution

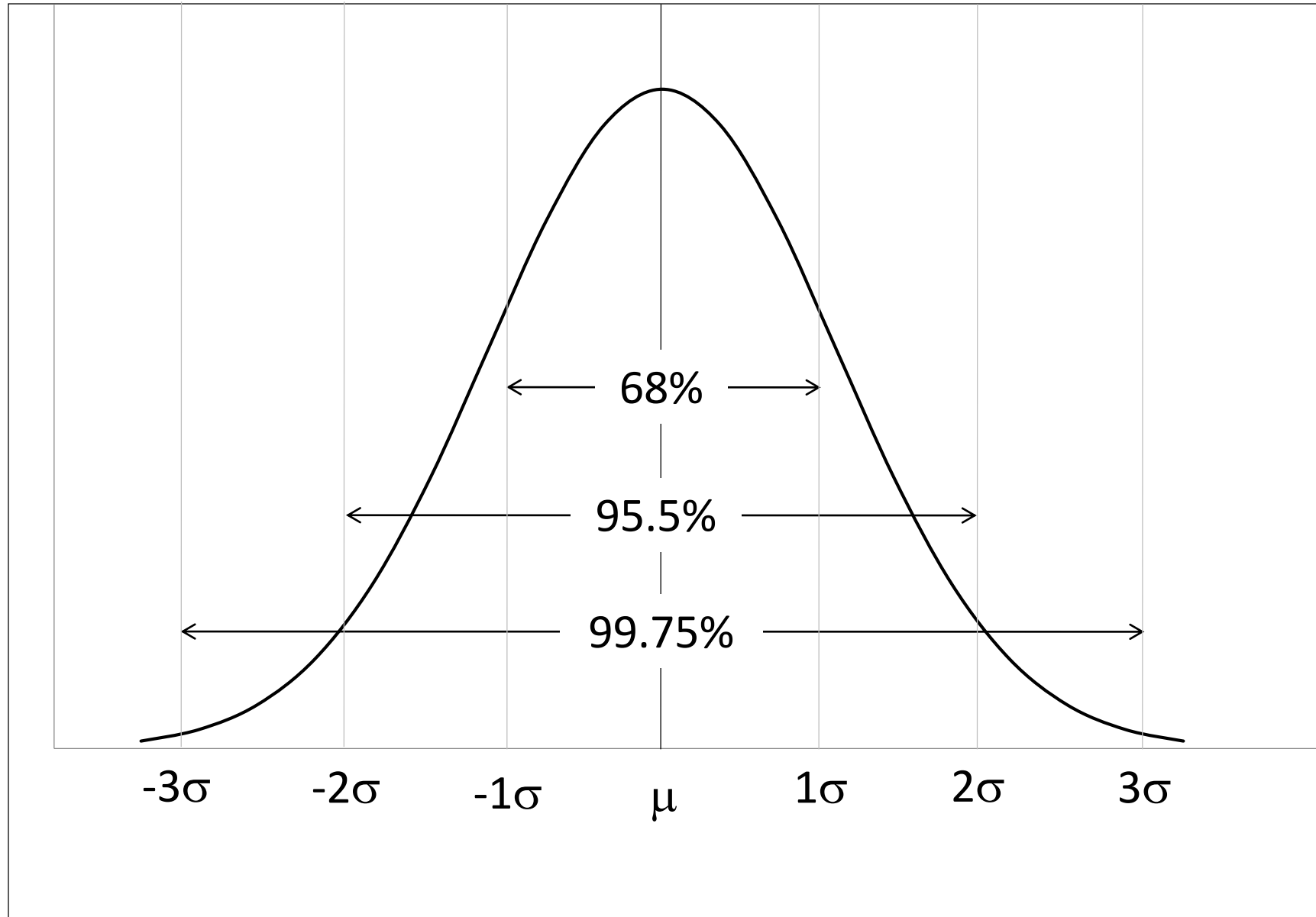


- The entire family of normal probability distributions is differentiated by its mean and standard distribution.
- The highest point of the distribution is the mean which is also the median and mode of the distribution.
- The mean can be any numeral +ve, 0 or -ve.
- Normal distribution is symmetric on y axis, bisected by mean.
- The tails of the curve extend to infinity.

# Normal Probability Distribution



# The Normal Curve



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

# Probability Calculations using Normal Curve

Using standard normal distribution we can find out probability that  $z$  assumes certain value.

$z$	Area = Probability
$< 1$	$0.3413 + 0.5 = 0.8413$
$> 2$	$0.5000 - 0.4772 = 0.0228$
$1 < z < 2$	$0.4772 - 0.3413 = 0.1359$
$< -2$	$= > +2 = 0.5000 - 0.4772 = 0.0228$
$-1 < z < -0.5$	$= 0.5 < z < 1 = 0.3413 - 0.1915 = 0.1498$
$-1 < z < 1.5$	$= z < 1 + z < 1.5 = 0.3413 + 0.4332 = 0.7745$
$z > 2.5$	
$0.5 < z < 2.5$	
$-1.5 < z < 1.5$	



# Using Standard Normal Curve for Computing Probability of X

Using standard normal distribution n we can find out probability that  $x \sim N(\mu, \sigma)$  assumes certain value.

Given  $\mu = 10$  and  $\sigma = 2$

x	z	
< 10	< 0	
< 5	< -2.5	
> 15	> 2.5	
$5 < x < 15$	$-2.5 < z < 2.5$	

# Exercises

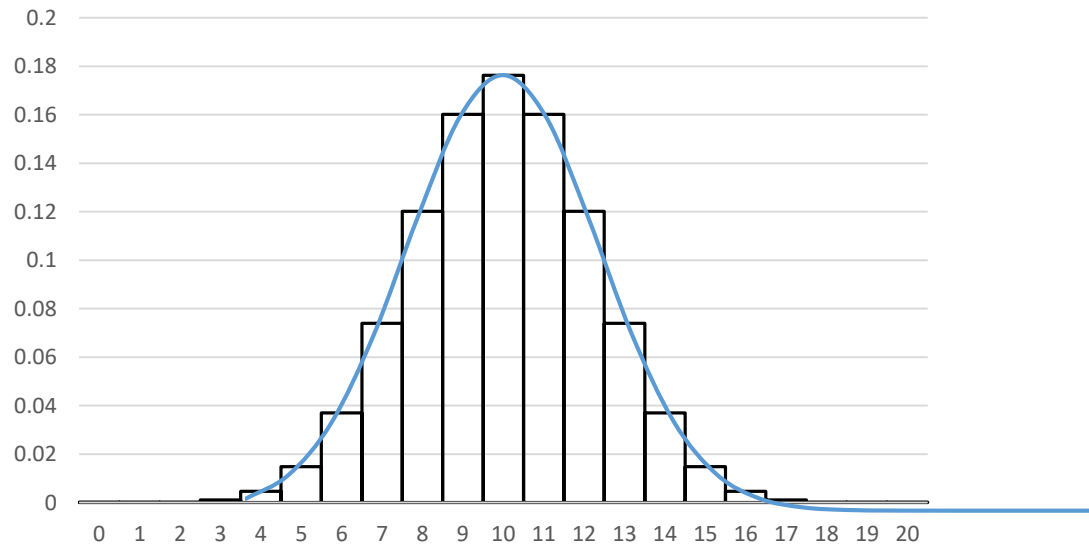
1. Bisleri produces water bottles with a mean of 1500 ml and SD of 30 ml and the process follows normal distribution. If filling less than 1450 or more than 1550 is considered defective, how many defectives may be expected in a batch of 1000 bottles?
2. The weight of breakfast cereal package follows a normal distribution with mean of 295 gm and SD of 25 gm. What is the probability that a randomly chosen package weighs
  - i)  $< 280$  gm
  - ii)  $> 350$  gm
  - iii) between 260 and 340 gm
3. A TV set is likely to fail if the supply voltage is  $< 180$  v or more than 260 v. In your locality the line voltage is normally distributed with a mean of 220v and SD of 20v. What is the probability of your TV breaking down?

# MRF

1. New tire mean mileage = 36500
2. Standard deviation = 5000
3. What is the probability that the tire will last  $> 40000$  km?
4. What mileage warranty the company should offer if warranty claims should not exceed 10%

# Normal Approximation to Binomial Distribution

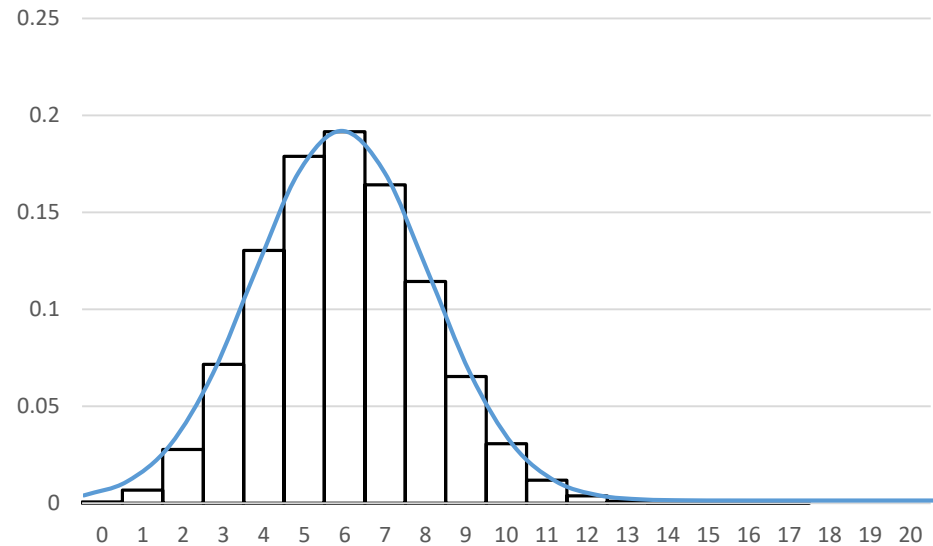
What is the probability of  $x$  successes in 20 trials with probability of success being 0.3



A binomial distribution may be approximated to normal distribution when  $n \cdot p \geq 5$ , The approximate normal distribution is given by  $\mu = np$  and  $\sigma = \sqrt{npq}$

# Normal Approximation to Poisson Distribution

What is the probability of  $x$  occurrences if average occurrences in an interval is 25?



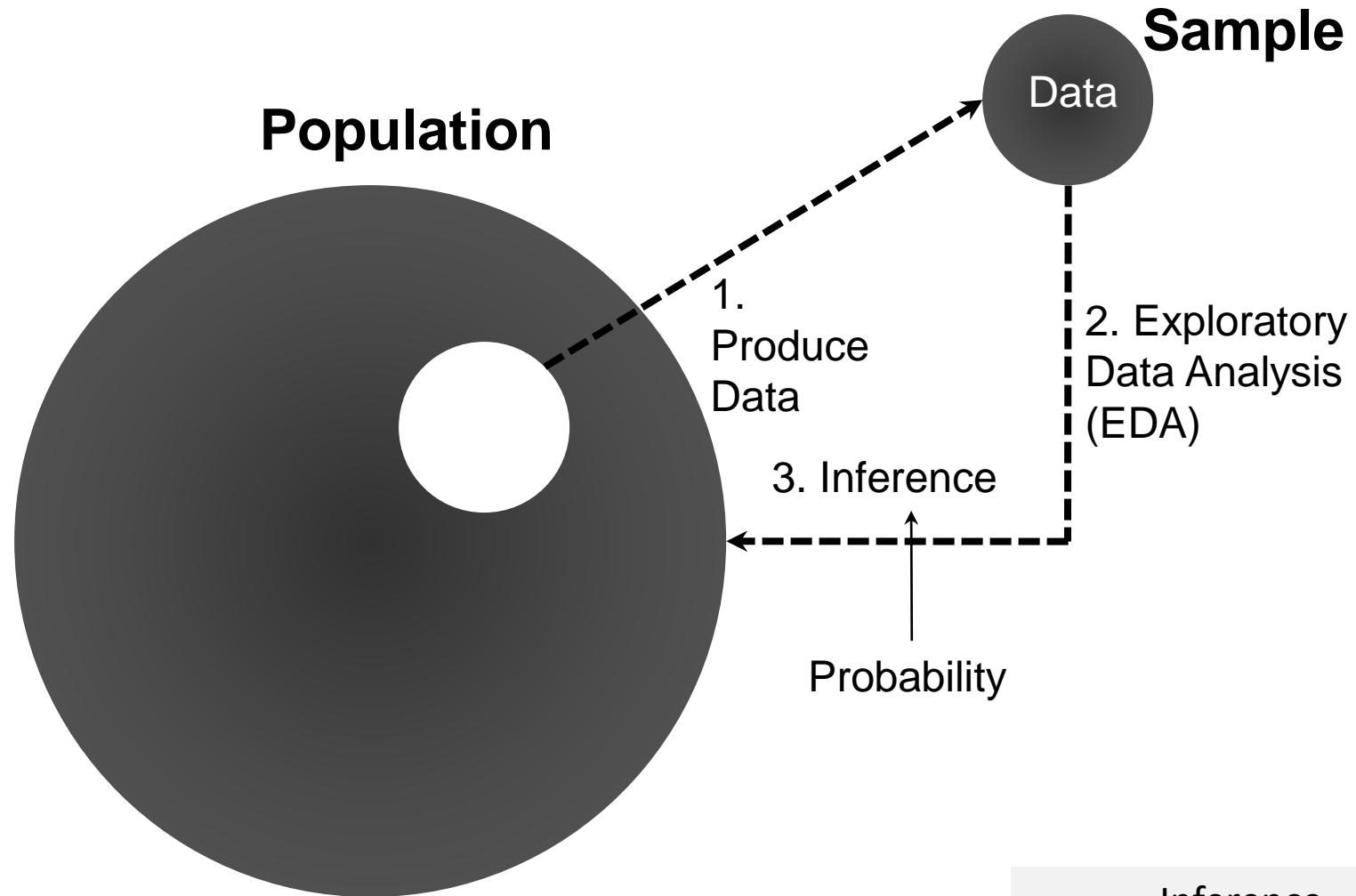
A distribution may be approximated to normal distribution when  $\lambda \geq 20$ . The approximate normal distribution is given by  $\mu = \lambda$  and  $\sigma = \sqrt{\lambda}$

# Sampling

# Sampling

- A population is the set of all the elements of interest in a study.
- A sample is a subset of population.
- When selected properly a sample is a representative of the population.
- i.e. the characteristics of sample provide an estimate (approximation) of population characteristics.
- These characteristics namely mean, standard deviation, variance, median (and a host of others) in the context of population are called parameters and in the context of sample are called statistics.

# Statistics – The Big Picture



## Inference

1. Estimation
2. Hypothesis Testing



# Sampling Terminology

Term	Description
Population	Set of all items of interest. Size = N
Sample	A subset chosen from Population for study. Size = n
Sampling Unit	An indivisible unit of population liable to be included in the sample set.
Sampling Frame	An object that represents population and facilitates selection of sample set.
Sampling	Selection of n units from N units of population with or without replacement.
Sampling fraction (f)	The ratio $f = n / N$ .

# Simple Random Sampling

- Most important sampling method in which every unit of population has an equal chance of being selected.
- Every sample of size  $n$  has the same probability of being selected.
- If there are  $N$  units in population, then all samples of size  $n$  have the same probability of  $1/({}^N C_n)$  of getting selected.

Procedure for SRS:

1. Prepare a unique list of  $N$  units of population – sampling frame.
2. Select  $n$  units from the list at random.

Example:

A bank wants to study customer perception of quality of its services to savings account holders. It is decided to go for a sample study. Say there are 1000 account holders and we want to select 50 a/c holders for the study.

