PROBABILITY AND STATISTICS

SEMESTER-II

CSE&IT

R-18

CO1-Discuss the concepts of probability, conditional probability, Baye's theorem and random variables

CO2- Classify the probability distributions and study their properties

CO3- Understand the concepts of correlation and regression to the given data.

CO4- Apply testing of Hypothesis for sample means and sample proportions

CO5- Estimate the truth value of the statistical hypotheses by using small sample tests.

CONTENTS

- Random Variables
- Probability Distribution
- Multiple Random Variables
- Sampling Distribution
- Testing Of Hypothesis
- Large Sample Tests
- Small Sample Tests

TEXT BOOKS

- Higher Engineering Mathematics by Dr.B.S.Grewal,Khanna publishers
- Probability and Statistics for Engineering and Scientists by Sheldon M Ross,Academic press
- Operation Research by S.D.Sarma

REFERENCES

- Mathematics for Engineering by K.B.Datta and M.A.S.Srinivas, Cengage Publications
- Probability and Statistics by T.K.V.Iyengar & B.Krishna Gandhi Et
- Fundamentals of Mathematical Statistics by S C Gupta and V.K.Kapoor
- Probability and Statistics for Engineers and Scientists by Jay I Devore

UNIT-I

Single Random Variables and Probability Distribution

CLO1- Describe the basic concepts of probability

CLO2- Summarize the concept of conditional probability and estimate the probability of event using Baye's theorem.

CLO3- Analyze the concepts of discrete and continuous random variables, probability distributions, expectation and variance..

CLO4- Use the concept of random variables in real-time problem like graph theory, machine learning

.

Basic Concepts

- An **experiment** is the process by which an observation (or measurement) is obtained.
- **Experiment: Record an age**
- **Experiment: Toss a die**
- **Experiment: Record an opinion (yes, no)**
- **Experiment: Toss two coins**

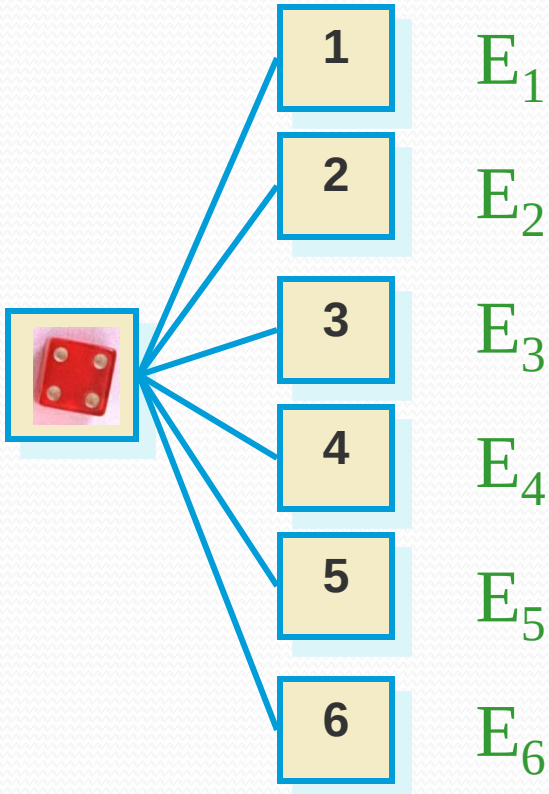
- A **simple event** is the outcome that is observed on a single repetition of the experiment.
  - The basic element to which probability is applied.
  - One and only one simple event can occur when the experiment is performed.

- A **simple event** is denoted by E with a subscript.

- Each simple event will be assigned a probability, measuring “how often” it occurs.
- The set of all simple events of an experiment is called the **sample space, S.**

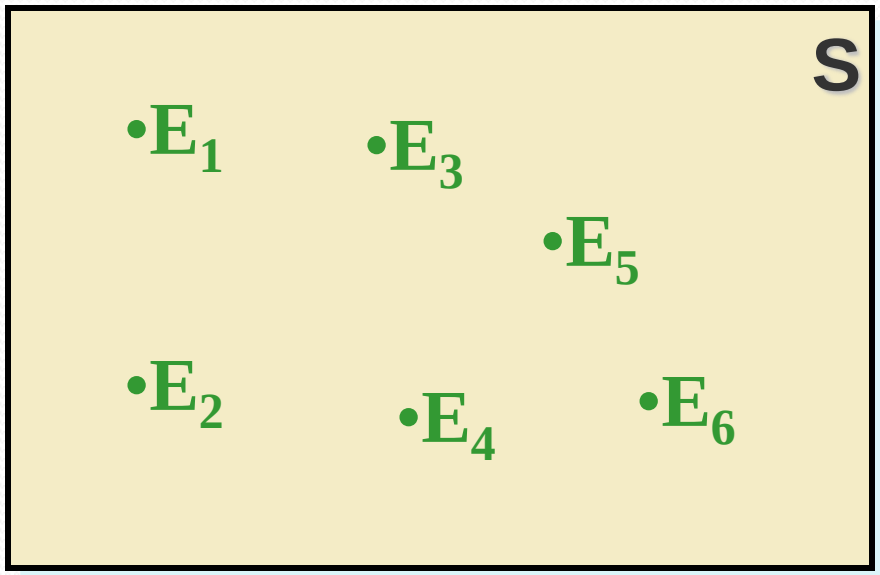
Example

- **The die toss:**
- Simple events:



Sample space:

$S = \{E_1, E_2, E_3, E_4, E_5, E_6\}$

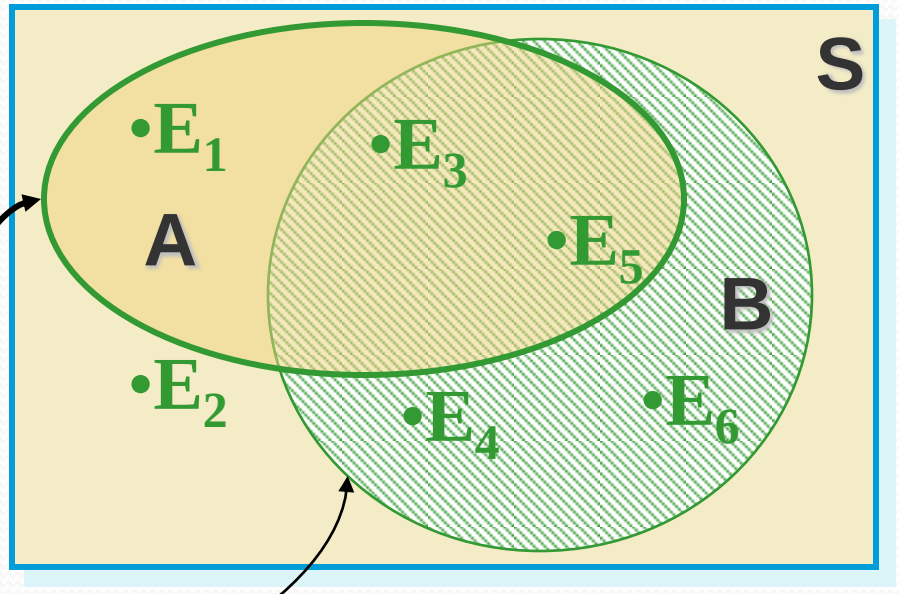


- An **event** is a collection of one or more **simple events.**

- **The die toss:**
  - –A: an odd number
  - –B: a number > 2

$A = \{E_1, E_3, E_5\}$

$B = \{E_3, E_4, E_5, E_6\}$



- Two events are **mutually exclusive** if, when one event occurs, the other cannot, and vice versa.

- **Experiment: Toss a die**
  - A: observe an odd number
  - B: observe a number greater than 2
  - C: observe a 6
  - D: observe a 3

Not Mutually Exclusive

Mutually Exclusive

B and C?
B and D?

- The probability of an event A measures "how often" we think A will occur. We write **P(A).**
- Suppose that an experiment is performed *n* times. The relative frequency for an event A is

$$\frac{\text{Number of times A occurs}}{n} = \frac{f}{n}$$

- If we let *n* get infinitely large,

$$P(A) = \lim_{n \to \infty} \frac{f}{n}$$

- P(A) must be between 0 and 1.
  - If event A can never occur, P(A) = 0. If event A always occurs when the experiment is performed, P(A) =1.
- The sum of the probabilities for all simple events in S equals 1.

•The **probability of an event A** is found by adding the probabilities of all the simple events contained in A.

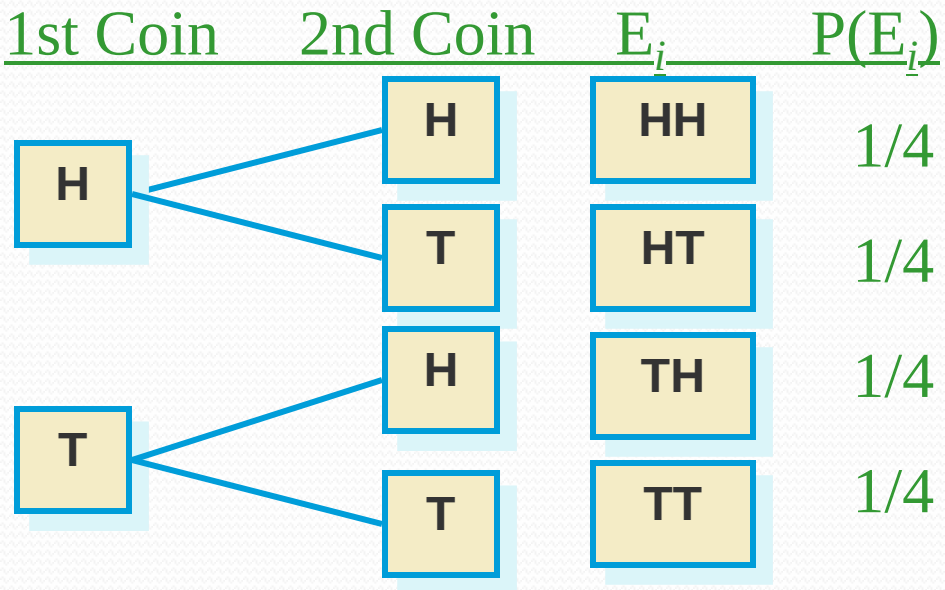
Finding Probabilities

- Probabilities can be found using
  - Estimates from empirical studies
  - Common sense estimates based on equally likely events.

- **Examples:**
  - Toss a fair coin. P(Head) = 1/2
  - 10% of the U.S. population has red hair. Select a person at random. P(Red hair) = .10

Example

- Toss a fair coin twice. What is the probability of observing at least one head?



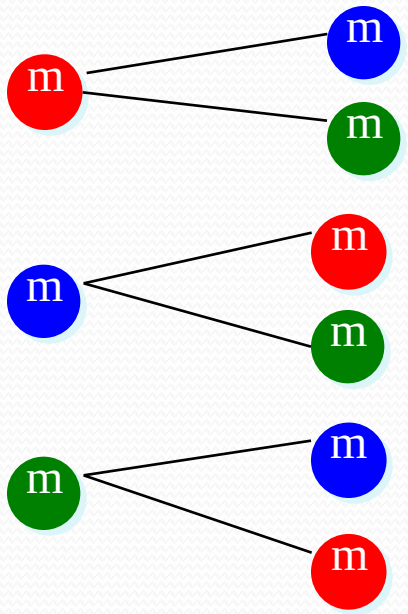
P(at least 1 head)

$= P(E_1) + P(E_2) + P(E_3)$

$= 1/4 + 1/4 + 1/4 = 3/4$

Example

- A bowl contains three M&Ms®, one red, one blue and one green. A child selects two M&Ms at random. What is the probability that at least one is red?



1st M&M	2nd M&M	E_i	$P(E_i)$
m	m	RB	1/6
m	m	RG	1/6
m	m	BR	1/6
m	m	BG	1/6
m	m	GB	1/6
m	m	GR	1/6

P(at least 1 red)

= P(RB) + P(BR)+ P(RG) + P(GR)

= 4/6 = 2/3

Counting Rules

- If the simple events in an experiment are **equally likely**, you can calculate

$$P(A) = \frac{n_A}{N} = \frac{\text{number of simple events in A}}{\text{total number of simple events}}$$

- You can use **counting rules** to find n_A and N.

The *mn* Rule

- If an experiment is performed in two stages, with m ways to accomplish the first stage and n ways to accomplish the second stage, then there are mn ways to accomplish the experiment.
- This rule is easily extended to k stages, with the number of ways equal to

$$n_1 n_2 n_3 \ldots n_k$$

Example: Toss two coins. The total number of simple events is:

$$2 \times 2 = 4$$

Examples

Example: Toss three coins. The total number of simple events is

$$2 \times 2 \times 2 = 8$$

Example: Toss two dice. The total number of simple events is:

$$6 \times 6 = 36$$

Example: Two M&Ms are drawn from a dish containing two red and two blue candies. The total number of simple eve

$$4 \times 3 = 12$$

Permutations

- The number of ways you can arrange ***n*** distinct objects, taking them ***r*** at a time is

$$P_r^n = \frac{n!}{(n-r)!}$$

where $n! = n(n-1)(n-2)\ldots(2)(1)$ and $0! \equiv 1$.

Example: How many 3-digit lock combinations can we make from the numbers 1, 2, 3, and 4?

The order of the choice is important!

$$P_3^4 = \frac{4!}{1!} = 4(3)(2) = 24$$

Combinations

- The number of distinct combinations of ***n*** distinct objects that can be formed, taking them ***r*** at a time is

$$C_r^n = \frac{n!}{r!(n-r)!}$$

Example: Three members of a 5-person committee must be chosen to form a subcommittee. How many different subcommittees could be formed?

The order of the choice is not important!

$$C_3^5 = \frac{5!}{3!(5-3)!} = \frac{5(4)(3)(2)1}{3(2)(1)(2)1} = \frac{5(4)}{(2)1} = 10$$

Example

- A box contains six M&Ms®, four red
- and two green. A child selects two M&Ms at random. What is the probability that exactly one is red?

The order of the choice is not important!

