



BITS Pilani
Pilani Campus

Descriptive Statistics

Akanksha Bharadwaj
Asst. Professor, BITS Pilani



BITS Pilani
Pilani Campus



SS ZG536, ADV STAT TECHNIQUES FOR ANALYTICS

Contact session 1



Books

No	Author(s), Title, Edition, Publishing House
----	---

T1	Probability and Statistics for Engineering and Sciences, 8 th Edition, Jay L Devore, Cengage Learning
T2	Applied Logistic Regression, Hosmer and Lemeshow, 3 rd Edition, Wiley
T3	Introduction to Time Series and Forecasting, Second Edition, Peter J <u>Brockwell</u> , Richard A Davis, Springer.

No	Author(s), Title, Edition, Publishing House
R1	Miller and Freund's Probability and statistics for Engineers, 8 th Edition, PHI
R2	Statistics for Business and Economics by Anderson, Sweeney and <u>Wiliams</u> , CENAGE learning



Session Plan

Contact hour-1, Module 1:

Time	Type	Description	References
Pre- CH-1	RL	RL – 1.1.1 (Data representation)	
During CH-1	CH-1	Discussion on data representation	T1:Chapter 1
Post-CH-1	HW	T1: Chapter 1	T1:Chapter 1
Lab			

Contact hour-2, Module 1

Time	Type	Description	References
Pre-CH-2	RL	RL – 1.1.2 (Data visualization)	
During CH-2	CH-2	Measures of Central Tendency, Measures of Variability	T1:Chapter 1
Post-CH-2	HW	T1: Chapter 1	T1:Chapter 1
Lab			



What is Statistics?

- a form of knowledge- a mode of arranging and stating facts which belong to various sciences (Lond. And Westn. Rev, 1838)
 - Science dealing with collection, analysis, interpretation, and presentation of masses of numerical data (Webster dictionary, 1966)
 - Science of collecting and analysing numerical data (Oxford dictionary, 1996)
-



Population Vs Sample

- An investigation will typically focus on a well-defined collection of objects constituting a **population** of interest
- When desired information is available for all objects in the population, we have what is called a **census**.
- Constraints on time, money, and other scarce resources usually make a census impractical or infeasible.
- Instead, a subset of the population—a **sample**—is selected in some prescribed manner.



Parameter Vs Statistic

- A descriptive measure of the population is called **parameter**
 - A descriptive measure of the sample is called **statistic**
-



Branches of statistics

-
- Descriptive statistics
 - Inferential statistics



Descriptive statistics

- If a business analyst is using data gathered on a group to describe and reach conclusions about the same group, the statistics are called descriptive statistics.
- Example- if an instructor produces statistics to summarize a class's examination efforts and uses those statistics to reach conclusions about that class only.



Inferential statistics

- If a researcher gathers data from a sample and uses the statistics generated to reach conclusions about the population from which the sample was taken
- Example- pharmaceutical research



BITS Pilani
Pilani Campus



Terminologies



Variable

- A **variable** is any characteristic whose value may change from one object to another in the population.
- E.g. age the patients, number of visits to a particular website , etc.



Types of variable

-
- **Categorical/ qualitative variables:**
 - Take category or label values and place an individual into one of several groups.
 - Each observation can be placed in only one category, and the categories are mutually exclusive.
 - **Quantitative variables:**
 - Take numerical values and represent some kind of measurement.

Example: Indian census data 2010



	State	Zip code	Family size	Annual income
1	U.P	201001	5	10,00,000
2	Delhi	110092	10	25,00,000
3	Gurgaon	122503	12	40,00,000
4	Delhi	110091	4	8,00,000
5	U.P	201003	2	2,00,000
6	Gurgaon	122004	1	5,00,000



Data Sets

- A **dataset** is a set of data identified with particular circumstances. Datasets are typically displayed in tables, in which rows represent individuals and columns represent variables.
- A **univariate** data set consists of observations on a single variable.
- **Bivariate** data sets have observations made on two variables
- **Multivariate** data arises when observations are made on more than one variable (so bivariate is a special case of multivariate)

Data Measurement



Nominal level-

- It is the lowest level of data measurement.
- Numbers representing nominal level data can be used only to classify or categorize
- Example- Employee ID



Data Measurement

Ordinal Level-

- In addition to nominal level capabilities, it can be used to rank or order objects
- The categories for each of these ordinal variables show order, but not the magnitude of difference between two adjacent points.
- Example- a supervisor can rank the productivity of employees from 1 to 5



Data Measurement

Interval level-

- In this distances between consecutive numbers have meaning and the data are always numerical.
- the distance between pairs of consecutive numbers is assumed to be equal.
- Zero is just another point on scale and does not mean the absence of phenomenon
- Example- temperature in Fahrenheit



Data Measurement

Ratio level-

- It has same properties as interval data, but ratio data have an absolute zero, and the ratio of two numbers is meaningful
- Example- height, weight, time, volume, production cycle time, etc.
- For instance, we know that someone who is forty years old is twice as old as someone who is twenty years old.
- There is a meaningful zero point – that is, it is possible to have the absence of age.



Exercise- Healthcare industry

The following type of questions are sometimes asked in the survey. These question will result in what level of data measurement

- How long ago were you released from the hospital?
- Which type of unit were you in for most of your stay?
 - Intensive care
 - Maternity care
 - Surgical unit
- How serious was your condition when you were first admitted to the hospital
 - Critical
 - Moderate
 - Minor
 - 1- it is a time measurement with absolute zero and is therefore ratio level measurement
 - 2- nominal data as it is used only to categorize
 - 3- it can be ranked by selection so ordinal data



BITS Pilani
Pilani Campus



Data Visualization

Pie charts and bar chart for a categorical data



- Refer to BMI data
- Both the pie chart and the bar chart help us visualize the distribution of a categorical variable

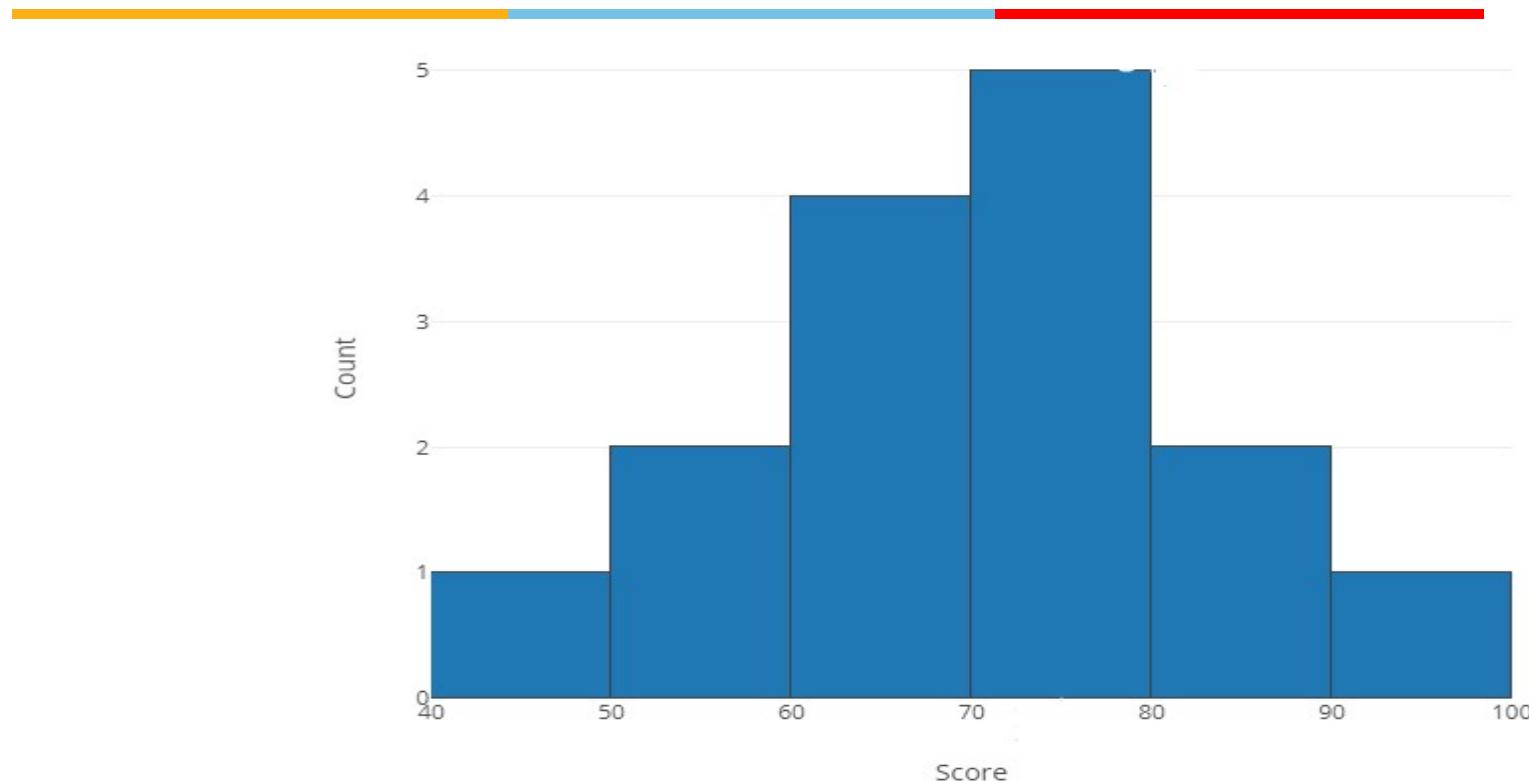
Histograms for quantitative data



Here are the scores in mathematics for 15 students:

78, 58, 65, 71, 57, 74, 79, 75, 87, 92, 81, 69, 66, 43, 63

Score	Count
[40-50]	1
[50-60]	2
[60-70]	4
[70-80]	5
[80-90]	2
[90-100]	1



Ques. What percentage of students earned less than a grade of 70 on the exam?

$$\frac{7}{15} * 100$$



Example

A survey was conducted to see how many video calls people made daily. The results are displayed in the table below:

Number of calls made	Frequency
1 – 3	10
4 - 7	7
8 – 11	4
12 - 15	1
16 - 19	1

Ques1. Tell how many of the people surveyed make less than 4 video calls daily? 16

Ques2. How many people were surveyed?

23

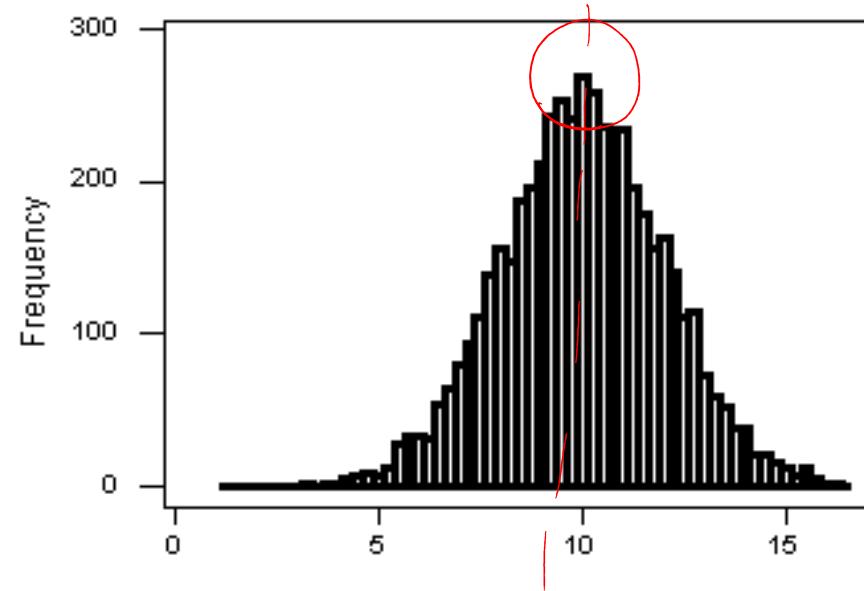


Shape of histograms

When describing the shape of a distribution, we should consider:

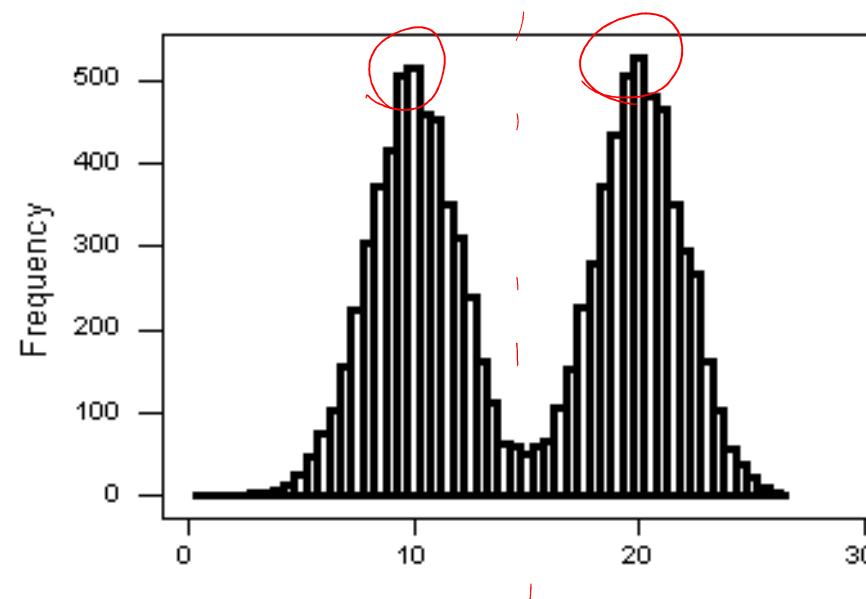
- **Symmetry/skewness** of the distribution.
- **Peakedness (modality)**—the number of peaks (modes) the distribution has.

Symmetric and single peaked distribution



<https://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/describing-distributions/>

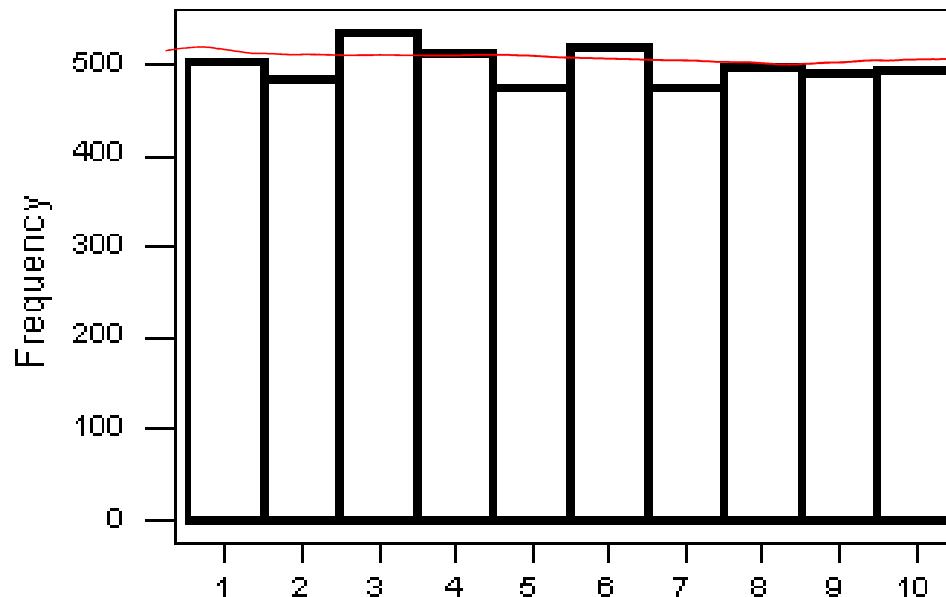
Symmetric and double peaked distribution



<https://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/describing-distributions/>

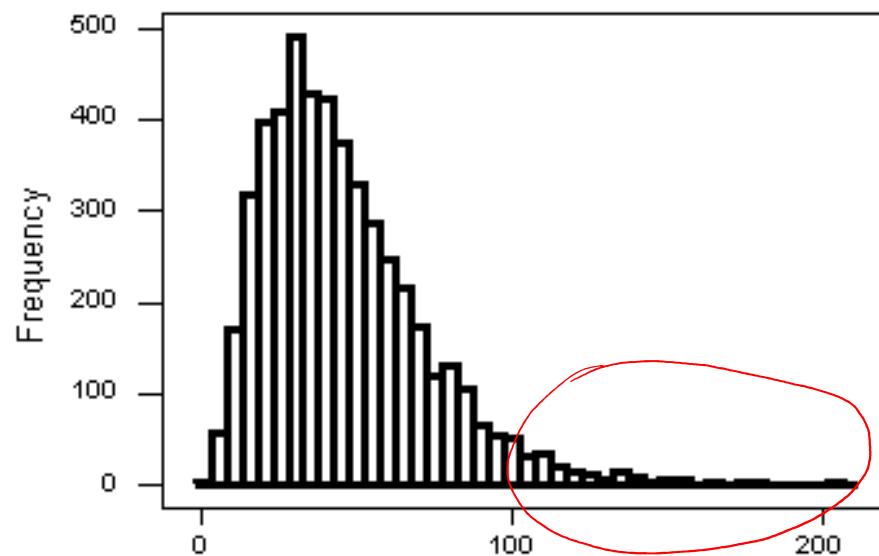


Symmetric and flat distribution

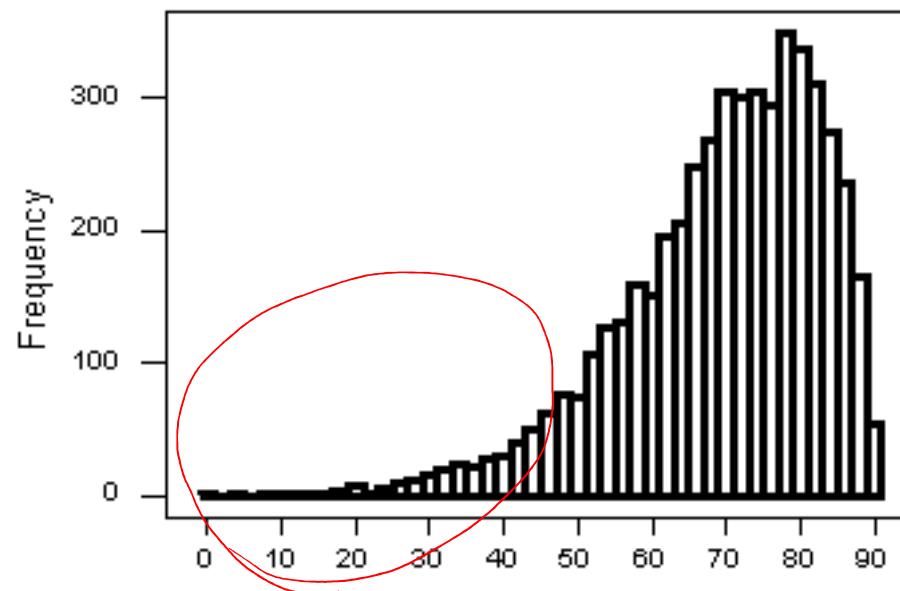


<https://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/describing-distributions/>

Right skewed distribution



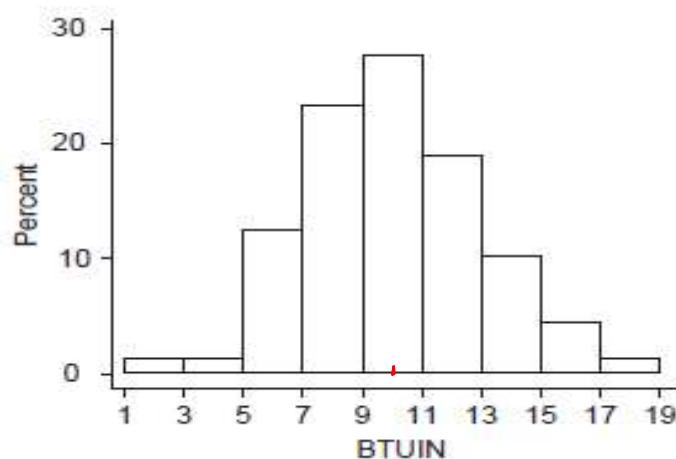
Left skewed distribution



Center



The center of the distribution is its **midpoint**—the value that divides the distribution so that approximately half the observations take smaller values, and approximately half the observations take larger values.



Histogram of the energy consumption data

Image: Book (Probability and statistics for the engineering and sciences by Devore)

Spread



The **spread** (also called **variability**) of the distribution can be described by the approximate range covered by the data. From looking at the histogram, we can approximate the smallest observation (min), and the largest observation (max), and thus approximate the range.

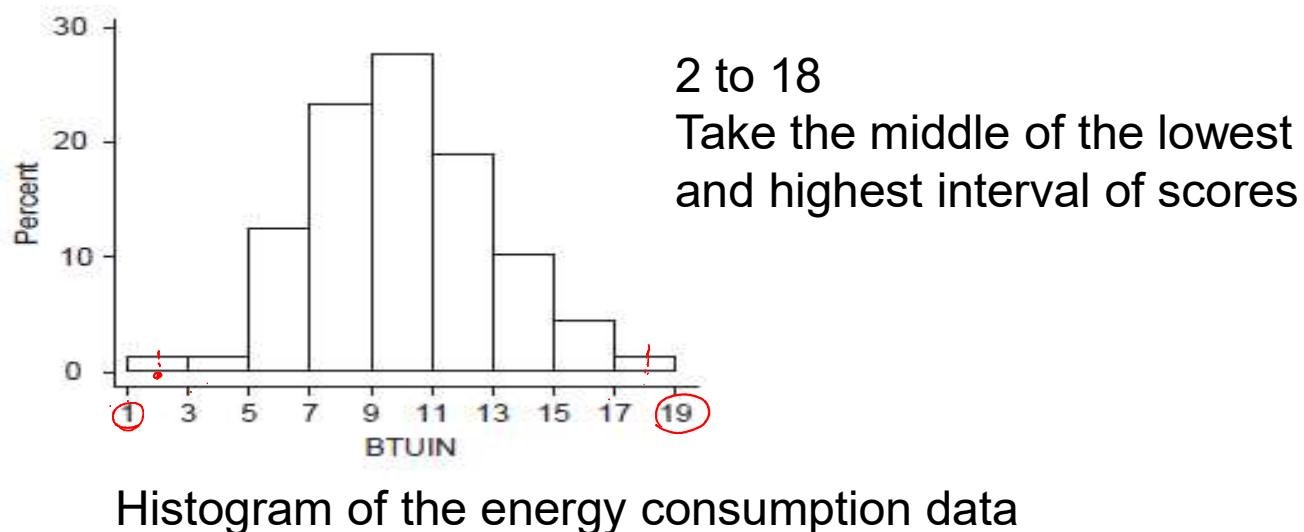


Image: Book (Probability and statistics for the engineering and sciences by Devore)



Stem and Leaf Plot

The stemplot (also called stem and leaf plot) is another graphical display of the distribution of quantitative data.

Separate each data point into a stem and leaf, as follows:

222 3
222 5
222 7

- The leaf is the right-most digit.
- The stem is everything except the right-most digit.
- So, if the data point is 54, then 5 is the stem and 4 is the leaf.
- If the data point is 5.35, then 5.3 is the stem and 5 is the leaf.

Example

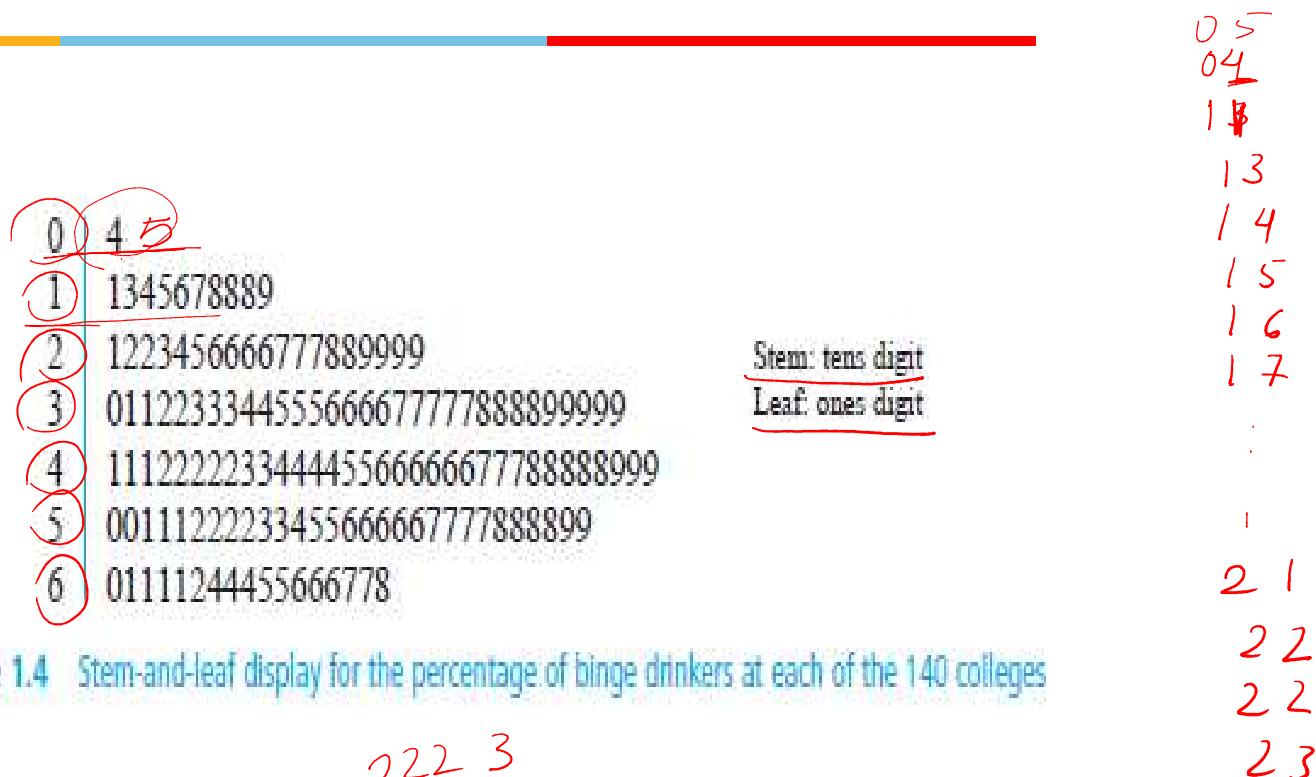


Figure 1.4 Stem-and-leaf display for the percentage of binge drinkers at each of the 140 colleges

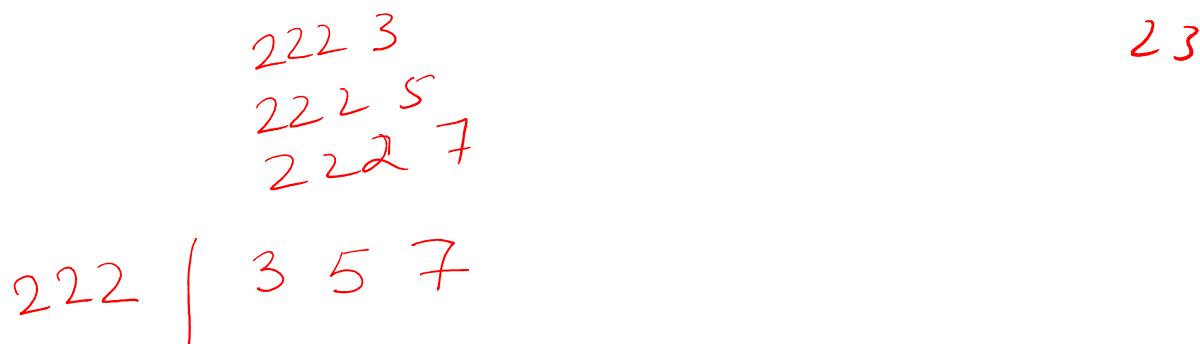


Image: Book (Probability and statistics for the engineering and sciences by Devore)



Dotplot

- Used to summarize a quantitative variable graphically. The dotplot, like the stemplot, shows each observation, but displays it with a dot rather than with its actual value.
- When a value occurs more than once, there is a dot for each occurrence, and these dots are stacked vertically.