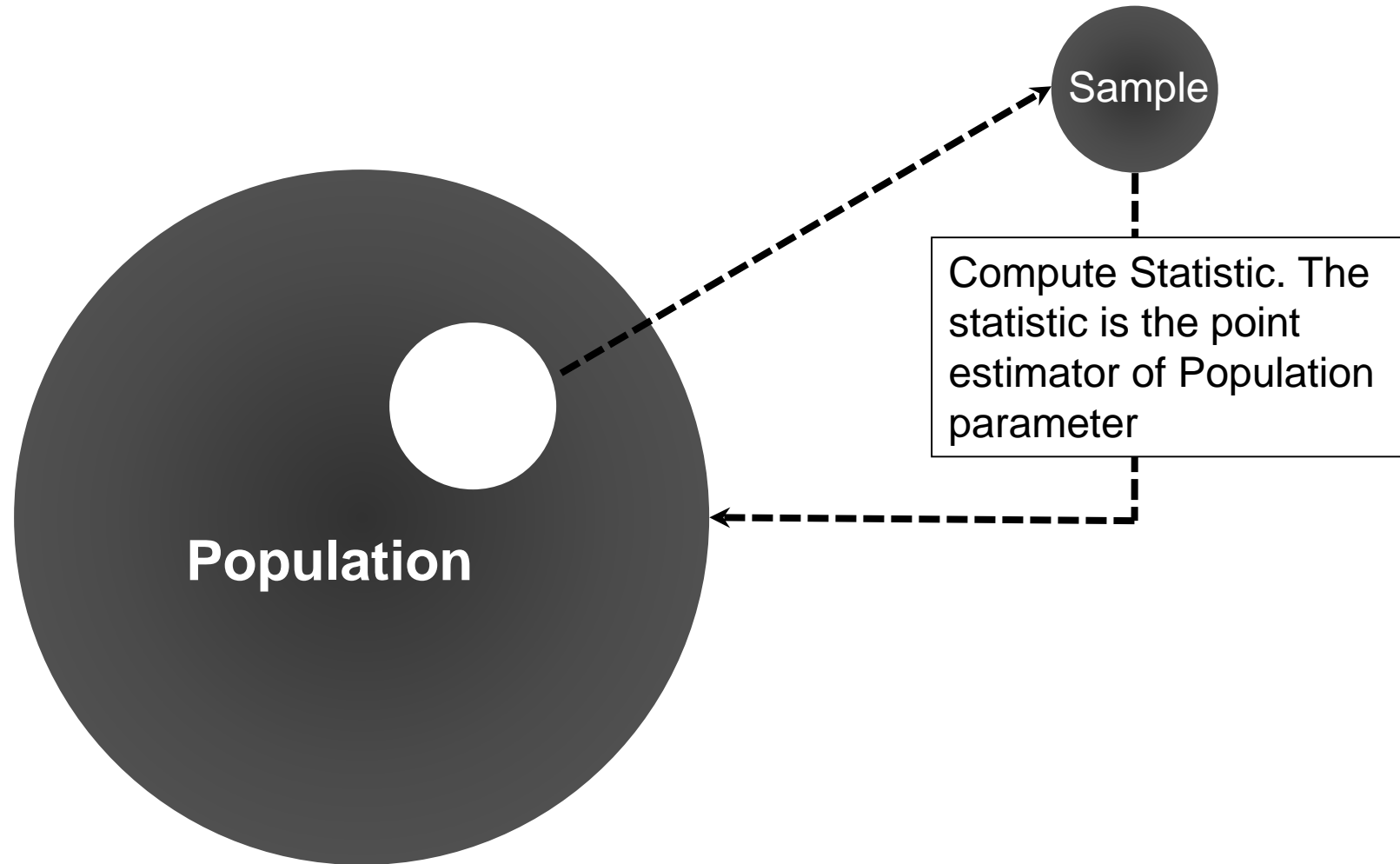
# Sampling Distribution

In sampling, the statistic of interest is a random variable.

- For example, the sample mean  $\bar{x}$  is a random variable.
- Repeatedly taking a sample of size of  $n$  is an experiment, we get different value of  $\bar{x}$  each time.
- So,  $\bar{x}$  has an expected value and variance associated with it.
- Like the mean ( $\bar{x}$ ), we may view every statistic of the sample as a random variable.
- Let us say an experiment consists of taking a sample of 4 students of this class and compute the average weight.
- If we take 10 samples we may get 10 values for  $\bar{x}$  as follows

$$\bar{x}_i = \{70, 67, 74, 66, 69, 75, 68, 71, 70, 66\}$$

# Point Estimation



# Point Estimation

Assume that a sample size of  $n$  was selected from infinite population using SRS. Identify each element in the sample as  $x_i$  where  $i = 1, 2, \dots, n$ .

Sample mean =  $\bar{x} = \sum_{i=1}^n x_i$

and

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

# Point Estimation Table

Population Parameter	Parameter Value	Sample Statistic	Point Estimate
$\mu$	51800	$\bar{x}$	51814
$\sigma$	4000	$s$	3347.8
$p$	0.6	$\bar{p}$	0.63

Sampling Error:

$ \bar{x} - \mu $	14
$ s - \sigma $	652.2
$ \bar{p} - p $	0.03

Sample 2:

$\bar{x}$	52669.7
$s$	4239
$p$	0.7

Sampling Distribution Simulation for EAI data

# Standard Error

- The sample statistic is an approximate estimator of population parameter it represents.
- $\bar{x}$  is an estimator of  $\mu$
- The error in estimation using  $\bar{x}$  to estimate  $\mu$  is given by standard deviation of  $\bar{x}$ .
- Therefore standard deviation of  $\bar{x}$  is called standard error. It represents the error we are committing in using  $\bar{x}$  as an estimator  $\mu$ .
- If this standard deviation is large we are less precise in estimating  $\mu$  and vice-versa.
- Even though we used  $\bar{x}$  for discussion above it equally applies to all statistics like median, standard deviation, correlation coefficient etc.

# Sampling Process

- Expected value  $E(\bar{x}) = \mu$
- Standard Deviation of  $\bar{x}$

Finite Population	Infinite Population
$\sigma_{\bar{x}} = \sqrt{\left(\frac{N - n}{N - 1}\right) \left(\frac{\sigma}{\sqrt{n}}\right)}$	$\sigma_{\bar{x}} = \left(\frac{\sigma}{\sqrt{n}}\right)$

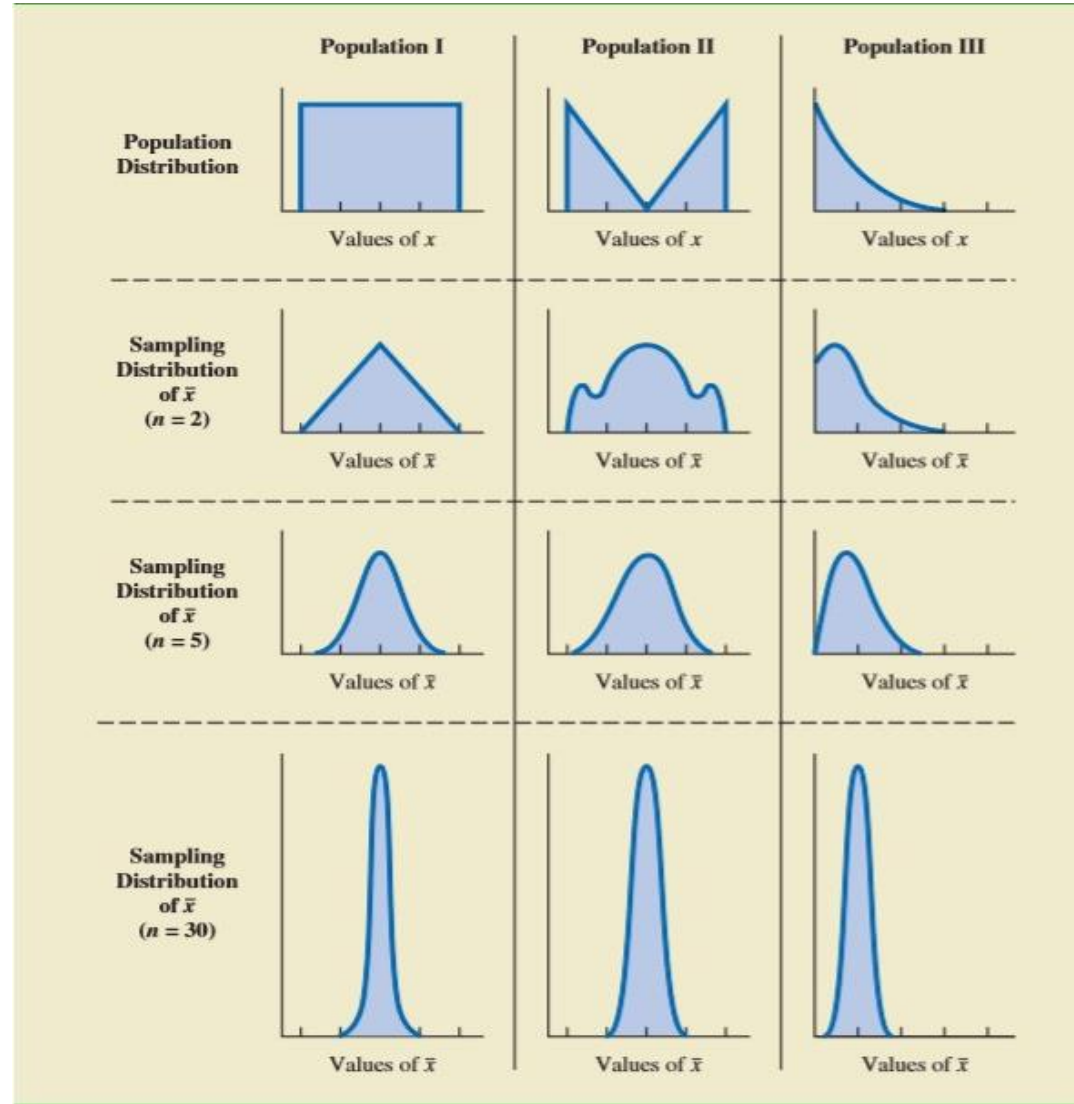
In general and in this course in particular it is assumed that  $n/N < 0.05$  and infinite population formula is used.



# The Mean and Standard Deviation of Sampling Distribution

- Expected value of  $\bar{x} = E(\bar{x}) = \mu$
- Standard Deviation of  $\bar{x} = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- When the population is normally distributed the sampling distribution follows normal distribution.
- When population distribution is not known we rely on **Central Limit Theorem** to approximate sampling distribution to normal distribution.
- *Central Limit theorem: In selecting simple random sample of size  $n$ , the sampling distribution of mean  $\bar{x}$  can be approximated by normal probability distribution as the sample size becomes large.*
- A sample size of 30 or more is considered large enough to use this approximation.

# Central Limit Theorem

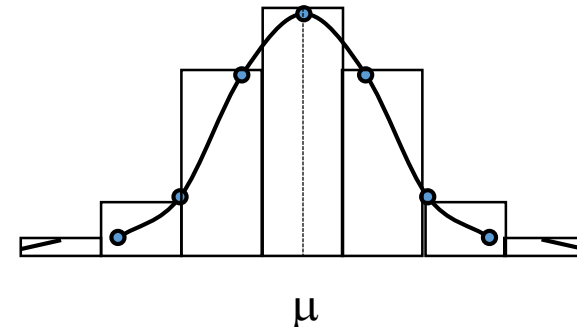


# Sampling Distribution

Sampling distribution of sample mean  $\bar{x}$  can be approximated to a normal distribution as the sample size becomes large.

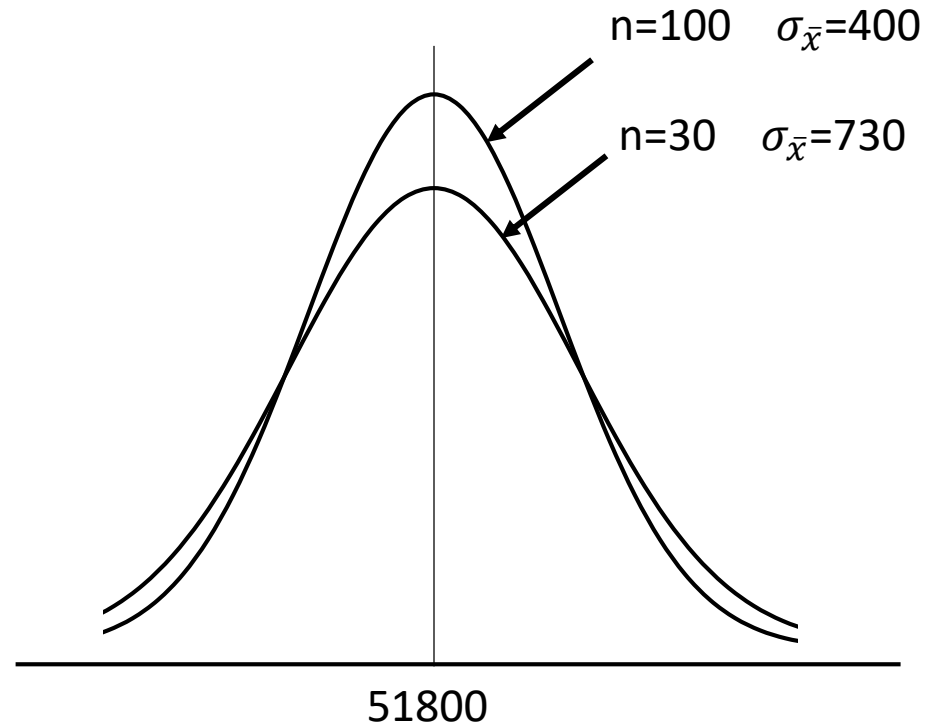
1. Sample size  $n \geq 30$  is considered as large for this purpose.
2. If condition 1 is satisfied we do not bother about shape population distribution and assume normality
3. If the population distribution is normal then sampling distribution is normal irrespective of sample size.
4. The mean and st. dev. Of sampling distribution are

$$\mu_{\bar{x}} = \mu \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$



# Effect of Sample Size on Sampling Distribution

Population  $\mu = 51800$ ; Population  $\sigma = 4000$



**For Margin = 500**

@30:  $z = \pm 0.68$  or probability =  $2 \times 0.2518 = 0.5036$

@100:  $z = \pm 1.25$  or probability =  $2 \times 0.3944 = 0.7888$

# Effect of Sample Size on Sampling Distribution's Mean

n	30	100
$E(\bar{x})$	51800	51800
$\sigma$	4000	4000
$\sigma_{\bar{x}}$	$4000/\sqrt{30} = 730.3$	$4000/\sqrt{100} = 400$
z	$500/730.3 = 0.68$	$500/400 = 1.25$
$P(z \leq 0.68)$	0.7517	.8944
$P(z \leq -0.68)$	0.2483	.1056
$P(-.68 \leq z \leq 0.68)$	0.5034	.7888

# Sampling Distribution for Sample Proportion

1. Nominal class variables are measured by proportions.
2. Proportions are relative frequency counts of class values/responses.
3. Do girls prefer medicine to engineering? Are overseas visitors safe in India? What proportion of people like Modi?
4. Here we would like to draw conclusions on population based on sample proportions.
5. If 70% of girls prefer medicine then  $P_{\text{MEDICINE}} = 0.7$
6. If 57% of voters like Modi then  $P_{\text{MODI}} = 0.57$
7. When we draw sample we may get different values for the proportion of same class.
8. The distribution of  $\bar{p}$  is the sampling distribution of proportion.