$C_2^6 = \frac{6!}{2!4!} = \frac{6(5)}{2(1)} = 15$ ways to choose 2 M & Ms.

$C_1^2 = \frac{2!}{1!1!} = 2$ ways to choose 1 green M & M.

$C_1^4 = \frac{4!}{1!3!} = 4$ ways to choose 1 red M & M.

$4 \times 2 = 8$ ways to choose 1 red and 1 green M&M.

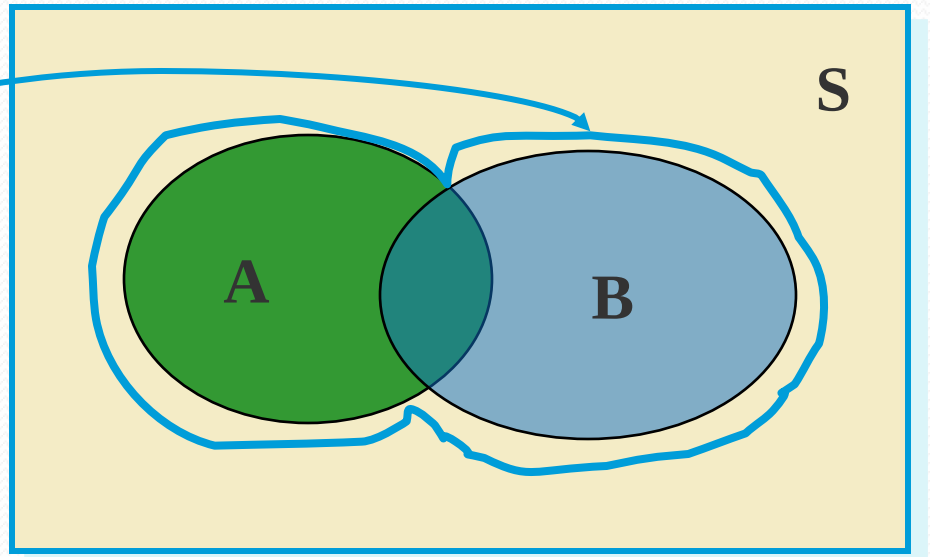
P(exactly one red) = 8/15

Event Relations

- The **union** of two events, A and B, is the event that either A **or** B **or both** occur when the experiment is performed. We write

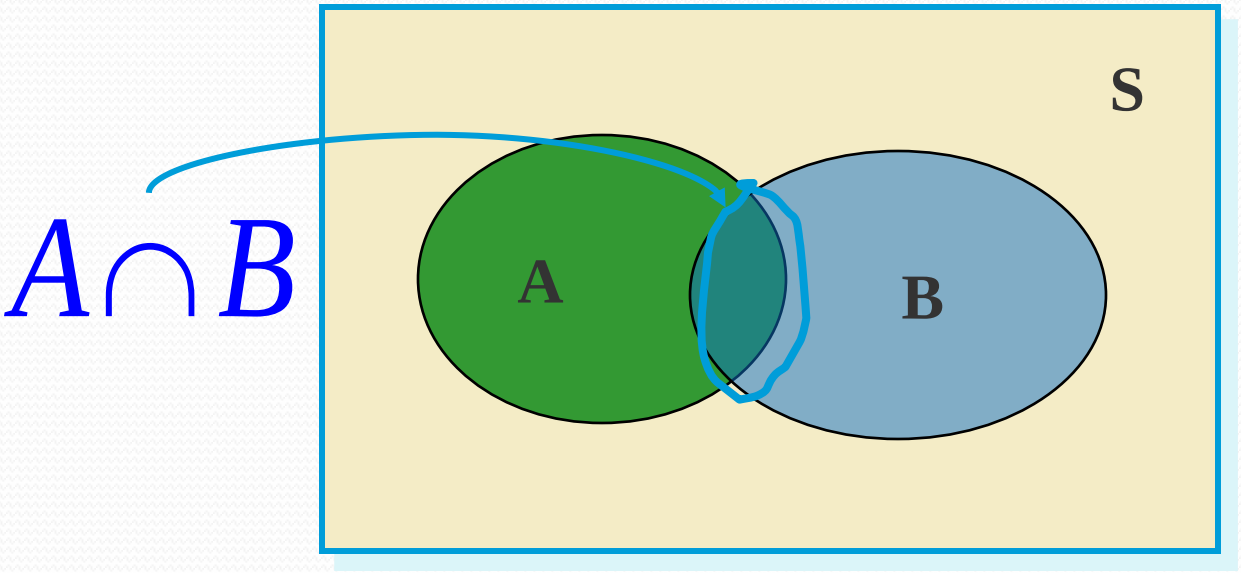
$$A \cup B$$

$A \cup B$



Event Relations

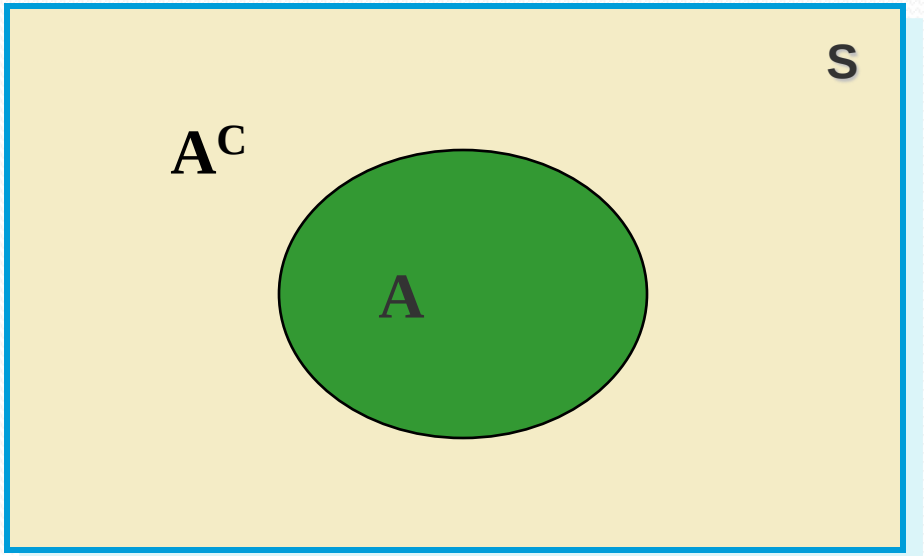
- The **intersection** of two events, **A** and **B**, is the event that both A **and** B occur when the experiment is performed. We write **A ∩ B.**



- If two events A and B are **mutually exclusive,** then P(**A ∩ B) = 0.**

Event Relations

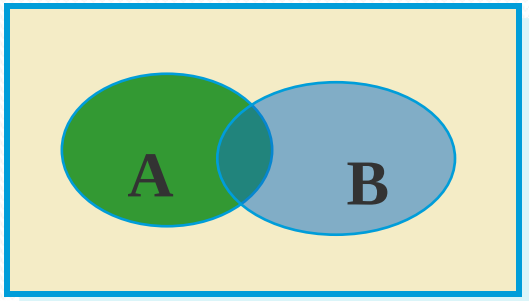
- The **complement** of an event **A** consists of all outcomes of the experiment that do not result in event A. We write A^C.



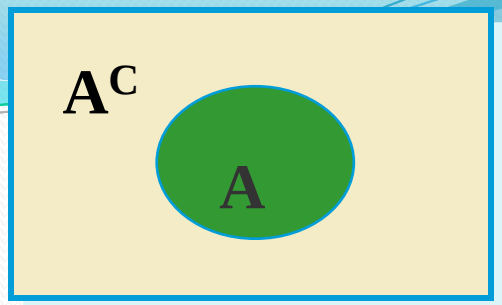
Calculating Probabilities for Unions and Complements

- There are special rules that will allow you to calculate probabilities for composite events.
- **The Additive Rule for Unions:**
- For any two events, **A** and **B**, the probability of their union, $\mathbf{P(A \cup B)}$, is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



Calculating Probabilities for Complements



- We know that for any event **A**:
  - $P(A \cap A^C) = 0$
- Since either **A** or A^C must occur,

  $P(A \cup A^C) = 1$
- so that $P(A \cup A^C) = P(A) + P(A^C) = 1$

$$P(A^C) = 1 - P(A)$$

Calculating Probabilities for Intersections

- In the previous example, we found $P(A \cap B)$ directly from the table. Sometimes this is impractical or impossible. The rule for calculating $P(A \cap B)$ depends on the idea of **independent and dependent events.**

Two events, **A** and **B**, are said to be **independent** if and only if the probability that event **A** occurs does not change, depending on whether or not event **B** has occurred.

Conditional Probabilities

- The probability that A occurs, given that event B has occurred is called the **conditional probability** of A given B and is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ if } P(B) \neq 0$$

"given"

Defining Independence

- We can redefine independence in terms of conditional probabilities:

Two events A and B are **independent** if and only if

$$P(A|B) = P(A) \quad \text{or} \quad P(B|A) = P(B)$$

Otherwise, they are **dependent**.

- Once you've decided whether or not two events are independent, you can use the following rule to calculate their intersection.

The Multiplicative Rule for Intersections

- For any two events, **A** and **B**, the probability that both **A** and **B** occur is

$$P(A \cap B) = P(A)\,P(B \text{ given that } A \text{ occurred}) = P(A)P(B|A)$$

- If the events **A** and **B** are independent, then the probability that both **A** and **B** occur is

$$P(A \cap B) = P(A)\,P(B)$$

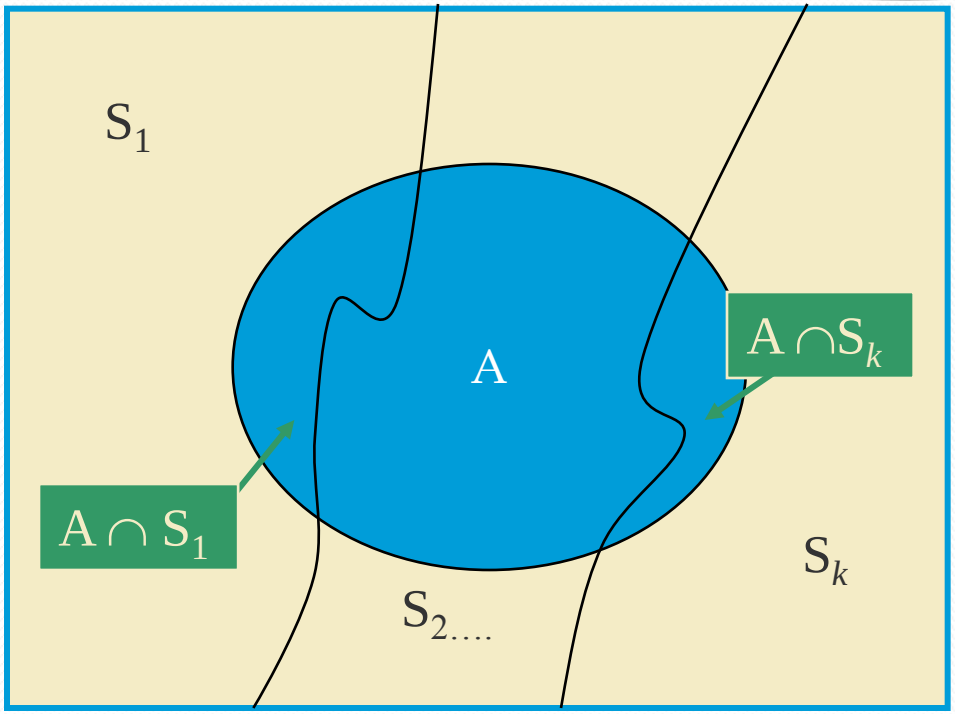
The Law of Total Probability

- Let S_1 , S_2 , S_3 ,…, S_k be mutually exclusive and exhaustive events (that is, one and only one must happen). Then the probability of another event A can be written as

$$P(A) = P(A \cap S_1) + P(A \cap S_2) + \ldots + P(A \cap S_k)$$

$$= P(S_1)P(A|S_1) + P(S_2)P(A|S_2) + \ldots + P(S_k)P(A|S_k)$$

The Law of Total Probability



$$P(A) = P(A \cap S_1) + P(A \cap S_2) + \ldots + P(A \cap S_k)$$

$$= P(S_1)P(A|S_1) + P(S_2)P(A|S_2) + \ldots + P(S_k)P(A|S_k)$$

Bayes' Rule

- Let S_1 , S_2 , S_3 ,…, S_k be mutually exclusive and exhaustive events with prior probabilities $P(S_1)$, $P(S_2)$,…,$P(S_k)$. If an event A occurs, the posterior probability of S_i, given that A occurred is

$$P(S_i \mid A) = \frac{P(S_i)P(A \mid S_i)}{\sum P(S_i)P(A \mid S_i)} \text{ for } i = 1, 2, \ldots k$$

Random Variables

- A quantitative variable *x* is a **random variable** if the value that it assumes, corresponding to the outcome of an experiment is a chance or random event.
- Random variables can be **discrete** or **continuous.**

- **Examples:**
  - ✓ *x* = SAT score for a randomly selected student
  - ✓ *x* = number of people in a room at a randomly selected time of day
  - ✓ *x* = number on the upper face of a randomly tossed die

Probability Distributions for Discrete Random Variables

- The **probability distribution for a discrete random variable *x*** resembles the relative frequency distributions we constructed in Chapter 1. It is a graph, table or formula that gives the possible values of *x* and the probability *p(x)* associated with each value.

We must have

$$0 \le p(x) \le 1 \text{ and } \sum p(x) = 1$$

Probability Distributions

- Probability distributions can be used to describe the population, just as we described samples in Chapter 1.
  - **Shape:** Symmetric, skewed, mound-shaped…
  - **Outliers:** unusual or unlikely measurements
  - **Center and spread:** mean and standard deviation. A population mean is called μ and a population standard deviation is called σ.