Example



A dot plot of 50 random values from 0 to 9.

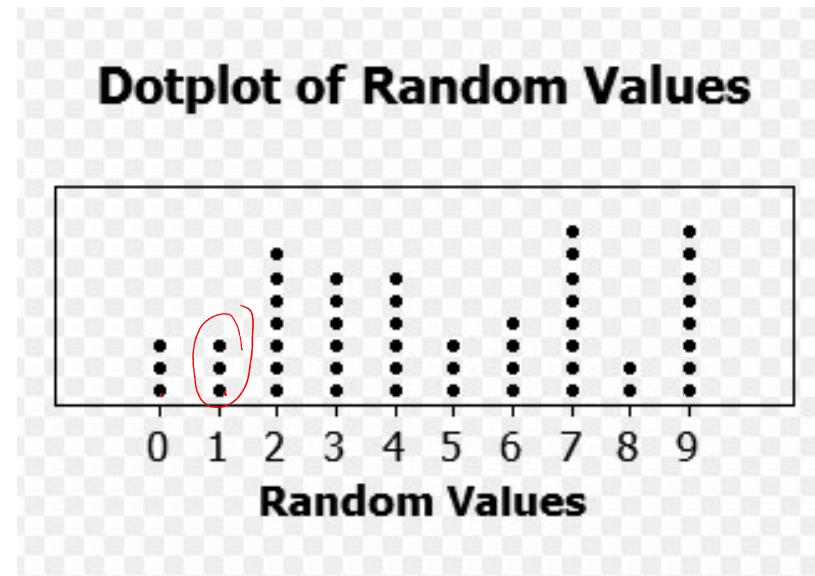


Image: Wikipedia



BITS Pilani
Pilani Campus



Measures of Location/Central Tendency: Ungrouped Data



Introduction

Here are the number of hours that 9 students spend on social media on a typical day:

11 6 7 5 2 8 11 12 15

Summarize the data using a single digit



Mean

Mean is the sum of the observations divided by the number of observations

11 6 7 5 2 8 11 12 15

Ques1. Mean is?

$$\underline{77/9 = 8.55}$$

Sample Mean	Population Mean
$\bar{x} = \frac{\sum x}{n}$	$\mu = \frac{\sum x}{N}$

where $\sum X$ is sum of all data values

N is number of data items in population

n is number of data items in sample



Mean

When to Use the Mean

- Sampling stability is desired.
- Other measures are to be computed such as standard deviation, coefficient of variation and skewness



Median

The median **M** is the midpoint of the **ordered** distribution.

Steps:

- Order the data from smallest to largest.
- Consider whether n , the number of observations, is even or odd.
 - If n is odd, the median M is the center observation in the ordered list. This observation is the one "sitting" in the $\underline{(n + 1) / 2}$ spot in the ordered list.
 - If n is even, the median M is the mean of the two center observations in the ordered list. These two observations are the ones "sitting" in the $\underline{n / 2}$ and $\underline{n / 2 + 1}$ spots in the ordered list.



Example

11 6 7 5 2 8 11 12 15

Ques1. Median is?

* 2 5 6 7 8 11 11 12 15

Location is $(9+1)/2 = 5^{\text{th}}$ element

So median is 8



Mode

the mode is the most commonly occurring value in a distribution.

11 6 7 5 2 8 11 12 15 12

Ques1. Mode is?

Mode will be 11

Ques2. What kind of distribution is formed by the data from the above 9 students?

Unimodal

Comparing the Mean and the Median



The mean is very sensitive to **outliers**, while the median is resistant to outliers?

TRUE or FALSE?

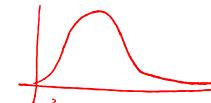
Use the below data to analyze

Data set A → 54 55 56 68 70 71 73

Data set B → 54 55 56 68 70 71 **730**

Comparing the Mean and the Median: Interpretations

- For symmetric distributions with no outliers: mean is approximately equal to median



- For skewed right distributions and/or datasets with high outliers: mean > median



- For skewed left distributions and/or datasets with low outliers: mean < median

Ques. The Current Population Survey conducted by the Census Bureau records the incomes of a large sample of Indian households each month. What will be the relationship between the mean and median of the collected data?

Ans. mean > median since data will be right skewed

The mean is an appropriate measure of center only for symmetric distributions with no outliers. In all other cases, the median should be used to describe the center of the distribution.

Quartiles

- Quartiles in statistics are values that divide your data into quarters.
- However, quartiles aren't shaped like pizza slices; Instead they divide your data into four segments according to where the numbers fall on the number line.

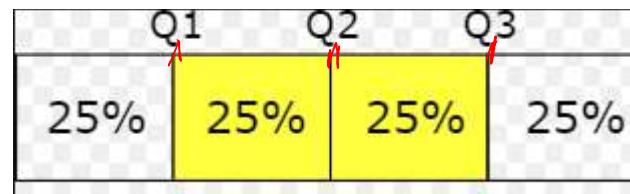
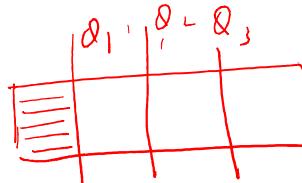


Image: Wikipedia

Quartiles



There are three quartiles: the first quartile (Q_1), the second quartile (Q_2), and the third quartile (Q_3).

- The first quartile (lower quartile, QL), is equal to the 25th percentile of the data.
- The second (middle) quartile or median of a data set is equal to the 50th percentile of the data
- The third quartile, called upper quartile (QU), is equal to the 75th percentile of the data.

Example



$$\left(\frac{n}{2}\right) \left(\frac{n}{2} + 1\right)$$

Ordered Data Set: 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

Q2 will be $(n+1)/2 = 12/2 = 6^{\text{th}}$ element i.e. **40**

Q1 will be $(15+36)/2 = 25.5$

Q3 will be $(42+43)/2 = 42.5$



BITS Pilani
Pilani Campus



Measures of variability: Ungrouped Data



Range

- The range is exactly the distance between the smallest data point (min) and the largest one (Max).
- Here are the number of hours that 9 students spend on social media on a typical day:

11 6 7 5 2 8 11 12 15

Ques. Range for above case is ?

2 5 6 7 8 11 11 12 15

Range is $15 - 2 = 13$



Inter-Quartile Range (IQR)

- While the range quantifies the variability by looking at the range covered by *ALL* the data,
- The IQR measures the variability of a distribution by giving us the range covered by the *MIDDLE* 50% of the data.
- The middle 50% of the data falls between Q1 and Q3, and therefore: $\text{IQR} = \text{Q3} - \text{Q1}$
- The IQR should be used as a measure of spread of a distribution only when the median is used as a measure of center.



Example

Ordered Data Set: 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

Q2 will be $(n+1)/2 = 12/2 = 6^{\text{th}}$ element i.e. **40**

Q1 will be $(15+36)/2 = \mathbf{25.5}$

Q3 will be $(42+43)/2 = \mathbf{42.5}$

Ques. IQR range will be?

$$Q3 - Q1 = 42.5 - 25.5 = 17$$

Using the IQR to Detect Outliers



- The IQR is used as the basis for a rule of thumb for identifying outliers.
- An observation is considered a suspected outlier if it is:
 - below $Q1 - 1.5(IQR)$ or
 - above $Q3 + 1.5(IQR)$



Boxplot

-
- The boxplot graphically represents the distribution of a quantitative variable by visually displaying the five-number summary and any observation that was classified as a suspected outlier using the $1.5(IQR)$ criterion.

Steps

- The central box spans from Q1 to Q3.
- A line in the box marks the median
- Lines go from the edges of the box to the smallest and largest observations that are not classified as outliers.

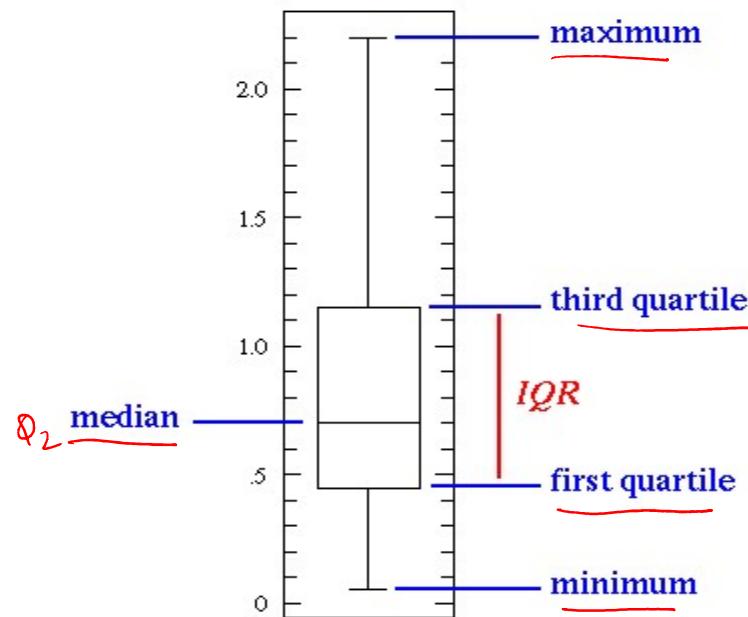
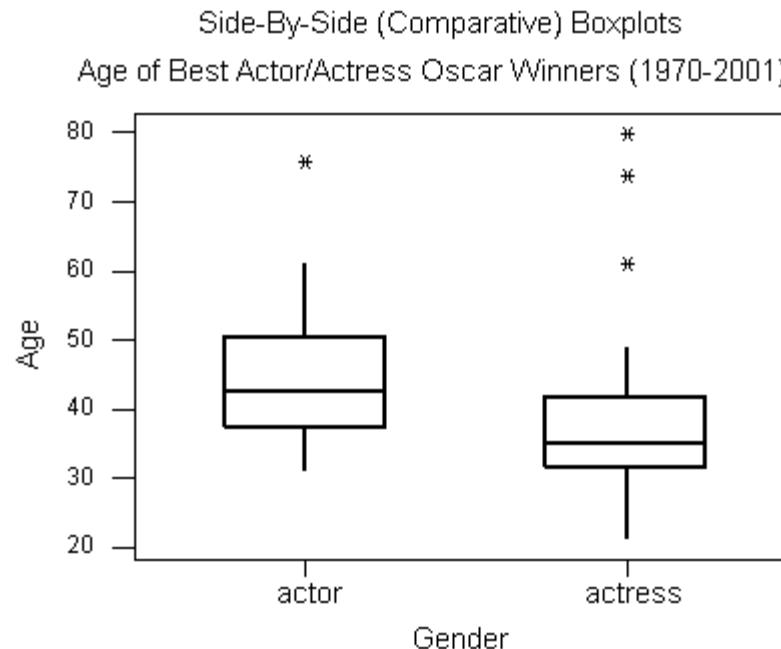


Image: google

Side by side Box plot

- Boxplots are most useful when presented side-by-side for comparing and contrasting distributions from two or more groups.



- Actors: min = 31, Q1 = 37.25, M = 42.5, Q3 = 50.25, Max = 76
- Actresses: min = 21, Q1 = 32, M = 35, Q3 = 41.5, Max = 80

Variance

- It is the average of squared deviations about the arithmetic mean for a set of numbers

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

Population Variance

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

Sample Variance

degree of freedom

Image: google



Standard Deviation

-
- The idea behind the standard deviation is to quantify the spread of a distribution by measuring how far the observations are from their mean.
 - The standard deviation gives the average (or typical distance) between a data point and the mean.
 - It is the square root of variance



Example

Here are the number of hours that 9 students spend on social media on a typical day:

11 6 7 5 2 8 11 12 15

Find the standard deviation.

Mean = 8.55

- Deviations from the mean

11-8.55, 6-8.55, 7-8.55, 5-8.55, 2-8.55, 8-8.55, 11-8.55, 12-8.55, 15-8.55

2.45, -2.55, -1.55, -3.55, -6.55, -0.55, 2.45, 3.45, 6.45



Example

11 6 7 5 2 8 11 12 15

Square of each of the deviation

6.0025, 6.5025, 2.4025, 12.6025, 42.9025, 0.3025, 6.0025,
11.9025, 41.6025

- Average the square deviations by adding them up, and dividing by $n - 1$

16.277

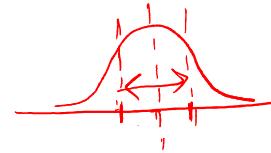
This average of the squared deviations is called the variance of the data.

- The SD of the data is the square root of the variance

4.034

The Empirical Rule

Consider a symmetric mound-shaped distribution, the following rule applies:



- Approximately 68% of the observations fall within 1 standard deviation of the mean.
- Approximately 95% of the observations fall within 2 standard deviations of the mean.
- Approximately 99.7% of the observations fall within 3 standard deviations of the mean.

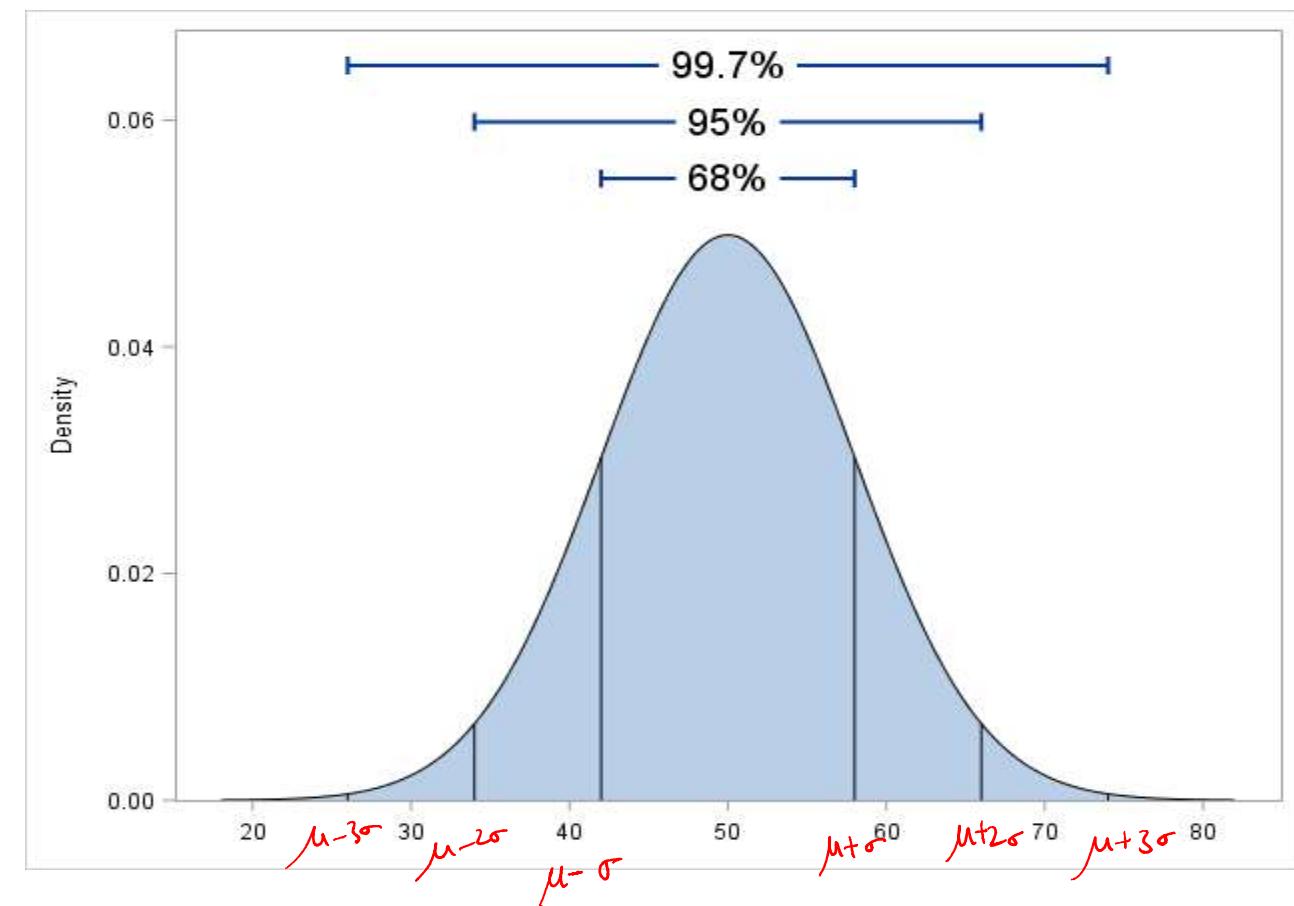


Image: google



Coefficient of variation

- It is the ratio of the standard deviation to the mean expressed in percentage and denoted by CV

CV for a population:

$$CV = \frac{\sigma}{\mu} * 100\%$$

CV for a sample:

$$CV = \frac{s}{\bar{x}} * 100\%$$



BITS Pilani
Pilani Campus



Measures of central tendency: Grouped Data



Mean

Here M_i represents class mid-point

$$\mu_{\text{grouped}} = \frac{\sum fM}{N} = \frac{\sum fM}{\sum f} = \frac{f_1M_1 + f_2M_2 + \dots + f_iM_i}{f_1 + f_2 + \dots + f_i}$$

where

i = the number of classes

10 - 20

f = class frequency

20 - 30

N = total frequencies

30 - 40

Median

$$\text{Median} = L + \frac{\frac{N}{2} - cf_p}{f_{med}}(W)$$

where:

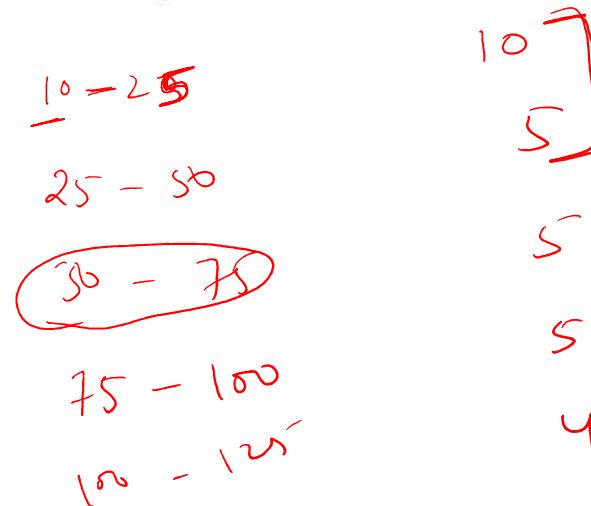
L = the lower limit of the median class interval

cf_p = a cumulative total of the frequencies up to but not including the frequency of the median class

f_{med} = the frequency of the median class

W = the width of the median class interval

\underline{N} = total number of frequencies



Example

(HW)



- Frequency Distribution of 60 Years of Unemployment Data for Canada (Grouped Data)

Class Interval	Frequency	Cumulative Frequency
1-under 3	4	4
3-under 5	12	16
5-under 7	13	29
7-under 9	19	48
9-under 11	7	55
11-under 13	5	60



Mode

- The mode for grouped data is the class midpoint of the modal class. The modal class is the class interval with the greatest frequency.

Ques. What will be the mode for the previous example data?

Class Interval	Frequency	Cumulative Frequency
1–under 3	4	4
3–under 5	12	16
5–under 7	13	29
7–under 9	19	48
9–under 11	7	55
11–under 13	5	60



BITS Pilani
Pilani Campus



Measures of variability: Grouped Data



Population Variance and Standard deviation

FORMULAS FOR
POPULATION VARIANCE
AND STANDARD DEVIATION
OF GROUPED DATA

Original Formula Computational Version

$$\sigma^2 = \frac{\sum f(M - \mu)^2}{N}$$
$$\sigma = \sqrt{\sigma^2}$$

$$\sigma^2 = \frac{\sum fM^2 - \frac{(\sum fM)^2}{N}}{N}$$

where:

f = frequency

M = class midpoint

N = $\sum f$, or total frequencies of the population

μ = grouped mean for the population

Ques. Calculate Variance and Standard deviation for previous example (HW)



Sample Variance and Standard deviation

FORMULAS FOR SAMPLE VARIANCE AND STANDARD DEVIATION OF GROUPED DATA

Original Formula

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

$$s = \sqrt{s^2}$$

Computational Version

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

where:

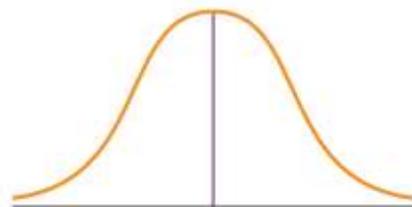
f = frequency

M = class midpoint

n = $\sum f$, or total of the frequencies of the sample

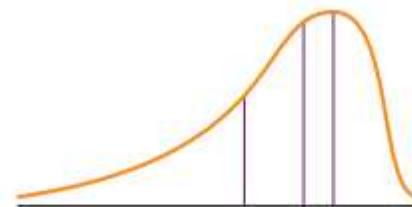
\bar{x} = grouped mean for the sample

Relation between mean, median and mode



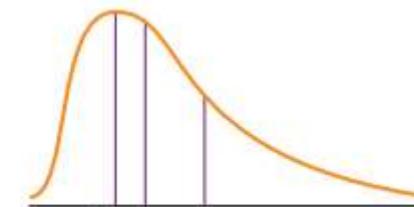
Mean
Median
Mode

(a)
**Symmetric distribution
(no skewness)**



Mean
Median
Mode

(b)
**Negatively
skewed**



Mode
Median
Mean

(c)
**Positively
skewed**

Questions





References

-
- Probability and Statistics for Engineering and Sciences, 8th Edition, Jay L Devore, Cengage Learning
 - Applied Business Statistics, Ken Black
 - [http://www2.isye.gatech.edu/~jeffwu/presentations/datas
cience.pdf](http://www2.isye.gatech.edu/~jeffwu/presentations/datas%20science.pdf)
 - [https://magazine.amstat.org/blog/2015/10/01/asa-
statement-on-the-role-of-statistics-in-data-science/](https://magazine.amstat.org/blog/2015/10/01/asa-statement-on-the-role-of-statistics-in-data-science/)
 - [https://link.springer.com/article/10.1007/s41060-018-
0102-5](https://link.springer.com/article/10.1007/s41060-018-0102-5)
 - [https://link.springer.com/article/10.1007/s42081-018-
0009-3#Sec2](https://link.springer.com/article/10.1007/s42081-018-0009-3#Sec2)



BITS Pilani
Pilani Campus

Probability

Akanksha Bharadwaj
Asst. Professor, BITS Pilani



BITS Pilani
Pilani Campus

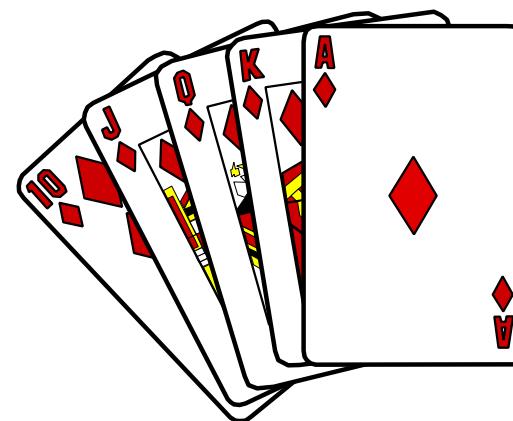


SS ZG536, ADV STAT TECHNIQUES FOR ANALYTICS Contact Session 2



Agenda

-
- Experiments, Counting Rules, and Assigning Probabilities
 - Events and Their Probability
 - Some Basic Relationships of Probability
 - Conditional Probability
 - Bayes' Theorem



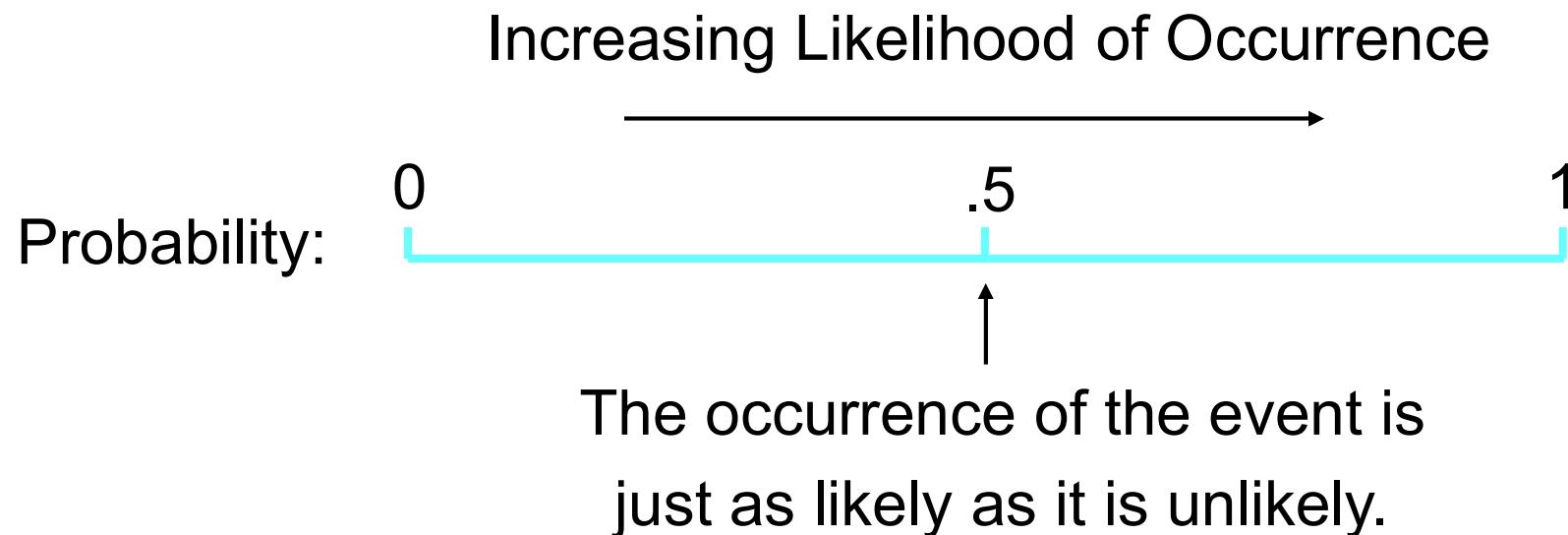
Probability



- The term probability refers to the study of randomness and uncertainty
- One way to think of probability is that it is the **likelihood** that something will occur.
- **Notation:** $P(A)$ is the probability that event A will occur

$$P(A) = \frac{\text{no. of times event occurred}}{\text{total no. of events.}}$$

Probability as a Numerical Measure of the Likelihood of Occurrence





Sample Space

- The **sample space** of an experiment, denoted by S , is the set of all possible outcomes of that experiment
- Example: If we examine three fuses in sequence and note the result of each examination, then an outcome for the entire experiment is any sequence of N's and D's of length 3, where N represents not defective, D represents defective
- $S = \{\underline{NNN}, \underline{NND}, \underline{NDN}, \underline{NDD}, \underline{DNN}, \underline{DND}, \underline{DDN}, \underline{DDD}\}$

for 2 coins :-

$$S = \{ HH, HT, TH, TT \}$$

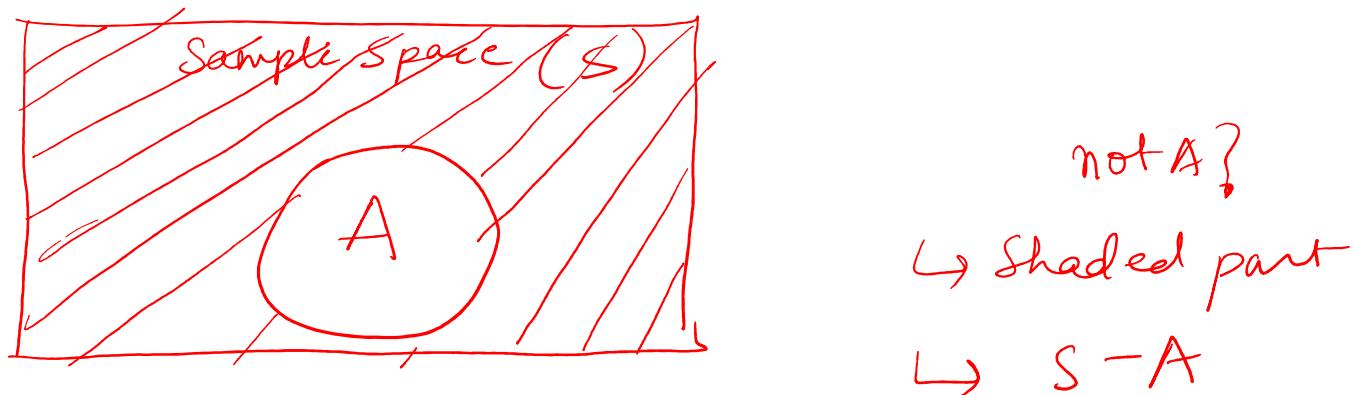


Event

- An event is a set of outcomes of the experiment. This includes the *null* (empty) set of outcomes and the set of *all* outcomes.