# Statistics for Proportions

Expected value of sample proportion =  $E(\bar{p}) = p$

Standard Deviation:

Finite Population	Infinite Population
$\sigma_{\bar{p}} = \sqrt{\left(\frac{N - n}{N - 1}\right) \left(\frac{p(1 - p)}{n}\right)}$	$\sigma_{\bar{p}} = \sqrt{\left(\frac{p(1 - p)}{n}\right)}$

The sampling distribution of sample proportion  $\bar{p}$  follows normal distribution when the following conditions are met.

1.  $np > 5$
2.  $n(1-p) > 5$

Expected value of sample proportion =  $E(\bar{p}) = p$

Standard deviation of  $\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$

# Examples

1.  $p = 0.6$  and  $n = 30$ . What is the probability of obtaining  $\bar{p}$  within  $\pm 0.05$  of  $p$ . Calculate same for sample size 100.
2. A sample of 20 girls out of 30 have indicated preference for medicine. Historically it is known that 70% of girls prefer medicine. Calculate  $\bar{p}$  and  $\sigma_{\bar{p}}$ . What is the probability of obtaining  $\bar{p}$  value between 0.65 and 0.75.



# Effect of Sample Size on Sampling Distribution Proportions

n	30	100
$E(\bar{p})$	0.6	0.6
$\sigma$	0.49	0.49
$\sigma_{\bar{p}}$	$0.49/\sqrt{30} = 0.0894$	$0.49/\sqrt{100} = 0.049$
z	$0.05/0.0894 = 0.56$	$0.05/0.049 = 1.02$
$P(z \leq 0.56)$	0.7123	.8461
$P(z \leq -0.56)$	0.2877	.1539
$P(-.56 \leq z \leq 0.56)$	0.4246	.6922

Estimation

# Properties of Point Estimators

- In the preceding slides we saw that sample mean and sample standard deviation provide one value (not a range of values) to estimate population mean and standard deviations.
- These are called point estimators.
- The important properties of point estimators are
  1. Unbiasedness: A sample statistic  $\hat{\theta}$  is an unbiased estimator if  $E(\hat{\theta}) = \theta$ . For a biased estimator the amount of bias =  $|E(\hat{\theta}) - \theta|$
  2. Efficiency: The point estimator with smallest standard deviation for a given sample size is said to be most efficient.
  3. Consistency: A point estimator is said to be consistent if it becomes more precise as the sample size  $n$  increases.

# Interval Estimation of Population Mean

$\bar{x}$  is a point estimator of  $\mu$ . It does not tell us anything about margin of error we are committing in the process of estimation.

If we like to have an idea of error associated with a point estimator we need to use process of interval estimation.

We know that the distribution sample mean has an expected value of  $\mu$  and standard deviation of  $\sigma/\sqrt{n}$ .

If we take a large sample and know the population standard deviation we can determine the sample standard deviation.

With knowledge of point estimation and sampling distribution we can make statement about precision of our estimate.

# Interval Estimation of Population Mean

A point estimate is a single value that provides an approximate value to corresponding population parameter.

The amount of approximation is indicated by margin of error.

The information about value of population parameter along with margin of error is called interval estimate.

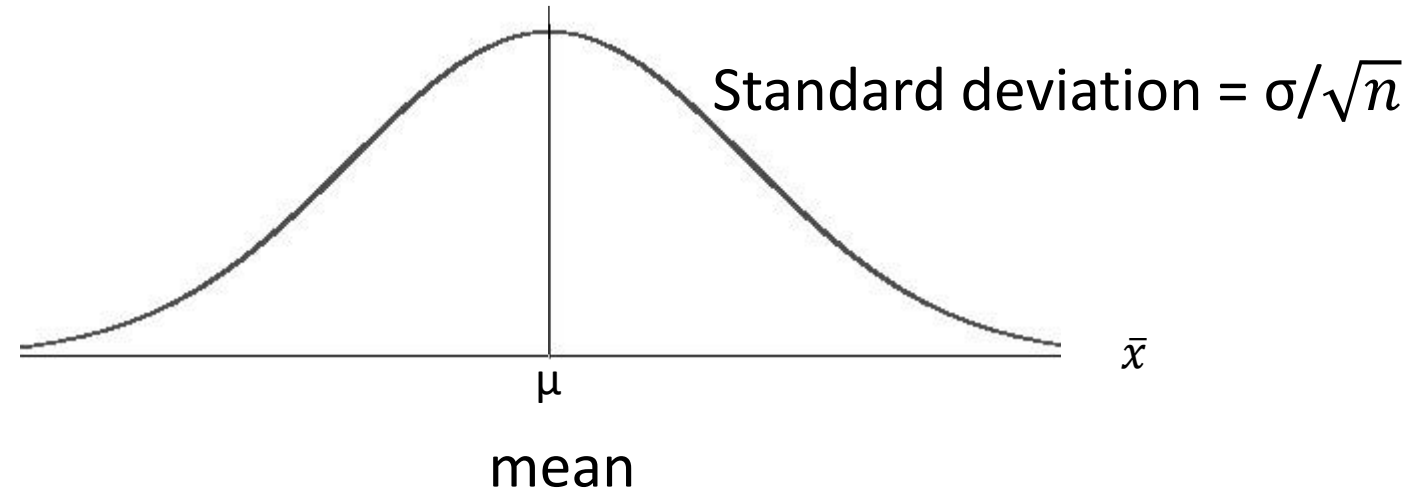
Interval Estimate = Point Estimate  $\pm$  Margin of Error.

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

$$p = \bar{p} \pm \text{Margin of Error}$$

# Interval Estimation

Sampling Distribution of  $\bar{x}$



If the population  $\sigma$  is 20 and sample size is 100 then the standard deviation (standard error) of sampling distribution is  $20/\text{sqrt}(100) = 2$ .

# Interval Estimation for Population Mean – Large Sample

Sample size  $\geq 30$ , therefore CLT applies. Normality assumed.

Assume  $\sigma$  known.

Sample size = 100, Sample mean  $\bar{x} = 82$ ,  $\sigma = 20$ .

For a standard normal distribution 95% of observations lie within  $\pm 1.96$  standard deviations.

Since  $\bar{x}$  follows normal distribution 95% of values of  $\bar{x}$  lie within  $\pm 1.96 \sigma_{\bar{x}}$ .

Since  $\sigma_{\bar{x}} = 2$ , 95% of  $\bar{x}$  values lie within  $\mu \pm 3.92$

# Confidence Interval for $\mu$

$$\mu \pm ME = \bar{x}$$

Therefore

$$\mu = \bar{x} \mp ME$$

Confidence Interval for

$$CI(\mu) = \bar{x} \mp ME$$

Calculate Confidence Interval of  $\mu$  for the previous example.

The size of the confidence interval depends on and increases with confidence level.



## Interval Estimate for Population Mean ( $n \geq 30$ )

$$\bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Where

$(1-\alpha)$  is confidence coefficient.

Generally confidence level is expressed in % terms and confidence coefficient is expressed as a fraction.  $\alpha$  is called level of significance.

To construct a 95% confidence interval

C. Coeff =  $(1 - \alpha) = 0.95$ ;  $\alpha = 0.05$ ;  $\alpha/2 = 0.025$

$$CI = \bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

For the case given in example

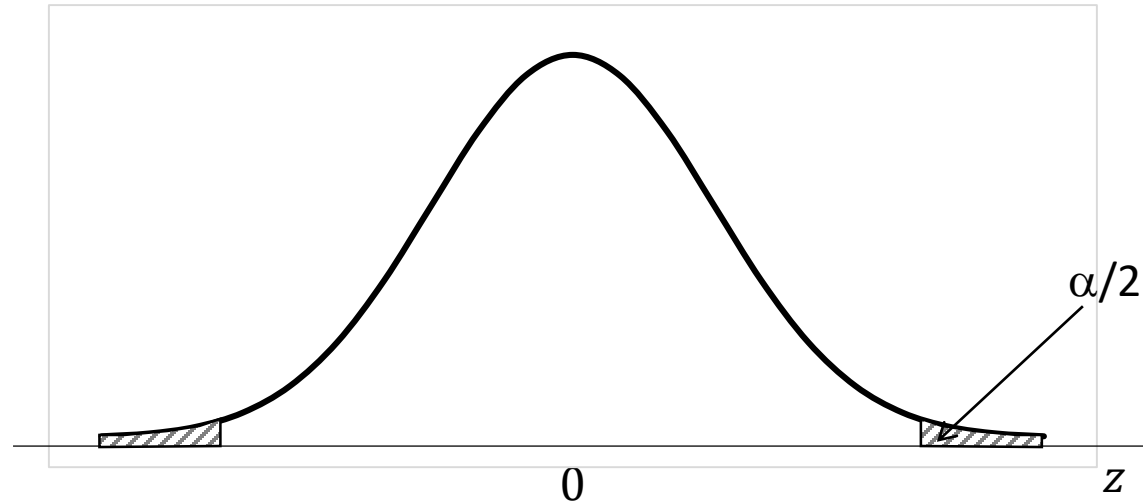
$$CI = 82 \pm 1.96 \times 3.92 = 78.08 \text{ to } 85.92$$

# Interval Estimation of Population Mean

CL = Confidence Coefficient =  $(1 - \alpha) = 95\%$

$z_{\alpha/2}$  = is the z value providing an area of  $\alpha/2$  in the upper tail

Margin of Error = Confidence Interval =  $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$



Confidence Level	$\alpha$	$\alpha/2$	$z_{\alpha/2}$
90%	0.10	0.05	1.645
95%	0.05	0.025	1.960
99%	0.01	0.005	2.576

# Example

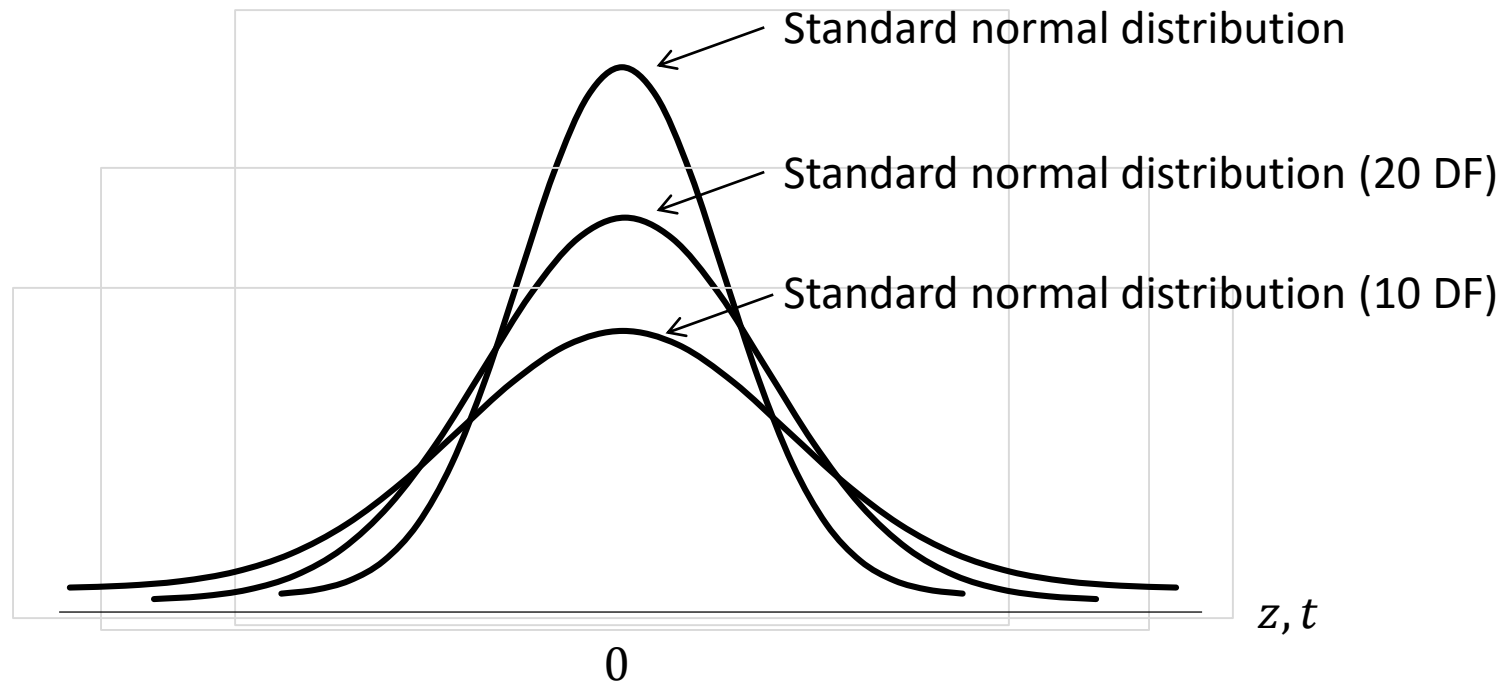
1. An online mail order company instituted a survey to assess the satisfaction of its customers. The overall score is out of 100 points. The company obtained a rating of 82 from a sample of 100 customers. It is known from historical records that the overall rating has a standard deviation of 20. What is the 95% confidence interval for overall score?
2. The score obtained by students in an aptitude test is distributed normally with standard deviation 12. A sample of 16 students produced a mean of 60. What is the 90% confidence interval for the population mean score?
3. Rework problem 1 with 95% confidence interval.
4. Rework problem 2 with 99% confidence interval.

# When $\sigma$ is Unknown Use $s$ to Estimate $\sigma$

1. The credit card balances for 85 US households yielded an average balance of \$5900 and a standard deviation of 3058. At 95% confidence interval estimate the population mean credit card balance.
2. A company is planning to buy a machine that can save significant labour hours. The labour hours saved by the machine follows a normal distribution. A sample of 36 observations indicated that on the average it saves 2200 labour hours and that there is a 50% chance that the average labour hours saved falls above 2400 or below 2000.
  - a. What is the population standard deviation?
  - b. What is the standard error in estimating the mean?
  - c. What is the 95% confidence interval for the mean?

# Small Sample Cases T- Distribution

When we have a sample of size  $< 30$ , population follows normal distribution and standard deviation is unknown, in order to estimate population mean we use student “t” distribution with  $n-1$  degrees of freedom.