The Mean and Standard Deviation

- Let x be a discrete random variable with probability distribution $p(x)$. Then the mean, variance and standard deviation of x are given as

$$\text{Mean}: \mu = \sum xp(x)$$

$$\text{Variance}: \sigma^2 = \sum (x-\mu)^2 p(x)$$

$$\text{Standard deviation}: \sigma = \sqrt{\sigma^2}$$

Example

- Toss a fair coin 3 times and record x the number of heads.

x	$p(x)$	$xp(x)$	$(x-\mu)^2p(x)$
0	1/8	0	$(-1.5)^2(1/8)$
1	3/8	3/8	$(-0.5)^2(3/8)$
2	3/8	6/8	$(0.5)^2(3/8)$
3	1/8	3/8	$(1.5)^2(1/8)$

$$\mu = \sum xp(x) = \frac{12}{8} = 1.5$$

$$\sigma^2 = \sum (x-\mu)^2 p(x)$$

$$\sigma^2 = .28125 + .09375 + .09375 + .28125 = .75$$

$$\sigma = \sqrt{.75} = .688$$

UNIT-II

Probability Distributions

CLO5- Determine the binomial distribution to find mean and variance.

CLO6- Understand the phenomena of real-time problem like sick versus healthy by using Binomial distribution.

CLO7- Determine the Poisson distribution to find mean and variance.

CLO8- Understand the phenomena of real-time problem of predicting soccer scores by using Poisson distribution.

CLO9- Illustrate the inferential methods relating to the means of normal distributions.

CLO10-Describe the mapping of normal distribution in real-world problem to analyze the stock market

Introduction

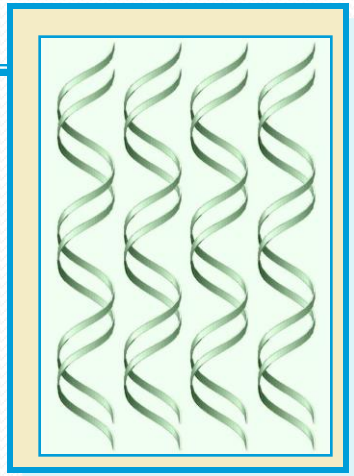
- Discrete random variables take on only a finite or countably number of values.
- Three discrete probability distributions serve as models for a large number of practical applications:

✓The **binomial** random variable

✓The **Poisson** random variable

The Binomial Random Variable

- Many situations in real life resemble the coin toss, but the coin is not necessarily fair, so that P(H) ≠ 1/2.



- Example: A geneticist samples 10 people and counts the number who have a gene linked to Alzheimer's disease.

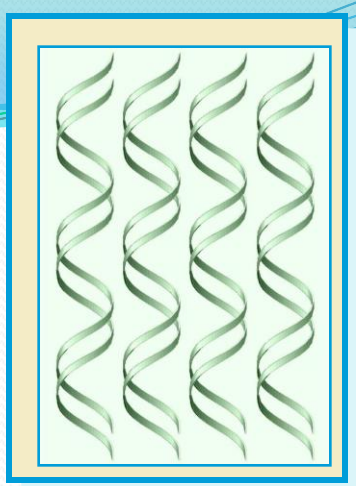
- Coin: Person
- Head: Has gene
- Tail: Doesn't have gene
- Number of tosses: $n = 10$
- P(H): P(has gene) = proportion in the population who have the gene.

The Binomial Experiment

1. The experiment consists of ***n* identical trials.**
2. Each trial results in **one of two outcomes**, success (S) or failure (F).
3. The probability of success on a single trial is p and **remains constant** from trial to trial. The probability of failure is $q = 1 - p$.
4. The trials are **independent**.
5. We are interested in ***x*, the number of successes in *n* trials.**

Binomial or Not?

- Very few real life applications satisfy these requirements exactly.



- Select two people from the U.S. population, and suppose that 15% of the population has the Alzheimer's gene.
  - For the first person, p = P(gene) = .15
  - For the second person, $p \approx$ P(gene) = .15, even though one person has been removed from the population.

The Binomial Probability Distribution

- For a binomial experiment with n trials and probability p of success on a given trial, the probability of k successes in n trials is

$$P(x=k) = C_k^n p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k} \text{ for } k = 0,1,2,\ldots n.$$

Recall $C_k^n = \dfrac{n!}{k!(n-k)!}$

with $n! = n(n-1)(n-2)\ldots(2)1$ and $0! \equiv 1$.

The Mean and Standard Deviation

- For a binomial experiment with n trials and probability p of success on a given trial, the measures of center and spread are:

$$\text{Mean: } \mu = np$$

$$\text{Variance: } \sigma^2 = npq$$

$$\text{Standard deviation: } \sigma = \sqrt{npq}$$

Cumulative Probability Tables

You can use the cumulative probability tables to find probabilities for selected binomial distributions.

- ✓Find the table for the correct value of n.
- ✓Find the column for the correct value of p.
- ✓The row marked "k" gives the cumulative probability, $P(x \leq k) = P(x = 0) + \ldots + P(x = k)$

The Poisson Random Variable

- The Poisson random variable *x* is a model for data that represent the number of occurrences of a specified event in a given unit of time or space.

- Examples:
  - The number of calls received by a switchboard during a given period of time.
  - The number of machine breakdowns in a day
  - The number of traffic accidents at a given intersection during a given time period.

The Poisson Probability Distribution

- ***x*** is the number of events that occur in a period of time or space during which an average of **μ** such events can be expected to occur. The probability of *k* occurrences of this event is

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

For values of $k = 0, 1, 2, \ldots$ The mean and standard deviation of the Poisson random variable are

Mean: μ

Standard deviation:

$$\sigma = \sqrt{\mu}$$

Cumulative Probability Tables

You can use the cumulative probability tables to find probabilities for selected Poisson distributions.

✓Find the column for the correct value of μ.

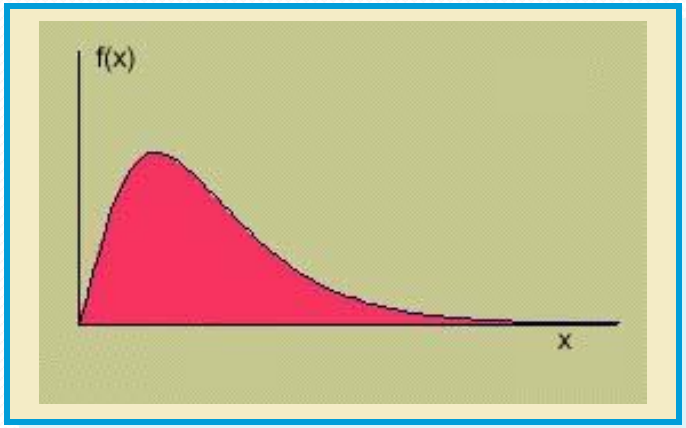
✓The row marked "k" gives the cumulative probability, $P(x \leq k) = P(x = 0) + \ldots + P(x = k)$

Continuous Random Variables

- Continuous random variables can assume the infinitely many values corresponding to points on a line interval.
- **Examples:**
  - Heights, weights
  - length of life of a particular product
  - experimental laboratory error

Continuous Random Variables

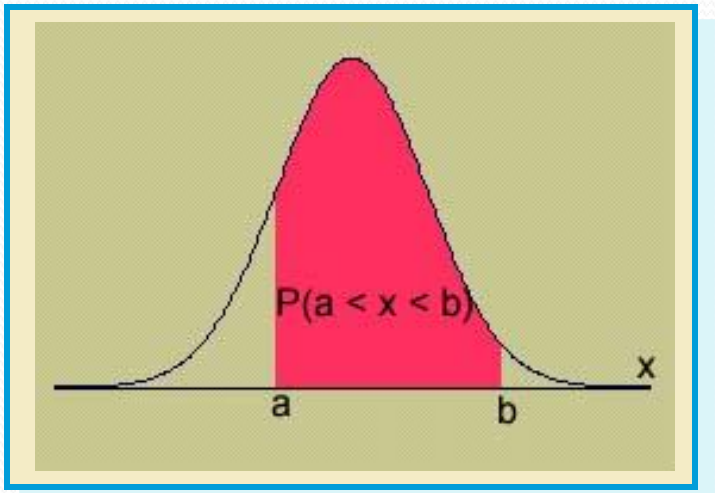
- A **smooth curve** describes the probability distribution of a continuous random variable.



•The depth or density of the probability, which varies with *x*, may be described by a mathematical formula *f* (*x*), called the probability distribution or probability density function for the random variable *x*.

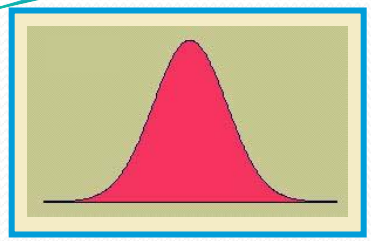
Properties of Continuous Probability Distributions

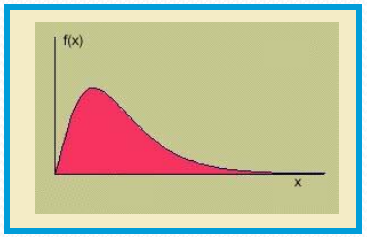
- The area under the curve is equal to **1.**
- P(a ≤ *x* ≤ b) = **area under the curve** between a and b.



- There is no probability attached to any single value of *x*. That is, **P(*x* = a) = 0.**

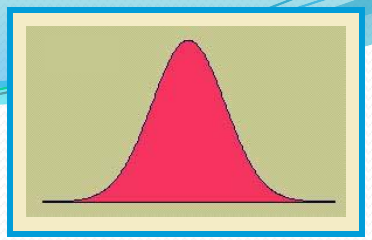
Continuous Probability Distributions





- There are many different types of continuous random variables
- We try to pick a model that
  - Fits the data well
  - Allows us to make the best possible inferences using the data.
- One important continuous random variable is the **normal random variable**.

The Normal Distribution



- The formula that generates the normal probability distribution is:

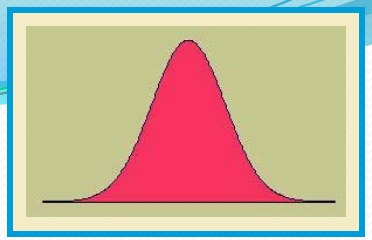
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for} -\infty < x < \infty$$

$e = 2.7183 \qquad \pi = 3.1416$

μ and σ are the population mean and standard deviation.

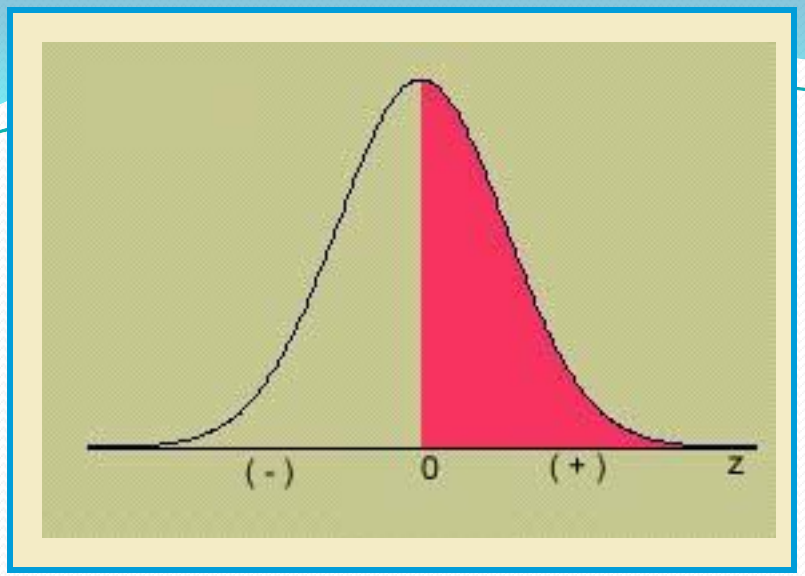
- The shape and location of the normal curve changes as the mean and standard deviation change.

The Standard Normal Distribution



- To find P(a < x < b), we need to find the area under the appropriate normal curve.
- To simplify the tabulation of these areas, we **standardize** each value of x by expressing it as a z-score, the number of standard deviations σ it lies from the mean μ.

$$z = \frac{x - \mu}{\sigma}$$



The Standard Normal (z) Distribution

- Mean = 0; Standard deviation = 1
- When $x = \mu$, $z = 0$
- Symmetric about $z = 0$
- Values of z to the left of center are negative
- Values of z to the right of center are positive
- Total area under the curve is 1.

Finding Probabilities for the General Normal Random Variable

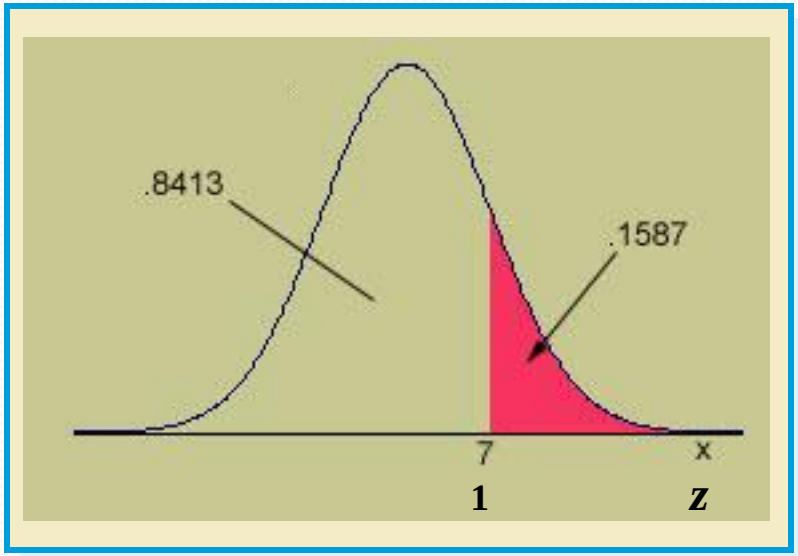
✓To find an area for a normal random variable *x* with mean μ and standard deviation σ, *standardize or rescale* the interval in terms of *z.*

✓Find the appropriate area using Table 3.

Example: *x* has a normal distribution with $\mu = 5$ and $\sigma = 2$. Find $P(x > 7)$.