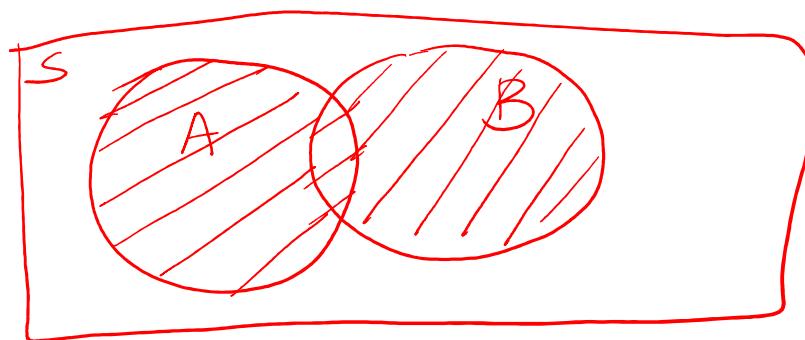
Complement of an Event

- The complement of event A is defined to be the event consisting of all sample points that are not in A .
- The complement of A is denoted by $\underline{A^c}$ or \underline{A}'
- Venn diagram can illustrate the concept of a complement



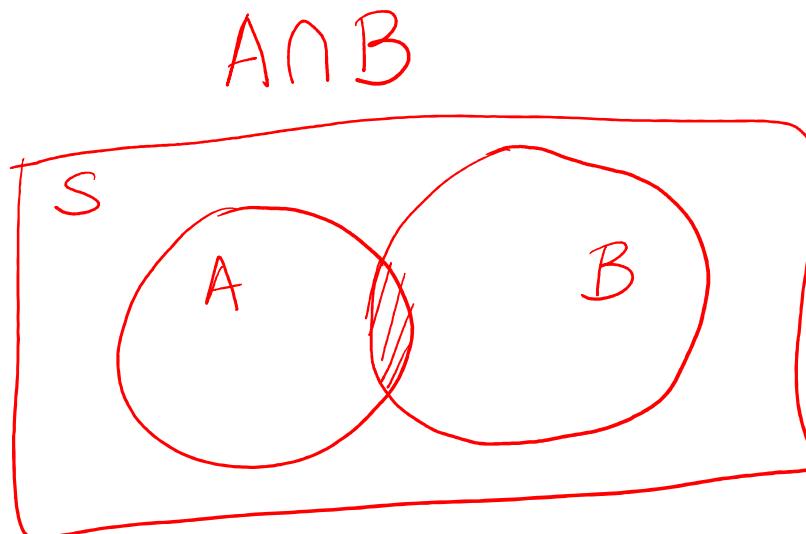
Union of Two Events

- The union of events A and B is the event containing all sample points that are in A or B or both.
- The union is denoted by $\underline{A \cup B}$
- The union of A and B is:



Intersection of Two Events

- The intersection of events A and B is the set of all sample points that are in both A and B .
- The intersection of A and B is the area of overlap





Set Theory Example

- For the experiment in which the number of pumps in use at a single six-pump gas station is observed, let $A = \{0, 1, 2, \underline{3}, \underline{4}\}$, $B = \{\underline{3}, \underline{4}, 5, 6\}$, and $C = \{1, 3, 5\}$.

Then, $S = \{0, 1, 2, 3, 4, 5, 6\}$

$$A' = \underline{\underline{S - A}} = \{5, 6\}$$

$$A \cup B = \{0, 1, 2, 3, 4, 5, 6\}$$

$$A \cup C = \{0, 1, 2, 3, 4, 5\}$$

$$A \cap B = \{3, 4\}$$

$$A \cap B$$

$$A \cap C = \{1, 3\}$$



Disjoint events

- Two events that cannot occur at the same time are called **disjoint** or **mutually exclusive**.
- The idea of **disjoint events** is about whether or not it is possible for the events to occur at the same time

eg:- flipping of coin

event A = Head

event B = tail

A & B are disjoint

$$A \cap B = ?$$

$$A \cap B = \emptyset$$

Non-disjoint events



- Two events can occur at the same time
- For non-disjoint events A and B

$$A \cap B \neq \emptyset / \text{null}$$



Independent events

- The idea of **independent events** is about whether or not the events affect each other in the sense that the occurrence of one event affects the probability of the occurrence of the other
- The event of getting a head on the first toss of a coin is independent of getting a head on the second toss.



Dependent events

- The occurrence of one event gives information about the occurrence of the other
- Suppose we have 5 blue marbles and 5 red marbles in a bag. We pull out one marble, which may be blue or red. Now there are 9 marbles left in the bag.
- The probability that the second marble will be red depends on first outcome

Basic properties of probability



- For any event A , $P(A) \geq 0$
- $P(S) = 1$
- If A_1, A_2, A_3, \dots is an infinite collection of disjoint events, then $P(A_1 \cup A_2 \cup A_3 \dots) = \sum_{i=1}^{\infty} P(A_i)$
- Complement rule: $P(\text{not } A) = 1 - \underline{P(A)}$;





Determining Probability

These three represent distinct conceptual approaches to the study of probability theory.

1. Classical approach
2. Relative frequency approach
3. Subjective probability



Classical method

- **Classical** methods are used for games of chance, such as flipping coins, rolling dice, spinning spinners, roulette wheels, or lotteries.
- When probabilities are assigned based on laws and rules, the method is referred to as the classical method of assigning probabilities.

CLASSICAL METHOD OF ASSIGNING PROBABILITIES

where

$$P(E) = \frac{n_e}{N}$$

N = total possible number of outcomes of an experiment

n_e = the number of outcomes in which the event occurs out of N outcomes



Example

- Example: Each traditional (cube-shaped) die has six sides, marked in dots with the numbers 1 through 6.
- On a "fair" die, these numbers are equally likely to end up face-up when the die is rolled.
- Thus, $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = ?$

$\frac{1}{6}$

Limitation of Classical definition



The classical definition of probability has certain drawbacks and fails at times in different situations as described below:

- This emphasizes that the events must be equally likely. Thus, it fails when various outcomes of a trial are not equally likely.
- For example if a die is biased that gives numbers greater than 3 more often than the numbers less than 3, then the occurrence of numbers on the die is not equally probable.

Limitation of Classical definition



- It does not consider those situations that are unlikely but that could conceivably happen.
 - Like the occurrence of a coin landing on its edge or our room burning down while watching TV etc., which are extremely unlikely but not impossible.
-



Random experiment

- Term "**random experiment**" is used to describe any action whose outcome is not known in advance.



Relative frequency

- To estimate the probability of event A, written $P(A)$, we may repeat the random experiment many times and count the number of times event A occurs.

PROBABILITY BY RELATIVE
FREQUENCY OF
OCCURRENCE

$$\frac{\text{Number of Times an Event Occurred}}{\text{Total Number of Opportunities for the Event to Occur}}$$

- This is also called the **relative frequency of event A**.
- Relative frequency of occurrence is not based on rules or laws but on what has occurred in the past.



Exercise

- The random experiment is rolling a fair die once.
- The sample space of all possible outcomes in this case this is $S = \{1, 2, 3, 4, 5, 6\}$.
- If we had following events in past 6,5,1,4,2,4,6,3,1,1,2,4
- We are interested in a particular type of outcome, which is represented by event E—getting an even number. What is the probability for this based on the past outcomes?

$$P(4) = 3/12$$

$$P(2) = 2/12$$

$$P(6) = 2/12$$

$$\begin{aligned} P(\text{even}) &= P(4) + P(2) \\ &\quad + P(6) \\ &= 7/12 \end{aligned}$$

Limitations of relative frequency



1. The condition of an experiment may not remain the same in long series of trials
 2. The relative frequency may not attain a unique value in spite of a large number of trials
-



Subjective Probability

- It is an estimate that reflects a person's opinion, or best guess about whether an outcome will occur.
- These are values (between 0 and 1 or 0 and 100%) assigned by individuals based on how likely they think events are to occur.
- Example: The probability of candidate winning in an election is based on opinion poll is 60%.



Exercise

- On the "Information for the Patient" label of a certain diabetes medicine it is claimed when taking this medication
- there is a 10% chance of experiencing sleeping problems (denote this event by \underline{S}) $P(S) = 0.1$
- there is a 29% chance of experiencing headaches (denote this event by H), and $P(H) = 0.29$
- - there is a 35% chance of experiencing at least one of these two side effects (denote this event by A) $0.35 = P(A)$

Ques. What is the probability that a patient taking this drug will not experience insomnia?

Ques. The probability of "the patient will experience neither of the two side effects"?



$$P(\text{Sleep problem}) = 0.1 \quad (P(S))$$

$$P(\text{Headache}) = 0.29 \quad / \quad P(H)$$

$$P(\text{Atleast 1}) = 0.35 \quad / \quad P(A)$$

Q1. $P(\text{not } S) = 1 - 0.1 = 0.9$

Q2. $P(\text{neither of 2 problems}) = 1 - 0.35$
 $= 0.65$



Exercise

- The sales manager of an e-commerce company says that 80% of those who visit their website for the first time do not buy any mobile. If a new customer visits the website ~~first time~~^{first time}, what is the probability that the customer would buy mobile

$$\begin{aligned} P(\text{visits \& buy}) &= 1 - 0.8 \\ &= 0.2 \end{aligned}$$



Exercise

-
- A woman's pocket contains two dollars and two pennies. She randomly extracts one of the coins and, after looking at it, replaces it before picking a second coin.
 - Let Q₁ be the event that the first coin is a dollar and Q₂ be the event that the second coin is a dollar.
 - **Are Q₁ and Q₂ independent events? Why?**

Yes because of replacement of coin



Exercise

- A woman's pocket contains two dollars and two pennies. She randomly extracts one of the coins, and **without placing** it back into her pocket, she picks a second coin. As before, let Q1 be the event that the first coin is a dollar, and Q2 be the event that the second coin is a dollar.
- Are Q1 and Q2 independent events? Why?

No
not replacing



Exercise

- 40% people have blood group as B, 30% have blood group as A, 25% have blood group as O and 5% have blood group as AB. From a **large population** three people are randomly selected. Let,
- $P(B_1)$ =probability that first person is of blood group B
- $P(B_2)$ =probability that second person is of blood group B
- $P(B_3)$ =probability that third person is of blood group B
- Are $P(B_1)$, $P(B_2)$ and $P(B_3)$ independent?

Yes

large population & random selection

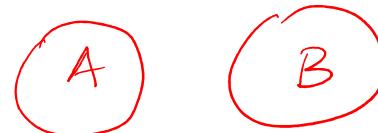
$P(A \text{ or } B)$

$P(A \cup B)$



- For disjoint events

$$P(A \text{ and } B) = 0$$



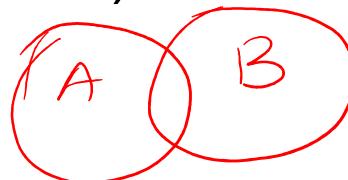
$P(A \text{ or } B) = P(\text{event A occurs or event B})$

i.e, $P(A \text{ or } B) = \underline{P(A)} + \underline{P(B)}$ (**addition rule**)

- For non-disjoint events

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

(General Addition Rule)





Exercise

- In a city it was observed that 80% of the families owns a two wheeler and 43% owns a car. Those who owns both are 38%. If a family is selected at random what is the probability that they own either a two wheeler or a car.

$$P(A) = \text{owns two wheeler} = 0.8$$

$$P(B) = \text{owns a car} = 0.43$$

$$P(A \cap B) = 0.38 \quad \text{non-disjoint events}$$

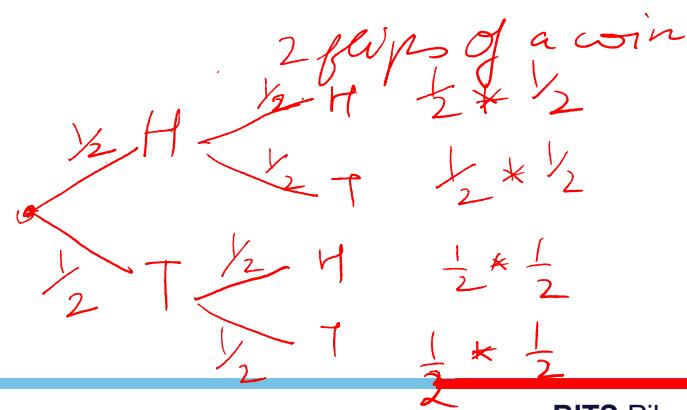
$$\begin{aligned} P(A \text{ or } B) / P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.8 + 0.43 - 0.38 \\ &= 0.85 \end{aligned}$$

P(A and B)

- $P(A \text{ and } B)$, the probability that both events A and B occur.
- $P(A \text{ and } B) = P(\text{event A occurs and event B occurs})$
- So, if events **A and B are disjoint**, then (by definition)

$$P(A \text{ and } B) = 0$$
- For **non-disjoint** events, $P(A \text{ and } B)$ **is not equal to 0**
- For **independent** events

$$P(A \text{ and } B) = P(A) * P(B)$$





Exercise

- A fair coin is tossed 10 times. Which of the following two outcomes is more likely?

(a) $\underbrace{H \underline{H} \underline{H} \underline{H} \underline{H} \underline{H} \underline{H} \underline{H} \underline{H} \underline{H}}_{\text{10 times}}$ independent events ? Yes
 $\left(\frac{1}{2}\right)^{10}$

(b) THTTTTHTTT $\left(\frac{1}{2}\right)^{10}$



Exercise (HW)

- A quiz consists of 10 multiple-choice questions, each with 4 possible answers, only one of which is correct. A student makes an independent random guess to answer each of the 10 questions. What is the probability that the student gets at least one question right?

$$P(\text{correct}) = 0.25$$

$$P(\text{not correct}) = 0.75$$



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 $n_1 n_2 n_3 \dots n_k$
 - **Example:** Toss two coins. The total number of simple events is:
 $2 \times 2 = 4$
-

Sampling from a Population with Replacement



- Suppose in a lottery six numbers are drawn from the digits 0 through 9, with replacement
- (digits can be reused). How many different groupings of six numbers can be drawn?
- N is the population of 10 numbers (0 through 9) and n is the sample size, six numbers.

$$N^n = (\underline{10})^6$$

- That is, a million six-digit numbers are available!
- Here N is the population size and n is the sample size

Permutation

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

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

$$\begin{aligned} {}^4 P_3 &= 4 \times 3 \times 2 \times 1 \times 0! \\ &= 24 \end{aligned}$$

where $n! = n(n-1)(n-2)\dots(2)(1)$ 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!

123, 321, 132 ...

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

Permutation



- **Example:** A lock consists of five parts and can be assembled in any order. A quality control engineer wants to test each order for efficiency of assembly. How many orders are there?

The order of the choice is important!

$${}^5 P_5 = \frac{5!}{0!} = 5(4)(3)(2)(1) = 120$$

Combination: Sampling from a Population Without Replacement

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

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

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!

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

$$P_1 P_2 P_3 = P_2 P_1 P_3 = P_3 P_2 P_1, \dots$$

Combination



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

The order of the choice is not important!

$$^6 C_2 = \frac{6!}{2!4!} = \frac{6(5)}{2(1)} = 15$$

ways to choose 2 balls.

$$^4 C_1 = \frac{4!}{1!3!} = 4$$

ways to choose 1 red ball.

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

$$P(\text{exactly one red}) = \frac{8}{15}$$

$$^2 C_1 = \frac{2!}{1!1!} = 2$$

ways to choose 1 green ball.

Exercise



- Suppose that there were 120 students in the classroom, and that they could be classified as follows. Calculate $P(A \cup B)$.



	Brown hair	Not Brown hair
Male	20	40
Female	30	30

$$P(A \text{ and } B) / P(A \cap B) = ? = 30/120$$

A: brown hair ✓
P(A) = 50/120
B: female ✓
P(B) = 60/120

$$P(A \cup B) = ?$$

$$\begin{aligned} P(A \cup B) / P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{50}{120} + \frac{60}{120} - \frac{30}{120} = 80/120 \end{aligned}$$



Exercise

- Suppose that there were 120 students in the classroom, and that they could be classified as follows. Calculate $P(A \cup B)$.

Disjoint events

	Brown hair	Not Brown hair
Male	20	40
Female	30	30

A: male with brown hair
 $P(A) = 20/120$

B: female with brown hair
 $P(B) = 30/120$

$$P(A \cup B) = ?$$

$$P(A \cup B) = P(A) + P(B) = 50/120$$

Probability table

- Complete the following table using $P(D) = 0.95$, $P(B) = 0.40$, D and B are independent $\rightarrow P(D \text{ and } B) = 0.38$

$$\begin{aligned} P(\text{not } B) &= ? \\ &= 1 - 0.4 \\ &= 0.6 \end{aligned}$$

	B	not B	Total
D	0.38	x	0.95
not D	0.02	y	0.05
Total	0.4	0.6	1

$$\begin{aligned} P(\text{not } D) &= ? \\ &= 1 - 0.95 \\ &= 0.05 \end{aligned}$$

$$P(B \text{ and not } D) = ? = 0.4 - 0.38 = 0.02$$

$$\begin{aligned} P(D \text{ and not } B) &= x = ? = 0.95 - 0.38 \\ y &=? = 0.6 - x \quad \text{or} \quad 0.05 - 0.02 \\ &= 0.03 \end{aligned}$$



Conditional probability

- Conditional probability is denoted $P(E_1 | E_2)$
- This expression is read: the probability that E_1 will occur given that E_2 is known to have occurred.
- Conditional probabilities involve knowledge of some prior information.
- The information that is known or given is written to the right of the vertical line in the probability statement.



Example

- An example of conditional probability is the probability that a person owns a Chevrolet given that she owns a Ford.
- This conditional probability is only a measure of the proportion of Ford owners who have a Chevrolet—not the proportion of total car owners who own a Chevrolet.

Conditional Probability



If A and B are any two events in S, and $P(B) \neq 0$, the conditional probability of A given B is

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

Similarly, if A and B are any two events in S, and $P(A) \neq 0$, the conditional probability of B given A is

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

Conditional Probability



$$P(A|B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(B)P(A|B) \quad \dots (1)$$

and $P(B|A) = \frac{P(A \cap B)}{P(A)} \Rightarrow P(A \cap B) = P(A)P(B|A)$

dependent events / independent event
P(A and B) $\rightarrow P(A \cap B) = P(A)*P(B)$ $\dots (2)$

$$\therefore P(A \cap B) = P(B)P(A|B) = P(A)P(B|A) \quad \dots (3)$$



Exercise

- A manufacturer of airplane parts knows from the past experience that the probability is 0.80 that an order will be ready for shipment on time, and it is 0.72 that an order will be ready for shipment and will also be delivered on time. What is the probability that such an order will be delivered on time given that it was ready for shipment on time?

$$P(R) = 0.8$$

$$P(R \cap D) = 0.72$$

$$P(D|R) = \frac{P(R \cap D)}{P(R)} = \frac{0.72}{0.8}$$

$$= 0.9$$

Exercise: Marginal Probabilities



Consider the example on Sex wise blood group distribution

Blood group	Male	Female	Total
O	20	20	40
A	17	18	35
B	8	7	15
AB	5	5	10
Total	50	50	100

What is the probability of a person selected randomly will have blood group A? $P(A) = 35/100$



Solution

Marginal probabilities appear on the “margins” of a probability table. It is probability of single outcome

Blood group	Male	Female	Total	Row probabilities
O	20	20	40	40/100
A	17	18	35	35/100
B	8	7	15	15/100
AB	5	5	10	10/100
Total	50	50	100	1
Column probabilities	50/100	50/100	1	

Marginal, Union, Joint, and Conditional Probabilities



Marginal	Union	Joint	Conditional
$P(X)$	$P(X \cup Y)$	$P(X \cap Y)$	$P(X Y)$
The probability of X occurring	The probability of X or Y occurring	The probability of X and Y occurring	The probability of X occurring given that Y has occurred
Uses total possible outcomes in denominator	Uses total possible outcomes in denominator	Uses total possible outcomes in denominator	Uses subtotal of the possible outcomes in denominator

Example

What is the probability that a person selected has blood group B given that he is male?

Blood group	Male	Female	Total
O	20	20	40
A	17	18	35
B	8	7	15
AB	5	5	10
Total	50	50	100

Probability
to be
computed

Given
information

$$P(B | \text{Male}) = \frac{P(B \cap \text{Male})}{P(\text{Male})} = \frac{8/100}{50/100}$$

8/50 = 0.16 ???

How ???

Example

What is the probability that a person selected is male given that his blood group is B?

Blood group	Male	Female	Total
O	20	20	40
A	17	18	35
B	8	7	15
AB	5	5	10
Total	50	50	100

Given information

Probability to be computed

$$P(\text{Male} | B) = \frac{P(\text{Male} \cap B)}{P(B)}$$

$$8/15 = 0.53 ???$$

How ???



Rule of total Probability

Theorem (Rule of total probability):

If B_1, B_2, \dots, B_n are mutually exclusive events of which one must occur, and A is a common event among them then

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(B) * P(A|B)$$

↑
conditional
prob.

$$\begin{aligned} P(A) &= \sum_{i=1}^n P(B_i)P(A|B_i) \\ &= (\underline{A \cap B_1}) + (\underline{A \cap B_2}) + (\underline{A \cap B_3}) \dots \dots \end{aligned}$$



Example

- Suppose for instance, that an assembly plant receives its voltage regulators from three suppliers,
- 60% from supplier B1, 30% from supplier B2, and 10% from B3.
- If 95% of voltage regulators from B1, 80% from B2 and 65% from B3 perform according to specification,
- **what we would like to know is the probability that any one voltage regulator received by the plant will perform according to specifications**

$P(A)$ = performing according to specification

Solution

- If A denotes the event that a voltage regulator received by the plant performs according to specifications and B₁, B₂, and B₃ are the events that it comes from the respective suppliers, we can write

$$A = A \cap [B_1 \cup B_2 \cup B_3]$$

$$= (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3)$$

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3)$$

- since A ∩ B₁, A ∩ B₂, A ∩ B₃ are mutually exclusive. By using general multiplication rule we get

$$P(A) = P(B_1) \cdot P(A|B_1) + P(B_2) \cdot P(A|B_2) + P(B_3) \cdot P(A|B_3)$$

$$\begin{aligned} P(A) &= (0.60) + (0.95) + (0.30)(0.80) + (0.10)(0.65) \\ &= 0.875 \end{aligned}$$

Bayes' Theorem

- Bayes' rule is a formula that extends the use of the law of conditional probabilities to allow revision of original probabilities with new information.

$$\frac{P(A|B_i)}{P(Y)} = P(X_i|Y) = \frac{P(X_i) \cdot P(Y|X_i)}{P(X_1) \cdot P(Y|X_1) + P(X_2) \cdot P(Y|X_2) + \dots + P(X_n) \cdot P(Y|X_n)}$$

$P(A|B_i)$

$P(X_i \cap Y)$

$P(Y)$

- The numerators of Bayes' rule and the law of conditional probability are the same
- This denominator is sometimes referred to as the “total probability formula.”



Example

Example 1:

In a certain assembly plant, three machines, B1, B2, and B3 make 30%, 45%, and 25% respectively of the products. It is known from the past experience that 2%, 3% and 2% of the products made by each machine respectively, are defective.

- (i) Suppose a finished product is randomly selected, what is the probability that it is defective? $P(\text{Defective})$
- (ii) If a product chosen randomly is found defective, what is the probability that it was made by machine B3?

$$P(B3 \mid \text{defective})$$



Solution

Let, A: event that the product is defective

B1: an event that product made by machine B1

B2: an event that product made by machine B2

B3: an event that product made by machine B3

$$P(B1) = 0.3, P(B2) = 0.45, P(B3) = 0.25$$

$$P(A|B1) = 0.02, P(A|B2) = 0.03, P(A|B3) = 0.02$$

$$P(A \text{ and } B1) = P(B1) \times P(A|B1) = 0.3 \times 0.02 = 0.006$$

$$P(A \text{ and } B2) = 0.45 \times 0.03 = 0.0135$$

$$P(A \text{ and } B3) = 0.25 \times 0.02 = 0.005$$

✓ $P(A) = 0.006 + 0.0135 + 0.005 = 0.0245$ (Rule of total probability)

$$P(B3|A) = P(A \text{ and } B3)/P(A) = 0.005/0.0245 = 0.204$$

(Bayes' Theorem)



Example

Example 2:

Four technicians regularly make repairs when breakdowns occur on an automated production line. Janet, who services **20%** of the breakdowns, makes an incomplete repair 1 time in 20; Tom, who services **60%** of the breakdowns makes an incomplete repair 1 time in 10; Georgia, who services **15%** of the breakdowns, makes an incomplete repair 1 time in 10; and Peter, who services **5%** of the breakdowns, makes an incomplete repair 1 time in 20.