# T- Distribution Table

Degrees of Freedom	Area in Upper Tail					
	.2	.1	.05	.025	.01	.005
1	1.3764	3.0777	6.3138	12.7062	31.8205	6.3138
2	1.0607	1.8856	2.9200	4.3027	6.9646	2.9200
3	.9785	1.6377	2.3534	3.1824	4.5407	2.3534
4	.9410	1.5332	2.1318	2.7764	3.7469	2.1318
5	.9195	1.4759	2.0150	2.5706	3.3649	2.0150
6	.9057	1.4398	1.9432	2.4469	3.1427	1.9432
7	.8960	1.4149	1.8946	2.3646	2.9980	1.8946
8	.8889	1.3968	1.8595	2.3060	2.8965	1.8595
9	.8834	1.3830	1.8331	2.2622	2.8214	1.8331
10	.8791	1.3722	1.8125	2.2281	2.7638	1.8125
11	.8755	1.3634	1.7959	2.2010	2.7181	1.7959
12	.8726	1.3562	1.7823	2.1788	2.6810	1.7823
13	.8702	1.3502	1.7709	2.1604	2.6503	1.7709
14	.8681	1.3450	1.7613	2.1448	2.6245	1.7613
15	.8662	1.3406	1.7531	2.1314	2.6025	1.7531

# T- Distribution Table

Degrees of Freedom	Area in Upper Tail					
	.2	.1	.05	.025	.01	.05
16	.8647	1.3368	1.7459	2.1199	2.5835	1.7459
17	.8633	1.3334	1.7396	2.1098	2.5669	1.7396
18	.8620	1.3304	1.7341	2.1009	2.5524	1.7341
19	.8610	1.3277	1.7291	2.0930	2.5395	1.7291
20	.8600	1.3253	1.7247	2.0860	2.5280	1.7247
21	.8591	1.3232	1.7207	2.0796	2.5176	1.7207
22	.8583	1.3212	1.7171	2.0739	2.5083	1.7171
23	.8575	1.3195	1.7139	2.0687	2.4999	1.7139
24	.8569	1.3178	1.7109	2.0639	2.4922	1.7109
25	.8562	1.3163	1.7081	2.0595	2.4851	1.7081
26	.8557	1.3150	1.7056	2.0555	2.4786	1.7056
27	.8551	1.3137	1.7033	2.0518	2.4727	1.7033
28	.8546	1.3125	1.7011	2.0484	2.4671	1.7011
29	.8542	1.3114	1.6991	2.0452	2.4620	1.6991
30	.8538	1.3104	1.6973	2.0423	2.4573	1.6973

# Small Sample Cases T- Distribution

A computer aided training program is expected to reduce the s/w project completion time. A sample of 20 project completion times is given. Develop a 95% Confidence interval for the mean completion time.

Employee	Time	Employee	Time	Employee	Time	Employee	Time
1	52	6	59	11	54	16	42
2	44	7	50	12	42	17	48
3	55	8	54	13	60	18	55
4	44	9	62	14	62	19	57
5	45	10	46	15	43	20	56



# Interval Estimation of Population Proportion

CL = Confidence Coefficient =  $(1 - \alpha) = 95\%$

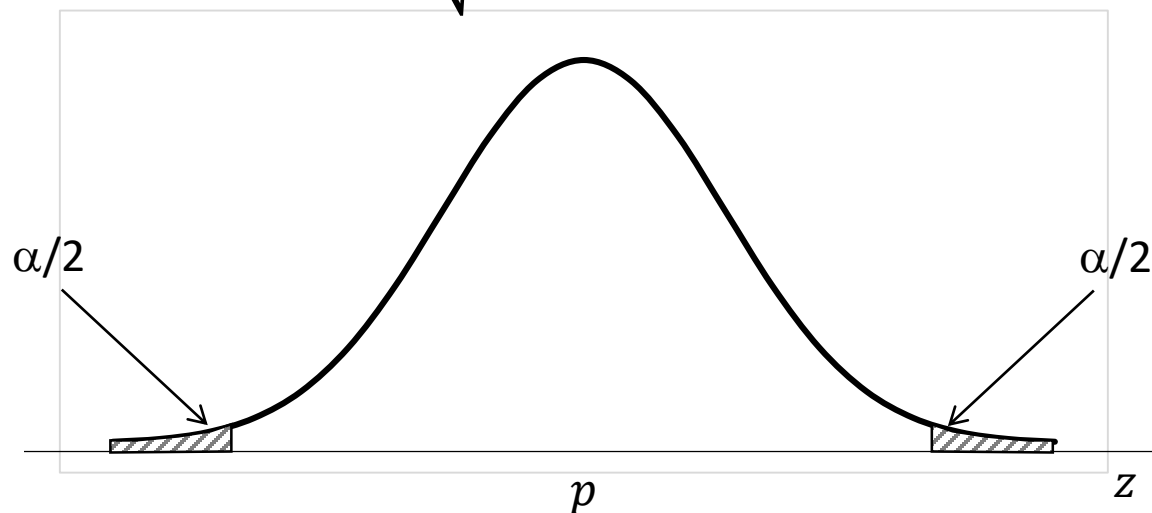
$z_{\alpha/2}$  = is the z value providing an area of  $\alpha/2$  in the upper tail

Margin of Error = Confidence Interval =  $\bar{p} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Point estimator of  $p = \bar{p}$

$$\sigma_p = \sqrt{p(1 - p)}$$

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



# Example

A hotel is interested in assessing the satisfaction level of its customers with respect to the waiting time to be seated on arrival. 902 customers were surveyed and 397 of them were satisfied with that service.

1. What is the point estimate for proportion of satisfied customers?
2. What is critical value for  $z_{\alpha/2}$  at 95% confidence level?
3. Estimate confidence interval

# Sample Size Determination

The quantity  $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$  is called margin of error.

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

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

$$n = (z_{\alpha/2})^2 \frac{\sigma^2}{E^2}$$

If we know confidence level, standard deviation and desired margin of error we can calculate the required sample size  $n$ .

# Sample Size Determination

The HR director of an IT firm wants to estimate average entry level salary for executives. He commissions a survey and wants the mean salary estimated should be within Rs.100. The surveyor from previous survey knows that the standard deviation of entry level salaries is Rs. 500. What should be sample size

$$n = (z_{\alpha/2})^2 \frac{\sigma^2}{E^2}$$

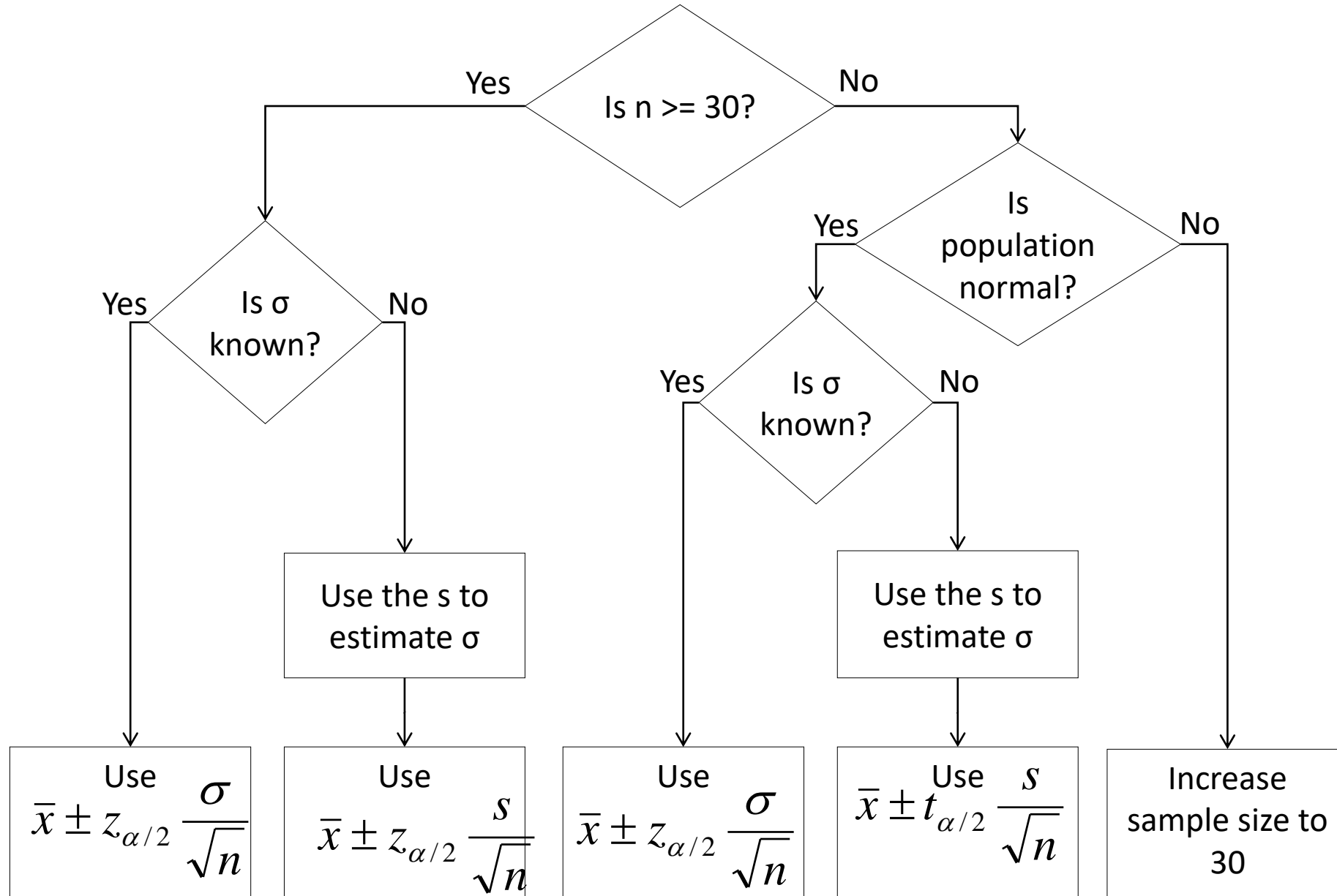
$$n = (1.96)^2 \frac{500^2}{100^2} = 96 \text{ (approx.)}$$

# Sample Size Determination

In 2012, NDTV survey of 814 adults indicated that 562 adults expect Pranab Mukherjee to win the Presidential election. What proportion of Indian adults thought Pranab would win? What was the margin of error? Use 95% confidence level.

Distractors argued that the margin of error and confidence intervals were too big. If you want to limit the error to 1% at 99% confidence what should be the sample size?

# Interval Estimation Procedure



# Confidence Interval Examples

Sample Size	$\bar{X}$	s	$\sigma$	C L	$Z_{\alpha/2}$	$t_{\alpha/2}$	CI	$\mu$
64	90		1.6	95%				
100	82		20	99%				
64	105	3.5		90%				
10	40	10		95%				

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$



# Hypothesis Testing

# What is Hypothesis Testing?

- In the unit on Estimation we saw that a sample can be used to calculate the population parameter within a specified degree of confidence.
- The same principle is carried forward in this unit to test a statement about population, given sample data.
- In Hypothesis Testing we make a tentative assumption about population parameter. This assumption expressed as statement is called **Null Hypothesis ( $H_0$ )**.
- We then define an opposite statement called **Alternate Hypothesis( $H_a$ )**.
- Only one, either  $H_0$  or  $H_a$  can be true. The Hypothesis testing procedure involves collecting data to ascertain which one of the two is true.

# Hypothesis Testing

Test Type	Description	Ho / Ha
Testing Research Hypothesis	R&D dept has developed a new product. The present performance is 24. The new product has a higher performance.	Ho: $\mu \leq 24$
		Ha: $\mu > 24$
Validating a claim	A Pepsi can carries a label - contents: 300 ml (Customer claims otherwise)	Ho: $\mu \geq 300$
		Ha: $\mu < 300$
Testing Decision making situations	The mean length of a supplied part is 2 inches (Deviation from this is not acceptable)	Ho: $\mu = 2$
		Ha: $\mu \neq 2$

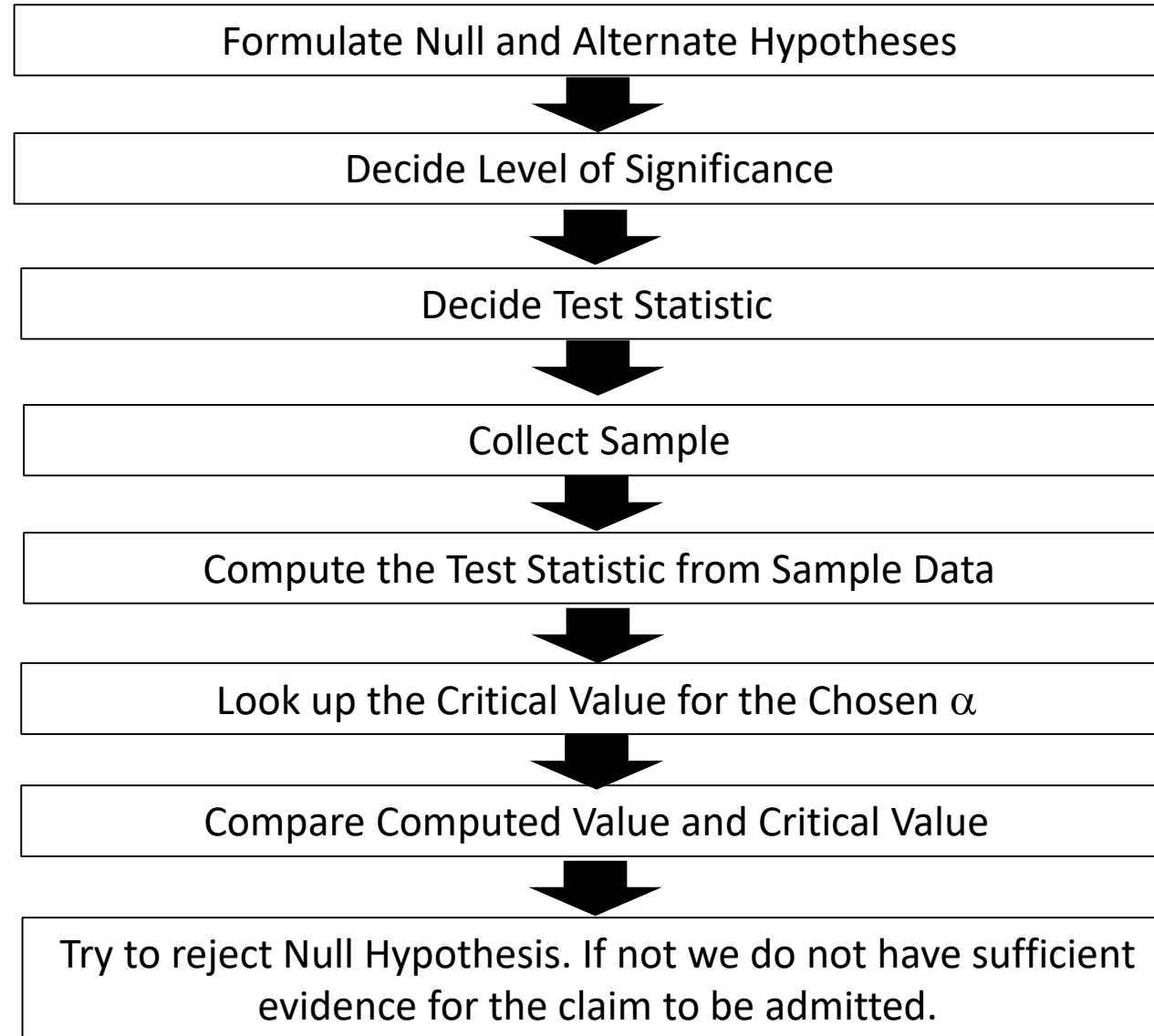
# Example Statements

- Bank: Average waiting time at its ATM is 15 minutes.
- LIC: Average commission of its agents is ₹ 15000.
- MTC: On the average 1000 people travel on route no 19B.
- Tata: Nano model gives mileage better than 24 kmpl.
- Complan: Children drinking Complan grow faster.
- Call-Center: It receives 15 calls per person hour.
- HLL: Pureit gives water with TDS of 400 ppm.
- Business School: Students earn average CTC ₹ 4 lakhs/annum.
- Pharma: Crocin contains 500mg of paracetamol.
- No. of school dropouts has reduced.
- LED bulbs have a life of more than 18000 hours