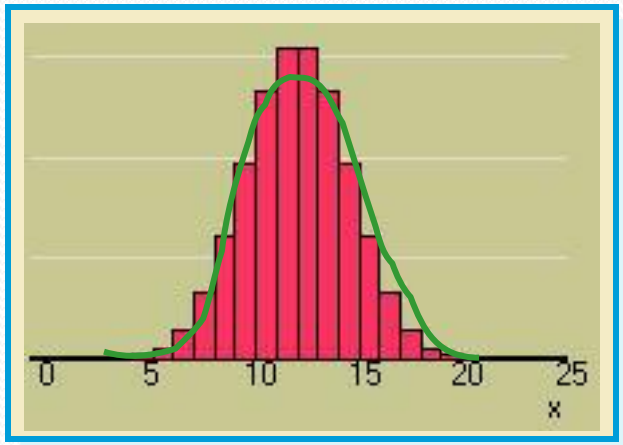
$$P(x > 7) = P(z > \frac{7-5}{2})$$

$$= P(z > 1) = 1 - .8413 = .1587$$



The Normal Approximation to the Binomial

- We can calculate binomial probabilities using
  - The binomial formula
  - The cumulative binomial tables
  - Java applets
- When n is large, and p is not too close to zero or one, areas under the normal curve with mean np and variance npq can be used to approximate binomial probabilities.



UNIT-III

CORRELATION AND REGRESSION

CLO 11	Demonstrate the concept of correlation for a Bivariate data .
CLO 12	Calculate the Karl Pearson's correlation coefficient for the given data
CLO 13	Calculate the Spearman's rank correlation coefficient for the given data.
CLO 14	Estimate the linear regression for the given data
CLO 15	Understand the phenomena of real-time problem like stock price and interest rates by using the concepts of correlation and regression.

Independence and Covariance

- Two random variables **X** and **Y** are said to be independent if
  - Discrete

$$p_{ij} = p_{i+} p_{+j} \quad \text{for all values } i \text{ of } X \text{ and } j \text{ of } Y$$

  - Continuous

$$f(x, y) = f_X(x) f_Y(y) \quad \textit{for all } x \textit{ and } y$$

  - How is this independency different from the independence among events?

Independence and Covariance

- Covariance

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E((X-E(X))(Y-E(Y))) \\ &= E(XY)-E(X)E(Y)\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E((X-E(X))(Y-E(Y))) \\ &= E(XY-XE(Y)-E(X)Y+E(X)E(Y)) \\ &= E(XY)-E(X)E(Y)-E(X)E(Y)+E(X)E(Y) \\ &= E(XY)-E(X)E(Y)\end{aligned}$$

  - May take any positive or negative numbers.
  - Independent random variables have a covariance of zero
  - What if the covariance is zero?

Independence and Covariance

- Example 19 (Air conditioner maintenance)

$$E(X) = 2.59, \qquad E(Y) = 1.79$$

$$E(XY) = \sum_{i=1}^{4}\sum_{j=1}^{3} ijp_{ij}$$

$$= (1\times 1\times 0.12) + (1\times 2\times 0.08)$$

$$+ \cdots + (4\times 3\times 0.07) = 4.86$$

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$$

$$= 4.86 - (2.59\times 1.79) = 0.224$$

Independence and Covariance

- Correlation:

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

  - Values between -1 and 1, and independent random variables have a correlation of zero

Independence and Covariance

- Example 19: (Air conditioner maintenance)

$$\mathrm{Var}(X) = 1.162, \quad \mathrm{Var}(Y) = 0.384$$

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

$$= \frac{0.224}{\sqrt{1.162 \times 0.384}} = 0.34$$

- What if random variable X and Y have linear relationship, that is,

$$Y = aX + b \quad a \neq 0$$

where

$$Cov(X,Y) = E[XY] - E[X]E[Y]$$

$$= E[X(aX+b)] - E[X]E[aX+b]$$

$$= aE[X^2] + bE[X] - aE^2[X] - bE[X]$$

$$= a(E[X^2] - E^2[X]) = aVar(X)$$

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{aVar(X)}{\sqrt{Var(X)a^2Var(X)}}$$

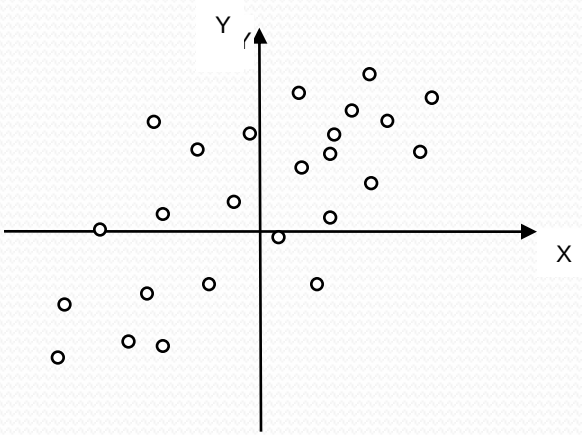
That is, Cov(X,Y)=1 if a>0; -1 if a<0.

The relationship between *x* and *y*

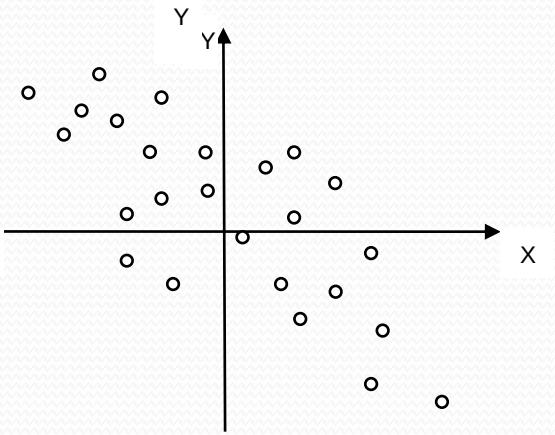
- Correlation: is there a relationship between 2 variables?
- Regression: how well a certain independent variable predict dependent variable?

- CORRELATION $\neq$ CAUSATION
  - In order to infer causality: manipulate independent variable and observe effect on dependent variable

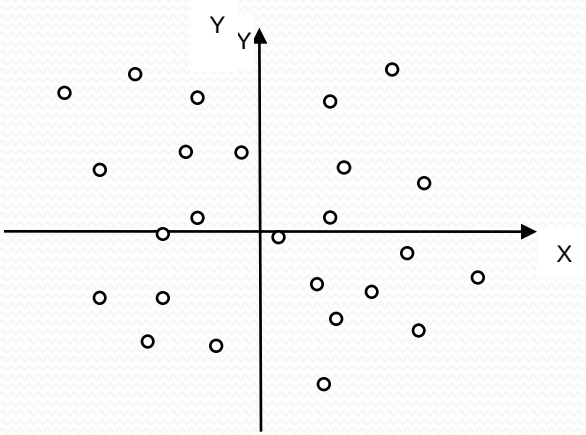
Scattergrams



Positive correlation



Negative correlation



No correlation

Variance vs Covariance

- *First, a note on your sample:*
  - *If you're wishing to assume that your sample is representative of the general population (RANDOM EFFECTS MODEL), use the degrees of freedom (n – 1) in your calculations of variance or covariance.*
  - *But if you're simply wanting to assess your current sample (FIXED EFFECTS MODEL), substitute n for the degrees of freedom.*

Variance vs Covariance

- Do two variables change together?

Variance:

• Gives information on variability of a single variable.

$$S_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Covariance:

• Gives information on the degree to which two variables vary together.

• Note how similar the covariance is to variance: the equation simply multiplies x's error scores by y's error scores as opposed to squaring x's error scores.

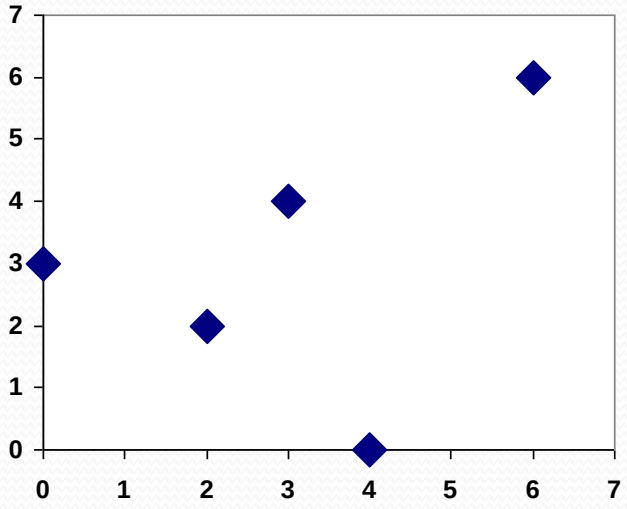
$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Covariance

$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- When X↑ and Y↑ : cov (x,y) = pos.
- When X↓ and Y↑ : cov (x,y) = neg.
- When no constant relationship: cov (x,y) = 0

Example Covariance



x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
0	3	-3	0	0
2	2	-1	-1	1
3	4	0	1	0
4	0	1	-3	-3
6	6	3	3	9
$\bar{x} = 3$	$\bar{y} = 3$			$\sum = 7$

$$\mathrm{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}))}{n-1} = \frac{7}{4} = 1.75$$

What does this number tell us?

Problem with Covariance:

- The value obtained by covariance is dependent on the size of the data's standard deviations: if large, the value will be greater than if small… *even if the relationship between x and y is exactly the same in the large versus small standard deviation datasets.*

Example of how covariance value relies on variance

	High variance data				Low variance data		
Subject	**x**	**y**	**x error * y error**		**x**	**y**	**X error * y error**
1	101	100	2500		54	53	9
2	81	80	900		53	52	4
3	61	60	100		52	51	1
4	51	50	0		51	50	0
5	41	40	100		50	49	1
6	21	20	900		49	48	4
7	1	0	2500		48	47	9
Mean	51	50			51	50	
Sum of x error * y error :			7000		**Sum of x error * y error :**		28
Covariance:			**1166.67**		**Covariance:**		**4.67**

Solution: Pearson's r

- **Covariance does not really tell us anything**
  - ***Solution: standardise this measure***
- **Pearson's R: standardises the covariance value.**
- **Divides the covariance by the multiplied standard deviations of X and Y:**

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{s_x s_y}$$

Pearson's R continued

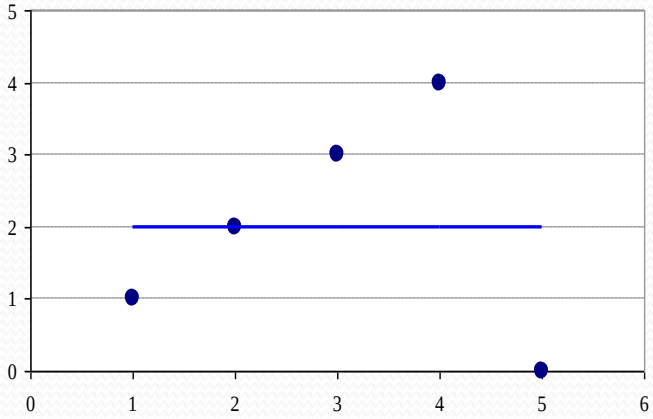
$$\operatorname{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \rightarrow r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

$$\downarrow$$

$$r_{xy} = \frac{\sum_{i=1}^{n} Z_{x_i} * Z_{y_i}}{n-1}$$

Limitations of r

- When r = 1 or r = -1:
  - We can predict y from x with certainty
  - all data points are on a straight line: y = ax + b
- r is actually $\hat{r}$
  - r = true r of whole population
  - $\hat{r}$ = estimate of r based on data
- r is very sensitive to extreme values:



Regression

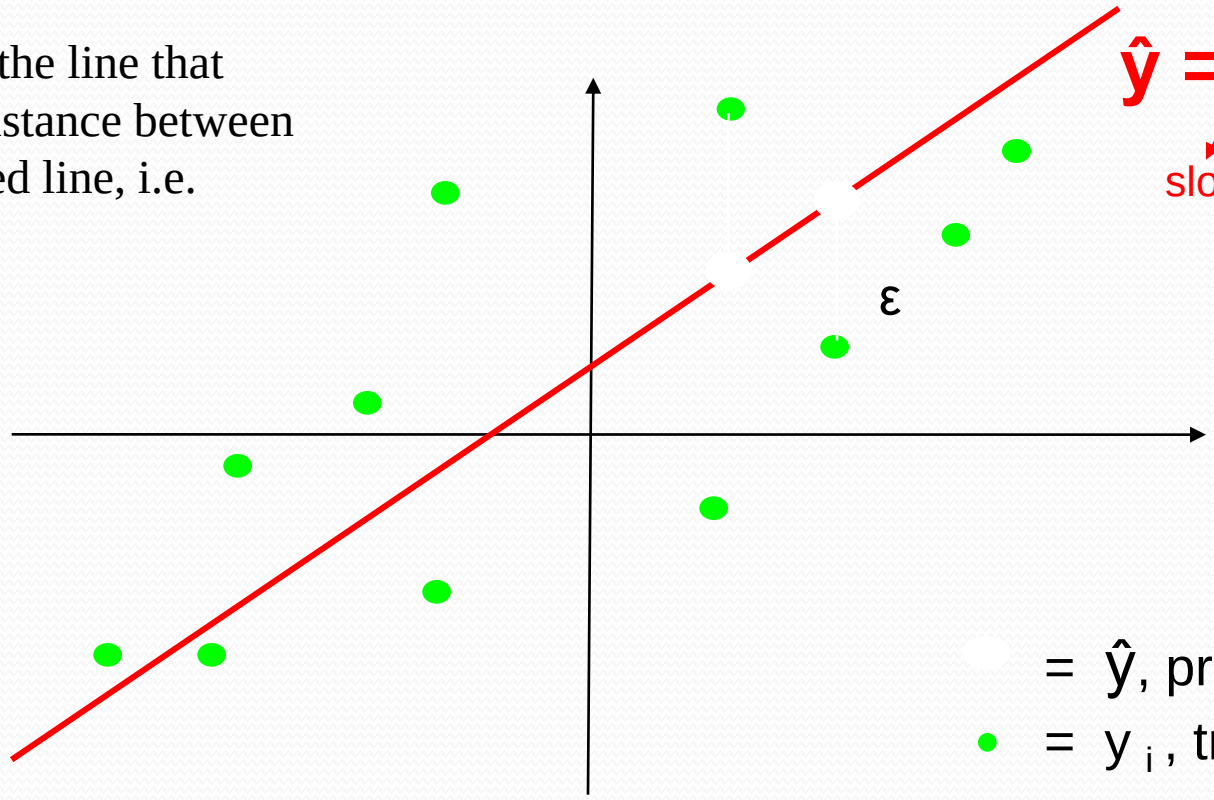
- Correlation tells you if there is an association between x and y but it doesn’t describe the relationship or allow you to predict one variable from the other.