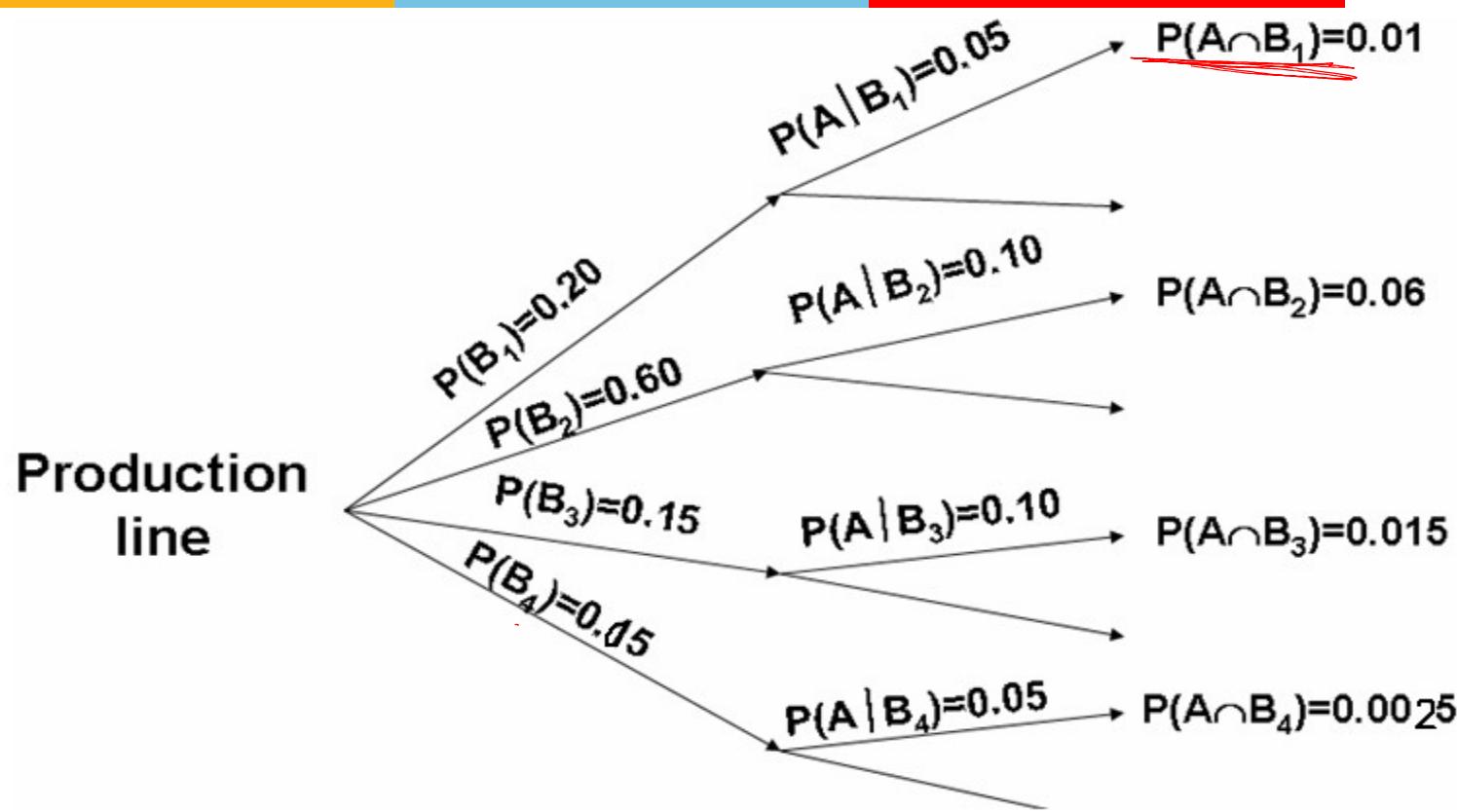
For the next problem with the production line diagnosed as being due to an initial repair that was incomplete, what is the probability that this initial repair was made by Janet?.



Solution

Let, A: be the event that the initial repair was incomplete,
B₁: an event that the initial repair was made by Janet,
B₂: an event that the initial repair was made by Tom,
B₃: an event that the initial repair was made by Georgia,
B₄: an event that the initial repair was made by Peter.

Solution



✓ $P(A) = 0.01 + 0.06 + 0.015 + 0.0025 = 0.0875$ (Rule of total probability)
 $P(B_1|A) = P(A \text{ and } B_1)/P(A) = 0.01/0.0875 = 0.1142$ (Bayes' Theorem)

Exercise

(HW)



- Using different rules of probability how can we determine if events A and B are independent?
- Hint: use conditional probability and multiplication rule



Solution

A and B are independent if

- $P(A)=P(A|B)$
- $P(B)=P(B|A)$
- $P(A|B)=P(A| \text{ not } B)$
- $P(\text{A and B})= P(\text{A}) * P(\text{B})$



Exercise (HW)

- An individual has 3 different mail accounts. Most of her messages, in fact 70% come into account #1, whereas 20% come into account #2 and the remaining 10% into account #3.
 - Of the messages into account #1, only 1% are spam whereas the corresponding for accounts # 2 and # 3 are 2% and 5% respectively.
 - What is the probability that a randomly selected message is spam?
-



References

- Probability and Statistics for Engineering and Sciences, 8th Edition, Jay L Devore, Cengage Learning
- Applied Business Statistics, Ken Black



BITS Pilani
Pilani Campus

Random Variable

Akanksha Bharadwaj
Asst. Professor, BITS Pilani



SS ZG536, ADV STAT TECHNIQUES FOR ANALYTICS

Contact Session 3

Random variable

- A random variable can be thought of as a function that associates exactly one of the possible numerical outcomes to each trial of a random experiment.
- However, that number can be the same for many of the trials.
- The set of possible values is called the **Sample Space**.
- A Random Variable is given a capital letter, such as **X** or **Z**.
- Any random variable whose only possible values are 0 and 1 is called a **Bernoulli random variable**.

True / False

Example

- Consider an experiment in which 9-volt batteries are tested until one with an acceptable voltage (S) is obtained. The sample space is $\{S, FS, FFS, FFFS, \dots\}$. Define a random variable X by
- $X = \text{the number of batteries}$ tested before the experiment terminates

$$X = ?$$

1, 2, 3, 4, 5, ..., ∞ or
num of
batteries
available

Example

- Consider the random experiment of flipping a coin twice. The sample space of possible outcomes is $S = \{ \text{HH}, \text{HT}, \text{TH}, \text{TT} \}$.
- Now, let's define the variable X to be the number of heads that the random experiment will produce.

$$X = ? \quad 0, 1, 2$$

- If the outcome is HH , we have two heads, so the value for X is 2.
- If the outcome is HT , we got one head, so the value for X is 1.
- If the outcome is TH , we again got one head, so the value for X is 1.
- Lastly, if the outcome is TT , we got zero heads, so the value for X is 0
- As the definition suggests, X is a quantitative variable that takes the possible values of 0, 1, or 2.

What is the probability that X will be 2?

$$\frac{1}{4}$$

Example

- Assume we choose a 13 year old boy at random and record his exact weight. The average weight for a 13-year-old boy is between 75 and 145 pounds, so the sample space here is $S = \{ \text{All the numbers in the interval } 75-145 \}$.
- We'll define X to be the weight of a 13 year old boy. Here X can take any value between 75 and 145.
- What is the probability that X will be more than 120?



Difference between examples

- What is the difference between the random variables in these examples?
- In the first example of coins, X has three distinct possible values: 0, 1, and 2. You can list them.
- In contrast, in the second example, X takes any value in the interval 75-145, and thus the possible values of X cover an infinite range of possibilities, and cannot be listed.

Types of random variable

- A random variable values are a list of distinct values, is called a **discrete random variable**.
- A random variable that can take any value in an interval, is called a **continuous random variable**.
- A good rule of thumb is that **discrete** random variables are things we **count**, while **continuous** random variables are things we **measure**.

Probability distribution

- For a random variable X , the list of possible values and probabilities is called the **probability distribution or probability mass function(pmf)**.
- Now, let's define the variable X to be the number of heads that the random experiment will produce with 2 coins.
- The **probability distribution of the random variable X** is easily summarized in a table:

$$\begin{array}{c} \text{HH, HT, TH, TT} \\ X = 0, 1, 2 \end{array}$$

X	0	1	2
$P(X)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Example: Two dice are tossed

- The Random Variable is $X = \text{The sum of the scores on the two dice}.$
- Let's make a table of all possible values:

		1st Die					
		1	2	3	4	5	6
2nd Die	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

- Sample Space is $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

Solution

$$P(X = 2)$$

$$\frac{1}{36}$$

$$P(X = 6)$$

$$\frac{5}{36}$$

$$P(X = 9)$$

$$\frac{4}{36}$$

$$P(X = 11)$$

$$\frac{2}{36}$$

$$P(5 \leq X \leq 8) = P(X=5) + P(X=6) + P(X=7) + P(X=8)$$

$$\frac{4}{36} + \frac{5}{36} + \frac{6}{36} + \frac{5}{36} = \frac{20}{36}$$

Example

- Consider a group of five potential blood donors— a , b , c , d , and e —of whom only a and b have type O^+ blood. Five blood samples, one from each individual, will be typed in random order until an O^+ individual is identified. Let the rv $Y =$ the number of typings necessary to identify an O^+ individual. Then the pmf of Y is

$$P(Y=1) = 2/5$$

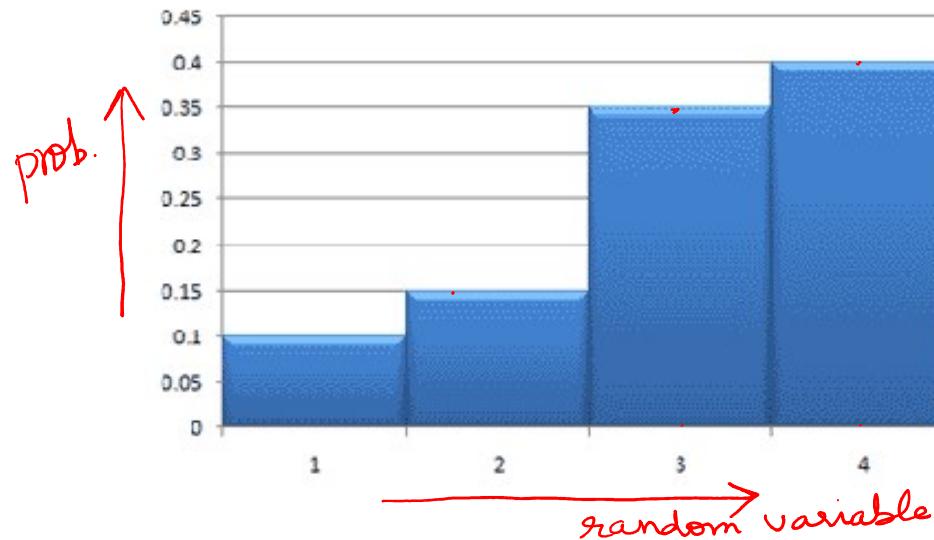
$$P(Y=2) = \frac{3}{5} * \frac{2}{4}$$

$$P(Y=4) = \frac{3C_3}{5C_3} * \frac{2C_1}{2C_1}$$

$$P(Y=3) = \frac{3C_2}{5C_2} * \frac{2C_1}{3C_1}$$

Probability distribution histogram

- The horizontal axis represents the range of all possible values of the random variable, and the vertical axis represents the probabilities of those values.



- The sum of the areas of all of the rectangles is the same as the sum of all of the probabilities.
- Therefore, the total area = 1.

Image: google

Valid Probability Model

- Alex is playing cricket. Various possible scenarios for possibility of catch on the next two balls are given below. Is it a valid model?

—
—

?

Scenarios	Probability
Miss both the catch	0.3
Miss one catch	0.4
Miss none	0.2

No
here \sum probabilities $\neq 1$

Exercise

- Six lots of components are ready to be shipped by a certain supplier. The number of defective components in each lot is as follows:

<i>Lot</i>	1	2	3	4	5	6
<i>Number of defectives</i>	0	2	0	1	2	0

- One of these lots is to be randomly selected for shipment to a particular customer. Let X be the number of defectives in the selected lot. What are the possible values of X and $P(X)$

Solution

$$P(X=0) = 3/6$$

$$P(X=1) = 1/6$$

$$P(X=2) = 2/6$$

Cumulative distribution function

- The cumulative distribution function (cdf) $F(x)$ of a discrete rv variable X with pmf $p(x)$ is defined for every number x by

$$F(x) = P(X \leq x) = \sum_{y:y \leq x} p(y)$$

eg:- HH, HT, TH, TH
 $X = 0, 1, 2$

$$F(1) = P(X \leq 1) = P(X=0) + P(X=1)$$

Example

- A store carries flash drives with either 1 GB, 2 GB, 4 GB, 8 GB, or 16 GB of memory. The accompanying table gives the distribution of Y = the amount of memory in a purchased drive:

y	1	2	4	8	16
$p(y)$.05	.10	.35	.40	.10

Solution

y	1	2	4	8	16
$p(y)$.05	.10	.35	.40	.10

$$F(1) = 0.05$$

$$F(2) = 0.05 + 0.10 = 0.15$$

$$F(4) = P(Y \leq 4) = P(Y = 1 \text{ or } 2 \text{ or } 4) = p(1) + p(2) + p(4) = \underline{.50}$$

$$F(8) = P(Y \leq 8) = p(1) + p(2) + p(4) + p(8) = .90$$

$$F(16) = P(Y \leq 16) = 1$$

Expected Value of X

- Let X be a discrete rv with set of possible values D and pmf $p(x)$. The **expected value** or **mean value** of X , denoted by $E(X)$ or μ_X or just μ , is

$$E(X) = \mu_X = \sum_{x \in D} (x \cdot p(x))$$

eg:-

X	0	1	2
$p(x)$	y_1	y_2	y_4

$$\begin{aligned}
 E(X) &= 0 \times y_1 + 1 \times y_2 + 2 \times y_4 \\
 &= 1
 \end{aligned}$$

Exercise

- Just after birth, each newborn child is rated on a scale called the Apgar scale. The possible ratings are 0, 1, . . . , 10, with the child's rating determined by color, muscle tone, respiratory effort, heartbeat, and reflex irritability (the best possible score is 10). Let X be the Apgar score of a randomly selected child born at a certain hospital during the next year, and suppose that the pmf of X is

x	$p(x)$
0	.002
1	.001
2	.002
3	.005
4	.02
5	.04
6	.18
7	.37
8	.25
9	.12
10	.01

- What is the expected value of X ?

$$\begin{aligned} \sum x \cdot p(x) \\ = 7.15 \end{aligned}$$

Exercise

- Pizza point delivers only one kind of pizza, which is sold for Rs150, and costs the pizza point Rs50 to make. The pizza point has the following policy regarding delivery: if the pizza takes longer than half an hour to arrive, there is no charge. Let the random variable X be the pizza point's gain for any one pizza.
- Experience has shown that delivery takes longer than half an hour only 10 percent of the time. Find the mean gain per pizza, μ_X .

Solution

- We first need to establish its probability distribution—the possible values and their probabilities.
- The random variable X has two possible values: either the pizza costs them Rs50 to make and they sell it for Rs150, in which case X takes the value $150 - 50 = \text{Rs}100$, or it costs them Rs50 to make and they give it away, in which case X takes the value $0-50 = -\text{Rs}50$.
- The probability of the latter case is given to be 10 percent, or .1, so using complements, the former has probability .9. Here, then is the probability distribution of X:

X	+100	-50
P(X=x)	.9	.1

- So, $\mu_X = (100)(.9) + (-50)(.1) = +85$
- In the long run, the pizza point gains an average of Rs85 per pizza delivered.

Expected Value of a Function

- If the rv X has a set of possible values D and pmf $p(x)$, then the expected value of any function $h(X)$,

$$E[h(X)] = \sum_D h(\underline{x}) \cdot p(x)$$

Exercise

- The cost of a certain vehicle diagnostic test depends on the number of cylinders X in the vehicle's engine. Suppose the cost function is given by $\underline{h(X)=20+3X+.5X^2}$. Since, X is a random variable, so is $Y=h(X)$. The pmf of X is as follows:

$\checkmark X$	4	6	8
$\checkmark p(x)$.5	.3	.2

$$\begin{aligned}
 h(4) &= 20 + 3 \times 4 + 0.5(4)^2 = 40 \\
 h(6) &= 20 + 3 \times 6 + 0.5(6)^2 = 56 \\
 h(8) &= 20 + 3 \times 8 + 0.5(8)^2 = 76
 \end{aligned}$$

- Calculate $E(h(X))$

$$\begin{aligned}
 E(h(X)) &= 40 \times 0.5 + 56 \times 0.3 + 76 \times 0.2 \\
 &= 20 + 16.8 + 15.2 \\
 &= 52
 \end{aligned}$$

Rules of Expected Value

Two special cases of the proposition yield two important rules of expected value.

1. For any constant a , $E(a\bar{X}) = a^*E(\bar{X})$
2. For any constant b , $E(\bar{X}+b) = E(\bar{X})+b$

-

Variance and standard deviation of X

Let X have pmf $p(x)$ and expected value μ . Then the variance of X , denoted by $V(X)$ or σ_X^2 , or just σ^2 , is

$$V(X) = \sum_D (x - \underline{\mu})^2 \cdot \underline{p(x)} = E[(X - \mu)^2]$$

The standard deviation (SD) of X is

$$\sigma_X = \sqrt{\underline{\sigma_X^2}}$$

Exercise

- A library has an upper limit of 6 on the number of videos that can be checked out to an individual at one time. Consider only those who check out videos, and let X denote the number of videos checked out to a randomly selected individual. The pmf of X is as follows:

x	1	2	3	4	5	6
$p(x)$.30	.25	.15	.05	.10	.15

$$E(X) = 1 \times 0.3 + 2 \times 0.25 + 3 \times 0.15 + 4 \times 0.05 + 5 \times 0.10 + 6 \times 0.15 = 2.85$$

- Calculate variance and standard deviation

$$\begin{aligned} V(X) &= (1 - 2.85)^2 \times 0.3 + (2 - 2.85)^2 \times 0.25 + (3 - 2.85)^2 \times 0.15 + (4 - 2.85)^2 \\ &\quad \times 0.05 + (5 - 2.85)^2 \times 0.10 + (6 - 2.85)^2 \times 0.15 = 3.2275 \end{aligned}$$

$$\sigma_X = \sqrt{3.2275} = 1.8$$

Continuous Variables

A random variable X is continuous if

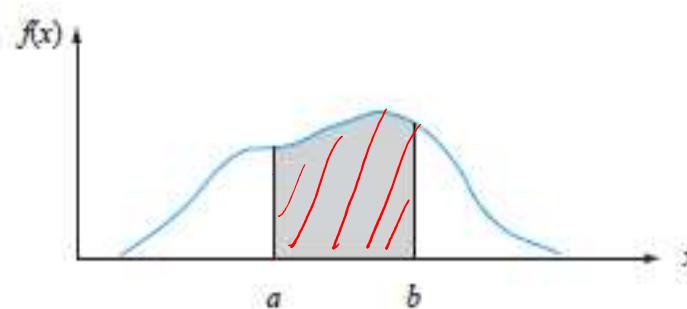
- Possible values comprise either a single interval on the number line or a union of disjoint intervals, and
- $P(X=c) = 0$ for any number c that is a possible value of X .

PDF of continuous variable

- Let X be a continuous rv. Then a **probability distribution** or **probability density function** (pdf) of X is a function $f(x)$ such that for any two numbers a and b with $a \leq b$,

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

- That is, the probability that X takes on a value in the interval $[a, b]$ is the area above this interval and under the graph of the density function, as illustrated



$P(a \leq X \leq b) =$ the area under the density curve between a and b .

Legitimate pdf

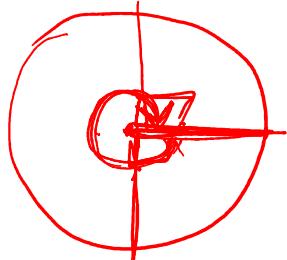
For $f(x)$ to be a legitimate pdf, it must satisfy the following two conditions:

1. $\underline{f(x)} \geq 0$ for all x

2. $\int_{-\infty}^{\infty} f(x) dx = \text{area under the entire graph of } f(x)$
 $= 1$

Example

- Consider the reference line connecting the valve stem on a tire to the center point, and let X be the angle measured clockwise to the location of an imperfection. One possible pdf for X is



$$f(x) = \begin{cases} \frac{1}{360} & 0 \leq x < 360 \\ 0 & \text{otherwise} \end{cases}$$

- The probability that the angle is between 90 degree and 180 degree is ?

$$P(90 \leq X \leq 180) = \int_{90}^{180} \frac{1}{360} dx = \frac{x}{360} \Big|_{x=90}^{x=180} = \frac{1}{4} = .25$$

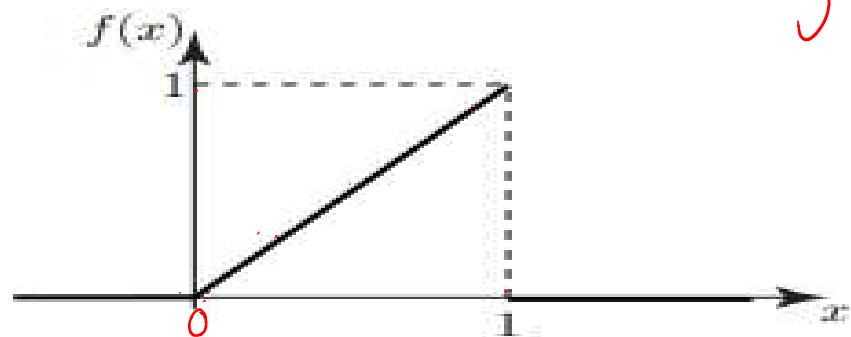
- The probability that the angle of occurrence is within 90 degree of the reference line is ?

$$0 \leq x \leq 90 \quad 270 \leq x \leq 360$$

$$\int_{270}^{360} \frac{1}{360} dx + \int_0^{90} \frac{1}{360} dx = \left[\frac{x}{360} \right]_0^{90} + \left[\frac{x}{360} \right]_{270}^{360} = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \text{ or } 0.5$$

Exercise

Ques. Is this a valid pdf?



$$f(x) = x \quad 0 \leq x \leq 1 \\ = 0 \quad \text{otherwise}$$

Solution

$f(x) = x$ for $0 \leq x \leq 1$ and 0 elsewhere

$f(x) \geq 0$ for all x

But,

$$\int_0^1 f(x) dx = \int_0^1 x dx = \left[\frac{x^2}{2} \right]_0^1 = \frac{1}{2}$$

Not a valid pdf

Exercise

$$\int x^n dx = \frac{x^{n+1}}{n+1}$$

Is this a valid pdf?

$$\int c dx = cx$$

$$f(x) = \begin{cases} x^2 - 4x + \frac{10}{3}, & 0 \leq x \leq 3 \\ 0, & \text{elsewhere} \end{cases}$$

$$f(0) = 10/3$$

$$f(1) = 1 - 4 + \frac{10}{3} = -\frac{9+10}{3} = \frac{1}{3}$$

$$f(2) = 4 - 8 + \frac{10}{3} = -\frac{12+10}{3} = -\frac{2}{3}$$

$$f(3) = 9 - 12 + \frac{10}{3} = \frac{1}{3}$$

Solution

$f(x)$ is not ≥ 0 for all values of x

It is < 0 for $x=2$ and ~~$x=3$~~

$$\begin{aligned}
 & \int_0^3 \left(x^2 - 4x + \frac{10}{3} \right) dx \\
 &= \left[\frac{x^3}{3} - \frac{4x^2}{2} + \frac{10}{3}x \right]_0^3 \\
 &= \frac{9 \times 3^2}{3} - \frac{4 \times 9}{2} + \frac{10}{3} \times 3 = 1
 \end{aligned}$$

Cumulative distribution function

The **cumulative distribution function** $F(x)$ for a continuous rv X is defined for every number x by

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(y) dy$$

For each x , $F(x)$ is the area under the density curve to the left of x . This is illustrated in Figure, where $F(x)$ increases smoothly as x increases.

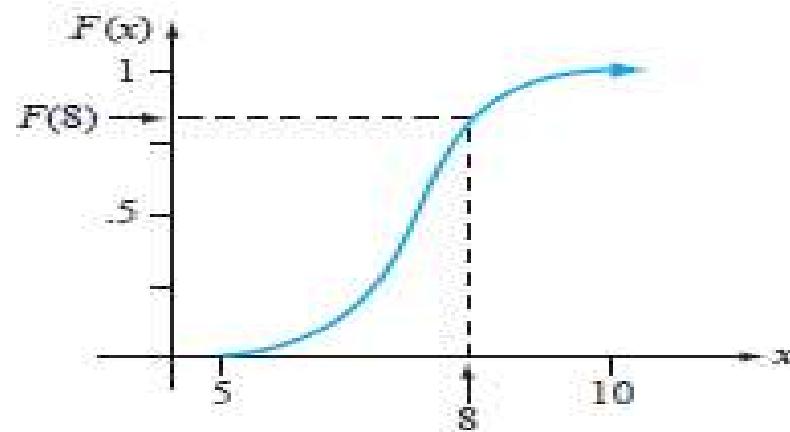
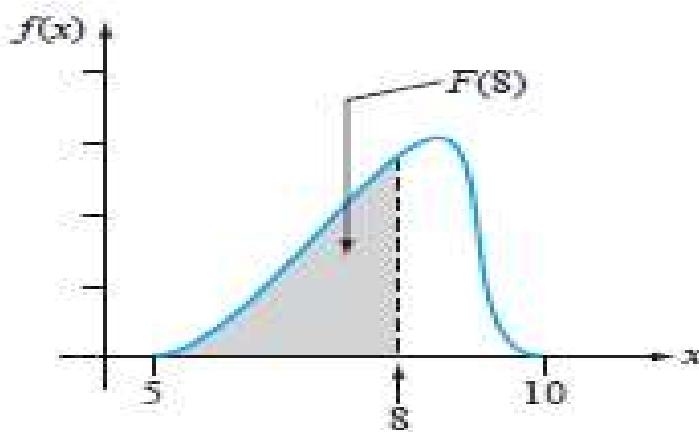


Figure 4.5 A pdf and associated cdf

Using $F(x)$ to Compute Probabilities

Let X be a continuous rv with pdf $f(x)$ and cdf $F(x)$. Then for any number a ,

$$P(X > a) = 1 - F(a)$$

and for any two numbers a and b with $a < b$,

$$P(a \leq X \leq b) = F(b) - F(a)$$

Example

- Suppose the pdf of the magnitude X of a dynamic load on a bridge (in newtons) is given by

$$\underline{f(x)} = \begin{cases} \frac{1}{8} + \frac{3}{8}x & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

- For any number x between 0 and 2, $F(X)$ is

$$\begin{aligned}
 F(x) &= \int_{-\infty}^x f(x) dx \\
 &= \int_0^2 \left(\frac{1}{8} + \frac{3}{8}x \right) dx = \left[\frac{x}{8} + \frac{3}{8} \times \frac{x^2}{2} \right]_0^2 \\
 &= \left(\frac{2}{8} + \frac{3}{8} \times \frac{4}{2} \right) = 1
 \end{aligned}$$

Example

The probability that the load is between 1 and 1.5 is

$$\begin{aligned}
 P(1 \leq X \leq 1.5) &= \int_1^{1.5} f(x) dx \\
 &= \underbrace{F(1.5)} - \underbrace{F(1)} \\
 &= \left[\frac{1}{8}x(1.5) + \frac{3}{16}(1.5)^2 \right] - \left[\frac{1}{8}x(1) + \frac{3}{16}(1)^2 \right] \\
 &= 19/64
 \end{aligned}$$

The probability that the load exceeds 1 is

$$\begin{aligned}
 P(X > 1) &= 1 - F(1) \\
 &= 1 - \left[\frac{1}{8} + \frac{3}{16}(1)^2 \right] \\
 &= 1 - \frac{5}{16} = \frac{11}{16}
 \end{aligned}$$



Expected Values

$p(x) = pmf$

$f(x) = pdf$

$F(x) = \text{Cumulative prob. dis.}$

The **expected or mean value** of a continuous rv X with pdf $f(x)$ is

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

—

Example

- The pdf of weekly gravel sales X was

$$f(x) = \begin{cases} \frac{3}{2}(1 - x^2) & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- Calculate mean value of X

$$\int_{-\infty}^{\infty} x f(x) dx$$

$$\begin{aligned}
 &= \frac{3}{2} \int_0^1 x (1-x^2) dx = \frac{3}{2} \left[\frac{x^2}{2} - \frac{x^4}{4} \right]_0^1 \\
 &= \frac{3}{2} \left[\frac{1}{2} - \frac{1}{4} \right] = 3/8
 \end{aligned}$$

Variance and standard deviation

The **variance** of a continuous random variable X with pdf $f(x)$ and mean value μ is

$$\sigma_X^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx = E[(X - \mu)^2]$$

$$V(X) = E(X^2) - [E(X)]^2$$

The **standard deviation** (SD) of X is

$$\sigma_X = \sqrt{V(X)}.$$

Example

① A random variable X has the following probability function

(i) find the value of K

(ii) Mean

(iii) Variance

(iv) $P(X > 3)$ (v) $P(1 < X \leq 5)$

x	1	2	3	4	5	6
$P(x)$	K	$3K$	$5K$	$7K$	$9K$	$11K$

Solution

(i) Work by defn of discrete random variable

$$\Rightarrow \sum_{i=1}^n p(x_i) = 1$$

$$\Rightarrow p(x=1) + p(x=2) + \dots + p(x=6) = 1,$$

$$\therefore k + 3k + 5k + 7k + 9k + 11k = 36k = 1$$

$$k = \frac{1}{36}$$

$$E(x) = \sum x p(x) = 1 \cdot \frac{1}{36} + 2 \cdot \frac{3}{36} + 3 \cdot \frac{5}{36} + 4 \cdot \frac{7}{36}$$

$$+ 5 \cdot \frac{9}{36} + 6 \cdot \frac{11}{36} = 4.47$$

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

$$\Rightarrow \sigma^2 = E(x^2) - [E(x)]^2$$

$$\Rightarrow \sigma^2 = \sum x^2 p(x) - [E(x)]^2$$

$$\sigma^2 = \frac{1}{36} + 4 \cdot \frac{3}{36} + 9 \cdot \frac{5}{36} + 16 \cdot \frac{7}{36} + 25 \cdot \frac{9}{36} + 36 \cdot \left(\frac{11}{36}\right) - (4.47)^2$$

$$\sigma^2 = \frac{791}{36} - (4.47)^2 = 1.99$$

$$\therefore \sigma^2 = 1.99$$

(iv) $P(X \geq 3)$

$$P(X \geq 3) = P(X=3) + P(X=4) + P(X=5) \\ + P(X=6)$$

$$= \frac{5}{36} + \frac{7}{36} + \frac{9}{36} + \frac{11}{36} = \frac{32}{36} = \frac{8}{9}$$

$$P(X \geq 3) = \frac{8}{9}$$

(v) $P(1 < X \leq 5)$ = $P(X=2) + P(X=3)$

$$+ P(X=4) + P(X=5)$$

$$= \frac{3}{36} + \frac{5}{36} + \frac{7}{36} + \frac{9}{36} = \frac{24}{36} = \frac{2}{3}$$

③ If p.d.f $f(x) = kx^3$ in $1 \leq x \leq 3$
elsewhere.