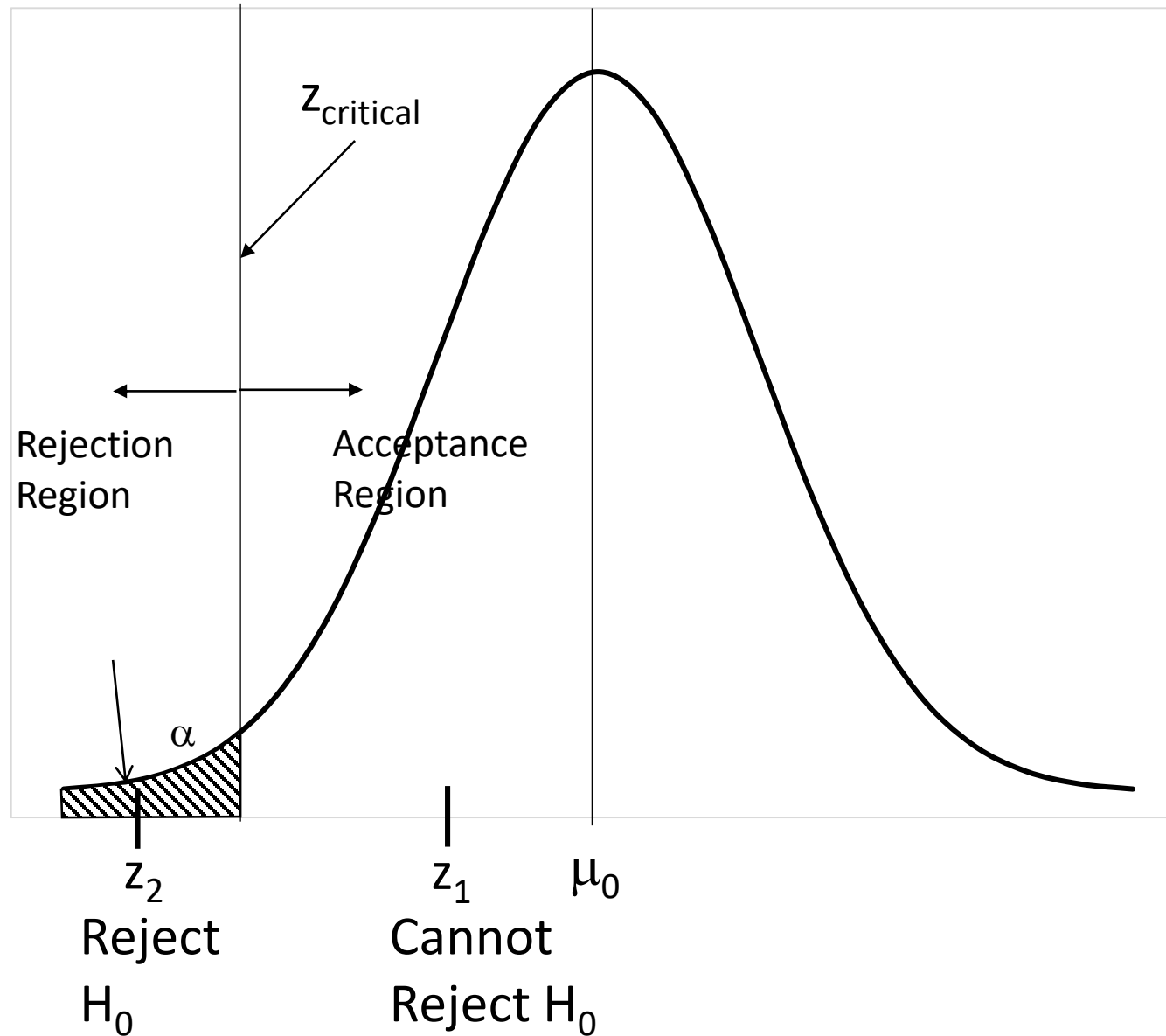
# Example Statements Continued

- More people support life.
- This year our company made more profit than last year (₹ 121cr)
- The performance of girls and boys is different.
- The new drug is better in controlling BP.
- Analytics course improves career prospects.

# Hypothesis Testing Procedure



# Hypothesis Testing



# Type Errors

		Population Condition	
		$H_0$ is True	$H_0$ is False
Conclusion	Accept $H_0$	Correct Conclusion	Type II Error
	Reject $H_0$	Type I Error	Correct Conclusion



# Summary of Forms of Null and Alternate Hypothesis

$$H_o: \mu \leq \mu_0$$

$$H_o: \mu \geq \mu_0$$

$$H_o: \mu = \mu_0$$

$$H_a: \mu > \mu_0$$

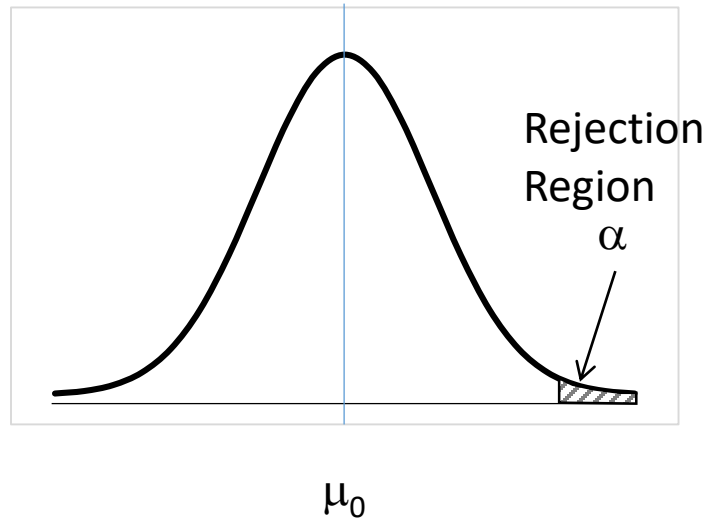
$$H_a: \mu < \mu_0$$

$$H_a: \mu \neq \mu_0$$

# One Tailed Tests

$$H_o: \mu \leq \mu_0$$

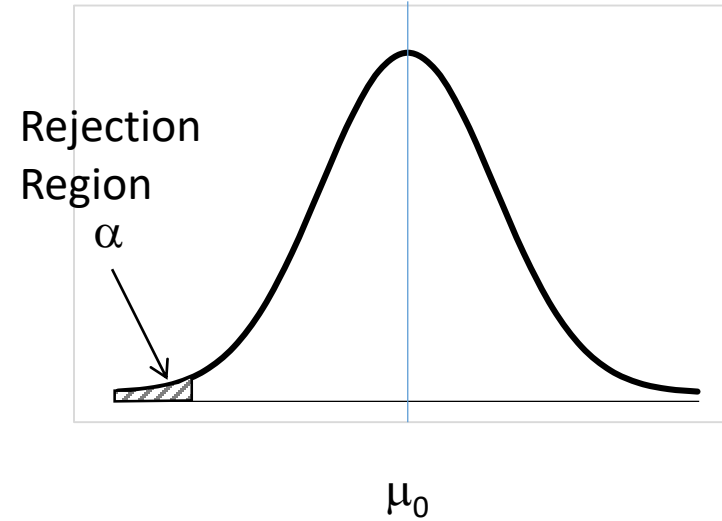
$$H_a: \mu > \mu_0$$



Right tail test

$$H_o: \mu \geq \mu_0$$

$$H_a: \mu < \mu_0$$

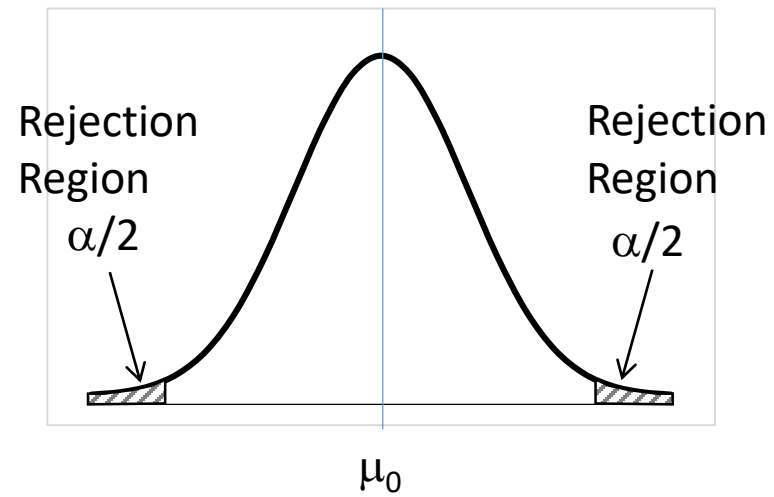


Left tail test

# Two Tailed Test

$$H_0: \mu = \mu_0$$

$$H_a: \mu \neq \mu_0$$

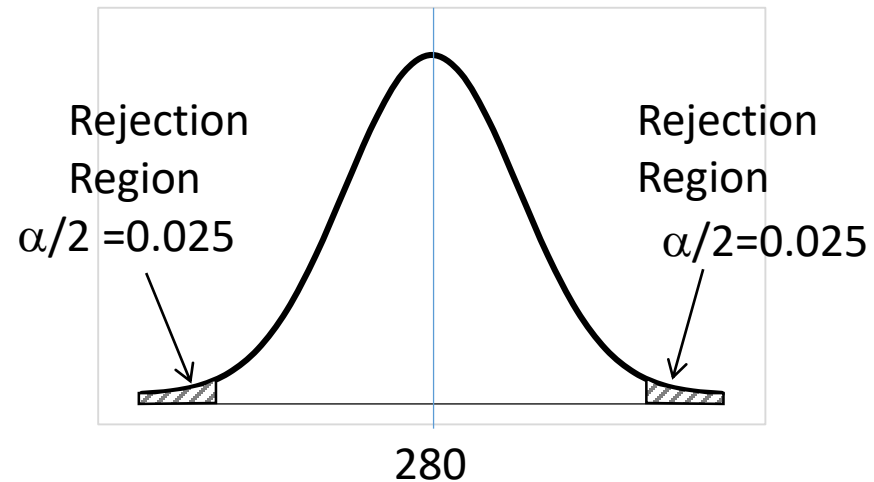


## Two Tailed Test

The weight of a drug has to be critically controlled. If less the drug may not be effective, if more it may have some side effects. The manufacturer of the drug wants ascertain that the weight of the capsule is 280 mg at 95% confidence.

$$H_o: \mu = 280(\mu_o)$$

$$H_a: \mu \neq 280(\mu_o)$$



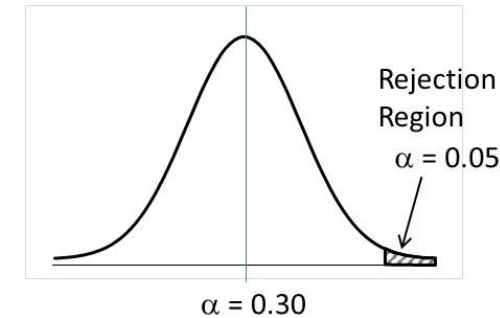
CL	95%	$\sigma$ (known)	18
Number of Tails	2	$z = (278.5 - 280) / (\sigma / \sqrt{n})$	-0.75
$\alpha$	0.05	$Z_{crit}$	-1.96
$\mu_o$	280	p value	.2266
$\bar{x}$	278.5	Decision: Cannot reject $H_o$	

# Hypothesis Testing – Proportions

A marketing manager wants to introduce a new product. A blind comparison test is made with a sample size of 200 to assess customer acceptance of the product. The product will be launched only if favorable response is 30% plus. The product found favor with 64 respondents. Should the product be launched?

$$H_o: p \leq .3 (\mu_0)$$

$$H_a: p > .3 (\mu_0)$$



C L	95%	$\sigma$ (est)	0.458
Number of Tails	1	$z = (.32 - .3)/(.458/\sqrt{200})$	0.617
$\alpha$	0.05	$Z_{crit}$	1.65
$\mu_0$	0.3	p value	0.27
$\bar{p}$	0.32	Decision: Cannot reject $H_o$	

# Hypothesis Testing – Small Sample $\sigma$ Estimated

A purchase manager wants to check the diameter of aluminum die cast sourced from external supplier. The required diameter is 33 and is normally distributed. A sample was taken and is observed to have

$$n = 8$$

$$\bar{x} = 31.5$$

$$s = 1.3$$

Should the purchase manager accept the lot at 90% confidence?

$$H_o: \mu = 33(\mu_0)$$

$$H_a: \mu \neq 33(\mu_0)$$

$$\sigma_{\bar{x}} = 1.3/\sqrt{8} = .46$$

$$t = (31.5 - 33)/0.46 = -3.26$$

$$t_{\text{crit}} = \text{TINV}(0.05, 7) = -2.36$$

Decision: Reject  $H_o$

# Usual Z & t Values

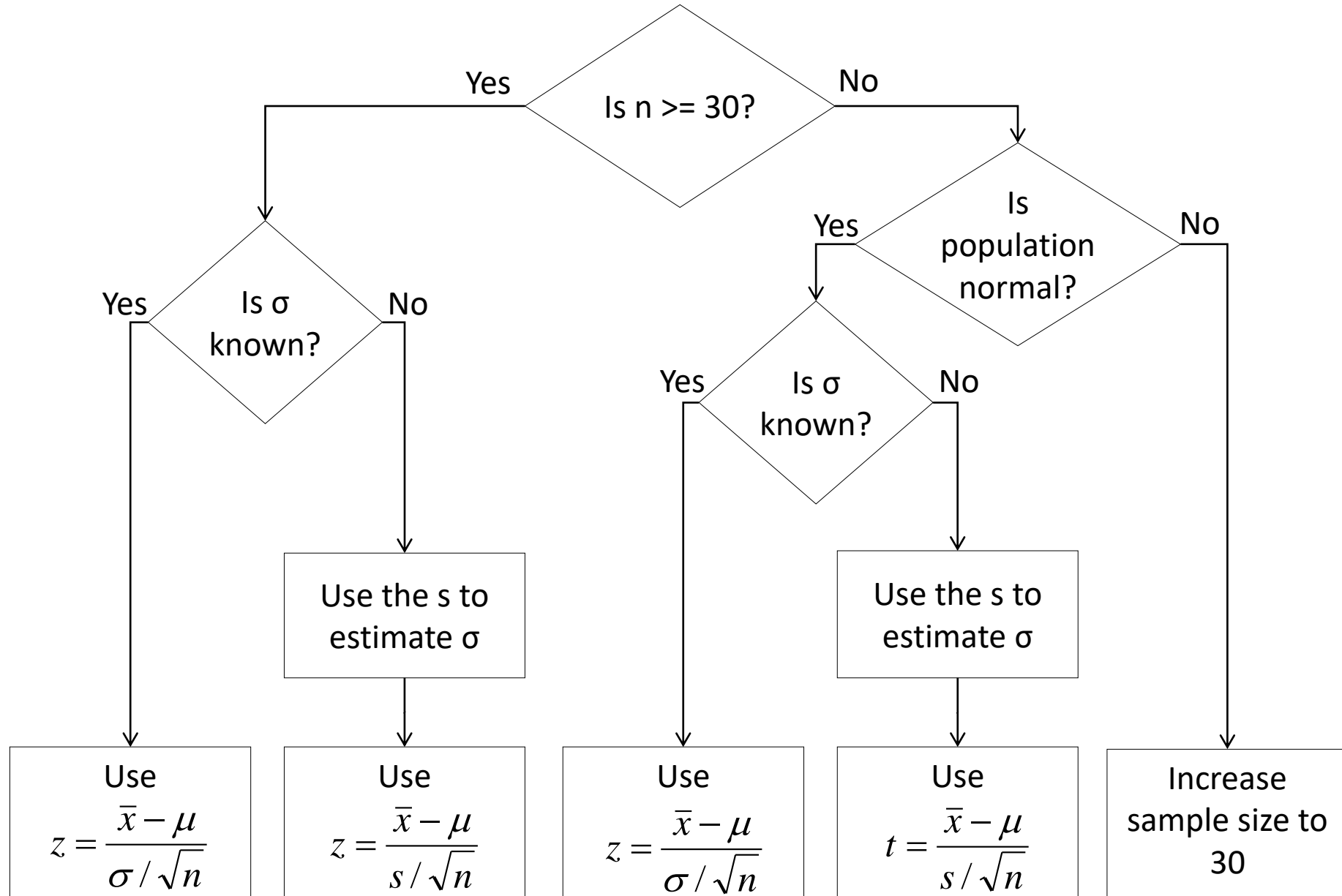
<b>Confidence Level</b>	<b>No of Tails</b>	<b>Level of Significance</b>	<b>Z<sub>critical</sub></b>	<b>t value (10 df)</b>	<b>t value (30 df)</b>
<b>90%</b>	<b>1</b>	<b>0.10</b>	<b>1.28</b>	<b>1.37</b>	<b>1.31</b>
	<b>2</b>	<b>0.05</b>	<b>1.65</b>	<b>1.81</b>	<b>1.69</b>
<b>95%</b>	<b>1</b>	<b>0.05</b>	<b>1.65</b>	<b>1.81</b>	<b>1.69</b>
	<b>2</b>	<b>0.025</b>	<b>1.96</b>	<b>2.23</b>	<b>2.04</b>
<b>99%</b>	<b>1</b>	<b>0.01</b>	<b>2.33</b>	<b>2.76</b>	<b>2.46</b>
	<b>2</b>	<b>0.005</b>	<b>2.58</b>	<b>3.17</b>	<b>2.75</b>