- To do this we need REGRESSION!

Best-fit Line

- Aim of linear regression is to fit a straight line, $\hat{y} = ax + b$, to data that gives best prediction of y for any value of x

- This will be the line that minimises distance between data and fitted line, i.e. the residuals



Least Squares Regression

- To find the best line we must minimise the sum of the squares of the residuals (the vertical distances from the data points to our line)

Model line: $\hat{y} = ax + b$ a = slope, b = intercept

Residual (ε) = $y - \hat{y}$

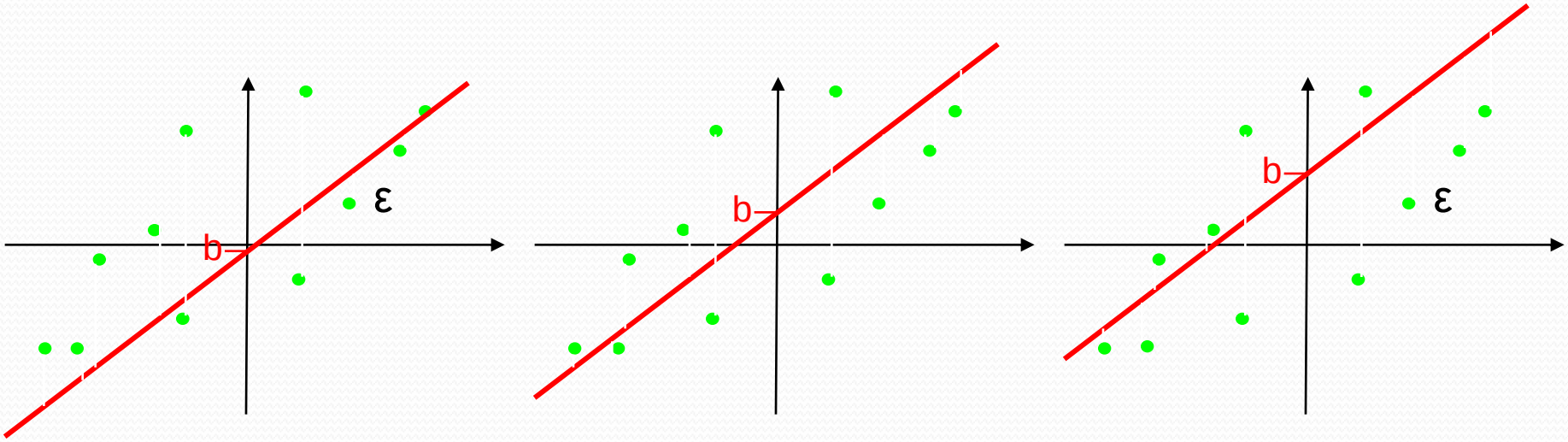
Sum of squares of residuals = $\Sigma (y - \hat{y})^2$

- we must find values of a and b that minimise

$$\Sigma (y - \hat{y})^2$$

Finding b

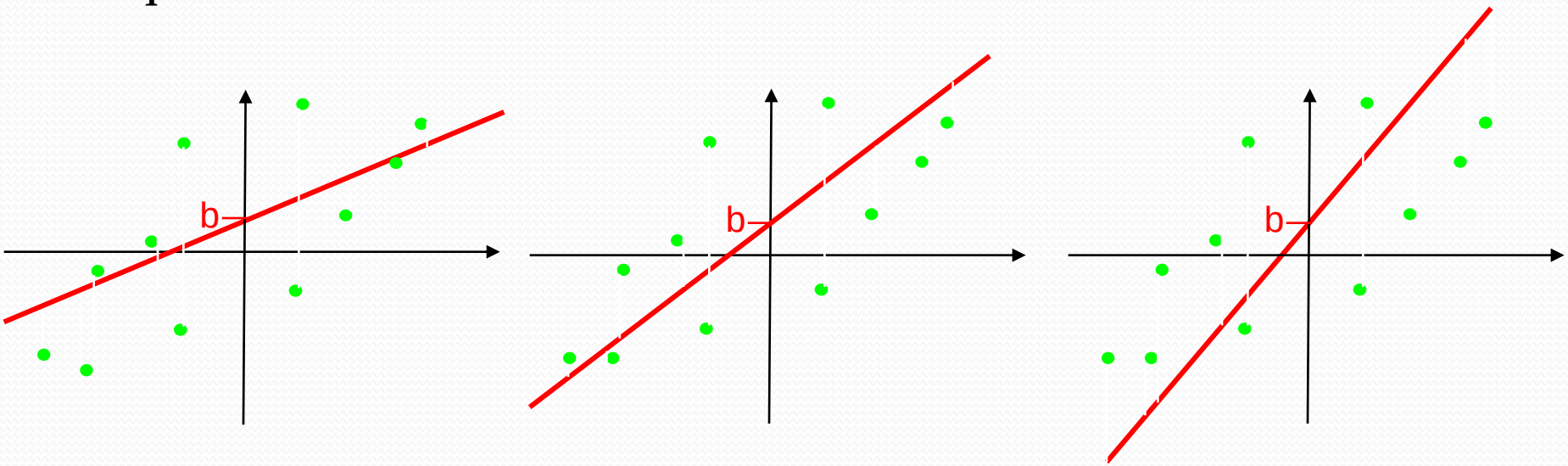
- First we find the value of b that gives the min sum of squares



- Trying different values of b is equivalent to shifting the line up and down the scatter plot

Finding a

- Now we find the value of a that gives the min sum of squares



- Trying out different values of a is equivalent to changing the slope of the line, while b stays constant

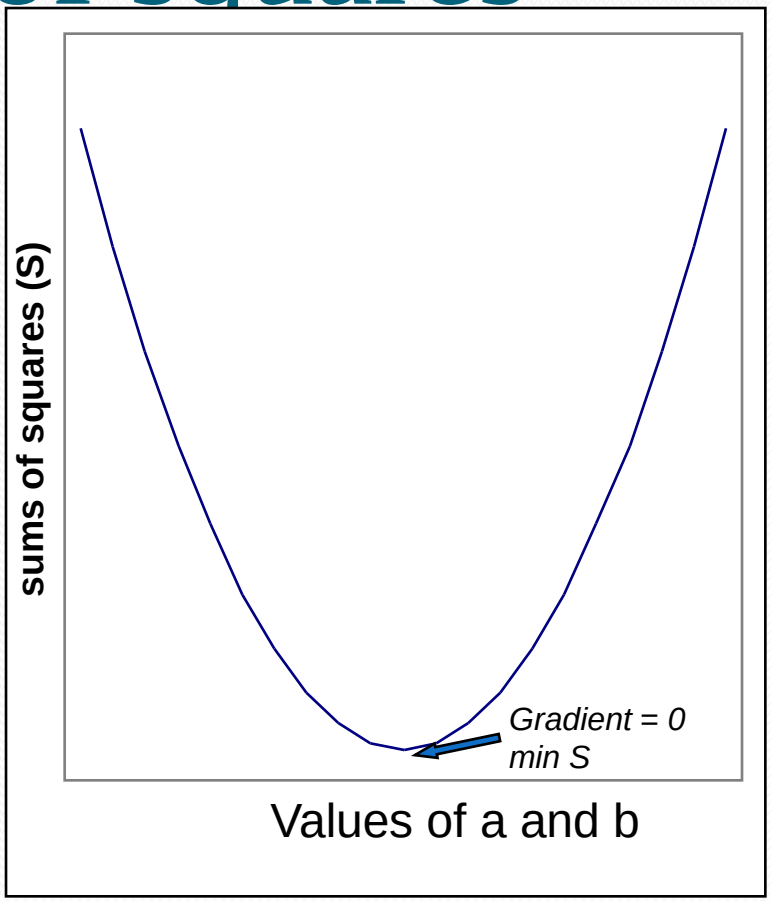
Minimising sums of squares

- Need to minimise $\Sigma(y-\hat{y})^2$
- $\hat{y} = ax + b$
- so need to minimise:

$$\Sigma(y - ax - b)^2$$

- If we plot the sums of squares for all different values of a and b we get a parabola, because it is a squared term

- So the min sum of squares is at the bottom of the curve, where the gradient is zero.



The maths bit

- The min sum of squares is at the bottom of the curve where the gradient = 0

- So we can find a and b that give min sum of squares by taking partial derivatives of $\Sigma(y - ax - b)^2$ with respect to a and b separately

- Then we solve these for 0 to give us the values of a and b that give the min sum of squares

The solution

- Doing this gives the following equations for a and b:

$$a = \frac{r\, s_y}{s_x}$$

r = correlation coefficient of x and y
s_y = standard deviation of y
s_x = standard deviation of x

- From you can see that:
  - A low correlation coefficient gives a flatter slope (small value of a)
  - Large spread of y, i.e. high standard deviation, results in a steeper slope (high value of a)
  - Large spread of x, i.e. high standard deviation, results in a flatter slope (high value of a)

The solution cont.

- Our model equation is $\hat{y} = ax + b$
- This line must pass through the mean so:

$$\bar{y} = a\bar{x} + b \quad \Rightarrow \quad b = \bar{y} - a\bar{x}$$

- We can put our equation for a into this giving:

$$b = \bar{y} - \frac{r\, s_y}{s_x}\, \bar{x}$$

r = correlation coefficient of x and y
s_y = standard deviation of y
s_x = standard deviation of x

- The smaller the correlation, the closer the intercept is to the mean of y

Back to the model

$$\hat{y} = ax + b = \frac{r\,s_y}{s_x}\,x + \bar{y} - \frac{r\,s_y}{s_x}\,\bar{x}$$

a b a

Rearranges to: $\hat{y} = \frac{r\,s_y}{s_x}\,(x - \bar{x}) + \bar{y}$

a

- If the correlation is zero, we will simply predict the mean of y for every value of x, and our regression line is just a flat straight line crossing the x-axis at y
- But this isn't very useful.
- We can calculate the regression line for any data, but the important question is how well does this line fit the data, or how good is it at predicting y from x

How good is our model?

- Total variance of y: $$s_y^2 = \frac{\sum(y - \bar{y})^2}{n - 1} = \frac{SS_y}{df_y}$$

- Variance of predicted y values ($\hat{y}$): $$s_{\hat{y}}^2 = \frac{\sum(\hat{y} - \bar{y})^2}{n - 1} = \frac{SS_{pred}}{df_{\hat{y}}}$$

  This is the variance explained by our regression model

- Error variance: $$s_{error}^2 = \frac{\sum(y - \hat{y})^2}{n - 2} = \frac{SS_{er}}{df_{er}}$$

  This is the variance of the error between our predicted y values and the actual y values, and thus is the variance in y that is NOT explained by the regression model

How good is our model cont.

- Total variance = predicted variance + error variance

$$s_y^2 = s_{\hat{y}}^2 + s_{er}^2$$

- Conveniently, via some complicated rearranging

$$s_{\hat{y}}^2 = r^2 s_y^2$$

$$\Downarrow$$

$$r^2 = s_{\hat{y}}^2 / s_y^2$$

- so r^2 is the proportion of the variance in y that is explained by our regression model

How good is our model cont.

- Insert $r^2 s_y^2$ into $s_y^2 = s_{\hat{y}}^2 + s_{er}^2$ and rearrange to get:

$$s_{er}^2 = s_y^2 - r^2 s_y^2$$
$$= s_y^2 (1 - r^2)$$

- From this we can see that the greater the correlation the smaller the error variance, so the better our prediction

Is the model significant?

- i.e. do we get a significantly better prediction of y from our regression equation than by just predicting the mean?

- F-statistic:

$$F_{(df_{\hat{y}}, df_{er})} = \frac{s_{\hat{y}}^2}{s_{er}^2} \quad \overset{\textit{complicated rearranging}}{=\ldots\ldots=} \quad \frac{r^2 (n - 2)^2}{1 - r^2}$$

- And it follows that:

(because $F = t^{2)}$) $$t_{(n-2)} = \frac{r (n - 2)}{\sqrt{1 - r^2}}$$

So all we need to know are r and n

General Linear Model

- Linear regression is actually a form of the General Linear Model where the parameters are a, the slope of the line, and b, the intercept.

$$y = ax + b + \varepsilon$$

- A General Linear Model is just any model that describes the data in terms of a straight line

Multiple regression

- Multiple regression is used to determine the effect of a number of independent variables, x_1, x_2, x_3 etc, on a single dependent variable, y
- The different x variables are combined in a linear way and each has its own regression coefficient:

$$y = a_1x_1 + a_2x_2 + \ldots\ldots + a_nx_n + b + \varepsilon$$

- The a parameters reflect the independent contribution of each independent variable, x, to the value of the dependent variable, y.
- i.e. the amount of variance in y that is accounted for by each x variable after all the other x variables have been accounted for

SPM

- Linear regression is a GLM that models the effect of one independent variable, x, on ONE dependent variable, y
- Multiple Regression models the effect of several independent variables, x_1, x_2 etc, on ONE dependent variable, y
- Both are types of General Linear Model
- GLM can also allow you to analyse the effects of several independent x variables on several dependent variables, y_1, y_2, y_3 etc, in a linear combination
- This is what SPM does and all will be explained next week!

UNIT-IV

Sampling Distribution and Testing of Hypothesis

Introduction

- Parameters are numerical descriptive measures for populations.
  - For the normal distribution, the location and shape are described by μ and σ.
  - For a binomial distribution consisting of *n* trials, the location and shape are determined by ***p***.
- Often the values of parameters that specify the exact form of a distribution are unknown.
- You must rely on the sample to learn about these parameters.