Find the value of k and find the

probability between $x = \frac{1}{2}$ and $x = \frac{3}{2}$

$$\text{Sol: } - \int_{-\infty}^{\infty} f(x) dx = 1$$

$$\Rightarrow \int_{-\infty}^{-1} f(x) dx + \int_{-1}^3 f(x) dx + \int_3^{\infty} f(x) dx = 1$$

$$\Rightarrow 0 + \int_{-1}^3 kx^3 dx + 0 = 1$$

$$\Rightarrow k \left[\frac{x^4}{4} \right]_{-1}^3 = 1 \Rightarrow \frac{k}{4} [3^4 - 1] = 1$$

$$\Rightarrow \frac{k}{4} [81 - 1] = 1 \Rightarrow \frac{k}{4} (80) = 1$$

$$\Rightarrow k = \frac{4}{80} \Rightarrow \boxed{k = \frac{1}{20}}$$

$$P\left(\frac{1}{2} \leq x \leq \frac{3}{2}\right) = \int_{\frac{1}{2}}^{\frac{3}{2}} f(x) dx = \int_{\frac{1}{2}}^{\frac{3}{2}} kx^3 dx$$

Homework

For the variable X with pdf, find E(X) and V(X)

$$f(x) = \begin{cases} \frac{1}{2}x, & 0 \leq x \leq 2 \\ 0, & \text{elsewhere} \end{cases}$$

Solution

$$\mathbb{E}(X) = \int_0^2 \frac{1}{2}x.x dx = \left[\frac{1}{6}x^3 \right]_0^2 = \frac{8}{6} = \frac{4}{3}.$$

$$\mathbb{E}(X^2) = \int_0^2 \frac{1}{2}x.x^2 dx = \left[\frac{1}{8}x^4 \right]_0^2 = 2.$$

$$\begin{aligned}\mathbb{V}(X) &= \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2 \\ &= 2 - \frac{16}{9} = \frac{2}{9}.\end{aligned}$$



References

- Probability and Statistics for Engineering and Sciences, 8th Edition, Jay L Devore, Cengage Learning
- Applied Business Statistics, Ken Black



BITS Pilani
Pilani Campus

Probability Distribution

Akanksha Bharadwaj
Asst. Professor, BITS Pilani



SS ZG536, ADV STAT TECHNIQUES FOR ANALYTICS

Contact Session 4

Binomial Random Variable

- A binary random variable
- e.g., head or tail in each toss of a coin; defective or not defective light bulb
- Generally called “success” and “failure”
- Probability of success is p, probability of failure is $1 - p$

Binomial Probability Distribution

An experiment for which Conditions 1–4 are satisfied is called a **binomial experiment**.

1. The experiment consists of a sequence of n smaller experiments called *trials*, where n is fixed in advance of the experiment.
2. Each trial can result in one of the same two possible outcomes, *Head or Tails*
3. The trials are **independent**, so that the outcome on any particular trial does not influence the outcome on any other trial.
4. The probability of success $P(S)$ is constant from trial to trial; we denote this probability by p .

$$P(\text{Heads}) = p$$

↑
success

Example

- The same coin is tossed successively and independently n times. We arbitrarily use S to denote the outcome H (heads) and F to denote the outcome T (tails)

$n = 3$

possible outcomes ?

$$2^3 = 8$$

HHH

HHT

HTH

HTT

THH

THT

TTH

TTT

if $n = 5$

then
possible
cases

$$\textcircled{b} 2^5$$

Exercise

- Suppose a certain city has 50 licensed restaurants, of which 15 currently have at least one serious health code violation and the other 35 have no serious violations. There are five inspectors, each of whom will inspect one restaurant during the coming week. The name of each restaurant is written on a different slip of paper, and after the slips are thoroughly mixed, each inspector in turn draws one of the slips without replacement.

- Is it a binomial experiment?**

HCV → possible outcomes
no HCV



NO trials are not independent

◦ ; w/o replacement

Solution

$$P(S \text{ on first trial}) = \frac{35}{50} = .70$$

and

$$\begin{aligned} P(S \text{ on second trial}) &= P(SS) + P(FS) \\ &= P(\text{second } S \mid \text{first } S) P(\text{first } S) \\ &\quad + P(\text{second } S \mid \text{first } F) P(\text{first } F) \\ &= \frac{34}{49} \cdot \frac{35}{50} + \frac{35}{49} \cdot \frac{15}{50} = \frac{35}{50} \left(\frac{34}{49} + \frac{15}{49} \right) = \frac{35}{50} = .70 \end{aligned}$$

Similarly, it can be shown that $P(S \text{ on } t\text{th trial}) = .70$ for $t = 3, 4, 5$. However,

$$P(S \text{ on fifth trial} \mid SSSS) = \frac{31}{46} = .67$$

whereas

$$P(S \text{ on fifth trial} \mid FFFF) = \frac{35}{46} = .76$$

Exercise

- A certain state has 500,000 licensed drivers, of whom 400,000 are insured. A sample of 10 drivers is chosen without replacement. The i th trial is labeled S if the i th driver chosen is insured.
- Although this situation would seem identical to that of previous example, the important difference is that the size of the population being sampled is very large relative to the sample size.
- ***Is it a binomial experiment?***

Yes

insured
not insured

Let, 1st trial be success

$$P(S_{\text{on } 2^{\text{nd}}} \mid S_{\text{on } 1^{\text{st}}}) = \frac{3,99,999}{4,99,999} \approx 0.8$$



$$\vdots$$
$$P(S_{\text{on } 10^{\text{th}}} \mid S_{\text{on previous}} \text{ all trials}) = \frac{3,99,991}{4,99,991} = 0.7999 \\ \approx 0.8$$

$$P(\text{Success}) \text{ or } P(\text{ensured}) = 0.8$$

i.e. constant from trial to trial

- These calculations suggest that although the trials are not exactly independent, the conditional probabilities differ so slightly from one another that for practical purposes the trials can be regarded as independent with constant probability.
- Thus, to a very good approximation, the experiment is binomial with $n=10$ and $p=.8$

Example

- Suppose, for example, that $n=3$. Then there are eight possible outcomes for the experiment:

SSS, SSF, SFS, SFF, FSS, FSF, FFS, FFF

\uparrow *random variable (no. of success in each trial)*
 $X = \{0, 1, 2, 3\}$

- From the definition of X , $X(\text{SSF})=2$, $X(\text{SFF})=1$, and so on. Possible values for X in an n -trial experiment are $x=0, 1, 2, 3, \dots, n$.
- Because the pmf of a binomial rv X depends on the two parameters n and p , we denote the pmf by $b(x; n, p)$.

$$P(S) = p \quad P(F) = (1-p)$$

Table 3.1 Outcomes and Probabilities for a Binomial Experiment with Four Trials

Outcome	x	Probability	Outcome	x	Probability
SSSS	4	p^4	FSSS	3	$p^3(1 - p)$
SSSF	3	$p^3(1 - p)$	FSSF	2	$p^2(1 - p)^2$
SSFS	3	$p^3(1 - p)$	FSFS	2	$p^2(1 - p)^2$
SSFF	2	$p^2(1 - p)^2$	FSFF	1	$p(1 - p)^3$
SFSS	3	$p^3(1 - p)$	FFSS	2	$p^2(1 - p)^2$
SFSF	2	$p^2(1 - p)^2$	FFSF	1	$p(1 - p)^3$
SFFS	2	$p^2(1 - p)^2$	FFFS	1	$p(1 - p)^3$
SFFF	1	$p(1 - p)^3$	FFFF	0	$(1 - p)^4$

$$\begin{aligned}
 b(3; 4, p) &= P(FSSS) + P(SFSS) + P(SSFS) + P(SSSF) \\
 &= 4p^3(1 - p)
 \end{aligned}$$

Theorem

- Since the ordering of S 's and F 's is not important, the second factor in the previous equation is $p^x * (1-p)^{n-x}$ (e.g., the first x trials resulting in S and the last resulting in F).
- The first factor is the number of ways of choosing x of the n trials to be S 's—that is, the number of combinations of size x that can be constructed from n distinct objects (trials here).

$${}^n C_x \ p^x (1-p)^{n-x}$$

$$b(x; n, p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, 2, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

Example

- Each of six randomly selected cola drinkers is given a glass containing cola S and one containing cola F . The glasses are identical in appearance except for a code on the bottom to identify the cola. Suppose there is actually no tendency among cola drinkers to prefer one cola to the other. Then $p = P(\text{a selected individual prefers } S) = 0.5$. So, with $X = \text{the number among the six who prefer } S$

$$P(X=3) = ?$$

$$X = \{0, 1, 2, 3, \dots, 6\}$$

$$P = 0.5, n = 6$$

$${}^n C_x p^x (1-p)^{n-x}$$

$$\rightarrow {}^6 C_3 (0.5)^3 (1-0.5)^{6-3} = 20 \times (0.5)^6 = 0.3125$$

Exercise

- Suppose that 20% of all copies of a particular textbook fail a certain binding strength test. Let X denote the number among 15 randomly selected copies that fail the test. Then X has a binomial distribution with $n=15$ and $p=.2$
- What is the probability that at most 8 fail the test?

$$P(X \leq 8) = \sum_{x=0}^8 b(x, 15, 0.2)$$

$$= B(8, 15, 0.2)$$

cumulative prob.

from Binomial table got value

$$= \underline{\underline{0.999}}$$

Solution

- What is the probability that exactly 8 fail?

$$\begin{aligned}
 \underline{P(X=8)} &= P(X \leq 8) - P(X \leq 7) \\
 &= B(8, 15, 0.2) - B(7, 15, 0.2) \\
 &= 0.999 - 0.996 = 0.003
 \end{aligned}$$

- What is the probability that at least 8 fail?

$$\begin{aligned}
 P(X \geq 8) &= ? = 1 - P(X \leq 7) \\
 &= 1 - B(7, 15, 0.2) = 1 - 0.996 \\
 &= 0.004
 \end{aligned}$$

- What is the probability that fail is between 4 and 7 (inclusive)

$$\begin{aligned}
 P(X \leq 7) - P(X \leq 3) \\
 &= B(7, 15, 0.2) - B(3, 15, 0.2) \\
 &= 0.996 - 0.648 \\
 &= 0.348
 \end{aligned}$$

Definitions: Bernouilli

- **Bernouilli trial:** If there is only 1 trial with probability of success p and probability of failure $1-p$, this is called a Bernouilli distribution. (special case of the binomial with $n=1$)
- Probability of success:
$$P(X = 1) = \binom{1}{1} p^1 (1-p)^{1-1} = p$$
- Probability of failure:
$$P(X = 0) = \binom{1}{0} p^0 (1-p)^{1-0} = 1 - p$$



Characteristics of Bernouilli distribution

For Bernouilli ($n=1$)

$$\underline{E(X) = p}$$

$$\underline{\text{Var}(X) = p(1-p)}$$

Expected value and variance of Binomial Distribution

If X follows a binomial distribution with parameters n and p :

$$X \sim \text{Bin}(n, p)$$

Then:

$$\mu_x = E(X) = np$$

$$\sigma_x^2 = \text{Var}(X) = np(1-p) = npq \quad (\text{here } q=1-p)$$

$$\sigma_x = \text{SD}(X) = \sqrt{np(1-p)}$$

Variance Proof (optional!)

For $Y \sim \text{Bernoulli}(p)$

$$\left\{ \begin{array}{l} Y=1 \text{ if yes} \\ Y=0 \text{ if no} \end{array} \right.$$

$$\begin{aligned} Var(Y) &= E(Y^2) - E(Y)^2 \\ &= [1^2 p + 0^2 (1-p)] - [1p + 0(1-p)]^2 \\ &= p - p^2 \\ &= p(1-p) \end{aligned}$$

For $X \sim \text{Bin}(N, p)$

$$X = \sum_{i=1}^n Y_{\text{Bernoulli}}; Var(Y) = p(1-p)$$

$$= Var(X) = Var(\sum_{i=1}^n Y) = \sum_{i=1}^n Var(Y) = np(1-p)$$

Poisson distribution

Poisson distribution is for counts—if events happen at a constant rate over time, the Poisson distribution gives the probability of X number of events occurring in time T .

Poisson Mean and Variance

Mean

$$\mu = \lambda$$

For a Poisson random variable, the variance and mean are the same!

- Variance and Standard Deviation

$$\sigma^2 = \lambda$$

$$\sigma = \sqrt{\lambda}$$

where λ = expected number of hits in a given time period

Poisson Distribution, example

- The Poisson distribution models counts, such as the number of new cases of COVID that occur in women in Bangalore next month.
- The distribution tells you the probability of all possible numbers of new cases, from 0 to infinity.
- If $\underline{X} = \# \text{ of new cases next month}$ and $X \sim \underline{\text{Poisson}}(\lambda)$, then the probability that $\underline{X=k}$ (a particular count) is:

$$p(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Example: Poisson distribution

- Suppose that a rare disease has an incidence of 1 in 1000 person-years. Assuming that members of the population are affected independently, find the probability of k cases in a population of 10,000 (followed over 1 year) for $k=0,1,2$.
- The expected value (mean) = $\lambda = .001 * 10,000 = 10$
- 10 new cases expected in this population per year →

X $e^{-\lambda}$
 $\lambda = 10$

$$P(X = 0) = \frac{(10)^0 e^{-(10)}}{0!} = .0000454$$

$$P(X = 1) = \frac{(10)^1 e^{-(10)}}{1!} = .000454$$

$$P(X = 2) = \frac{(10)^2 e^{-(10)}}{2!} = .00227$$

more on Poisson...

- “Poisson Process” (rates)
- Note that the Poisson parameter λ can be given as the mean number of events that occur in a defined time period OR, equivalently, λ can be given as a rate, such as $\lambda=2/\text{month}$ (2 events per 1 month) that must be multiplied by $t=\text{time}$ (called a “Poisson Process”) →
- $X \sim \text{Poisson } (\lambda)$

$$P(X = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

$$\mathbb{E}(X) = \lambda t$$

$$\text{Var}(X) = \lambda t$$

Practice problems

1a. If the calls received on your mobile phone follow the Poisson distribution with a constant rate $\lambda=4$ calls per hour, what's the probability that, if you forget to turn your phone off in a 1.5 hour theater play, your phone rings during that time?

$$\lambda = 4 \text{ calls/hr}$$

$$\lambda t = 4 * 1.5 = 6$$

$$P(X \geq 1) = 1 - P(0) = 1 - \left[\frac{(6)^0 e^{-6}}{0!} \right] \approx 0.9$$

1b. How many ^{avg.}_^ phone calls do you expect to get during the play?

$$E(X) = \lambda t = 4 \times 1.5 = 6$$

Poisson distribution as limit

- In any binomial experiment in which n is large and p is small, $b(x;n,p)$ is approximately equal to $p(x;\mu)$, where $\mu = \underline{np}$.
- As a rule of thumb, this approximation can safely be applied if $n > 50$ and $\underline{np < 5}$.

Example

- If a publisher of nontechnical books takes great pains to ensure that its books are free of typographical errors, so that the probability of any given page containing at least one such error is .005 and errors are independent from page to page, what is the probability that one of its 400-page novels will contain exactly one page with errors? At most three pages with errors?

Solution

- With S denoting a page containing at least one error and F an error-free page, the number X of pages containing at least one error is a binomial rv with $n=400$ and $p=.005$, so $np=2$.

Solution

Binomial distribution Yes

$$n = 400 \quad p = 0.005$$

$$np = ? \quad 2$$

$\therefore n > 50$ & $np < 5$ we can apply poison dis.

$$P(X=1) \text{ Then } \mu = \lambda = np = 2 \cdot \frac{2^x e^{-2}}{x!} = 0.27$$

$$P(X \leq 3) = \sum_{x=0}^{3} \frac{e^{-2} \times 2^x}{x!} = 0.857$$

Table 3.2 Comparing the Poisson and Three Binomial Distributions

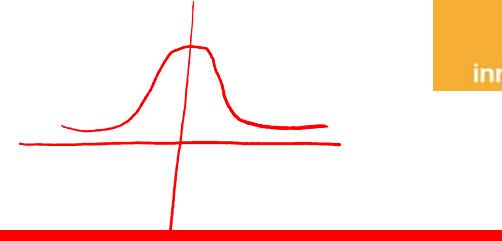
x	$n = 30, p = .1$	$n = 100, p = .03$	$n = 300, p = .01$	Poisson, $\mu = 3$
0	0.042391	0.047553	0.049041	0.049787
1	0.141304	0.147070	0.148609	0.149361
2	0.227656	0.225153	0.224414	0.224042
3	0.236088	0.227474	0.225170	0.224042
4	0.177066	0.170606	0.168877	0.168031
5	0.102305	0.101308	0.100985	0.100819
6	0.047363	0.049610	0.050153	0.050409
7	0.018043	0.020604	0.021277	0.021604
8	0.005764	0.007408	0.007871	0.008102
9	0.001565	0.002342	0.002580	0.002701
10	0.000365	0.000659	0.000758	0.000810

The Mean and Variance of X

- Since as $b(x;n,p) \rightarrow p(x;\mu)$ as $n \rightarrow \infty$, $p \rightarrow 0$, $np \rightarrow \mu$, the mean and variance of a binomial variable should approach those of a Poisson variable.
- These limits are $np \rightarrow \mu$ and $np(1 - p) \rightarrow \mu$

If X has a Poisson distribution with parameter μ , then $E(X) = V(X) = \mu$.

Normal distribution



- symmetric bell shape
- mean and median are equal; both located at the center of the distribution

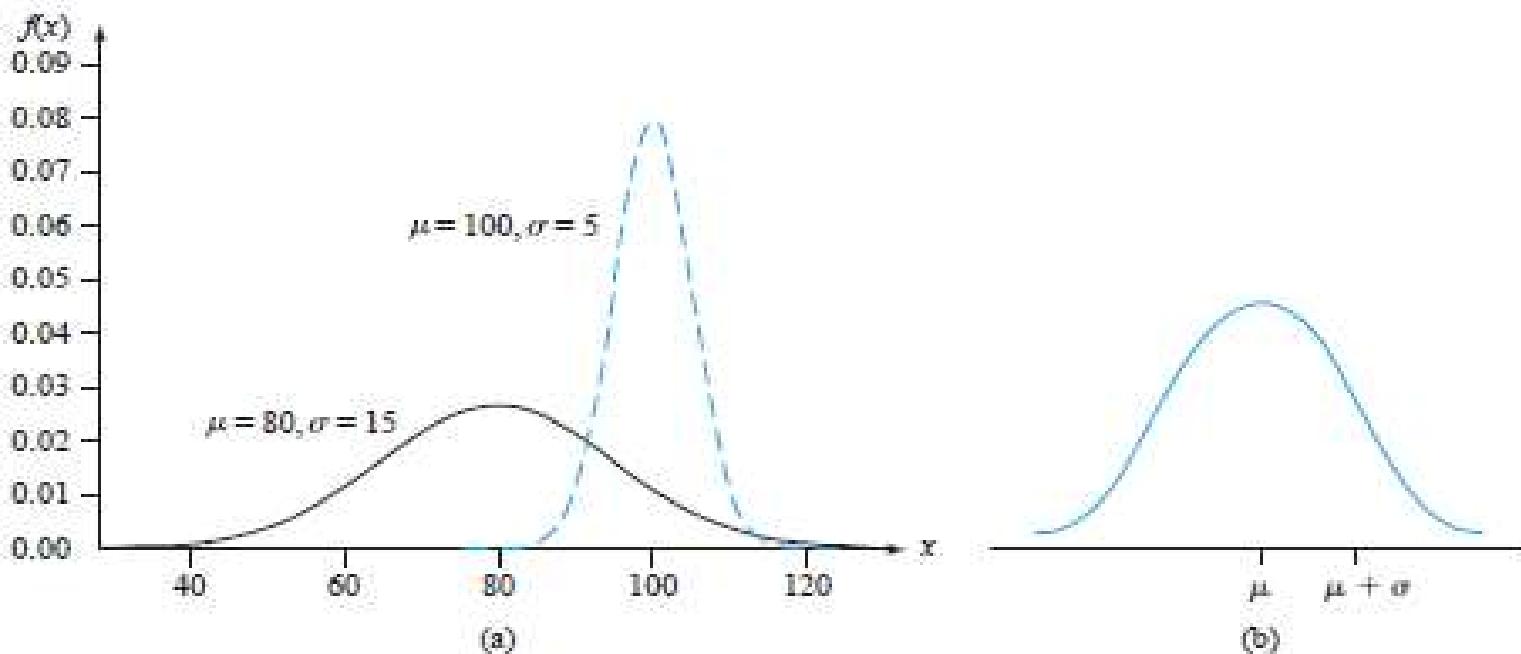


Figure 4.13 (a) Two different normal density curves (b) Visualizing μ and σ for a normal distribution

Observations of Normal Distributions

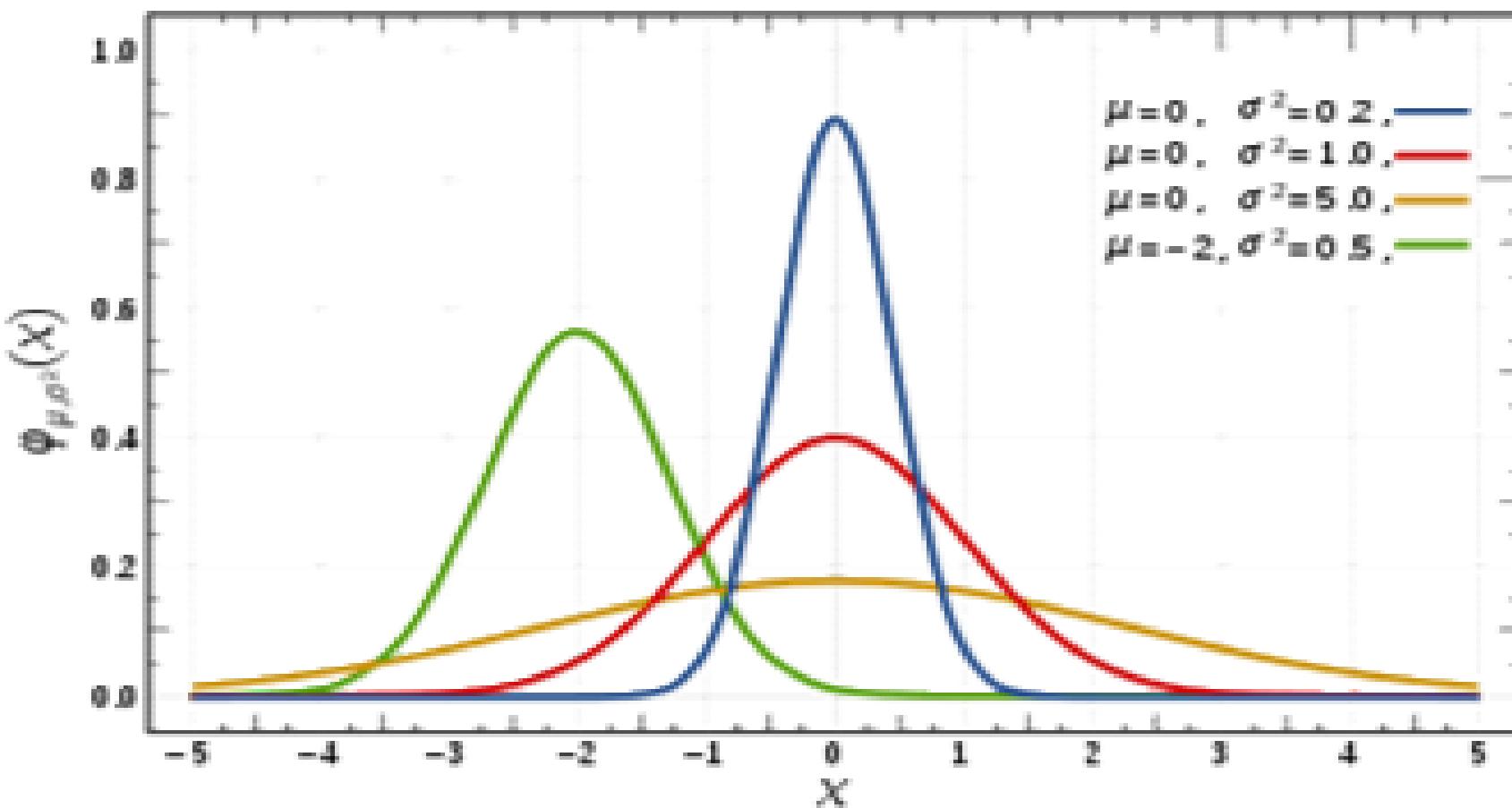


Image: google

Standard Deviation Rule for Normal Random Variables

In general, if X is a normal random variable, then the probability is

- 68% that X falls within 1σ of μ , that is, in the interval $\mu \pm \sigma$
- 95% that X falls within 2σ of μ , that is, in the interval $\mu \pm 2\sigma$
- 99.7% that X falls within 3σ of μ , that is, in the interval $\mu \pm 3\sigma$

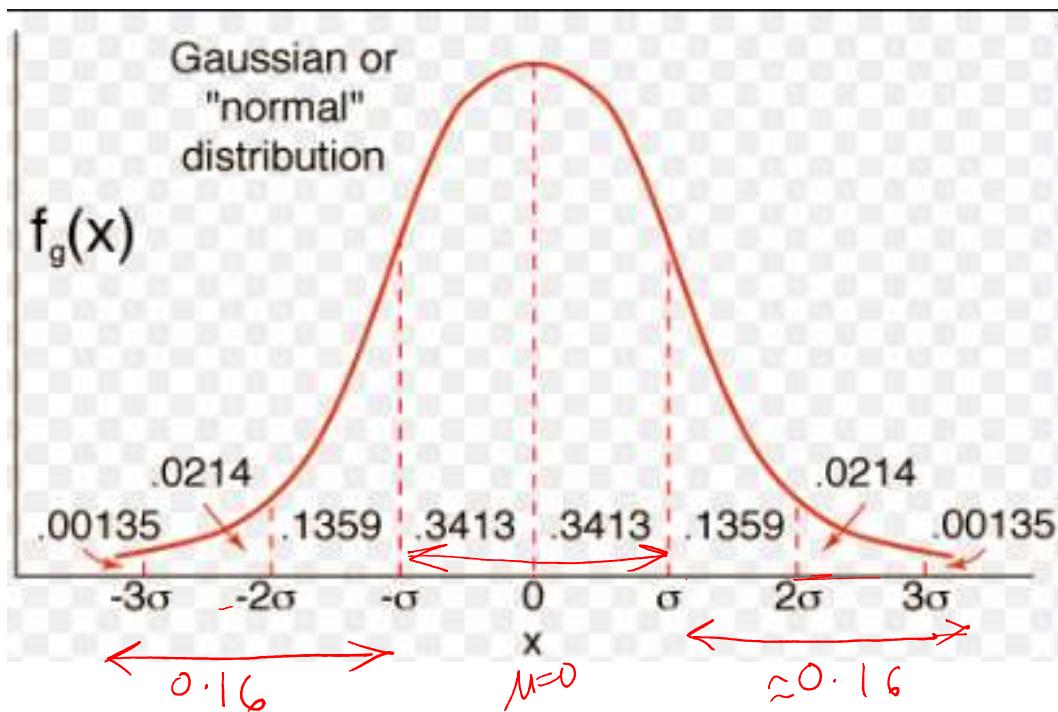
Using probability notation, we may write

$$0.68 = P(\mu - \sigma < X < \mu + \sigma)$$

$$0.95 = P(\mu - 2\sigma < X < \mu + 2\sigma)$$

$$0.997 = P(\mu - 3\sigma < X < \mu + 3\sigma)$$

Normal Distribution



- since 0.68 is the probability of being within 1 standard deviation of the mean,
- $(1 - .68) / 2 = 0.16$ is the probability of being further than 1 standard deviation below the mean (or further than 1 standard deviation above the mean).
- Likewise, $(1 - .95) / 2 = 0.025$ is the probability of being more than 2 standard deviations below (or above) the mean;
- $(1 - .997) / 2 = 0.0015$ is the probability of being more than 3 standard deviations below (or above) the mean.