# Tests for Population Mean

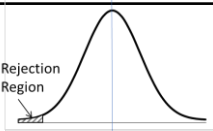
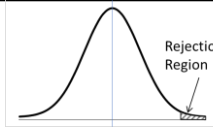
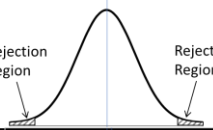
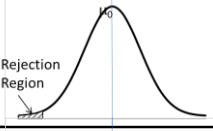
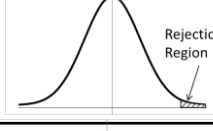
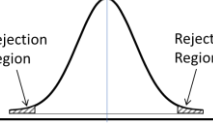
Type	Sample Size	Test Statistic
Population Mean	Large	$Z = \frac{(\bar{X} - \mu_0)}{\sqrt{(\sigma^2 / n)}}$
Population Mean	Small	$t = \frac{(\bar{X} - \mu_0)}{\sqrt{(s^2 / n)}}$
Population Proportion	Large	$Z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$



# Hypothesis Testing Procedure



# Hypothesis Tests in Python

Sample size	Statistic	Tails	Significance	For Critical Value		For p-Value	Rejection Region
				z or t < 0	z or t > 0		
Large ( $\geq 30$ )	z	Left	alpha	norm.ppf		norm.cdf	
		Right	alpha	norm.isf		1 - norm.cdf	
		2-Tail	alpha/2	norm.ppf	norm.isf	2*max(norm.cdf, 1 - norm.cdf)	
Small ( $< 30$ )	t	Left	alpha	t.ppf		t.cdf	
		Right	alpha	t.isf		1 - t.cdf	
		2-Tail	alpha/2	norm.ppf	norm.isf	2*max(t.cdf, 1 - t.cdf)	

$\mu_0$

# Bivariate Tests

# Bivariate Tests – Tests for Mean Difference

Type	Sample Size	Test Statistic	
Independent Sample	Large	$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)}}$	
Independent Sample	Small	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$	$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$
Dependent Sample	Large	$Z = \frac{\bar{D}}{\sigma / \sqrt{n}}$	$D = X_1 - X_2$
Dependent Sample	Small	$t = \frac{\bar{D}}{s / \sqrt{n}}$	$D = X_1 - X_2$

# Testing for Mean Difference of Two Population Independent Samples – Large Sample Case

- Since it is large sample, we use z statistic
- Let  $\mu_1$  be the mean of population 1,  $\mu_2$  be the mean of population 2 and  $\mu_1 - \mu_2$  is the population mean difference.
- Let  $\bar{x}_1$  be the mean of sample from population 1,  $\bar{x}_2$  be the mean of sample from population 2 and  $\bar{x}_1 - \bar{x}_2$  is the population mean difference.
- The point estimator of  $(\mu_1 - \mu_2)$  is  $(\bar{x}_1 - \bar{x}_2)$
- $(\bar{x}_1 - \bar{x}_2)$  is a RV with
  - Expected value =  $(\mu_1 - \mu_2)$  and
  - Standard Deviation =  $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
  - If  $\sigma$  is unknown estimate using sample SD.

# Testing for Mean Difference of Two Population Independent Samples – Small Sample Case

- Let  $\bar{x}_1 - \bar{x}_2$  be the population mean difference .
- In this case we use t statistic.
- Assuming equal variances for both populations
  - If population  $\sigma$  is known, use it
  - Otherwise we use pooled variance given by
- $$s^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1+n_2-2)}$$
- Sample SD  $s_{(\bar{x}_1 - \bar{x}_2)}^2 = \sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$

# Testing for Mean Difference of Two Population Matched Samples

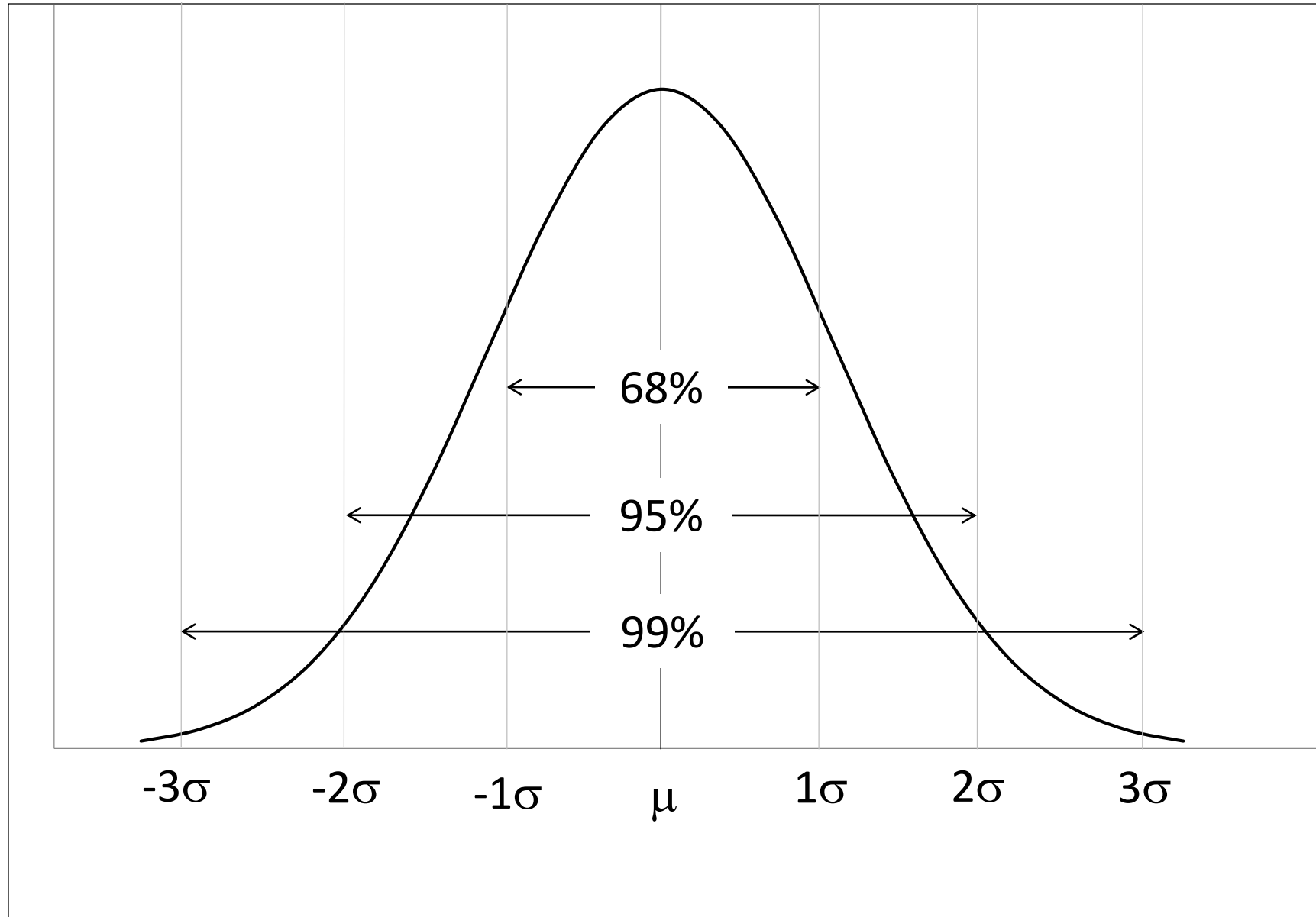
- In independent sample design, a sample of observations is used for treatment1 and another sample of observations is for treatment2.
- In the matched sample design same sample of observations is used for both the treatments.
- Then  $\mu_d$  is the mean of differences between treatment 1 and treatment 2.
- $\bar{d} = \frac{\sum d_i}{n}$
- $s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}}$

# Bivariate Tests – Exercises

1. A test was conducted to determine if there is any difference in the performance of boys and girls in the participants. Data given in accompanying excel sheet. Did boys and girls perform equally?
2. An aptitude test was conducted in which groups of engineering and accounting executives participated. Is there a significant difference in the performance the groups?
3. A company conducted a campaign to promote the sales of its products. The sales revenues for 10 of its retail outlets is given in the accompanying excel sheet. Is there a significant difference?



# The Normal Curve



# Chi Square & ANOVA

# Univariate Data

Univariate data has a single variable

Examples:

- Heights of people
- Profits of a company
- Temperature of a city

We used descriptive summaries, Histograms, Box plots to examine univariate data

# Bivariate Data

Bivariate data has two variables and studies relationship between them.

Examples:

- Relationship between age and height
- Relationship between profit and number of people employed.

The type of analysis used depends on the type of variables in the data set.

<b>Dependent Var</b>	<b>Independent Var</b>	<b>Technique</b>
Categorical	Categorical	Chi-Square
Continuous	Categorical	ANOVA
Continuous	Continuous	Regression

# Chi Square ( $\chi^2$ ) Test



## CONTENTS

- Important Terms.
- Introduction.
- Characteristics of Tests.
- Chi Square Distribution.
- Applications of Chi Square ( $\chi^2$ ) Test.
- Calculation of Chi-Square.
- Conditions for the application of the Tests.
- Examples.
- Limitations of the Test.

# IMPORTANT TERMS

## Parametric Tests

The tests in which the population constants like mean, standard deviation, standard error, correlation coefficient, proportion etc. are used and data tend to follow one assumed or established distributions such as normal, binomial, Poisson etc.

## Non Parametric Tests

The tests in which no constant of population is used. Data do not follow any specific distribution and no assumptions is made in these tests. e.g. To classify good better and best we do not allot arbitrary numbers or marks to each category.

## Hypothesis:

It is a definitive statement about population parameter.

# IMPORTANT TERMS

## Null Hypothesis ( $H_0$ ):

States that no association exists between the two cross-tabulated variables in the population and therefore the variables are statistically independents. E.g. if we want to compare two methods say method A and Method B for its superiority and if the assumption is that both methods are equally good then this assumption is called as NULL HYPOTHESIS.

## Alternate Hypothesis ( $H_1$ ):

Proposes that the two variables are related in the population. If we assume that from two methods , method A is superior than method B, then this assumption is called as alternate hypothesis.

# IMPORTANT TERMS

## Degrees of Freedom:

It denotes the extent of independence (freedom) enjoyed by a given set of observed frequencies. Suppose you are given a set of  $n$  observed frequencies which are subject to  $k$  independent constraints (restrictions) then

d.f. = (number of frequencies) – (number of independent constraints on them) in other terms,

$$df = (r-1)(c-1)$$

where

$r$  = the number of rows

$c$  = the number of columns



# IMPORTANT TERMS

## The Contingency Table:

When a table is prepared by enumeration of qualitative data by entering the actual frequencies, and if that table represents occurrences of two sets of events, that table is called the contingency table. (Latin Con-together, tangere – to touch). It is also called association table.

# Chi Square Test

# Test for Independence

- The chi-square ( $\chi^2$ ) test is an important test amongst several tests of significance developed by statisticians
- It was developed by Karl Pearson in 1900
- CHI SQUARE TEST is a non parametric test and is not based any assumption or distribution of any variable.
- The statistical test follows a specific distribution known as chi square distribution
- In general the test we use to measure the difference between what is observed and what is expected according to an assumed hypothesis is called the  $\chi^2$  test.

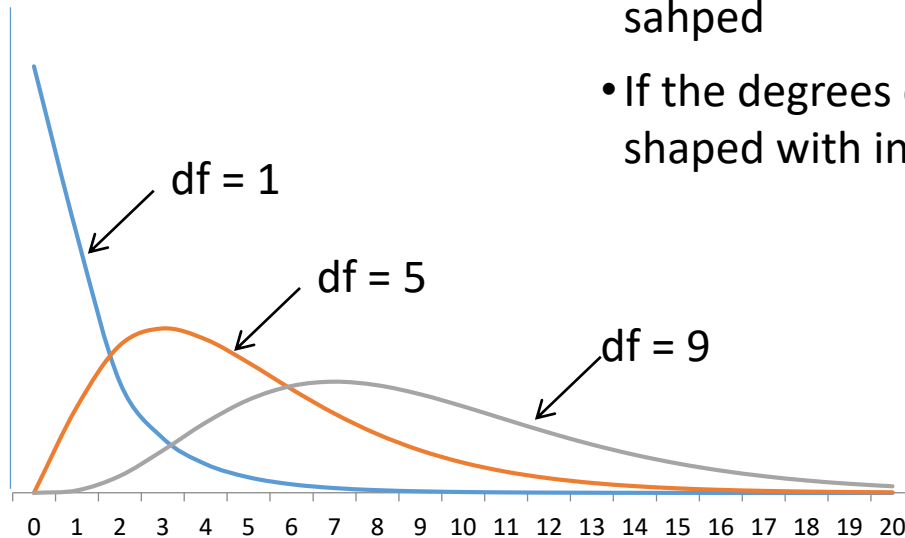
# IMPORTANT CHARACTERISTICS OF $\chi^2$ TEST

- This test ( as a non parametric test) is based on frequencies and not on parameters like mean and standard deviation.
- The test can also be applied to a complex contingency table with several classes and as such is a very useful test in research work.
- This test is an important non-parametric as no rigid assumptions are necessary in regard to the type of population, no need of parameter values and relatively less mathematical details are involved.

# IMPORTANT CHARACTERISTICS OF $\chi^2$ TEST

If  $X_1, \dots, X_n$  are  $n$  independent normal variates and each is distributed normally with mean zero and standard deviation unity then  $X_1^2 + X_2^2 + \dots + X_n^2 = \sum X_i^2$  is distributed as chi-square ( $\chi^2$ ) with  $n$  degrees of freedom (d.f.) where  $n$  is large. The Chi square curve for  $df = 1, 5, 9$  is as shown below:

- If degrees of freedom  $> 2$  the distribution is bell shaped
- If the degrees of freedom  $= 2$  the distribution is L shaped
- If the degrees of freedom  $< 2$  ( $> 0$ ) the distribution is L shaped with infinite ordinate at the origin



# APPLICATION OF CHI SQUARE TEST

This test can be used in

- Goodness of Fit of distribution
- Test for independence of attributes
- Test of homogeneity

# TEST OF GOODNESS OF FIT OF DISTRIBUTIONS

This test enables us to see how well does the assumed theoretical distribution (such as Binomial, Poisson or Normal Distributions ) fit to the observed value.