Sampling

Examples:

- A pollster is sure that the responses to his "agree/disagree" question will follow a binomial distribution, but p, the proportion of those who "agree" in the population, is unknown.
- An agronomist believes that the yield per acre of a variety of wheat is approximately normally distributed, but the mean μ and the standard deviation σ of the yields are unknown.

✓ If you want the sample to provide reliable information about the population, you must select your sample in a certain way!

Simple Random Sampling

- The **sampling plan** or **experimental design** determines the amount of information you can extract, and often allows you to measure the **reliability of your inference**.
- **Simple random sampling** is a method of sampling that allows each possible sample of size *n* an equal probability of being selected.

Types of Samples

- Sampling can occur in two types of practical situations:

1. Observational studies: The data existed before you decided to study it. Watch out for
   - ✓ Nonresponse: Are the responses biased because only opinionated people responded?
   - ✓ Undercoverage: Are certain segments of the population systematically excluded?
   - ✓ Wording bias: The question may be too complicated or poorly worded.

Types of Samples

- Sampling can occur in two types of practical situations:

2. Experimentation: The data are generated by imposing an experimental condition or treatment on the experimental units.
   - ✓ Hypothetical populations can make random sampling difficult if not impossible.
   - ✓ Samples must sometimes be chosen so that the experimenter believes they are representative of the whole population.
   - ✓ Samples must behave like random samples!

Other Sampling Plans

- There are several other sampling plans that still involve randomization:

1. Stratified random sample: Divide the population into subpopulations or strata and select a simple random sample from each strata.
2. Cluster sample: Divide the population into subgroups called clusters; select a simple random sample of clusters and take a census of every element in the cluster.
3. 1-in-k systematic sample: Randomly select one of the first k elements in an ordered population, and then select every k-th element thereafter.

Non-Random Sampling Plans

- There are several other sampling plans that do not involve randomization. They should NOT be used for statistical inference!

1. Convenience sample: A sample that can be taken easily without random selection.
   - People walking by on the street
2. Judgment sample: The sampler decides who will and won't be included in the sample.
3. Quota sample: The makeup of the sample must reflect the makeup of the population on some selected characteristic.
   - Race, ethnic origin, gender, etc.

Sampling Distributions

- Numerical descriptive measures calculated from the sample are called statistics.
- Statistics vary from sample to sample and hence are random variables.
- The probability distributions for statistics are called sampling distributions.
- In repeated sampling, they tell us what values of the statistics can occur and how often each value occurs.

Sampling Distributions

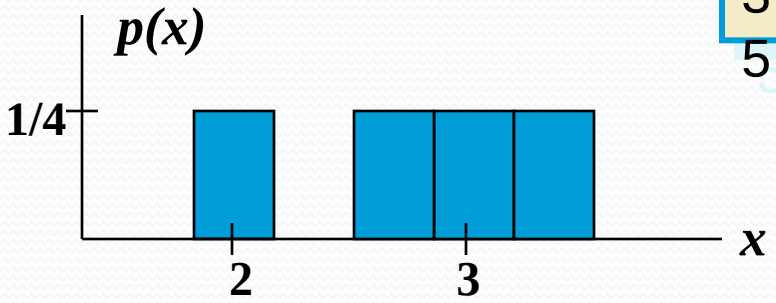
Definition: The sampling distribution of a statistic is the probability distribution for the possible values of the statistic that results when random samples of size *n* are repeatedly drawn from the population

Population: 3, 5, 2, 1

Draw samples of size *n* = 3 without replacement

Possible samples	$\bar{x}$
3, 5, 2	$10/3 = 3.33$
3, 5, 1	$9/3 = 3$
3, 2, 1	$6/3 = 2$
5, 2, 1	$8/3 = 2.67$

Each value of x-bar is equally likely, with probability 1/4



Sampling Distributions

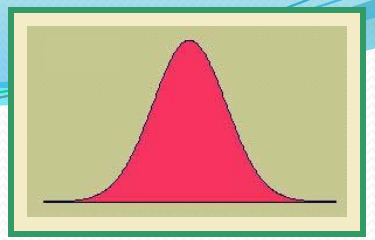
Sampling distributions for statistics can be

- ✓Approximated with simulation techniques
- ✓Derived using mathematical theorems
- ✓The Central Limit Theorem is one such theorem.

Central Limit Theorem: If random samples of *n* observations are drawn from a nonnormal population with finite μ and standard deviation σ , then, when *n* is large, the sampling distribution of the sample mean is approximately normally distributed, with mean μ and standard deviation $\sigma / \sqrt{n}$. The approximation becomes more accurate as *n* becomes large.

$\bar{x}$

Why is this Important?



✓The Central Limit Theorem also implies that the sum of *n* measurements is approximately normal with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$.

✓Many statistics that are used for statistical inference are sums or averages of sample measurements.

✓When *n* is large, these statistics will have approximately normal distributions.

✓This will allow us to describe their behavior and evaluate the reliability of our inferences.

How Large is Large?

If the sample is normal, then the sampling distribution of $\bar{x}$ will also be normal, no matter what the sample size.

When the sample population is approximately symmetric, the distribution becomes approximately normal for relatively small values of *n.*

When the sample population is skewed, the sample size must be at least 30 before the sampling distribution of $\bar{x}$ becomes approximately normal.

The Sampling Distribution of the Sample Mean

✓A random sample of size *n* is selected from a population with mean μ and standard deviation σ.

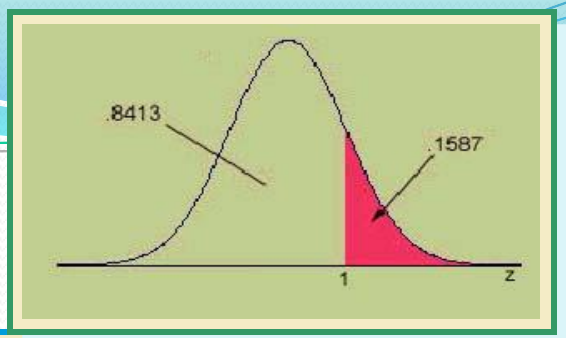
✓The sampling distribution of the sample mean $\bar{X}$ will have mean μ and standard deviation $\sigma / \sqrt{n}$.

✓If the original population is normal, the sampling distribution will be normal for any sample size.

✓If the original population is nonnormal, the sampling distribution will be normal when *n* is large.

The standard deviation of *x*-bar is sometimes called the STANDARD ERROR (SE).

Finding Probabilities for the Sample Mean



✓If the sampling distribution of $\bar{x}$ is normal or approximately normal, *standardize or rescale* the interval of interest in terms of

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

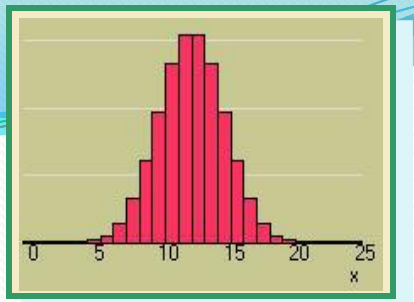
✓Find the appropriate area using Table 3.

Example: A random sample of size n = 16 from a normal distribution with μ = 10 and σ = 8.

$$P(\bar{x} > 12) = P\left(z > \frac{12 - 10}{8 / \sqrt{16}}\right)$$

$$= P(z > 1) = 1 - .8413 = .1587$$

The Sampling Distribution of the Sample Proportion

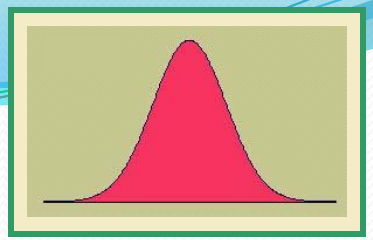


✓The Central Limit Theorem can be used to conclude that the binomial random variable *x* is approximately normal when *n* is large, with mean *np* and standard deviation .

✓The sample proportion, $\hat{p} = \frac{x}{n}$ is simply a *rescaling* of the binomial random variable *x*, dividing it by *n*.

✓From the Central Limit Theorem, the sampling distribution of $\hat{p}$ will also be approximately normal, with a *rescaled* mean and standard deviation.

The Sampling Distribution of the Sample Proportion



✓A random sample of size *n* is selected from a binomial population with parameter *p*.

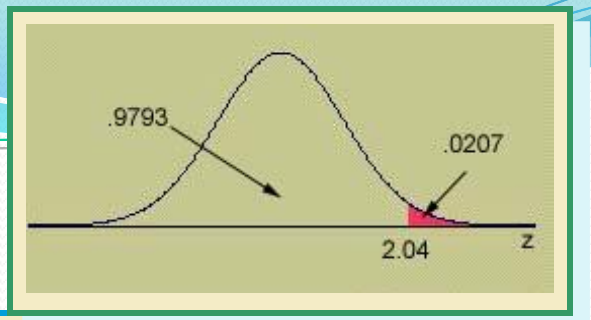
✓The sampling distribution of the sample proportion,

$$\hat{p} = \frac{x}{n}$$

✓will have mean *p* and standard deviation $\sqrt{\frac{pq}{n}}$

✓If *n* is large, and *p* is not too close to zero or one, the sampling distribution of $\hat{p}$ will be approximately normal.

Finding Probabilities for the Sample Proportion



✓If the sampling distribution of $\hat{p}$ is normal or approximately normal, *standardize or rescale* the interval of interest in terms of $z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{pq}{n}}}$

✓Find the appropriate area using Table 3.

Example: A random sample of size n = 100 from a binomial population with p = .4.

$$P(\hat{p} > .5) = P\left(z > \frac{.5 - .4}{\sqrt{\frac{.4(.6)}{100}}}\right)$$

$$= P(z > 2.04) = 1 - .9793 = .0207$$

Types of Inference

- **Estimation:**
  - Estimating or predicting the value of the parameter
  - "What is (are) the most likely values of μ or p?"
- **Hypothesis Testing:**
  - Deciding about the value of a parameter based on some preconceived idea.
  - "Did the sample come from a population with $\mu = 5$ or $p = .2$?"

Types of Inference

- **Examples:**
  - A consumer wants to estimate the average price of similar homes in her city before putting her home on the market.

Estimation: Estimate μ, the average home price.

–A manufacturer wants to know if a new type of steel is more resistant to high temperatures than an old type was.

Hypothesis test: Is the new average resistance, μ_N equal to the old average resistance, μ_O?

Types of Inference

- Whether you are estimating parameters or testing hypotheses, statistical methods are important because they provide:
  - **Methods for making the inference**
  - **A numerical measure of the goodness or reliability of the inference**

Definitions

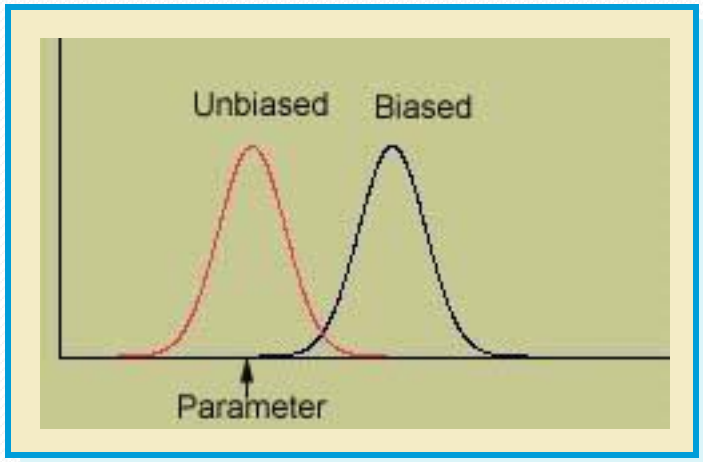
- An **estimator** is a rule, usually a formula, that tells you how to calculate the estimate based on the sample.
  - **Point estimation:** A single number is calculated to estimate the parameter.
  - **Interval estimation:** Two numbers are calculated to create an interval within which the parameter is expected to lie.

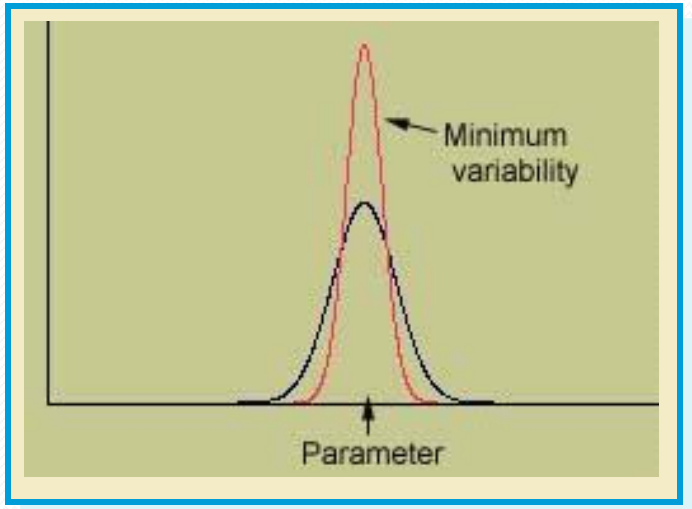
Properties of Point Estimators

- Since an estimator is calculated from sample values, it varies from sample to sample according to its **sampling distribution.**
- An **estimator** is **unbiased** if the mean of its sampling distribution equals the parameter of interest.
  - It does not systematically overestimate or underestimate the target parameter.

Properties of Point Estimators

- Of all the **unbiased** estimators, we prefer the estimator whose sampling distribution has the **smallest spread** or **variability**.





Measuring the Goodness of an Estimator

- The distance between an estimate and the true value of the parameter is the **error of estimation.**

The distance between the bullet and the bull's-eye.

- In this chapter, the sample sizes are large, so that our *unbiased* estimators will have normal dist

Because of the Central Limit Theorem.