Exercise

- Suppose that hair length of a randomly chosen female is a normal random variable with mean $\mu=11$ and standard deviation $\sigma=1.5$.

Ques1. What is the probability that a randomly chosen female will have hair length between 8 and 14 inches? $11 \pm 2 \times 1.5$ is 8 to 14 range

$$\mu \pm 2\sigma \rightarrow \text{prob?} \rightarrow 0.95$$

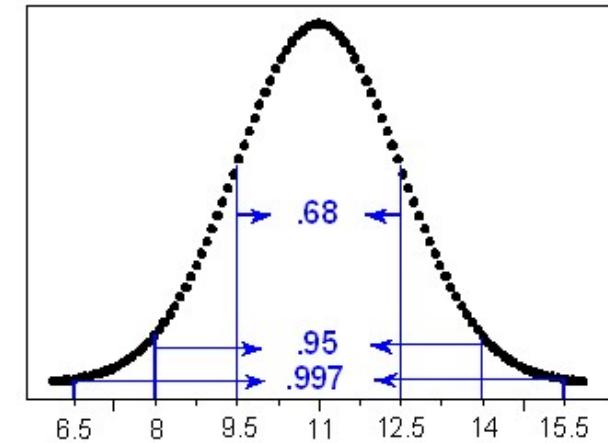
Ques2. A female is almost guaranteed (.997 probability) to have hair length between what two values?

$$\mu \pm 3\sigma = 11 \pm 2 \times 1.5 \text{ i.e. } 6.5 \text{ to } 15.5$$

Ques3. The probability is only 2.5% female will have hair length greater than how many inches?

95% of people/female have value b/w 8 to 14 inch.

2.5 of female have < 8 inch length & 2.5 have > 14 inch length.



The Normal Distribution

- The normal distribution is the most important one in all of probability and statistics.
- Many numerical populations have distributions that can be fit very closely by an appropriate normal curve.
- Examples include heights, weights, and other physical characteristics

A continuous rv X is said to have a **normal distribution** with parameters μ and σ (or μ and σ^2), where $-\infty < \mu < \infty$ and $0 < \sigma$, if the pdf of X is

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \quad -\infty < x < \infty \quad (4.3)$$



Standard Normal Distribution

- The normal distribution with parameter values $\mu=0$ and $\sigma=1$ is called the **standard normal distribution**.

Non-standard Normal Distributions

- Every unique pair of μ and σ values defines a different normal distribution
- Fortunately, a mechanism was developed by which all normal distributions can be converted into a single distribution: the z distribution.
- This process yields the **standardized normal distribution** (or curve).

$$z = \frac{x - \mu}{\sigma}, \quad \sigma \neq 0$$

Probabilities using z value

If X has a normal distribution with mean μ and standard deviation σ , then

$$Z = \frac{X - \mu}{\sigma}$$

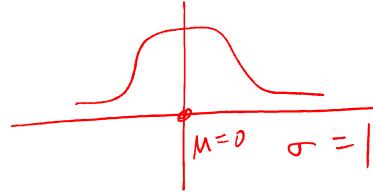
has a standard normal distribution. Thus

$$P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}\right)$$

$$= \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$$

$$P(X \leq a) = \Phi\left(\frac{a - \mu}{\sigma}\right) \quad P(X \geq b) = 1 - \Phi\left(\frac{b - \mu}{\sigma}\right)$$

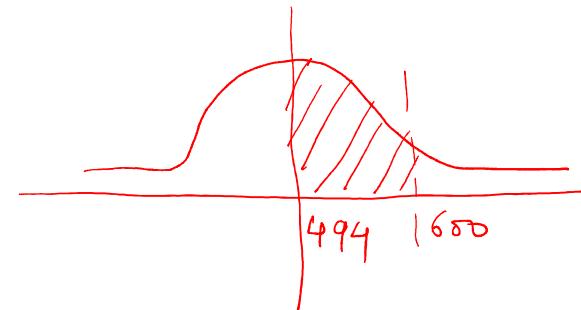
z - score



- A z score is the number of standard deviations that a value, x , is above or below the mean.
- If the value of x is less than the mean, the z score is negative;
- If the value of x is more than the mean, the z score is positive; and
- If the value of x equals the mean, the associated z score is zero.

Example

The Graduate Management Aptitude Test (GMAT), produced by the Educational Testing Service in Princeton, New Jersey, is widely used by graduate schools of business in the United States as an entrance requirement. Assuming that the scores are normally distributed, probabilities of achieving scores over various ranges of the GMAT can be determined. In a recent year, the mean GMAT score was 494 and the standard deviation was about 100. What is the probability that a randomly selected score from this administration of the GMAT is between 600 and the mean?

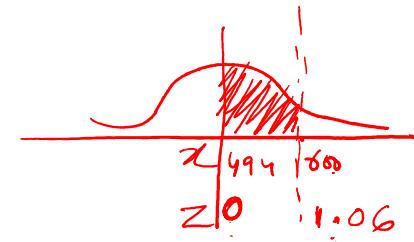
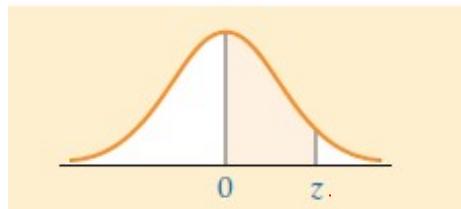


Solution

$$\text{for } x=600, \quad Z = \frac{x-\mu}{\sigma} = \frac{600 - 494}{100} = 1.06$$

$$P(494 \leq x \leq 600 | \mu = 494 \text{ and } \sigma = 100) = ?$$

area under curve or prob = 0.3554

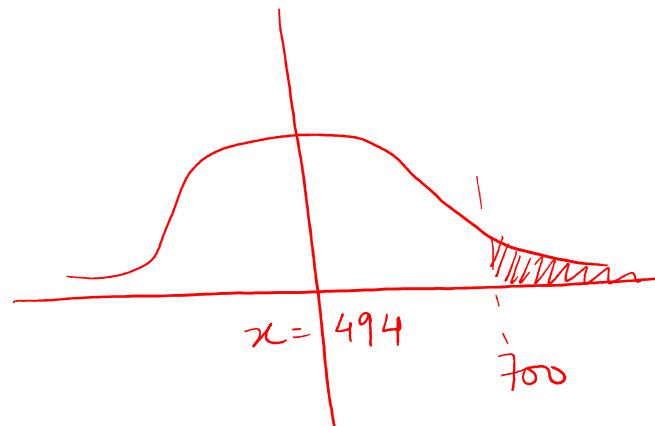


SECOND DECIMAL PLACE IN z

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

Exercise

What is the probability of obtaining a score greater than 700 on a GMAT test that has a mean of 494 and a standard deviation of 100? Assume GMAT scores are normally distributed.



Solution

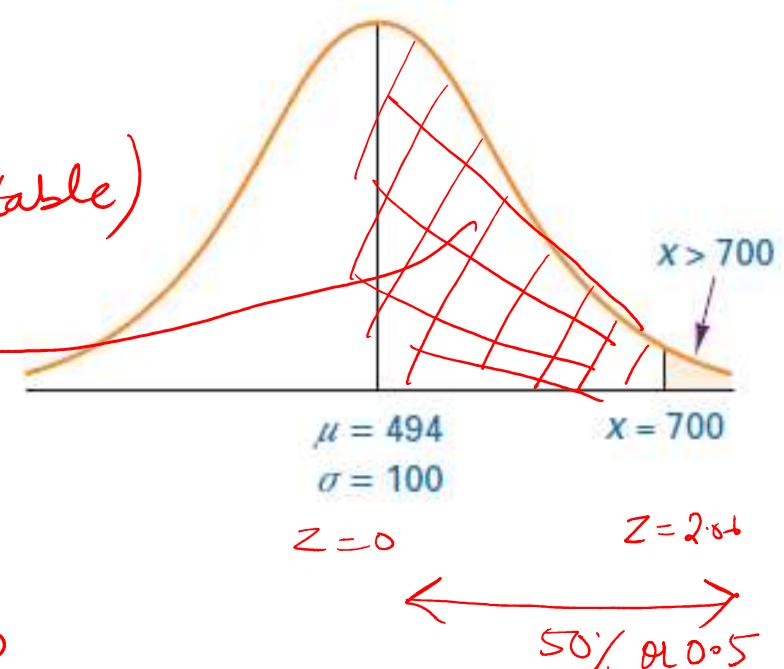
$$\text{for } x=700, z = \frac{x-\mu}{\sigma} = \frac{700-494}{100} = 2.06$$

for $z = 2.06$
 $\text{prob.} = ? = 0.4803$ (from z table)

Ans. $P(X > 700)$

$$= 0.5 - 0.4803$$

$$= 0.0197$$



Exercise

For the same GMAT examination, what is the probability of randomly drawing a score that is 550 or less?

$$\mu = 494$$

$$\sigma = 180$$

$$P(x \leq 550) = ?$$

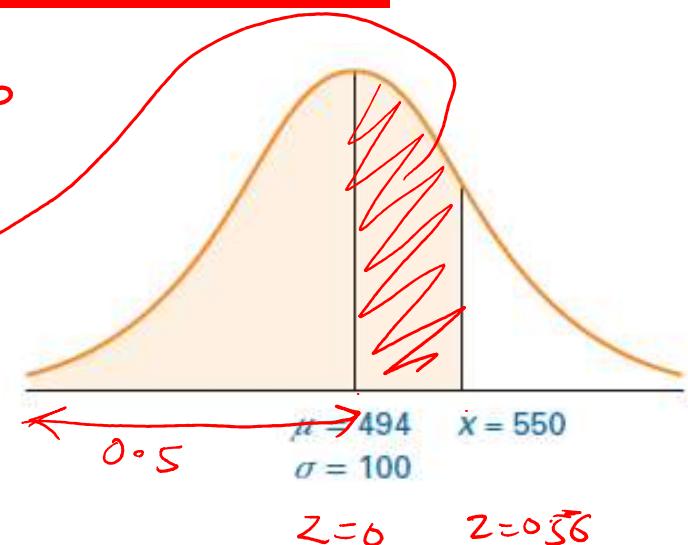
Solution

for $x = 550$

$$z = \frac{550 - 494}{100} = 0.56$$

for $z = 0.56$

$$\text{prob} = 0.2123$$



$$\begin{aligned} P(X \leq 550) &= 0.5 + 0.2123 \\ &= 0.7123 \end{aligned}$$

Exercise

What is the probability of randomly obtaining a score between 300 and 600 on the GMAT exam?

$$P(300 \leq X \leq 600)$$

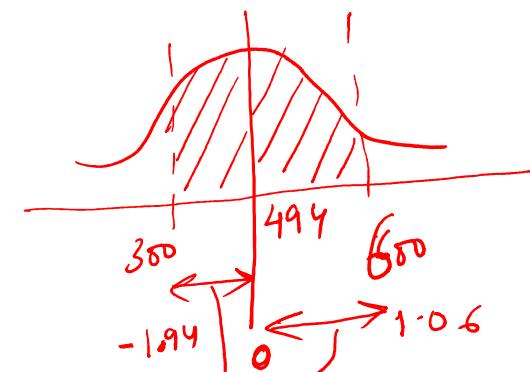
$$\mu = 494$$

$$\sigma = 100$$

Solution

$$\text{for } x = 300 \quad z = ? = \frac{300 - 494}{100} = -1.94$$

$$\text{for } x = 600 \quad z = \frac{x - \mu}{\sigma} = \frac{600 - 494}{100} = 1.06$$



for $z = 1.06$

prob. is 0.3554

for $z = -1.94$

prob. is 0.4738

$$\begin{aligned} \text{Ans} &= 0.3554 \\ &+ 0.4738 \\ &\hline \end{aligned}$$

$$0.8292$$



Exercise

(HW)

What is the probability of getting a score between 350 and 450 on the same GMAT exam?

Solution

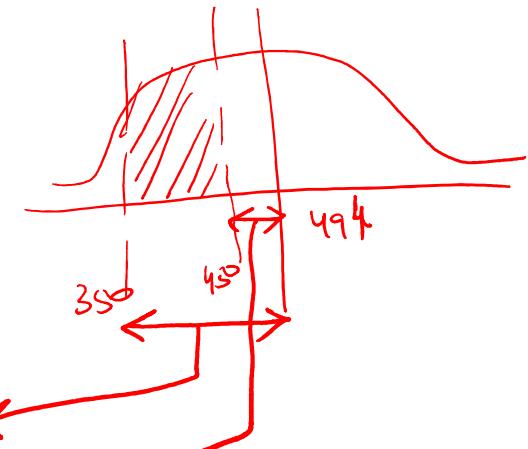
$$\text{for } x = 35^\circ, z = \frac{350 - 494}{100} = -1.44$$

$$\text{for } x = 45^\circ, z = \frac{450 - 494}{100} = -0.44$$

prob. for $z = -1.44$ is 0.4251

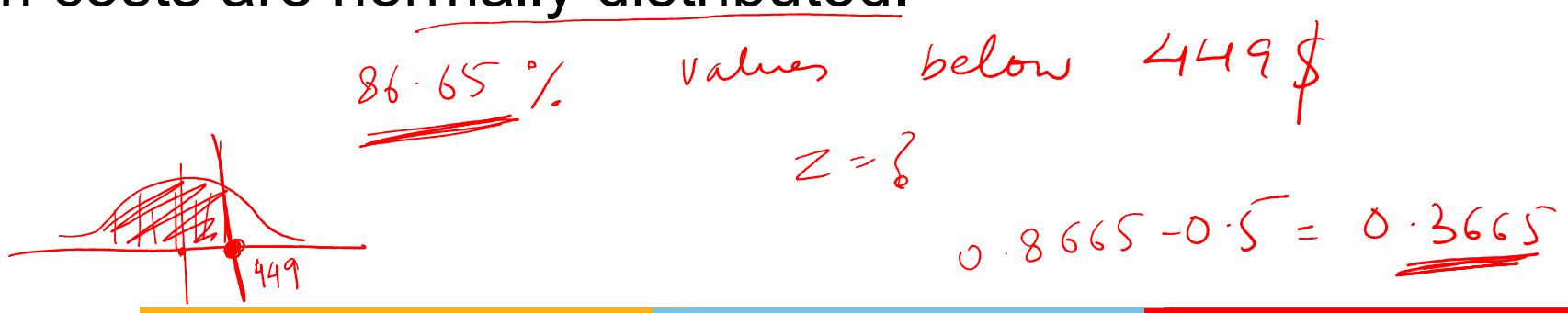
prob. for $z = -0.44$ is 0.1700

$$\begin{array}{r}
 \text{Ans} = 0.4251 \\
 - 0.1700 \\
 \hline
 0.2551
 \end{array}$$



Exercise

Runzheimer International publishes business travel costs for various cities throughout the world. In particular, they publish per diem totals, which represent the average costs for the typical business traveler including three meals a day in business-class restaurants and single-rate lodging in business-class hotels and motels. If 86.65% of the per diem costs in Buenos Aires, Argentina, are less than \$449 and if the standard deviation of per diem costs is \$36, what is the average per diem cost in Buenos Aires? Assume that per diem costs are normally distributed.



Solution

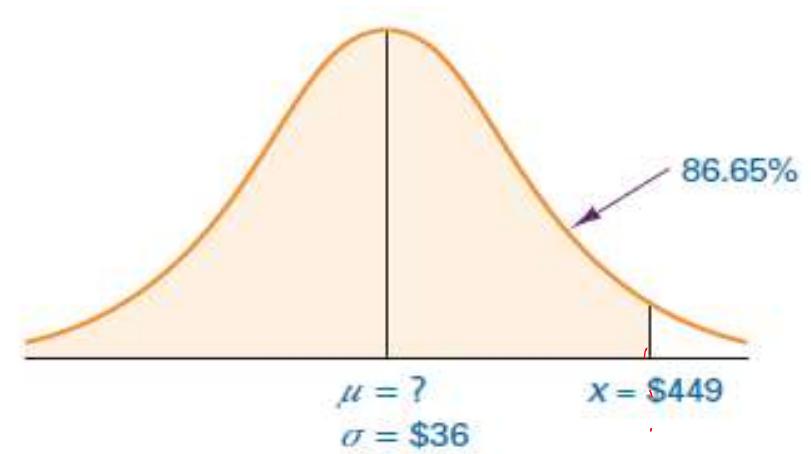
for area = 0.3665

z value is 1.11

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

$$1.11 = \frac{449 - \mu}{36}$$

$$\mu = 409.04 \$$$



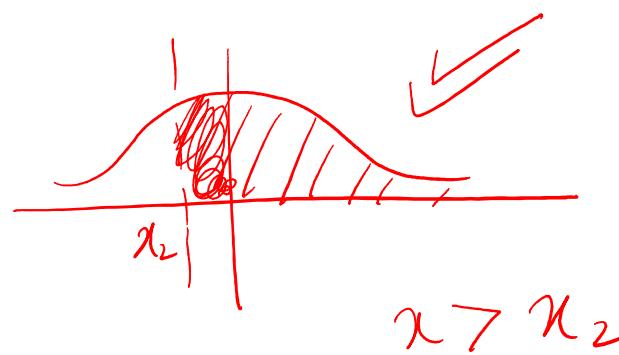
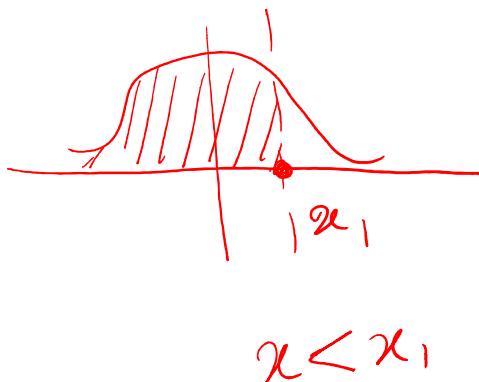
$$Z = 0$$

$$Z = 1.11$$

Exercise

(HW)

The U.S. Environmental Protection Agency publishes figures on solid waste generation in the United States. One year, the average number of waste generated per person per day was 3.58 pounds. Suppose the daily amount of waste generated per person is normally distributed, with a standard deviation of 1.04 pounds. Of the daily amounts of waste generated per person, 67.72% would be greater than what amount?



$$\begin{aligned}
 0.6772 - 0.5 \\
 = 0.1772 \\
 \uparrow \text{area} \\
 z = ?
 \end{aligned}$$

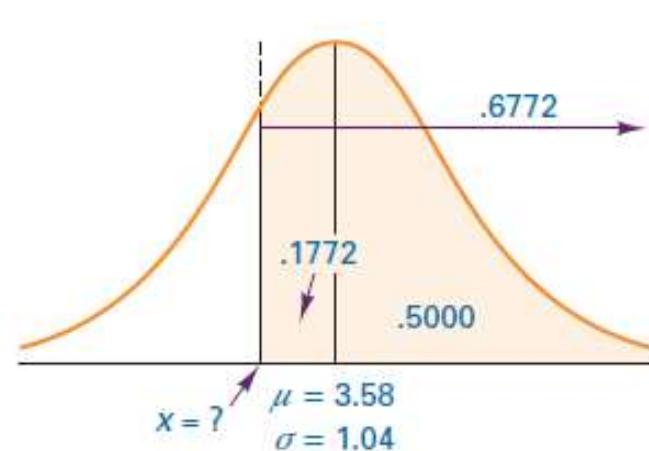
Solution

$$\text{area} = 0.1772$$

$$\downarrow$$

$$z = 0.46$$

left side of curve
will -ve.



$$-0.46 = \frac{x - 3.58}{1.04}$$

$$\underline{\underline{x = 3.10}}$$



BITS Pilani
Pilani Campus

Sampling and Estimation

Akanksha Bharadwaj
Asst. Professor, BITS Pilani



BITS Pilani
Pilani Campus



SS ZG536, ADV STAT TECHNIQUES FOR ANALYTICS

Contact Session 5



Quick Review of last session

Exercise

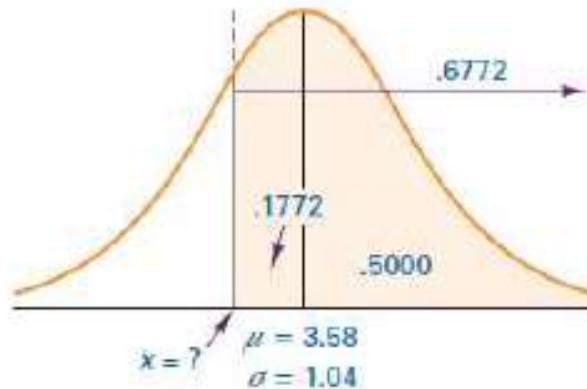
The U.S. Environmental Protection Agency publishes figures on solid waste generation in the United States. One year, the average number of waste generated per person per day was 3.58 pounds. Suppose the daily amount of waste generated per person is normally distributed, with a standard deviation of 1.04 pounds. Of the daily amounts of waste generated per person, 67.72% would be greater than what amount?

Solution

$$\text{area} = 0.1772$$

$$\downarrow \\ z = 0.46$$

left side of curve
z will -ve.



$$-0.46 = \frac{x - 3.58}{1.04}$$

$$\underline{\underline{x = 3.10}}$$

Approximate Binomial Distribution Problems



- As sample sizes become large, binomial distributions approach the normal distribution in shape regardless of the value of p .
- This phenomenon occurs faster (for smaller values of n) when p is near .50.
- To work a binomial problem by the normal curve requires a translation process.
- The first part of this process is to convert the two parameters of a binomial distribution, n and p , to the two parameters of the normal distribution, μ and σ .

$$\mu = n \cdot p \text{ and } \sigma = \sqrt{n \cdot p \cdot q}$$

continued

- After completion of this, a test must be made to determine whether the normal distribution is a good enough approximation of the binomial distribution:
Does the interval $\mu \pm 3\sigma$ lie between 0 and n ?
- For a normal curve approximation of a binomial distribution problem to be acceptable, all possible x values should be between 0 and n , which are the lower and upper limits, respectively, of a binomial distribution.
- Another rule of thumb for determining when to use the normal curve to approximate a binomial problem is that the approximation is good enough if both $np > 5$ and $nq > 5$.

Continuity correction factor

- It is used when you use a continuous probability distribution to approximate a discrete probability distribution. For example, when you want to use the normal to approximate a binomial.
- When you use a normal distribution to approximate a binomial distribution, you're going to have to use a continuity correction factor. **It's as simple as adding or subtracting .5 to the discrete x-value:** use the following table to decide whether to add or subtract.

If $P(X=n)$ use $P(n - 0.5 < X < n + 0.5)$

If $P(X > n)$ use $P(X > n + 0.5)$

If $P(X \leq n)$ use $P(X < n + 0.5)$

If $P(X < n)$ use $P(X < n - 0.5)$

If $P(X \geq n)$ use $P(X > n - 0.5)$

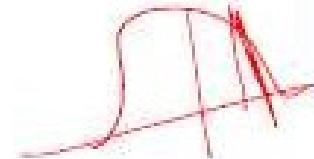
Example

Work the following binomial distribution problem by using the normal distribution.

$$P(x = 12 \mid n = 25 \text{ and } p = .40) = ?$$

Solution

$$n=25 \quad p=0.4$$



$$\checkmark \mu = np = 25 \times 0.4 = 10$$

$$\checkmark \sigma = \sqrt{npq} = \sqrt{25 \times 0.4 \times 0.6}$$

$$= 2.45$$

$$q = 1 - p$$

$$\boxed{\mu \pm 3\sigma = 10 \pm (3 \times 2.45)}$$

range is 2.65 to 17.35
this lies 0 to 25 (0 to 25)
yes

$$P(X=12) = P(11.5 < X < 12.5)$$

for 11.5

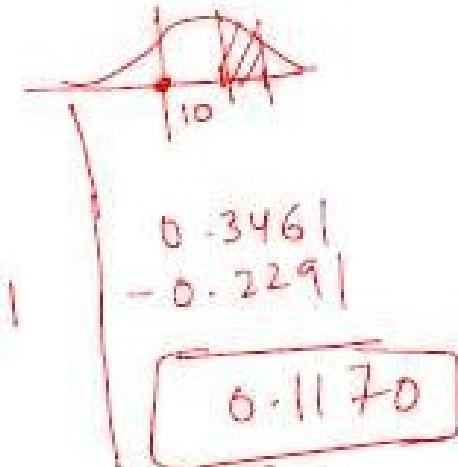
$$z = \frac{11.5 - 10}{2.45} = 0.61$$

 prob
 0.2291

for 12.5

$$z = \frac{12.5 - 10}{2.45} = 10.2$$

 prob
 0.3461



Exercise (HW)

Solve the following binomial distribution problem by using the normal distribution

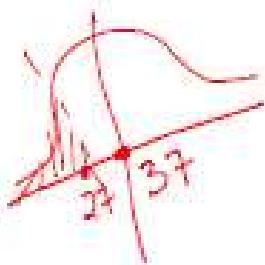
$$P(x < 27 | n = 100 \text{ and } p = .37) = ?$$

Solution

Homework



$$P(X < 27 \mid n = 100 \text{ & } p = 0.37) = ?$$



$$\mu = np = 37, \sigma = 4.83$$

$\because np > 5$ we can use normal distribution for this

$$\mu \pm 3\sigma = 37 \pm 14.49$$

so, range is 22.51 to 51.49

\therefore this lies b/w 0 to n i.e. 0 to 100

\therefore we can use normal distribution approximation.

$$P(X < 27) = P(X < 26.5) \quad [\text{Based on continuity correction factor}]$$

$$Z = \frac{26.5 - 37}{4.83} = -2.17$$

\hookrightarrow prob using z-table is 0.4850

$$\text{thus} = 0.5 - 0.4850 = 0.0150$$

continued

Had this problem been solved by using the binomial formula, the probabilities would have been the following.

x Value	Probability
26	.0059
25	.0035
24	.0019
23	.0010
22	.0005
21	.0002
20	<u>.0001</u>
$x < 27$.0131

The answer obtained by using the normal curve approximation (.0150) compares favorably to this exact binomial answer. The difference is only .0019.