The chi-square formula for goodness of fit is:

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

Where

$O_i$  = observed frequency

$E_i$  = expected theoretical frequency

If  $\chi^2(\text{calculated}) > \chi^2(\text{tabulated})$  with  $(n-1)$  d.f. then null hypothesis is rejected otherwise it is accepted.

And if null hypothesis is accepted then it can be concluded that the given distribution follows theoretical distribution.

# TEST OF INDEPENDENCE OF ATTRIBUTES

- This test enables us to explain whether or not two attributes are associated.
- For instance we may be interested in knowing whether a new medicine is effective in controlling fever or not,  $\chi^2$  test is useful.
- In such a situation we proceed to with the null hypothesis that the two attributes (viz. new medicine and control fever) are independent which means that the medicine is not effective in controlling the fever.
- $\chi^2$  (calculated)  $>$   $\chi^2$  (tabulated) at a certain level of significance for given degrees of freedom the null hypothesis is rejected i.e. the two variables are dependent (i.e. the new medicine is effective in controlling the fever and if  $\chi^2$  (calculated)  $<$   $\chi^2$  (tabulated) the null hypothesis is accepted i.e. the two variables are independent (i.e. the new medicine is not effective in controlling the fever.)
- When null hypothesis is rejected it can be concluded that there is a significant association between the two attributes.







# Cross Tabulation




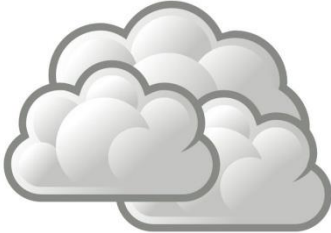
- Cross tabulation is a tabular summary of two variables

	Smoker	Non Smoker
Male		
Female		

# Cross Tabulation

<div><div>Boss Mood</div><div>Weather</div></div>		
	72%	28%
	72%	28%

# Cross Tabulation

<div><div>Boss Mood</div><div>Weather</div></div>		
	82%	18%
	60%	40%

# TEST OF HOMOGENITY

- This test can also be used to test whether the occurrence of events follow uniformity or not. E.g. the admission of patients in government hospital in all days of week is uniform or not can be tested with the help of chi square test.
- $\chi^2$  (calculated) <  $\chi^2$  (tabulated), then null hypothesis is accepted and it can be concluded that there is a uniformity in the occurrence of the events. (uniformity in the admission of patients throughout the week)

# CALCULATION OF CHI SQUARE

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

Where

O = observed frequency

E = expected frequency

If two distributions (observed and theoretical) are exactly alike  $\chi^2 = 0$ ;  
(but generally due to sampling errors,  $\chi^2$  is not equal to zero)

# STEPS INVOLVED IN CALCULATING $\chi^2$

1. Calculate the expected frequencies and the observed frequencies.

## Observed frequencies:

The cell frequencies actually observed in a contingency table.

## Expected frequencies:

The cell frequencies that would be expected in a contingency table if the two variables are statistically independent.

$$f_e = \frac{(\text{column total}) \times (\text{row total})}{N}$$

To obtain the expected frequencies for any cell in any cross tabulation in which the two variables are assumed independent multiply the row and column totals for that cell and divide the product by the total number of cases in the table.

2. Then  $\chi^2$  is calculated as

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

# CONDITIONS FOR APPLICATION OF $\chi^2$ TEST

1. The following conditions should be satisfied before  $\chi^2$  test can be applied:
2. Data must be in the form of frequencies
3. The frequency data must have precise numerical value and must be organized into categories or groups
4. Observations recorded and used are collected on a random basis
5. All the items in the sample must be independent
6. No group should contain very few items, say less than 5. In case where the frequencies are less than 5, regrouping is done by combining the frequencies of adjoining groups so that the new frequencies become greater than 5. (Some statisticians prefer 10, but 5 is ok)
7. The overall number of items must also be reasonable, should at least be 30.

# LIMITATIONS OF $\chi^2$ TEST

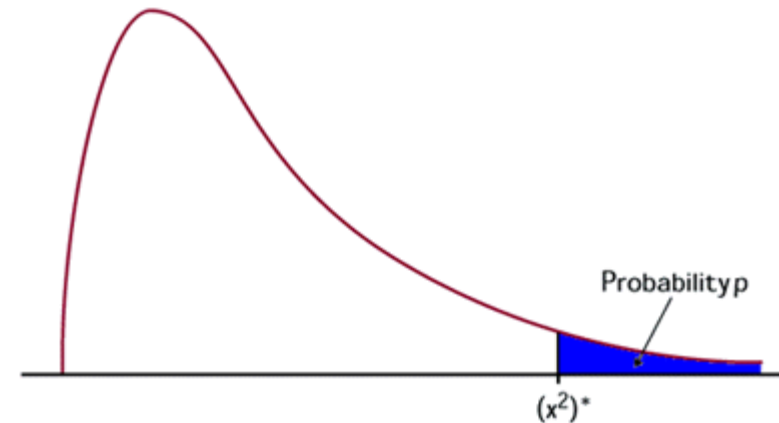
- The data must be from random sample
- The expected frequencies any cell must be  $\geq 5$ .
- The accuracy falls if the total sample size  $< 30$ .
- The test only detects presence or absence of association but not the strength of association
- The test does not tell the cause and effect.
- The sample observations are independent of each other.



# Chi Square ( $\chi^2$ ) Test

- The Chi Square test is a non-parametric Test
- It tests relationship between two nominal variables
- The statistic chi square is given by the equation and has a distribution as shown.

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



The conditions for using chi square

- Data must be counts
- The observations must be random and independent
- None of the cells must have a count less than 5

# Chi Square ( $\chi^2$ ) Test

Business Case:

A consumer survey was conducted for a brand of detergent. One question dealt with the influence income category on purchase intension. The income and purchase intent categories are shown below:

Income Category

LOWER

MIDDLE

UPPER

AFFLUENT

Intent Category

NO

NEUTRAL

YES

Class ↓/PI →	Yes	Neutral	No
Lower	20	30	20
Middle	30	60	25
Upper	6	8	3
Affluent	4	7	12

# Analysis

OBSERVED				
	YES	NEUTRAL	NO	TOTAL
LOWER	20	30	20	70
MIDDLE	30	60	25	115
UPPER	10	15	15	40
	60	105	60	225

EXPECTED				
	YES	NEUTRAL	NO	TOTAL
LOWER	18.67	32.67	18.67	70
MIDDLE	30.67	53.67	30.67	115
UPPER	10.67	18.67	10.67	40
	60	105	60	225

CHISQ = 4.74

CHISQ<sub>CRITICAL</sub> = 9.49

# Exercises

## Test of Goodness of Fit

In order to improve sales a company planned to redesign its product packaging. Five designs were made. In order to ascertain the most preferred design, a survey was conducted and results tabulated. Perform a chi square test to verify if there is preference to any one package.

Package Preference Data																			
D	E	A	E	A	D	C	A	A	E	E	C	D	B	E	D	C	C	E	A
E	E	E	A	E	D	C	B	E	E	D	E	B	B	C	D	D	A	D	B
B	E	B	B	B	E	E	B	B	E	B	D	B	B	B	D	B	D	B	D
B	B	D	E	A	C	B	C	D	B	C	E	D	B	B	E	C	B	B	E
C	C	B	D	E	C	A	B	D	A	C	C	A	D	A	A	C	B	A	A
B	B	A	E	E	C	A	C	D	D	E	D	C	A	A	E	D	A	C	A
B	D	D	C	B	D	C	A	B	E	E	B	B	E	C	C	A	B	A	B
D	B	E	B	B	B	D	B	C	A	B	C	E	A	C	C	C	A	D	B
E	D	A	C	A	C	B	D	B	E	C	A	D	E	A	E	A	B	C	E
B	A	B	C	E	C	A	C	B	C	D	A	C	D	C	B	D	B	C	A

# Exercises

## Test of Independence

A management institute wants to verify if there is an association between branch of management preferred by the students to their education background. The details of student background and branch selected by previous students is given below. Check if there is an association. The survey data is listed in the spreadsheet named chisq exercises.xls

# Exercises

## Goodness of Fit – Poisson Distribution

Food Mart Retail is considering redesigning their checkout counter layout in order to improve customer experience. In this context the engineers want to see if the customer arrivals follow Poisson distribution. A sample of 128 5-Minute periods are furnished. Verify if the observed values align with Poisson distribution.

Number of Customers	0	1	2	3	4	5	6	7	8	9
Observed Frequency	2	8	10	12	18	22	22	16	12	6

# Exercises

## Goodness of Fit – Normal Distribution

TCS recruits thousands of engineers every year. The Director HR would like to know if the test scores by the applicants follow a normal distribution. A sample of 50 scores are given below. Please verify whether the test scores are normally distributed.

EMPLOYEE TEST SCORE				
71	86	56	61	65
60	63	76	69	56
55	79	56	74	93
82	80	90	80	73
85	62	64	54	54
65	54	63	73	58
77	56	65	76	64
61	84	70	53	79
79	61	62	61	65
66	70	68	76	71

ANOVA



# ANOVA Introduction

- The term ANOVA stands for AAnalysis Of Variance. This technique is a part of the domain called experimental design.
- It was invented by Sir Ronald Fisher, considered as “Father of Inferential Statistics”
- It helps in establishing the Cause – Effect relation amongst the variables. It tests for equality of two or more populations.
- Used for comparing the means of two or more populations as an alternate to multiple t tests which inflates Type I error.
- Extensively used in analyzing the results in Regression Analysis.
- Plays a key role in the branch of statistics called Experimental Design.

# ANOVA Introduction

Analysis of Variance is a statistical procedure that can be used to test if significant differences exists in the means of different populations.

Consider three populations with means  $\mu_1$ ,  $\mu_2$ ,  $\mu_3$  respectively. Using sample results we can set up the following NULL and ALTERNATE hypotheses.

Test:

If  $\mu_1$ ,  $\mu_2$ ,  $\mu_3$  are means three populations then

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

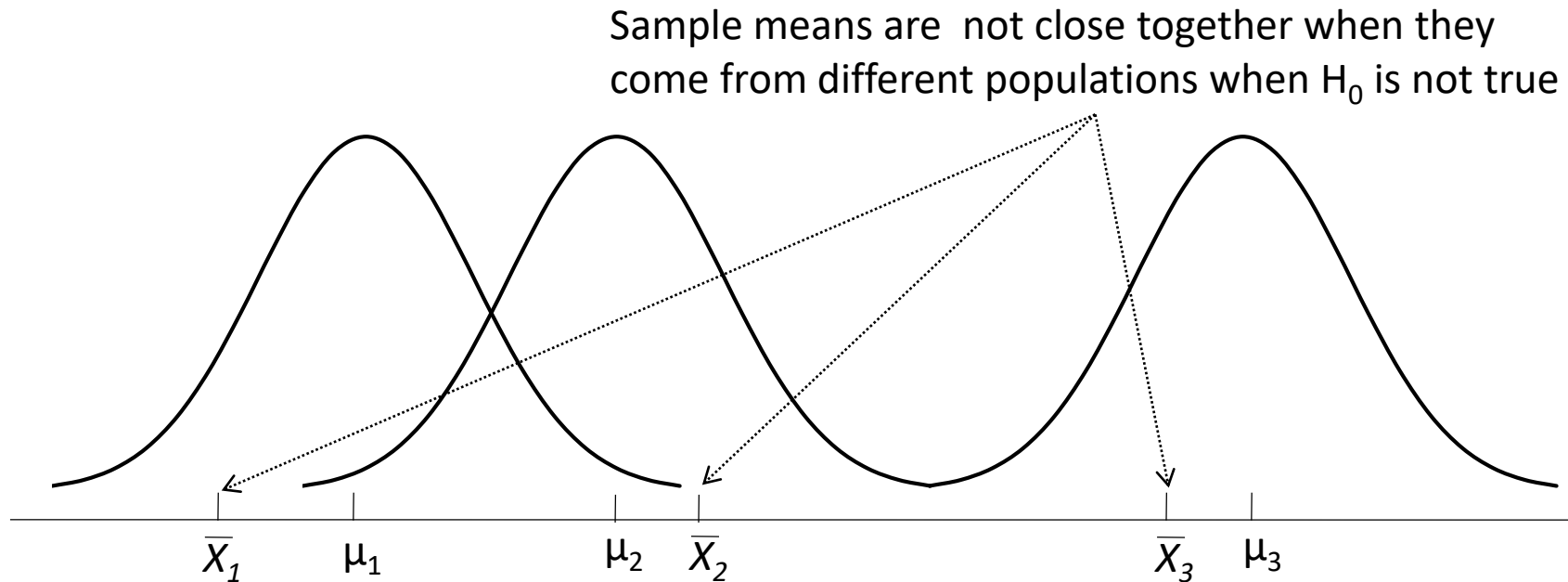
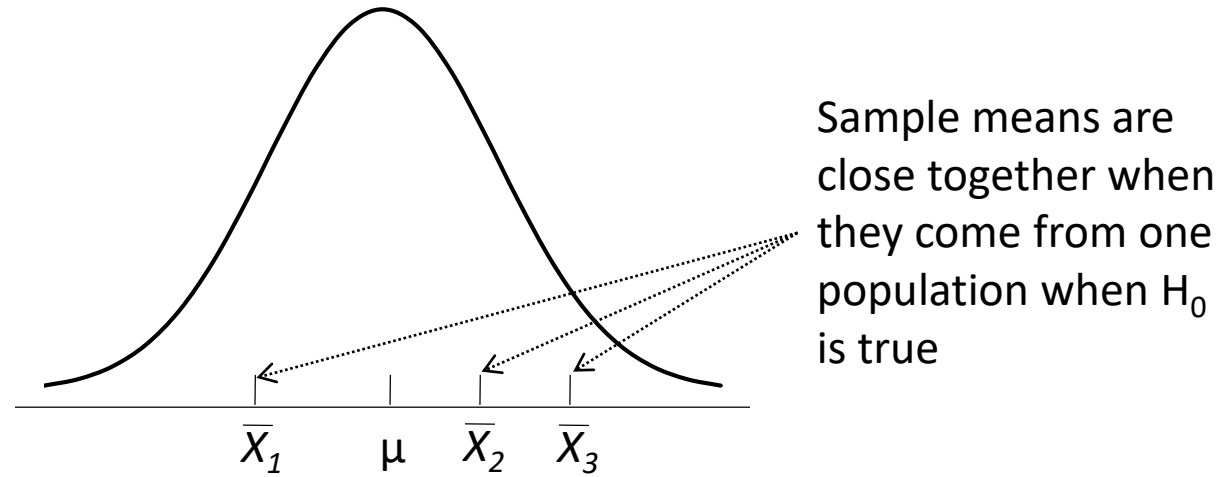
$$H_a : \mu_1 \neq \mu_2 \neq \mu_3$$

ANOVA is used to determine whether the observed differences in the sample means are large enough to reject the NULL hypothesis.

# ANOVA Assumptions

1. For each population, the response variable is normally distributed
2. The variance  $\sigma^2$  of response variable is same for all the distributions.
3. The observations are independent of each other.