Estimating Means and Proportions

- For a quantitative population,

Point estimator of population mean μ : $\bar{x}$

Margin of error $(n \geq 30)$: $\pm 1.96 \dfrac{s}{\sqrt{n}}$

- For a binomial population,

Point estimator of population proportion p : $\hat{p} = x/n$

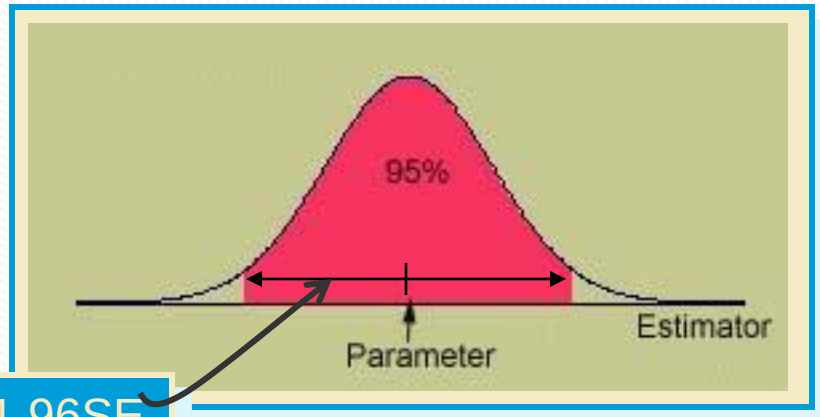
Margin of error $(n \geq 30)$: $\pm 1.96 \sqrt{\dfrac{\hat{p}\hat{q}}{n}}$

Interval Estimation

- Create an interval (a, b) so that you are fairly sure that the parameter lies between these two values.
- "Fairly sure" is means "with high probability", measured using the confidence coefficient,

Usually, $1-\alpha = .90, .95, .98, .99$

- Suppose $1-\alpha = .95$ and that the estimator has a normal distrib



Parameter ± 1.96SE

Confidence Intervals for Means and Proportions

- For a quantitative population,

Confidence interval for a population mean μ :

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

- For a binomial population,

Confidence interval for a population proportion p :

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

UNIT-IV

Large Sample Tests

CLO 16	Understand the fundamentals of hypothesis testing.
CLO 17	Calculate the value of test statistic for the data related to single mean and single proportion.
CLO 18	Calculate the value of test statistic for the data related to difference of means.
CLO 19	Calculate the value of test statistic for the data related to difference of proportions.
CLO 20	Summarize the concept of hypothesis testing to select the best means to stop the hazardous problems like smoking.

Test statistic for T.O.H. in several cases are

- Statistic for test concerning mean σ known

  $$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

- Statistic for large sample test concerning mean with σ unknown

  $$Z = \frac{\overline{X} - \mu_0}{S / \sqrt{n}}$$

- Statistic for test concerning difference between the means

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right)}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}}$$

under NH $H_0 : \mu_1 - \mu_2 = \delta$ against the AH, $H_1: \mu_1 - \mu_2 > \delta$ or $H_1: \mu_1 - \mu_2 < \delta$ or $H_1: \mu_1 - \mu_2 \neq \delta$

- Statistic for large samples concerning the difference between two means (σ_1 and σ_2 are unknown)

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right)}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}}$$

Statistics for large sample test concerning one proportion

- $Z = \frac{X - np_o}{\sqrt{np_0(1-p_0)}}$ under the N.H: H_o: $p = p_o$ against H_1: $p \neq p_o$ or $p > p_o$ or $P<P_o$
- **Statistic for test concerning the difference between two proportions**
- $Z= \frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$
- With $p = \frac{X_1 + X_2}{n_1 + n_2}$ under the NH : H_o: $p_1=p_2$ against the AH H_1:$p_1 < p_2$ or $p_1 > p_2$ or $p_1 \neq p_2$

Estimating the Difference between Two Means

- Sometimes we are interested in comparing the means of two populations.
  - The average growth of plants fed using two different nutrients.
  - The average scores for students taught with two different teaching methods.
- To make this comparison,

A random sample of size n_1 drawn from population 1 with mean μ_1 and variance σ_1^2.

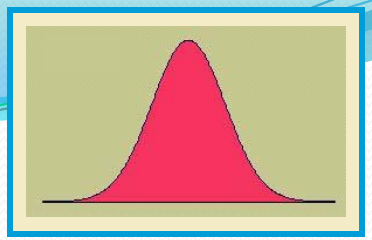
A random sample of size n_2 drawn from population 2 with mean μ_2 and variance σ_2^2.

Estimating the Difference between Two Means

- We compare the two averages by making inferences about μ_1-μ_2, the difference in the two population averages.
  - If the two population averages are the same, then μ_1-$\mu_2 = 0$.
  - The best estimate of μ_1-μ_2 is the difference in the two sample means,

$$\bar{x}_1 - \bar{x}_2$$

The Sampling Distribution of $\bar{x}_1 - \bar{x}_2$



1. The mean of $\bar{x}_1 - \bar{x}_2$ is $\mu_1 - \mu_2$, the difference in the population means.

2. The standard deviation of $\bar{x}_1 - \bar{x}_2$ is $\text{SE} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

3. If the sample sizes are large, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is approximately normal, and SE can be estimated as $\text{SE} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

Estimating μ_1-μ_2

- For large samples, point estimates and their margin of error as well as confidence intervals are based on the standard normal (z) distribution.

Point estimate for $\mu_1 - \mu_2 : \bar{x}_1 - \bar{x}_2$

Margin of Error : $\pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Confidence interval for $\mu_1 - \mu_2$:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Estimating the Difference between Two Proportions

- Sometimes we are interested in comparing the proportion of "successes" in two binomial populations.
  - The germination rates of untreated seeds and seeds treated with a fungicide.
  - The proportion of male and female voters who favor a particular candidate for governor.

A random sample of size n_1 drawn from binomial population 1 with parameter p_1.

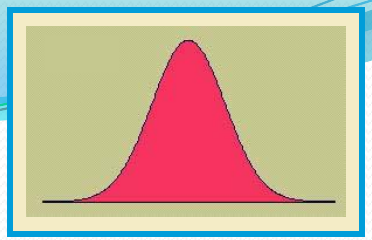
A random sample of size n_2 drawn from binomial population 2 with parameter p_2.

Estimating the Difference between Two Means

- We compare the two proportions by making inferences about p_1-p_2, the difference in the two population proportions.
  - If the two population proportions are the same, then p_1-$p_2 = 0$.
  - The best estimate of p_1-p_2 is the difference in the two sample proportions,

$$\hat{p}_1 - \hat{p}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2}$$

The Sampling Distribution of $\hat{p}_1 - \hat{p}_2$

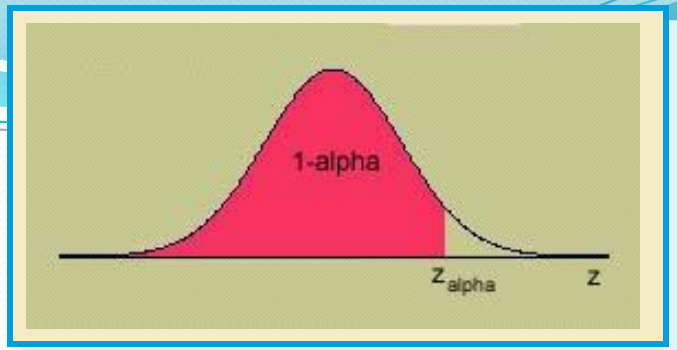


1. The mean of $\hat{p}_1 - \hat{p}_2$ is $p_1 - p_2$, the difference in the population proportions.

2. The standard deviation of $\hat{p}_1 - \hat{p}_2$ is $\text{SE} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$.

3. If the sample sizes are large, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal, and SE can be estimated as $\text{SE} = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$.

One Sided Confidence Bounds



- Confidence intervals are by their nature **two-sided** since they produce upper and lower bounds for the parameter.
- **One-sided bounds** can be constructed simply by using a value of z that puts α rather than $\alpha/2$ in the tail of the z distribution.

$$\text{LCB} : \text{Estimator} - z_{\alpha} \times (\text{Std Error of Estimator})$$

$$\text{UCB} : \text{Estimator} + z_{\alpha} \times (\text{Std Error of Estimator})$$

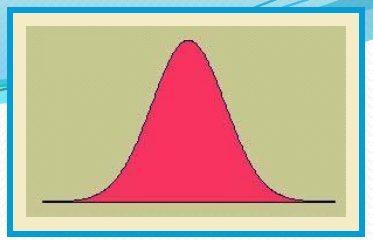
Parameter	Point Estimator	Margin of Error
μ	$\bar{x}$	$\pm 1.96\left(\frac{s}{\sqrt{n}}\right)$
p	$\hat{p} = \frac{x}{n}$	$\pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
$p_1 - p_2$	$(\hat{p}_1 - \hat{p}_2) = \left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right)$	$\pm 1.96\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$

IV. Large-Sample Interval Estimators

To estimate one of four population parameters when the sample sizes are large, use the following interval estimators.

Parameter	$(1-\alpha)100\%$ Confidence Interval
μ	$\bar{x} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$
p	$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$
$\mu_1 - \mu_2$	$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
$p_1 - p_2$	$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$

The Sampling Distribution of the Sample Mean



- When we take a sample from a normal population, the sample mean $\bar{x}$ has a normal distribution for any sample size *n*, and

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$\frac{\bar{x} - \mu}{s / \sqrt{n}} \text{ is not normal!}$$

- has a standard normal distribution.
- But if σ is unknown, and we must use *s* to estimate it, the resulting statistic **is not normal**.

UNIT-V

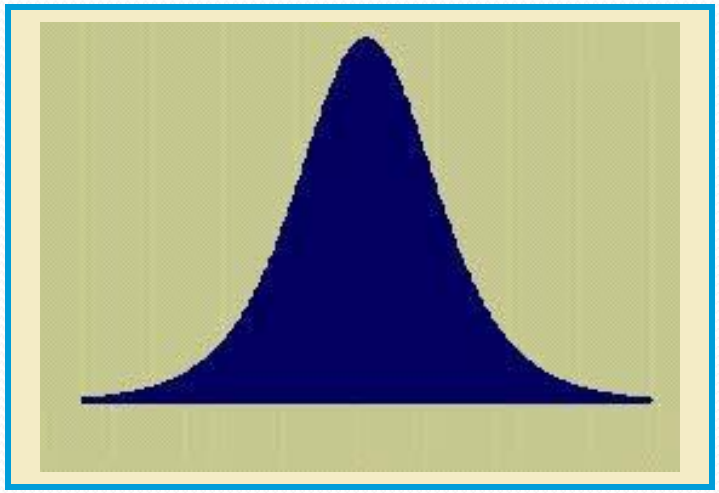
Small Sample Tests

CLO 21	Use Student t-test to predict the difference in sample means.
CLO 22	Apply F-test to predict the difference in sample variances.
CLO 23	Understand the characteristics between the samples using Chi-square test.

Student's t Distribution

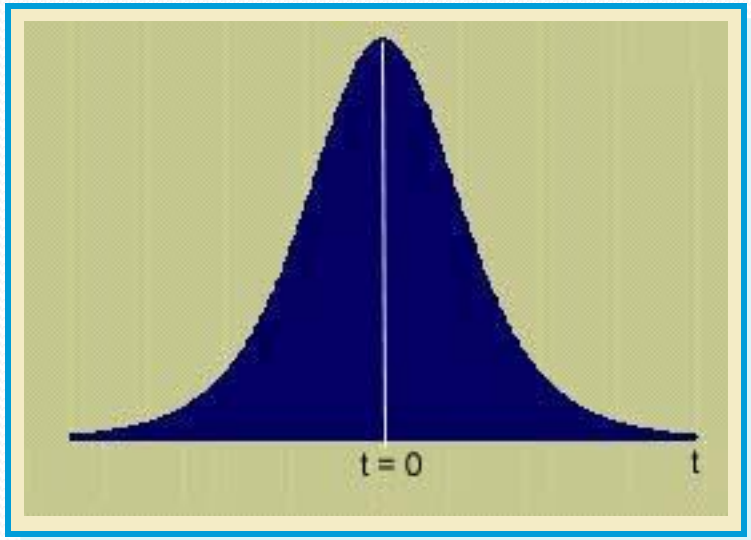
- Fortunately, this statistic does have a sampling distribution that is well known to statisticians, called the **Student's t distribution**, with ***n*-1 degrees of freedom.**

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



•We can use this distribution to create estimation testing procedures for the population mean μ.

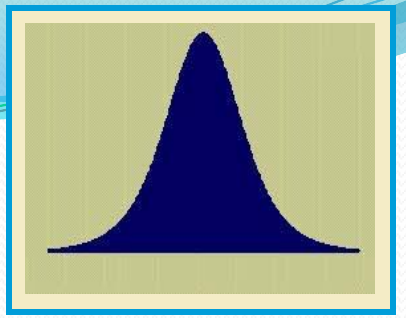
Properties of Student's *t*



- Mound-shaped and symmetric about 0.
- More variable than *z*, with "heavier tails"

- Shape depends on the sample size *n* or the **degrees of freedom, *n*-1.**
- As *n* increases the shapes of the *t* and *z* distributions become almost identical.

Small Sample Inference for a Population Mean μ



- The basic procedures are the same as those used for large samples. For a test of hypothesis:

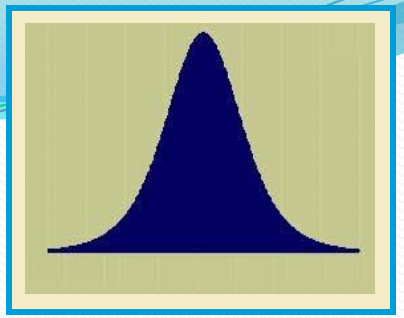
Test $H_0 : \mu = \mu_0$ versus H_a : one or two tailed

using the test statistic

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

using p - values or a rejection region based on a t - distribution with $df = n - 1$.