Sampling

- Sampling is widely used in business as a means of gathering useful information about a population.
- Data are gathered from samples and conclusions are drawn about the population as a part of the inferential statistics process
- A sample provides a reasonable means for gathering useful decision-making information that might be otherwise unattainable and unaffordable.

Reasons for Sampling

- Taking a sample instead of conducting a census offers several advantages
 1. The sample can save money.
 2. The sample can save time.
 3. For given resources, the sample can broaden the scope of the study.
 4. If accessing the population is impossible, the sample is the only option.

Random Versus Non-random Sampling

- In **random** sampling every unit of the population has the same probability of being selected into the sample.
- In **non-random** sampling not every unit of the population has the same probability of being selected into the sample.



Random Sampling

Simple random sampling

- With simple random sampling, each unit of the frame is numbered from 1 to N (where N is the size of the population).
- Next, a table of random numbers or a random number generator is used to select n items into the sample.
- A random number generator is usually a computer program that allows computer-calculated output to yield random numbers.

Stratified Random Sampling

- In this the population is divided into nonoverlapping **subpopulations** called **strata**.
- The researcher then extracts a random sample from each of the subpopulations.
- The main reason for using stratified random sampling is that it has the potential for reducing sampling error.
- With stratified random sampling, the potential to match the sample closely to the population is greater than it is with simple random sampling because portions of the total sample are taken from different population subgroups.

Systematic Random Sampling

- With systematic sampling, every kth item is selected to produce a sample of size n from a population of size N.
- The value of k, sometimes called the **sampling cycle**, can be determined by the following formula.

$$k = \frac{N}{n}$$

where

n = sample size

N = population size

k = size of interval for selection



Non-random Sampling

Convenience Sampling

- In convenience sampling, elements for the sample are selected for the convenience of the researcher.
- The researcher typically chooses elements that are readily available, nearby, or willing to participate
- The sample tends to be less variable than the population because in many environments the extreme elements of the population are not readily available.
- The researcher will select more elements from the middle of the population.

Quota Sampling

- It appears to be similar to stratified random sampling. Certain population subclasses, such as age group, gender, or geographic region, are used as strata.
- However, instead of randomly sampling from each stratum, the researcher uses a nonrandom sampling method to gather data from one stratum until the desired quota of samples is filled.

Snowball Sampling

- Another nonrandom sampling technique is **snowball sampling**, in which *survey subjects are selected based on referral from other survey respondents.*
- The researcher identifies a person who fits the profile of subjects wanted for the study.
- The researcher then asks this person for the names and locations of others who would also fit the profile of subjects wanted for the study.
- Through these referrals, survey subjects can be identified cheaply and efficiently, which is particularly useful when survey subjects are difficult to locate.

Sampling Error

-
- **Sampling error** occurs *when the sample is not representative of the population.*
 - When random sampling techniques are used to select elements for the sample, sampling error occurs by chance.



Sampling Variation

Population of Wages of employees of an organization

1861	2495	1000	2497	1865	791	2090	2637	1327	1678
1680	2858	795	2495	2496	2501	1160	1480	1860	2490
2090	2840	2490	2640	659	827	2646	2638	2643	868
1327	1866	1861	2486	2865	3011	2494	1489	1865	2855
2840	2499	2093	2660	1165	2600	2085	2640	2998	1861
2956	2495	2865	1865	3000	3019	1670	2858	2642	1680
3038	3000	1313	596	656	3240	590	2501	2485	3015
2092	1679	3024	2497	2825	2630	2070	2900	1861	2636
2495	2637	2497	1159	2640	3050	870	2896	2500	2638
926	2860	1481	875	2482	1860	2086	934	3200	2490

Select different samples of varied sizes

Sample 1

3000 2486 820 1678 2070 2638 2490 1865 1000 2090 596 3200

Sample 2

2840 2858 3000 2490 2998 3050 2070 2896 3200 2490 3280

Sample 3

2858 3240 2497 2865 656 2093 934 1861 868 795

Sample 4

2086 1000 2497 596 656 875 2085 934 1313

Sample 5

820 1313 3000 2640 596 2640 2600 2495 934 2500

Select different samples of varied sizes

Sample 6

2840 2499 1327 1861 2495 3024 3038 2497

Sample 7

2858 2490 868 1670 1480 2643 1480 1680 2085 2490

Sample 8

2495 2858 1861 2092 2499 3000 2660 1000 1679 926 2660

Sample 9

795 791 3200 2085 2638 2497 2486 1159 2640

Sample 10

3019 3240 3200 3050 3000 3015 2900 2896 2998

Compute sample mean of these samples

Sample No.	Sample size	Mean	SD
1	12	1994.42	843.23
2	11	2830.18	349.94
3	10	1866.70	988.57
4	9	1338.00	704.36
5	10	1953.80	920.44
6	8	2447.63	590.64
7	10	1974.40	638.05
8	11	2157.27	715.10
9	9	2032.33	891.53
10	9	3035.33	117.40
Overall	100	2162.24	732.26

Sampling Variability

- The term "sampling variability" refers to the fact that the statistical information from a sample (called a *statistic*) will vary as the random sampling is repeated.
- **Sampling variability will decrease as the sample size increases.**
- the samples must be randomly chosen, must be of the same size (not smaller than 30), and the more samples that are used, the more reliable the information gathered will be.



Sampling Distribution

Do you consider these sample means and sample SDs as variable?

If yes, should we not describe the distribution of these variables?

The distribution of the sample estimates is called sampling distribution

For example the distribution of sample means is called Sampling distribution of mean

Definition

The probability distribution of a statistic (sample estimate) is called sampling distribution.

The sampling distribution of a statistic depends on the

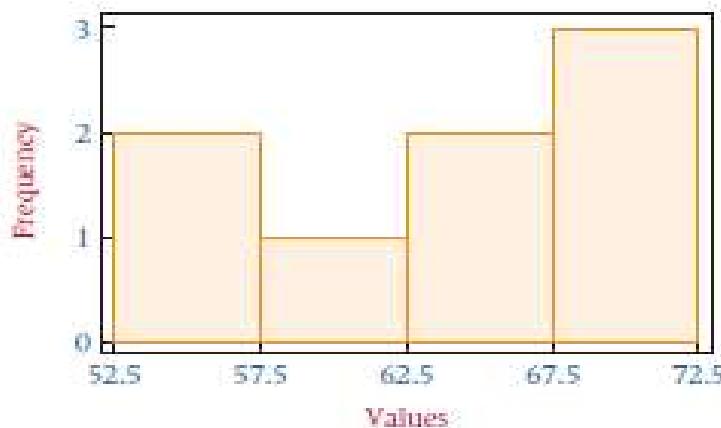
- distribution of the population,
- the size of the sample,
- and the method of sample selection.

Sampling Distribution Of \bar{x}

- The sample mean is one of the more common statistics used in the inferential process.
- The **distribution** of the values of the sample mean (\bar{x}) in repeated **samples** is called the **sampling distribution of \bar{x}**
- One way to examine the distribution possibilities is to take a population with a particular distribution, randomly select samples of a given size, compute the sample means, and attempt to determine how the means are distributed.

Example

- Suppose a small finite population consists of only $N = 8$ numbers:
54 55 59 63 64 68 69 70
- Using an Excel-produced histogram, we can see the shape of the distribution of this population of data.



- Suppose we take all possible samples of size $n = 2$ from this population with replacement.

Example

The result is the following pairs of data.

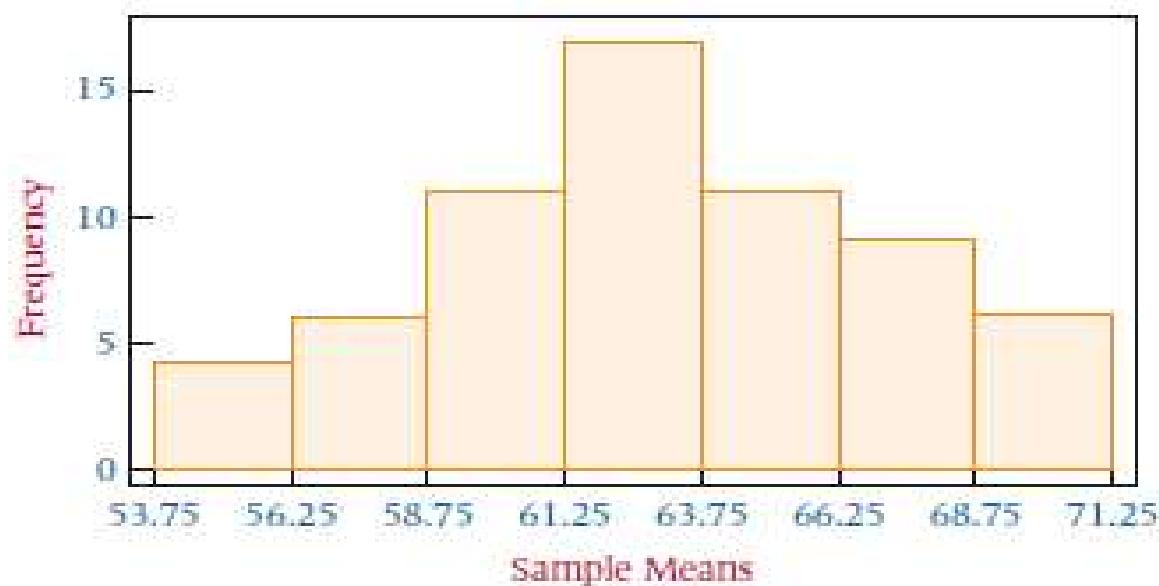
(54,54)	(55,54)	(59,54)	(63,54)
(54,55)	(55,55)	(59,55)	(63,55)
(54,59)	(55,59)	(59,59)	(63,59)
(54,63)	(55,63)	(59,63)	(63,63)
(54,64)	(55,64)	(59,64)	(63,64)
(54,68)	(55,68)	(59,68)	(63,68)
(54,69)	(55,69)	(59,69)	(63,69)
(54,70)	(55,70)	(59,70)	(63,70)
(64,54)	(68,54)	(69,54)	(70,54)
(64,55)	(68,55)	(69,55)	(70,55)
(64,59)	(68,59)	(69,59)	(70,59)
(64,63)	(68,63)	(69,63)	(70,63)
(64,64)	(68,64)	(69,64)	(70,64)
(64,68)	(68,68)	(69,68)	(70,68)
(64,69)	(68,69)	(69,69)	(70,69)
(64,70)	(68,70)	(69,70)	(70,70)

The means of each of these samples follow.

54	54.5	56.5	58.5	59	61	61.5	62
54.5	55	57	59	59.5	61.5	62	62.5
56.5	57	59	61	61.5	63.5	64	64.5
58.5	59	61	63	63.5	65.5	66	66.5
59	59.5	61.5	63.5	64	66	66.5	67
60	61.5	63.5	65.5	66	68	68.5	69
61.5	62	64	66	66.5	68.5	69	69.5
62	62.5	64.5	66.5	67	69	69.5	70

Example

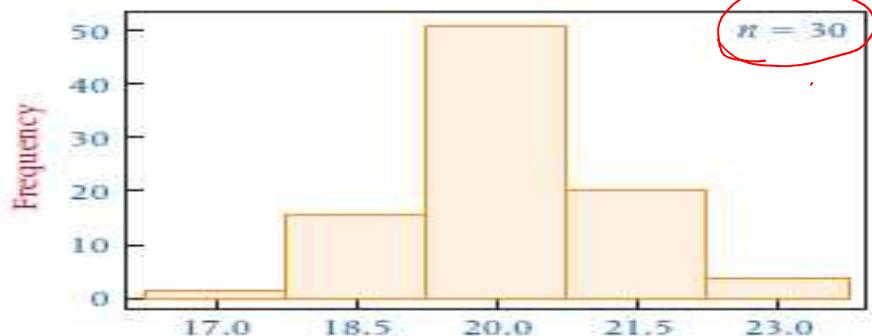
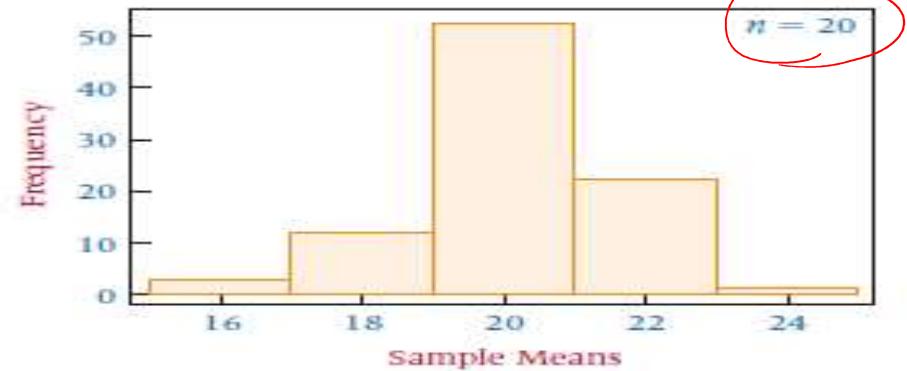
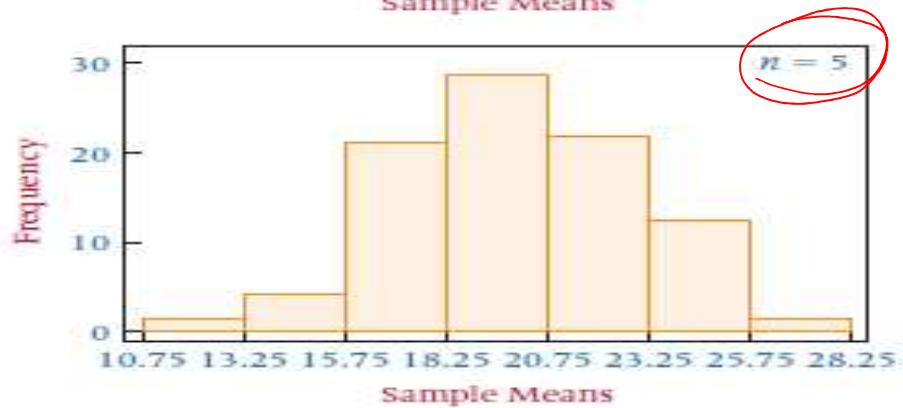
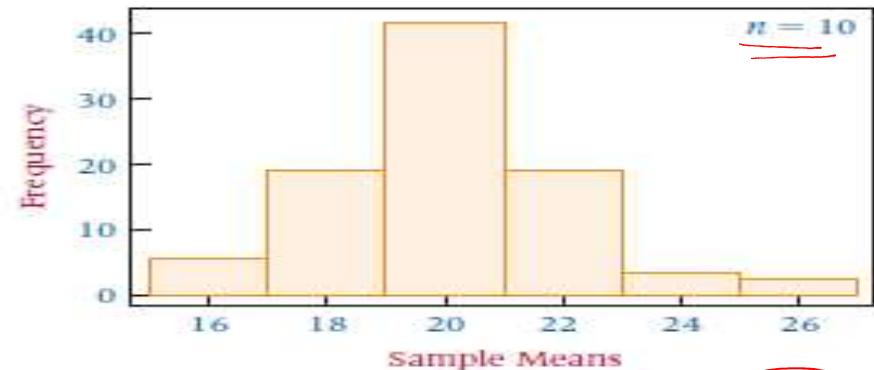
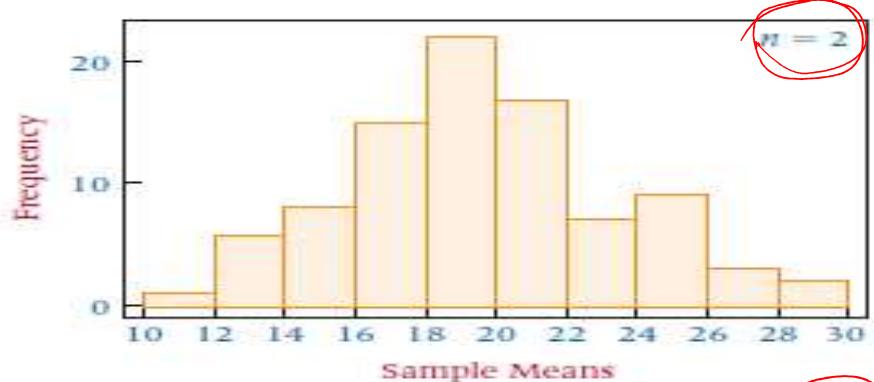
- Again using an Excel-produced histogram, we can see the shape of the distribution of these sample means.



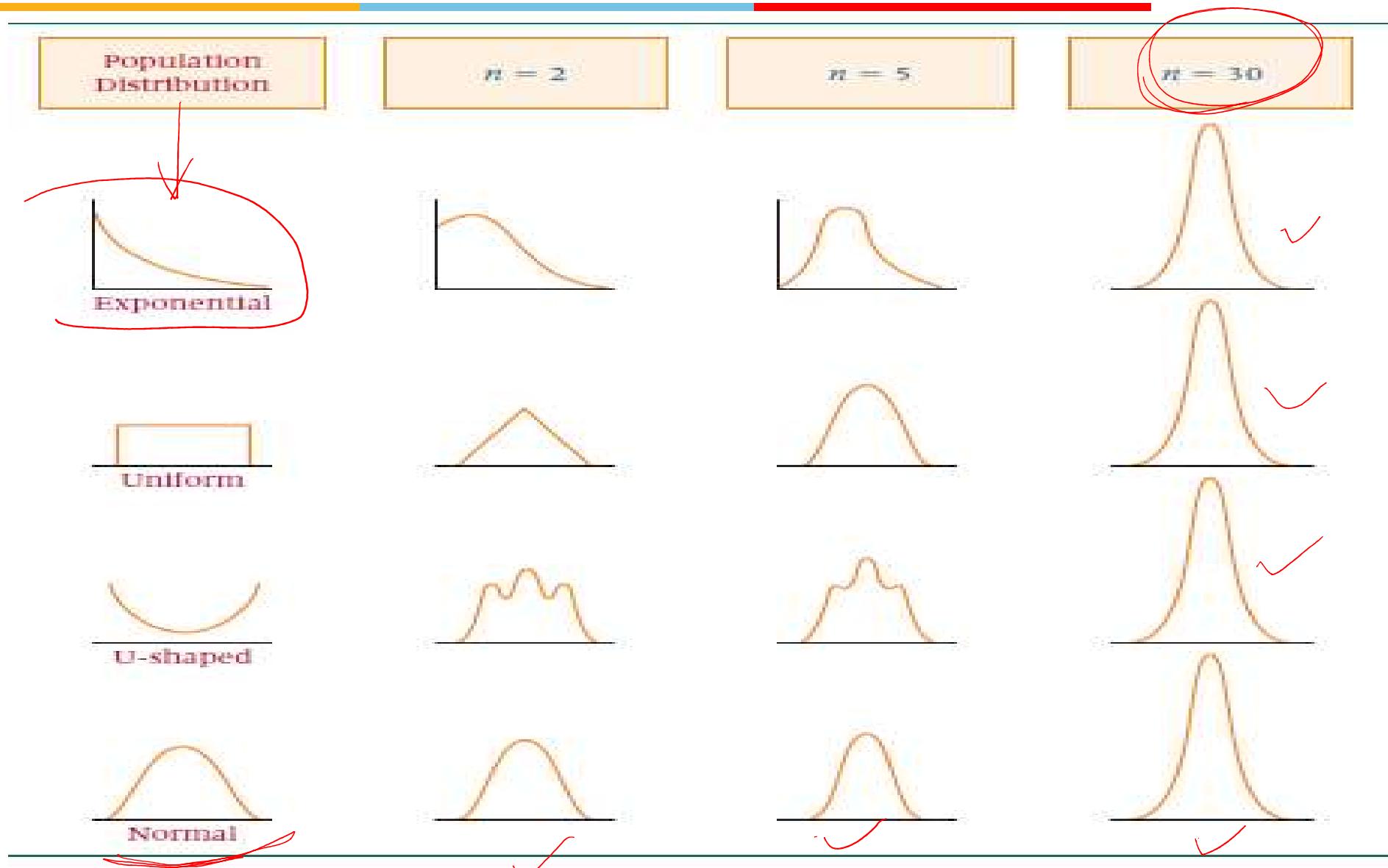
Conclusions

- Notice that the shape of the histogram for sample means is quite unlike the shape of the histogram for the population.
- The sample means appear to “pile up” toward the middle of the distribution and “tail off” toward the extremes.
- As sample sizes become much larger, the sample mean distributions begin to approach a normal distribution and the variation among the means decreases.

Sample Means from 90 Samples Ranging in Size from $n = 2$ to $n = 30$ from a Uniformly Distributed Population with $a = 10$ and $b = 30$



Shapes of the Distributions of Sample Means



Central Limit Theorem

- If samples of size n are drawn randomly from a population that has a mean of μ and a standard deviation of σ , the sample means, \bar{x} , are approximately normally distributed for sufficiently large sample sizes ($n \geq 30$) regardless of the shape of the population distribution.
- If the population is normally distributed, the sample means are normally distributed for any size sample.
- From mathematical expectation

$$\mu_{\bar{x}} = \mu$$

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

sample size

Z score for sample means

- The central limit theorem states that sample means are normally distributed regardless of the shape of the population for large samples and for any sample size with normally distributed populations.
- Thus, **sample means** can be **analyzed** by using **z scores**
- The formula to determine z scores for individual values from a normal distribution:
$$z = \frac{x - \mu}{\sigma}$$
- If sample means are normally distributed, the z score formula applied to sample means would be
$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$
- The standard deviation of the statistic of interest is $\sigma_{\bar{x}}$, sometimes referred to as the **standard error of the mean**.

Z score for sample means

- The researcher would randomly draw out all possible samples of the given size from the population, compute the sample means, and average them. This task is virtually impossible to accomplish in any realistic period of time.
- Similar activity for variation
- the mean of the sample means is the population mean.
- the standard deviation of the sample means is the standard deviation of the population divided by the square root of the sample
- Using central limit theorem:

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

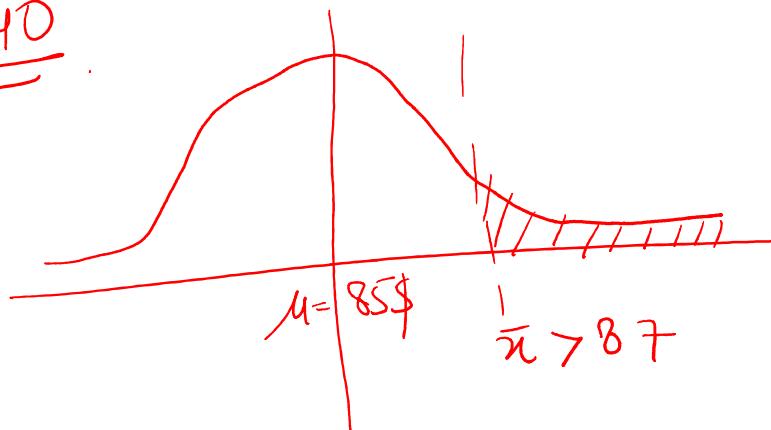

Example

Suppose the mean expenditure per customer at a tire store is \$85.00, with a standard deviation of \$9.00. If a random sample of 40 customers is taken, what is the probability that the sample average expenditure per customer for this sample will be \$87.00 or more?

$$P(\bar{x} \geq 87)$$

$$\underline{n=40}$$

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

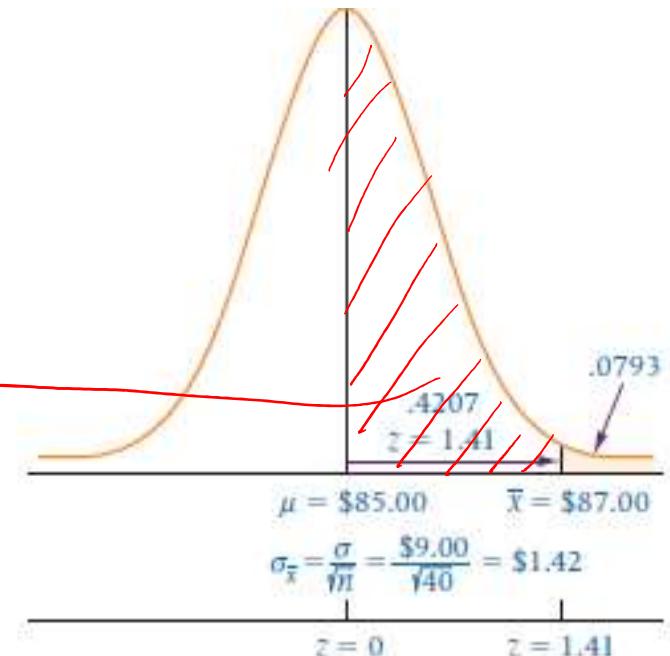


Solution

$$Z = \frac{87 - 85}{\sqrt{\frac{9}{40}}} = \frac{2}{1.42} \approx 1.41$$

area under curve for 1.41
= 0.4207

$$\begin{aligned} P(\bar{x} \geq 87) &= 0.5 - 0.4207 \\ &= \underline{\underline{0.0793}} \end{aligned}$$



Exercise

- Suppose that during any hour in a large department store, the average number of shoppers is 448, with a standard deviation of 21 shoppers. What is the probability that a random sample of 49 different shopping hours will yield a sample mean between 441 and 446 shoppers?

$$P(441 \leq \bar{X} \leq 446) = ?$$

$\therefore n \geq 30$ we can say sample means
are normally distributed

Solution

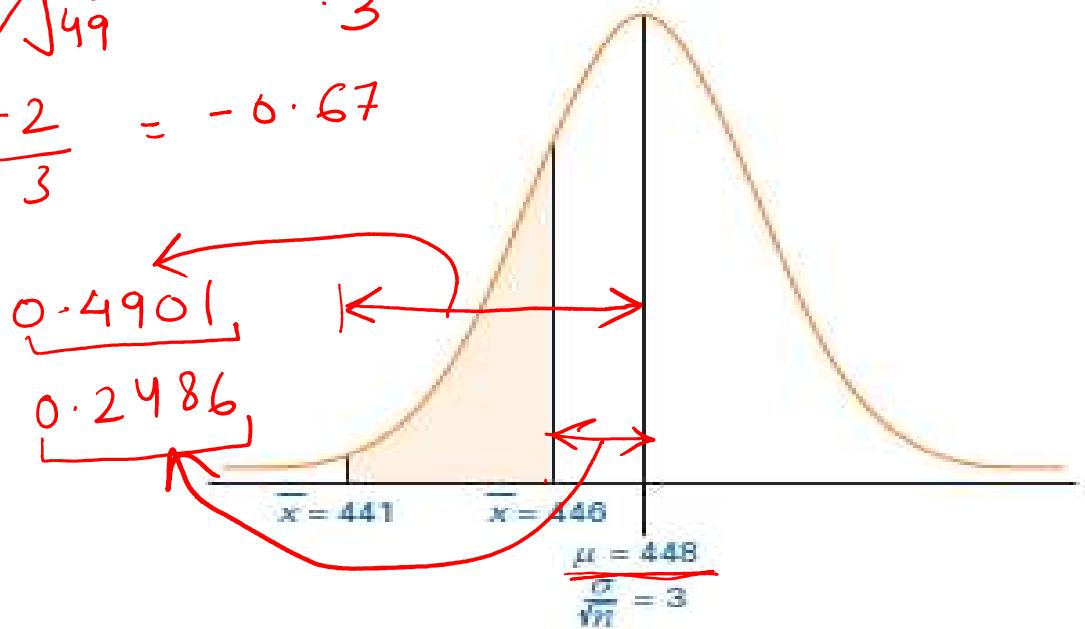
For this problem, $\mu = 448$, $\sigma = 21$, and $n = 49$. The problem is to determine $P(441 \leq \bar{x} \leq 446)$. The following diagram depicts the problem.

$$\text{for } 441, z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{441 - 448}{21/\sqrt{49}} = \frac{-7}{3} = -2.33$$

$$\text{for } 446, z = \frac{446 - 448}{21/\sqrt{49}} = \frac{-2}{3} = -0.67$$

prob. for $z = -2.33$ is

prob. for $z = -0.67$ is



$$\begin{aligned} \text{Ans} &= 0.4901 - 0.2486 \\ &= 0.2415 \end{aligned}$$

Sampling from a Finite Population

- The earlier example was based on the assumption that the population was infinitely or extremely large.
- In cases of a finite population, a *statistical adjustment can be made to the z formula for sample means*. The adjustment is called the **finite correction factor**
- Following is the z formula for sample means when samples are drawn from finite populations.

$$\sqrt{\frac{N-n}{N-1}}$$

population size
sample size

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}}$$

Rules for finite population

- As the size of the finite population becomes larger in relation to sample size, the finite correction factor approaches 1.
- In theory, whenever researchers are working with a finite population, they can use the finite correction factor.
- A rough rule of thumb for many researchers is that, if the sample size is **less than 5%** of the finite population size or $n/N < 0.05$, the finite correction factor does **not** significantly modify the solution.

$$\sqrt{\frac{N-n}{N-1}}$$

Exercise

A production company's 350 hourly employees average 37.6 years of age, with a standard deviation of 8.3 years. If a random sample of 45 hourly employees is taken, what is the probability that the sample will have an average age of less than 40 years?

$$n > 30$$

$$P(\bar{X} < 40)$$

$$\mu = 37.6$$

$$\sigma = 8.3$$

$$N = 350$$

$$n = 45$$

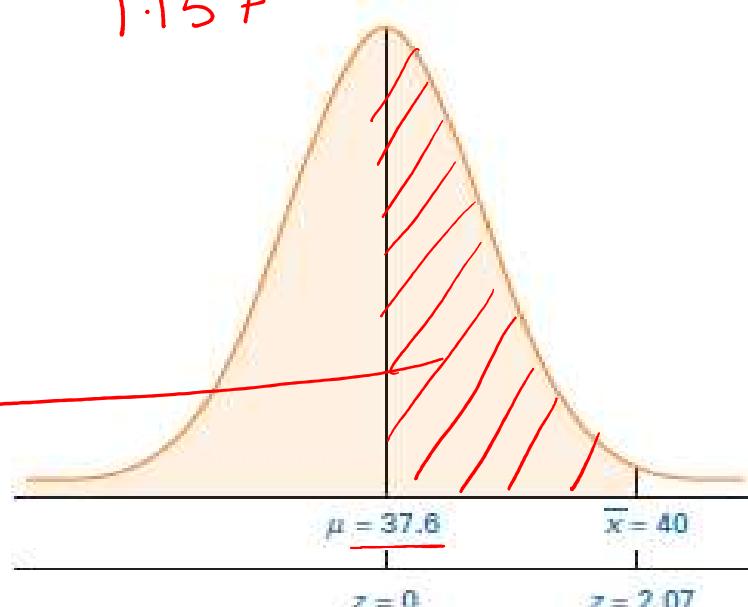
Solution

$$\text{for } \bar{x} = 40, Z = \frac{40 - 37.6}{\frac{8.3}{\sqrt{45}} \sqrt{\frac{350-45}{350-1}}} = \frac{2.4}{1.157} = 2.07$$

area under curve for 2.07

$$= 0.4808$$

$$\begin{aligned} \text{Ans} \rightarrow & \quad 0.5000 \\ & + 0.4808 \\ \hline & \underline{0.9808} \end{aligned}$$



Sampling Distribution Of Sample Proportion



- If research results in ***countable*** items such as how many people in a sample have a flexible work schedule, the sample proportion is often the statistic of choice.