# ANOVA Conceptual Overview



# ANOVA Conceptual Overview

- Consider 3 populations with means  $\mu_1, \mu_2, \mu_3$  respectively and we have drawn one sample of size  $n$  from each.
- If the means of the three populations are equal, we would expect sample means to be close together.
- More the sample means differ more is the evidence that population means differ.
- In other words if variability amongst the sample means is small it supports NULL hypothesis, otherwise it supports  $H_a$
- If null hypothesis is true, then we can consider that each sample came from same normal distribution.
- When  $H_0$  is false, we have three sampling distributions. The means are not close together and the variance  $S_{\bar{x}}^2$  will be larger causing  $\sigma^2$  to be larger.
- The ratio of between treatments estimate to within treatment estimate is called F ratio.
- Larger the F ratio, more is evidence that the populations differ.

# Sum of Squares in ANOVA

- In analysis of variance (ANOVA), the total sum of squares helps express the total variation that can be attributed to various factors. For example, you do an experiment to test the effectiveness of three laundry detergents.
- The total sum of squares (SST) = treatment sum of squares (SSTR) + sum of squares of the residual error (SSE)
- The treatment sum of squares is the variation attributed to, or in this case between, the laundry detergents. The sum of squares of the residual error is the variation attributed to the error.
- Converting the sum of squares into mean squares by dividing by the degrees of freedom lets you compare these ratios and determine whether there is a significant difference due to detergent. The larger this ratio is, the more the treatments affect the outcome.

# ANOVA Terminology

If there are  $n$  observations and  $k$  treatments

Total Sum of Squares (SST)  $= \sum_i \sum_j (x_{ij} - \bar{x})^2$

Total Degrees of Freedom  $= n - 1$

Sum of Squares due to Treatment (SSTR)  $= \sum_j n_j (\bar{x}_j - \bar{x})^2$

Treatment Degrees of Freedom  $= k - 1$

Mean Square due to Treatment (MSTR)  $= \frac{SSTR}{k-1}$

Sum of Squares due to Error (SSE)  $= \sum_i \sum_j (x_{ij} - \bar{x}_j)^2$

Error degrees of Freedom  $= n - k$

Mean Square due to Error (MSE)  $= \frac{\sum_j (n_j - 1)s_j^2}{n - k}$

# ANOVA Table

Analysis of Variance					
Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square Error	F	P
Treatment	$k - 1$	SSTR	$MSTR = SSTR/(k-1)$	$MSTR/MSE$	
Error	$n - k$	SSE	$MSE = SSE/(n-k)$		
Total	$n - 1$	SST			



# One Way ANOVA

## Business Case:

An operations manager wants to buy a new machine. He has been using 3 models and would like to go for one which produced minimum % no. of defects. He collected the following data:

% Number of defects produced		
Machine 1	Machine 2	Machine 3
15	10	17
14	14	12
20	9	14
15	10	15
16	11	12

## Decision:

1. Is there a significant performance difference between machines?
2. If Yes, Which is the best performer?

# ANOVA

ANOVA

DEFECTS

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	68.800	2	34.400	7.533	.008
Within Groups	54.800	12	4.567		
Total	123.600	14			

# Two Way ANOVA

## Business Case:

A super market has a chain of 5 stores. The General Manager is seriously concerned about quality of service at each stores. He conducted a survey on customer perceived quality on each day of the week.

Customer Rating on quality					
Week Day	S1	S2	S3	S4	S5
MON	79	81	74	77	66
TUE	78	86	89	97	86
WED	81	87	84	94	82
THU	80	83	81	88	83
FRI	70	74	77	89	68

## Decision:

1. Is there a significant quality difference among the stores?
2. Is day of the week influencing the quality of service?

# Two Way ANOVA

Tests of Between-Subjects Effects

Dependent Variable: RATING

Source	Type I Sum of Squares	df	Mean Square	F	Sig.
Model	166565.360	9	18507.262	1047.680	.000
DAY	166103.600	5	33220.720	1880.596	.000
STORE	461.760	4	115.440	6.535	.003
Error	282.640	16	17.665		
Total	166848.000	25			

a R Squared = .998 (Adjusted R Squared = .997)

# DESIGN OF EXPERIMENTS

# DESIGN OF EXPERIMENTS

- Statistical studies can be classified as observational or experimental.
- In observational studies, the variables of interest are identified and are only observed. The researcher does not attempt to influence or control the variables.
- A survey is a typical example of observational study.
- In experimental studies, one or more independent variables called factors are controlled and response variable(s) are studied.
- Experiments may have factors that are qualitative or quantitative or a mix of both.
- The levels of factor are called the treatments.
- An experiment is usually designed in such a way that sample observations are fully or randomly assigned to each level.

# EXAMPLES

1. A company investigated effect of selling price and type of promotion on sales. Three selling prices (Rs. 59, Rs. 60, and Rs. 64) were studied as well as two promotional campaigns TV ads and Newspaper ads. 12 communities (areas) were selected within two each were assigned to one of the following 6 combinations. What type of study is this

59 & TV	60 & TV	64 & TV	59 & NP	60 & NP	64 & NP
---------	---------	---------	---------	---------	---------

2. An analyst studied the effect of family income (1) Under Rs 20000, (2) Rs 20000 to Rs 40000, (3) Rs. 40000 to Rs. 80000 (4) Rs. 80000 and their life style above with respect to home appliance purchase i.e (TV, Smart Phone; Washing Machine, Two Wheeler; Car, Music System; House, Vacation Package). The analyst select 20 families for each of 16 possible combinations for further study. What type of study is this?

# IMPORTANT TERMS

- Response variable: The dependent variable of interest or being studied.
- Independent variable: The variable influencing the response variable.
- Factor: One or more independent variable, being controlled.
- Treatment: The level or quantity of factor maintained in the experiment. A factor will have multiple treatments.
- Experimental unit: The observation/entity assigned to each level of the factor.

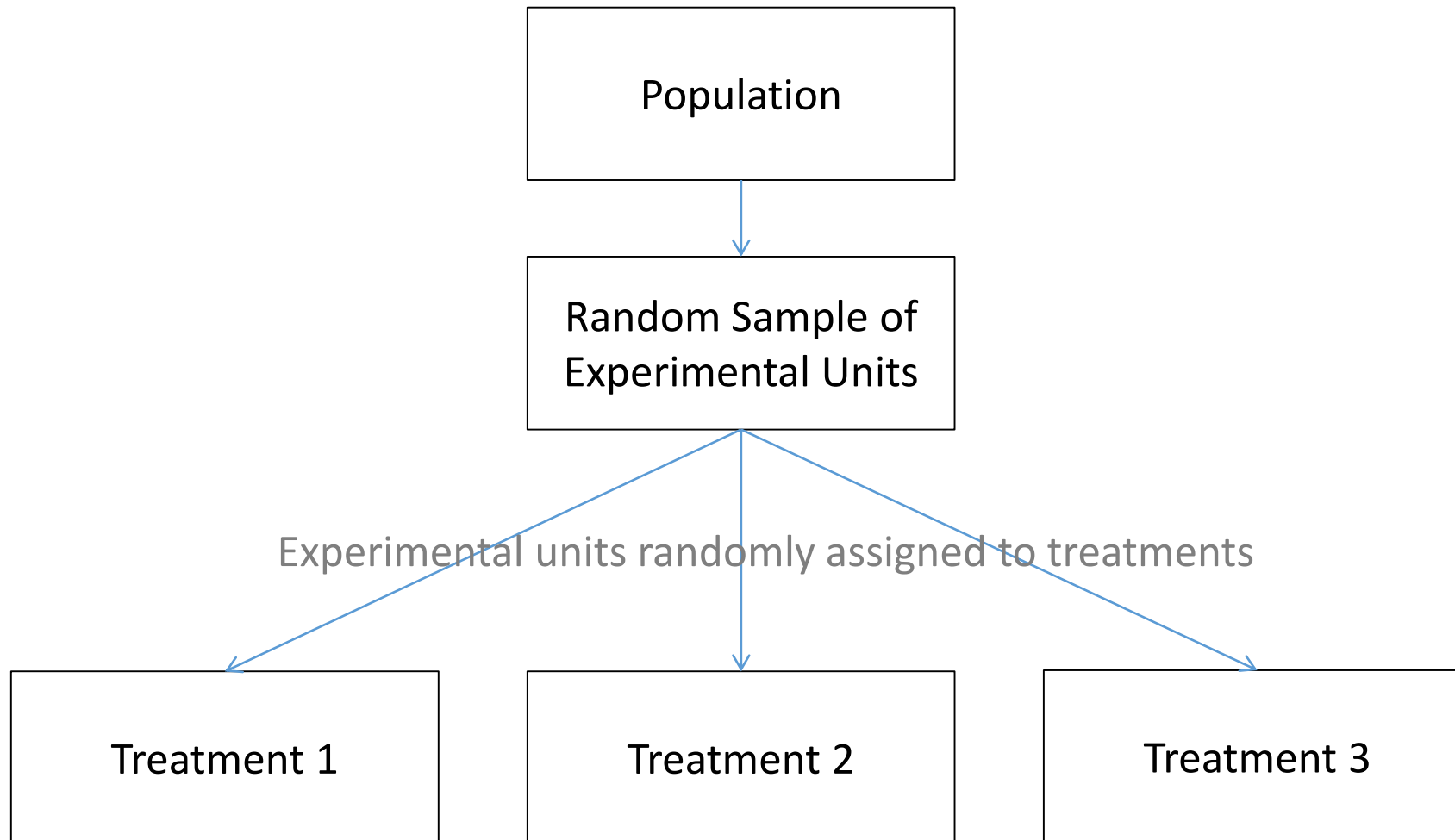


# TYPES OF EXPERIMENTAL DESIGNS

In this course we study three types of experimental designs.

- Completely Randomized Design.
- Randomized Block Design
- Factorial Design

# Completely Randomized Design



# Completely Randomized Design

NCP is considering a new process to assemble Carburetors for Aactiva model of 2-Wheelers. Three methods of assembly say Method-A, Method-B and Method-C are being considered. Managers want to select the Method which allows maximum units to be produced in given time. The designed an experiment in which 15 workers were randomly selected and 5 workers randomly assigned each of the process. The output per day from each worker is given below:

	Method		
Observation	A	B	C
1	58	58	48
2	64	69	57
3	55	71	59
4	66	64	47
5	67	68	49

# Randomized Block Design

- In completely Randomized design problems can arise extraneous factors (ones that are not considered in the experiment) can cause MSE term become large.
- In this case  $F (=MSTR/MSE)$  becomes small signaling no difference when in fact difference exists.
- In such situations Randomized Block Design is helpful.
- The purpose of this design is to control the extraneous sources and therefore limiting MSE term from becoming large.

In order to reduce the stress and fatigue in Air Traffic Controllers, several new designs were proposed. Of these 3 specific alternatives System A, System B and System C were selected. The stress induced by the system is measured in a follow-up interview. It is known people have different ability to handle stress. The problem is to check if the three systems differed significantly in inducing stress which also controlling the effect of ATC's ability to handle stress.

# Randomized Block Design

		Treatments		
		System A	System B	System C
Blocks	Controller 1	15	15	18
	Controller 2	14	14	14
	Controller 3	10	11	15
	Controller 4	13	12	17
	Controller 5	16	13	16
	Controller 6	13	13	13

# Randomized Block Design

		Treatments			Row or Block Totals	Block Means
		System A	System B	System C		
Blocks	Controller 1	15	15	18	48	$\bar{x} = 48/3 = 16$
	Controller 2	14	14	14	42	$\bar{x} = 42/3 = 14$
	Controller 3	10	11	15	36	$\bar{x} = 36/3 = 12$
	Controller 4	13	12	17	42	$\bar{x} = 42/3 = 14$
	Controller 5	16	13	16	45	$\bar{x} = 45/3 = 15$
	Controller 6	13	13	13	39	$\bar{x} = 39/3 = 13$
Column or Treatment Totals		81	78	93	252	$\bar{\bar{X}} = 252/18 = 14$
Treatment Means		$\bar{x} = 81/6 = 13.5$	$\bar{x} = 78/6 = 13$	$\bar{x} = 93/6 = 15.5$		

# Randomized Block Design

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Treatments	SSTR	$k - 1$	$MSTR = \frac{SSTR}{k - 1}$	$\frac{MSTR}{MSE}$
Blocks	SSBL	$b - 1$	$MSBL = \frac{SSBL}{b - 1}$	
Error	SSE	$(k - 1)(b - 1)$	$MSE = \frac{SSE}{(k - 1)(b - 1)}$	
Total	SST	$n_T - 1$		

# FACTORIAL EXPERIMENTS

- In completely random design and randomized block design we were focusing on one factor while controlling the effect of extraneous factors.
- In factorial experiments we want to draw conclusions about more than one factor.
- The name factorial is used because we use all the combinations of all the levels of the factors involved.
- If two factors A and B have levels a and b, we will collect data on a x b factor treatment combinations.



# ANOVA FOR FACTORIAL EXPERIMENTS

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F
Factor A	SSA	$a - 1$	$MSA = \frac{SSA}{a - 1}$	$\frac{MSA}{MSE}$
Factor B	SSB	$b - 1$	$MSB = \frac{SSB}{b - 1}$	$\frac{MSB}{MSE}$
Interaction	SSAB	$(a - 1)(b - 1)$	$MSAB = \frac{SSAB}{(a - 1)(b - 1)}$	$\frac{MSAB}{MSE}$
Error	SSE	$ab(r - 1)$	$MSE = \frac{SSE}{ab(r - 1)}$	
Total	SST	$nT - 1$		

# FACTORIAL EXPERIMENTS ILLUSTRATION

An American university wants to improve the performance of its students in the Graduate Management Admission Test (GMAT). GMAT scores range from 200 to 800 higher being better. The university is considering a preparation program that will improve the GMAT scores. Three programs are suggested.

- A 3-Hour review program (3H)
- A 1-Day training workshop (1D)
- A 10-Week intensive course. (10W)

The students hail from different backgrounds, namely Business, Engineering and Arts. The university wants to assess program, background and their combination that most influences GMAT scores. An experiment was conducted and results are given below.

# FACTORIAL EXPERIMENTS ILLUSTRATION

		FACTOR B: COLLEGE		
FACTOR A: PREPARATION PROGRAM		BUSINESS	ENGINEERING	ARTS
	3-H	500	540	480
		580	460	400
	1-D	460	560	420
		540	620	480
	10-W	560	600	480
		600	580	410

# FACTORIAL EXPERIMENTS ILLUSTRATION

	FACTOR B: COLLEGE				
	BUSINESS	ENGG	ARTS		
FACTOR A: PREP PROGRAM	3-H	500	540	480	2960 $\bar{x}_{1.} = 2960/3 = 493.33$
		580	460	400	
		$\bar{x}_{11} = 1080/2 = 540$	$\bar{x}_{12} = 1000/2 = 500$	$\bar{x}_{13} = 880/2 = 440$	
	1-D	460	560	420	3080 $\bar{x}_{2.} = 3080/6 = 513.33$
		540	620	480	
		$\bar{x}_{21} = 1000/2 = 500$	$\bar{x}_{22} = 1180/2 = 590$	$\bar{x}_{23.} = 900/2 = 450$	
	10-W	560	600	480	3230 $\bar{x}_{3.} = 3230/6 = 538.33$
		600	580	410	
		$\bar{x}_{31} = 1160/2 = 580$	$\bar{x}_{32} = 1180/2 = 590$	$\bar{x}_{33} = 890/2 = 445$	
COL TOTALS		3240	3360	3670	9720 $\bar{\bar{x}} = 9270/18 = 515$
FACTOR B MEANS		$\bar{x}_{.1} = 1160/6 = 540$	$\bar{x}_{.2} = 3360/6 = 560$	$\bar{x}_{.3} = 2670/6 = 445$	