Small Sample Inference for a Population Mean μ



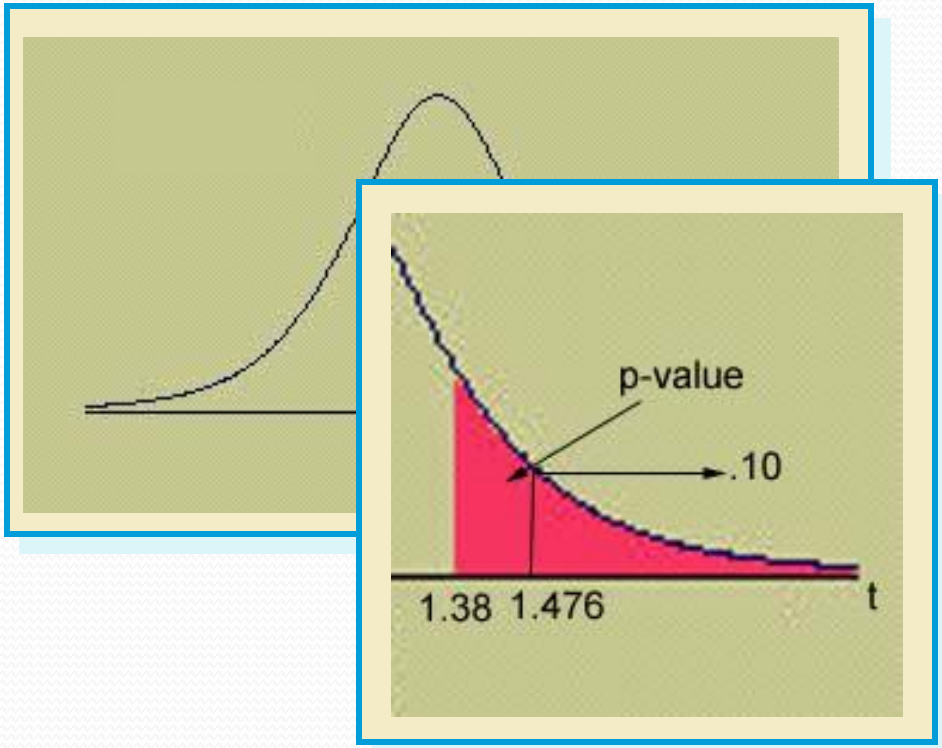
- For a 100(1–α)% confidence interval for the population mean μ:

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

where $t_{\alpha/2}$ is the value of t that cuts off area $\alpha/2$ in the tail of a t-distribution with $df = n - 1$.

Approximating the *p*-value

- You can only approximate the *p*-value for the test using Table 4.



df	$t_{.100}$	$t_{.050}$
1	3.078	6.314
2	1.886	2.920
3	1.638	2.353
4	1.533	2.132
5	1.476	2.015

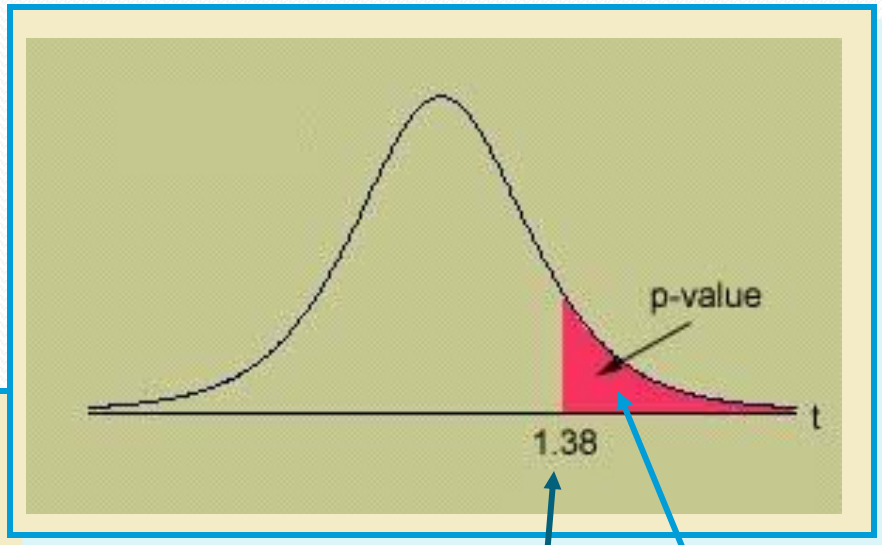
Since the observed value of $t = 1.38$ is smaller than $t_{.10} = 1.476$,

p-value > .10.

The exact *p*-value

- You can get the exact *p*-value using some calculators or a computer.

p-value = .113 which is greater than .10 as we approximated using Table 4.

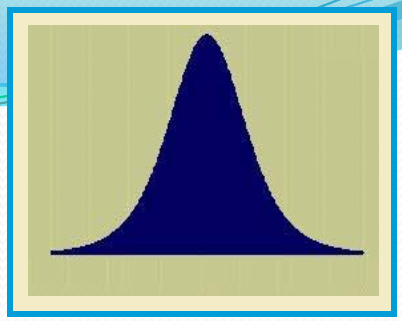


One-Sample T: Times

Test of mu = 15 vs > 15

Variable	N	Mean	StDev	SE Mean	95% Lower Bound	T	P
Times	6	19.1667	7.3869	3.0157	13.0899	1.38	0.113

Testing the Difference between Two Means



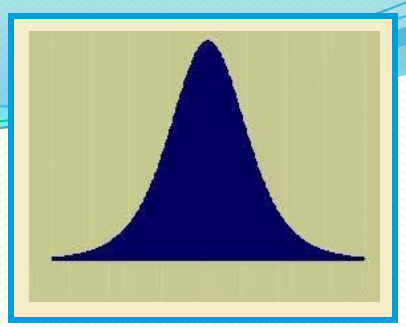
As in Chapter 9, independent random samples of size n_1 and n_2 are drawn from populations 1 and 2 with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2.

Since the sample sizes are small, the two populations must be normal.

- To test:
  - **H_0: $\mu_1 - \mu_2 = D_0$ versus H_a: one of three**

where D_0 is some hypothesized difference, usually 0.

Testing the Difference between Two Means

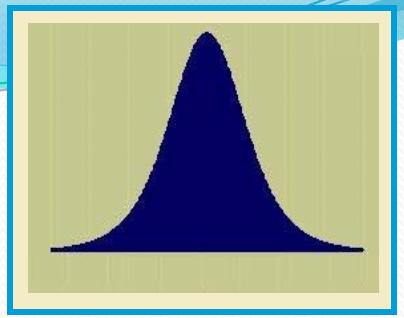


- The test statistic used in Chapter 9

$$z \approx \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- does not have either a *z* or a *t* distribution, and cannot be used for small-sample inference.
- We need to make one more assumption, that **the population variances, although unknown, are equal.**

Testing the Difference between Two Means



- Instead of estimating each population variance separately, we estimate the common variance with

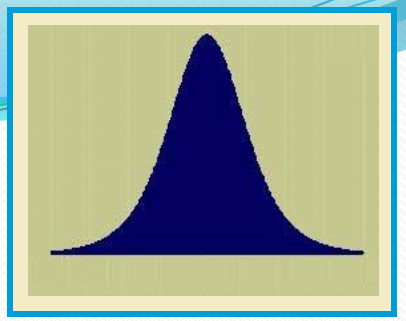
$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- And the resulting test statistic,

$$t = \frac{\bar{x}_1 - \bar{x}_2 - D_0}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

has a t distribution with n_1+n_2-2 degrees of freedom.

Estimating the Difference between Two Means



- You can also create a 100(1-α)% confidence interval for μ_1-μ_2.

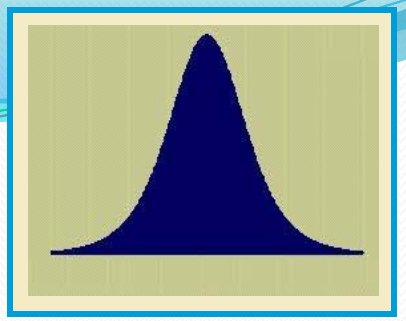
$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\text{with } s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Remember the three assumptions:

1. Original populations normal
2. Samples random and independent
3. Equal population variances.

Testing the Difference between Two Means



- How can you tell if the equal variance assumption is reasonable?

Rule of Thumb :

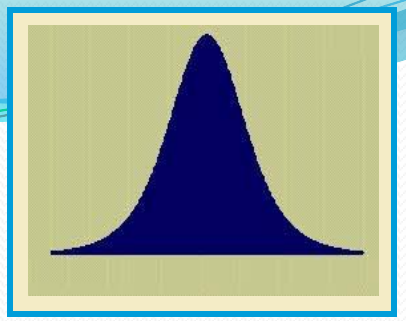
If the ratio, $\frac{\text{larger } s^2}{\text{smaller } s^2} \leq 3$,

the equal variance assumption is reasonable.

If the ratio, $\frac{\text{larger } s^2}{\text{smaller } s^2} > 3$,

use an alternative test statistic.

Testing the Difference between Two Means



- If the population variances cannot be assumed equal, the test statistic

$$t \approx \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

- has an approximate *t* distribution with degrees of freedom given above. This is most easily done by computer.

The Paired-Difference Test

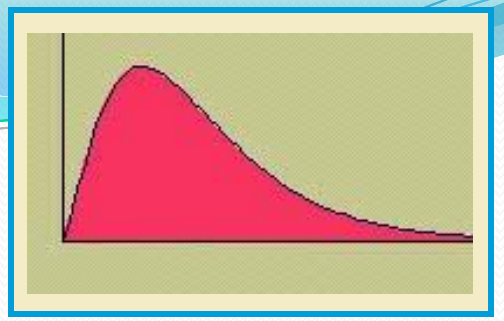
To test $H_0 : \mu_1 - \mu_2 = 0$ we test $H_0 : \mu_d = 0$ using the test statistic

$$t = \frac{\bar{d} - 0}{s_d / \sqrt{n}}$$

where n = number of pairs, $\bar{d}$ and s_d are the mean and standard deviation of the differences, d_i.

Use the p-value or a rejection region based on a t-distribution with $df = n - 1$.

Inference Concerning a Population Variance

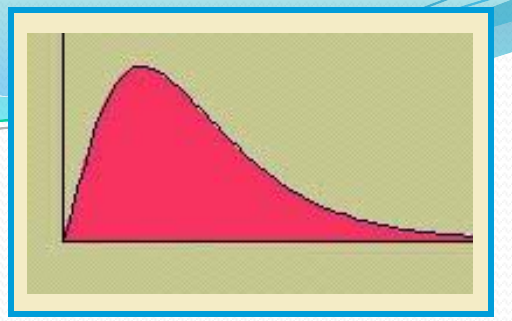


- Sometimes the primary parameter of interest is not the population mean μ but rather the population variance σ^2. We choose a random sample of size n from a normal distribution.
- The sample variance s^2 can be used in its standardized form:

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

- which has a Chi-Square distribution with n - 1 degrees of freedom.

Inference Concerning a Population Variance



To test $H_0 : \sigma^2 = \sigma_0^2$ versus H_a : one or two tailed

we use the test statistic

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$ with a rejection region based on

a chi-square distribution with $df = n-1$.

Confidence interval :

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

Inference Concerning Two Population Variances

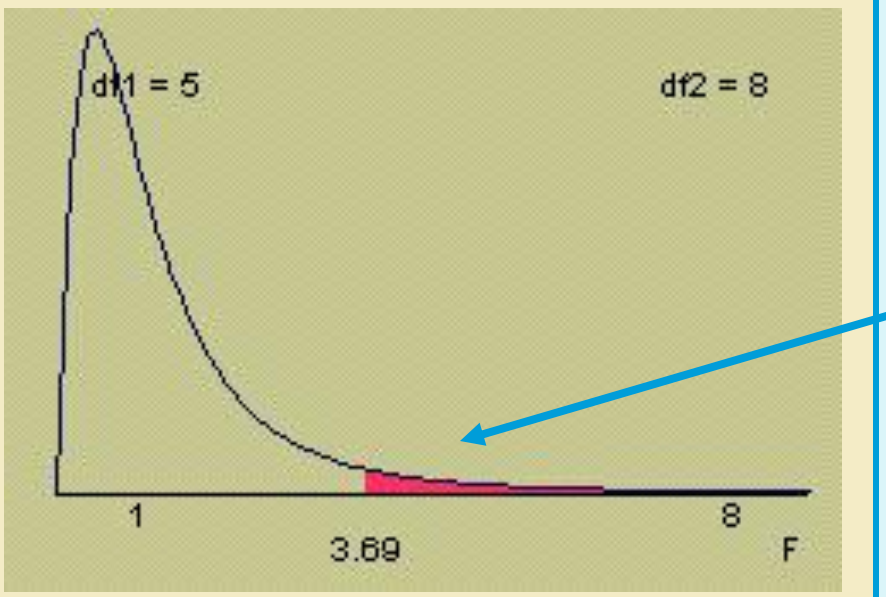
- We can make inferences about the ratio of two population variances in the form a ratio. We choose two independent random samples of size n_1 and n_2 from normal distributions.
- If the two population variances are equal, the statistic

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

- has an F distribution with $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$ degrees of freedom.

Inference Concerning Two Population Variances

- Table 6 gives only upper critical values of the F statistic for a given pair of df_1 and df_2.



For example, the value of F that cuts off .05 in the upper tail of the distribution with $df_1 = 5$ and $df_2 = 8$ is F =3.69.

Inference Concerning Two Population Variances

To test $H_0 : \sigma_1^2 = \sigma_2^2$ versus H_a : one or two tailed
we use the test statistic

$F = \dfrac{s_1^2}{s_2^2}$ where s_1^2 is the larger of the two sample variances.

with a rejection region based on an F distribution with $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$.

Confidence interval :

$$\frac{s_1^2}{s_2^2}\frac{1}{F_{df_1,df_2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2}F_{df_2,df_1}$$

Parameter	Test Statistic	Degrees of Freedom
μ	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$n - 1$
$\mu_1 - \mu_2$ (equal variances)	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$	$n_1 + n_2 - 2$
$\mu_1 - \mu_2$ (unequal variances)	$t \approx \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$	Satterthwaite's approximation
$\mu_1 - \mu_2$ (paired samples)	$t = \dfrac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$	$n - 1$
σ^2	$\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$	$n - 1$
σ_1^2/σ_2^2	$F = s_1^2/s_2^2$	$n_1 - 1$ and $n_2 - 1$