SAMPLE PROPORTION

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

where

x = number of items in a sample that have the characteristic

n = number of items in the sample

Example

- In a sample of 100 factory workers, 30 workers might belong to a union.
- The value of sample proportion for this characteristic, union membership, is

$$30/100 = .30$$

How does a researcher use the sample proportion in analysis?

- The central limit theorem applies to sample proportions in that the normal distribution approximates the shape of the distribution of sample proportions
- If $\underline{n \cdot p > 5}$ and $\underline{n \cdot q > 5}$ (p is the population proportion and $q = 1 - p$).
- The mean of sample proportions for all samples of size n randomly drawn from a population is p (the population proportion) and the **standard deviation of sample proportions** is $\sqrt{\frac{p \cdot q}{n}}$
- sometimes referred to as the **standard error of the proportion**

Z Formula For Sample Proportions

For $n^*p > 5$ and $n^*q > 5$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}}$$

where

\hat{p} = sample proportion

n = sample size

p = population proportion

$q = 1 - p$

Example

- Suppose 60% of the electrical contractors in a region use a particular brand of wire. What is the probability of taking a random sample of size 120 from these electrical contractors and finding that .50 or less use that brand of wire?

$$P(\hat{p} \leq 0.5)$$

$$np = 120 \times 0.6 = 72.0$$

$$nq = 120 \times 0.4 = 48$$

$$np > 5 \text{ & } nq > 5$$

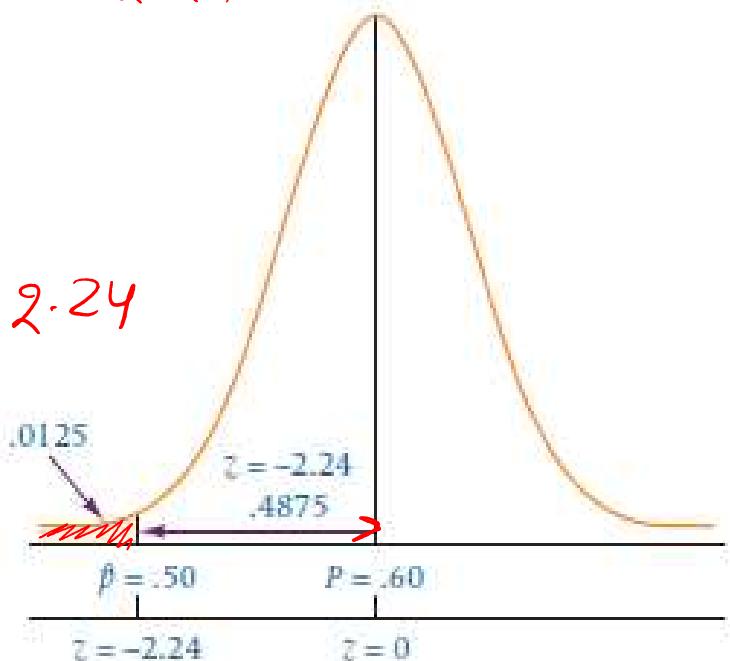
Solution

$$Z = \frac{\hat{P} - P}{\sqrt{\frac{Pq}{n}}}$$

for 0.5 , $Z = \frac{0.5 - 0.6}{\sqrt{\frac{0.6 \times 0.4}{120}}} = -2.24$

area under the curve / prob. of $Z = -2.24$

0.4875



$$\begin{aligned} P(\hat{P} \leq 0.5) &= 0.5 - 0.4875 \\ &= 0.0125 \end{aligned}$$

Exercise

If 10% of a population of parts is defective, what is the probability of randomly selecting 80 parts and finding that 12 or more parts are defective?

$$p = 0.1$$

$$n = 80$$

$$\hat{p} = ?$$

$$\frac{12}{80} = 0.15$$

$$P(\hat{p} \geq 0.15)$$

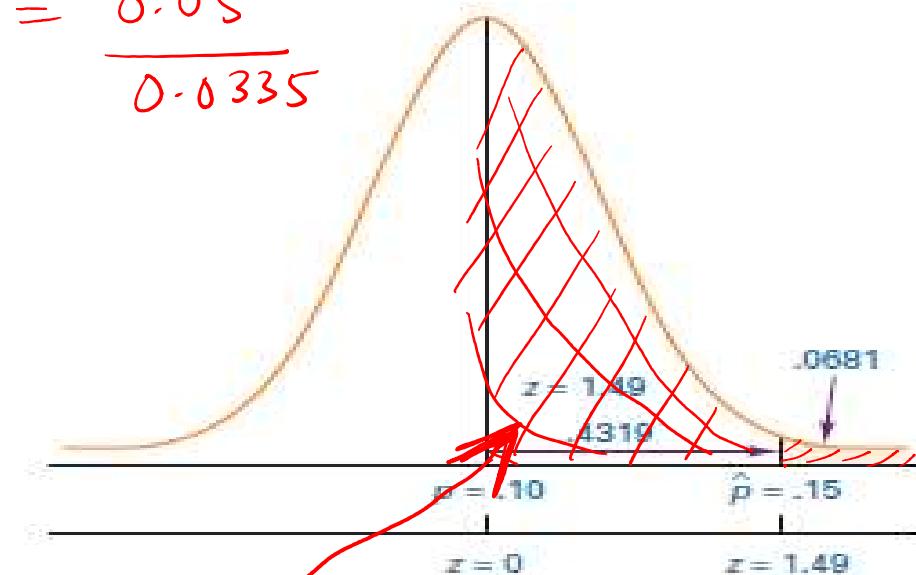
Solution

for $\hat{p} = 0.15$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.15 - 0.10}{\sqrt{\frac{0.1 \times 0.9}{80}}} = \frac{0.05}{0.0335} = 1.49$$

area under curve for 1.49

0.4319



$$\text{Ans} = 0.5 - 0.4319$$

$$= 0.0681$$



BITS Pilani
Pilani Campus

Hypothesis Testing

Akanksha Bharadwaj
Asst. Professor, CS/IS Department



Forms Of Statistical Inference

Three forms of statistical inference

- Point estimation
- Interval estimation
- Hypothesis testing



Estimating The Population Mean Using The Z Statistic



Point Estimate

- A **point estimate** is a statistic taken from a sample that is used to estimate a population parameter.
- A point estimate is only as good as the representativeness of its sample.
- If other random samples are taken from the population, the point estimates derived from those samples are likely to vary.



Interval Estimate

- Because of variation in sample statistics, estimating a population parameter with an interval estimate is often preferable to using a point estimate.
- An interval estimate (confidence interval) is a range of values within which the analyst can declare, with some confidence, the population parameter lies.

Central Limit Theorem

- z formula for sample means can be used if the population standard deviation is known when sample sizes are large, regardless of the shape of the population distribution, or for smaller sizes if the population is normally distributed.
- Rearranging this formula algebraically to solve for μ gives $\mu = \bar{x} - z \frac{\sigma}{\sqrt{n}}$,
- Because a sample mean can be greater than or less than the population mean, z can be positive or negative.
- Thus, μ can be $\bar{x} \pm \frac{z\sigma}{\sqrt{n}}$

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

(confidence interval formula)



Confidence Interval to Estimate μ

100(1 - α)% CONFIDENCE
INTERVAL TO ESTIMATE μ :
 σ KNOWN (8.1)

or

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

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

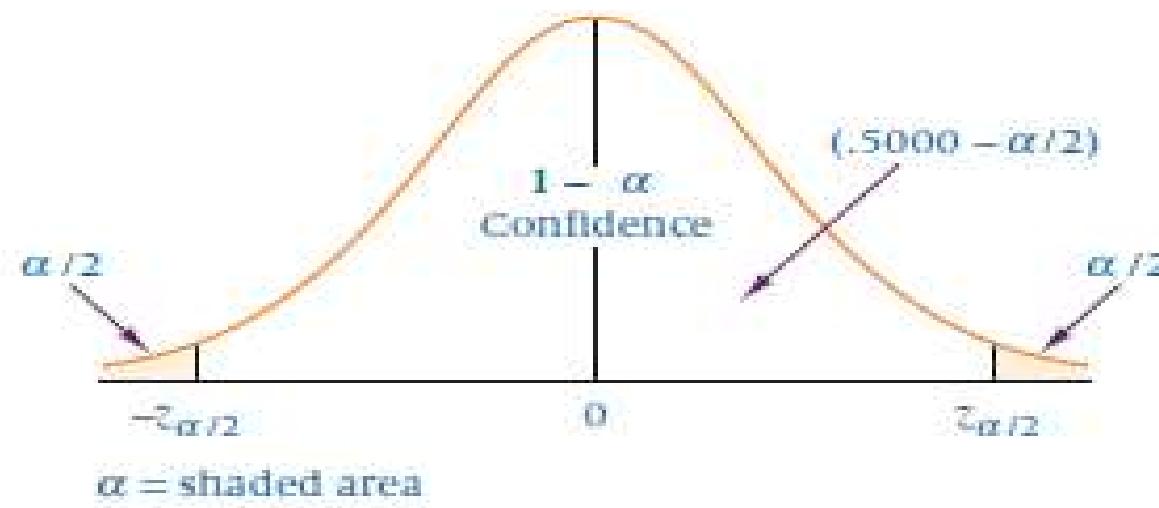
where

α = the area under the normal curve outside the confidence interval area

$\alpha/2$ = the area in one end (tail) of the distribution outside the confidence interval

- Alpha is the area under the normal curve in the tails of the distribution outside the area defined by the confidence interval
- The confidence interval formula (8.1) yields a range (interval) within which we feel with some confidence that the population mean is located.

Z Scores for Confidence Intervals in Relation to alpha



- Here we use α to locate the z value in constructing the confidence interval
- Because the standard normal table is based on areas between a z of 0 and $z_{\alpha/2}$,
- the table z value is found by locating the area of .5000 , which is the part of the normal curve between the middle of the curve and one of the tails.

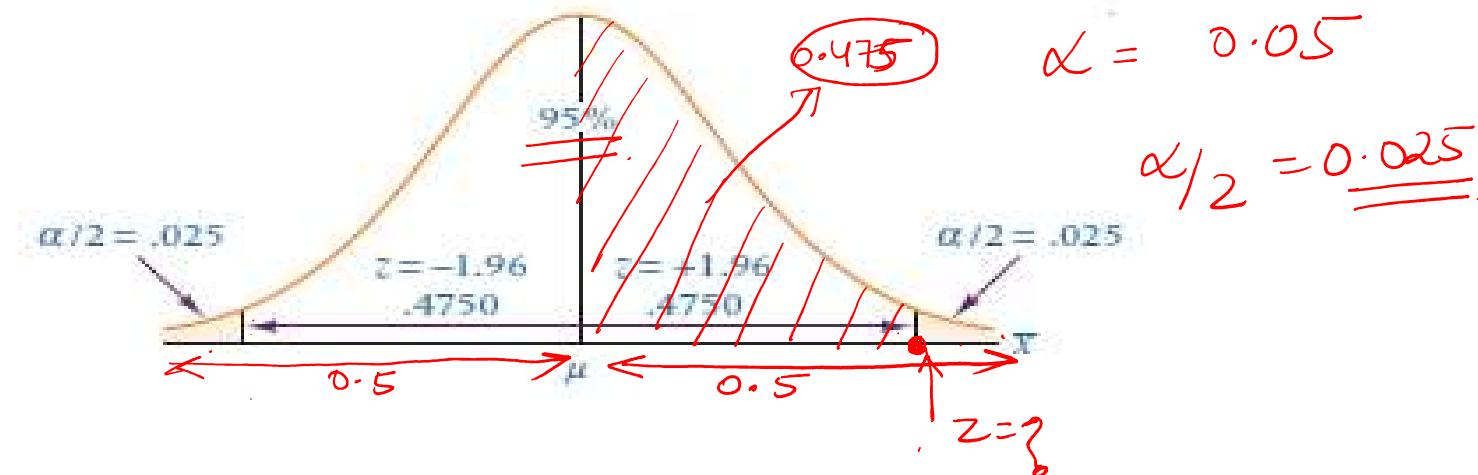
Distribution of Sample Means for 95% Confidence



$$I = 0.95 + \alpha$$

$$\alpha = 0.05$$

$$\alpha/2 = 0.025$$



- For 95% confidence, $\alpha = .05$ and $\alpha/2 = .025$.
- The value of $z_{\alpha/2}$ or $z_{.025}$ is found by looking in the standard normal table under $.5000 - .0250 = .4750$.
- This area in the table is associated with a z value of **1.96**.

Example

- In the cellular telephone company problem of estimating the population mean number of minutes called per residential user per month, from the sample of 85 bills it was determined that the sample mean is 510 minutes. Suppose past history and similar studies indicate that the population standard deviation is 46 minutes. Determine a 95% confidence interval.

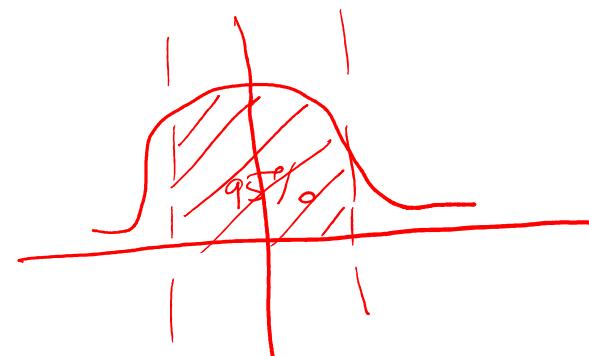
$$n = 85$$

$$\bar{x} = 510 \text{ min.}$$

$$\sigma = 46 \text{ min}$$

$$\text{for area} = 0.5 - 0.025 \\ = 0.475$$

$$\text{value of } z = \pm 1.96$$



$$1 - \alpha = 0.95$$

$$\alpha = 0.05$$

$$\alpha/2 = 0.025$$

Solution



$$\bar{x} - z \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z \frac{\sigma}{\sqrt{n}}$$

$$510 - 1.96 \times \frac{46}{\sqrt{85}} \leq \mu \leq 510 + 1.96 \times \frac{46}{\sqrt{85}}$$

$$500.22 \leq \mu \leq 519.78$$

Finite correction factor

Confidence Level	<i>z</i> Value
90%	1.645
95%	1.96
98%	2.33
99%	2.575

Values of *z* for Common
Levels of Confidence

CONFIDENCE INTERVAL TO
ESTIMATE μ USING THE
FINITE CORRECTION
FACTOR (8.2)

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Exercise

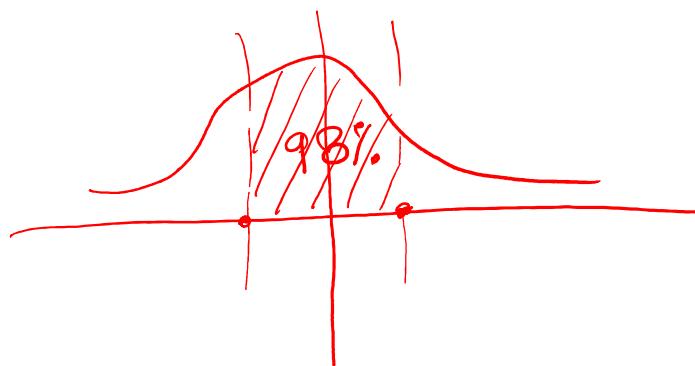
A study is conducted in a company that employs 800 engineers. A random sample of 50 engineers reveals that the average sample age is 34.3 years. Historically, the population standard deviation of the age of the company's engineers is approximately 8 years. Construct a 98% confidence interval to estimate the average age of all the engineers in this company.

$$\checkmark N = 800$$

$$\bar{x} = 34.3 \text{ yrs}$$

$$\checkmark n = 50$$

$$\sigma = 8 \text{ yrs.}$$





Solution

$$1 - \alpha = 0.98$$

$$\alpha = 0.02, \frac{\alpha}{2} = 0.01$$

area under curve from mean to z value
 $= 0.5 - 0.01 = 0.49$

for 0.49 value of z will be 2.33

$$34.3 - 2.33 \times \frac{8}{\sqrt{50}} \sqrt{\frac{800-50}{800-1}} \leq \mu \leq 34.3 + 2.33 \times \frac{8}{\sqrt{50}} \sqrt{\frac{750}{799}}$$

$$31.75 \leq \mu \leq 36.85$$



Estimating The Population Proportion

- Methods similar to those used earlier can be used to estimate the population proportion.
- The central limit theorem for sample proportions led to the following formula

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}}$$

- where $q = 1 - p$. Recall that this formula can be applied only when $n*p$ and $n*q$ are greater than 5.
- **for confidence interval purposes only and for large sample sizes—** is substituted for p in the denominator, yielding

$$z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}}$$



Confidence Interval To Estimate P

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$$

where:

\hat{p} = sample proportion

$\hat{q} = 1 - \hat{p}$

p = population proportion

n = sample size

In this formula, \hat{p} is the point estimate and $\pm z_{\alpha/2} \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$ is the error of the estimation.





Example

- A study of 87 randomly selected companies with a telemarketing operation revealed that 39% of the sampled companies used telemarketing to assist them in order processing. Using this information, how could a researcher estimate the population proportion of telemarketing companies that use their telemarketing operation to assist them in order processing?
- Use 95% confidence interval

$$n = 87 \\ \hat{p} = 0.39 \\ \hat{q} = 1 - 0.39 = 0.61$$

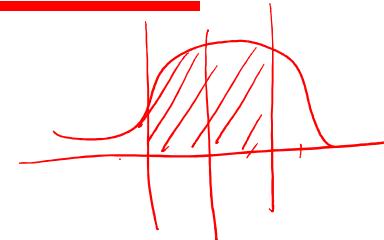
Solution



95% confidence interval

$$z = ?$$

$$z = \pm 1.96$$



$$0.39 - 1.96 \cdot \sqrt{\frac{0.39 \times 0.61}{87}} \leq p \leq 0.39 + 1.96 \cdot \sqrt{\frac{0.39 \times 0.61}{87}}$$

$$0.2875 \leq p \leq 0.4925$$



Exercise (HW)

- Coopers & Lybrand surveyed 210 chief executives of fast-growing small companies. Only 51% of these executives had a management succession plan in place. A spokesperson for Cooper & Lybrand said that many companies do not worry about management succession unless it is an immediate problem. However, the unexpected exit of a corporate leader can disrupt and unfocus a company for long enough to cause it to lose its momentum. Use the data given to compute a 92% confidence interval to estimate the proportion.

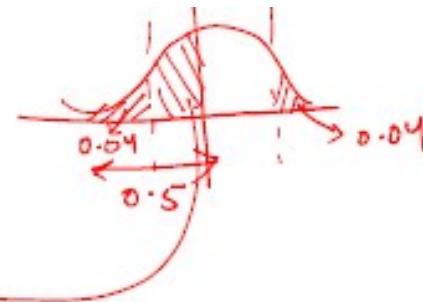
Solution

$$\hat{p} = 0.51$$
$$n = 210$$

$$1 - \alpha = 0.92$$
$$\therefore \alpha = 0.08$$
$$\alpha/2 = 0.04$$

$$0.5 - 0.04 \\ = 0.46$$

for 0.46 my $Z = \pm 1.75$



$$0.51 - 1.75 \sqrt{\frac{0.51 \times 0.49}{210}} \leq p \leq 0.51 + 1.75 \sqrt{\frac{0.51 \times 0.49}{210}}$$
$$0.45 \leq p \leq 0.57$$



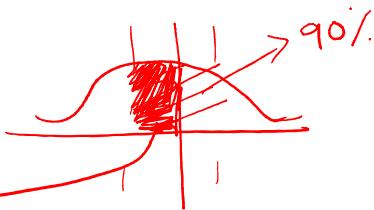
Exercise

- A clothing company produces men's jeans. The jeans are made and sold with either a regular cut or a boot cut. In an effort to estimate the proportion of their men's jeans market in Oklahoma City that prefers boot-cut jeans, the analyst takes a random sample of 212 jeans sales from the company's two Oklahoma City retail outlets. Only 34 of the sales were for boot-cut jeans. Construct a 90% confidence interval to estimate the proportion of the population in Oklahoma City who prefer boot-cut jeans.

$$n = 212$$

$$\hat{p} = \frac{34}{212} = 0.16$$

Solution



$$\begin{aligned}q &= 1 - p \\q &= 1 - 0.16 \\q &= 0.84\end{aligned}$$

$$1 = 0.9 + \alpha$$

$$\alpha = 0.1$$

$$\alpha/2 = 0.05$$

$$\text{area under curve.} = 0.5 - 0.05 = 0.45$$

for 0.45 value of $z = \pm 1.645$

$$0.16 - 1.645 \sqrt{\frac{0.16 \times 0.84}{212}} \leq p \leq 0.16 + 1.645 \sqrt{\frac{0.16 \times 0.84}{212}}$$

$$0.119 \leq p \leq 0.201$$



BITS Pilani
Pilani Campus



SS ZG536, ADV STAT TECHNIQUES FOR ANALYTICS

Contact Session 6



Need for testing of hypothesis

- Often the decisions are made based on samples estimates to generalize on population parameter (as described in sampling and estimation).
- In this process, there may be a difference between the estimate and the parameter



Need for testing of hypothesis

The following possibilities might arise due to sampling

$$|\text{Estimate} - \text{Parameter}| = \begin{cases} 0 \\ \text{Small} \\ \text{Large} \end{cases}$$

Case(i): If the difference is zero, it is called unbiased



Need for testing of hypothesis

Case(ii): If the difference is small, it may be due to chance or sampling error
(improper sampling technique used leads to sampling error)

Case(iii): If the difference is large, it may be a real one or due to sampling error
(improper sampling technique used leads to sampling error)

Hence, there is a need to test what type of difference is between estimate and parameter.

Statistical Hypothesis



A statement which is yet to be proved/ established or a statement on the parameter(s) of the Probability distribution to be tested

Null Hypothesis

Alternative Hypothesis

there is nothing new happening, the old theory is still true, the old standard is correct, and the system is in control.

states that the new theory is true, there are new standards, the system is out of control, and/or something is happening.

Hypothesis testing (Non-statistical)



A suspected criminal is produced before jury. The Jury has to decide whether the defendant is innocent or guilty.



Jury must decide between two hypotheses

The null hypothesis



H_0 : The defendant may be innocent

The alternative hypothesis



H_1 : The defendant may be guilty

or H_a

Hypothesis - Formulation



Judge 1



Judge 2

Suppose based on evidences, if we are interested in finding **proportion of false positivity** in the judgment of two Judges

Formulate the hypotheses

???

Hypothesis - Formulation

H_0



The proportion of false positive judgement between Judges may be same

$$H_0 : P_1 = P_2$$

H_1



The proportion of false positive judgement by Judge 1 may be lower than proportion of false positive judgement by Judge 2

$$H_1 : P_1 < P_2$$

Hypothesis - Formulation

H_1



The proportion of false positive judgement by Judge 1 may be more than proportion of false positive judgement by Judge 2

$$H_1 : P_1 > P_2$$

H_1



The proportion of false positive judgement between both Judges may be different

$$H_1 : P_1 \neq P_2$$



Example

- Suppose flour packaged by a manufacturer is sold by weight; and a particular size of package is supposed to average 40 ounces. Suppose the manufacturer wants to test to determine whether their packaging process is out of control as determined by the weight of the flour packages.
- The null hypothesis for this experiment is that the average weight of the flour packages is 40 ounces (no problem).
- The alternative hypothesis is that the average is not 40 ounces (process is out of control).

$$H_0: \mu = 40 \text{ oz.}$$

$$H_a: \mu \neq 40 \text{ oz.}$$



Exercise

- According to the Cancer Control and Prevention committee, the proportion of Indian adults age 25 or older who smoke is 0.12. A researcher suspects that the rate is lower among Indian adults 25 or older who have a bachelor's degree or higher education level.
 - **What is the null hypothesis in this case?**
 - The Proportion of smokers among Indian adults 25 or older who have a bachelor degree or higher is 0.12
 - **What is the alternative hypothesis in this case?**
 - The Proportion of smokers among Indian adults 25 or older who have a bachelor degree or higher is less than 0.12
-



Rejection and Non-rejection Regions

The possible statistical outcomes of a study can be divided into two groups:

1. Those that cause the rejection of the null hypothesis
2. Those that do not cause the rejection of the null hypothesis.

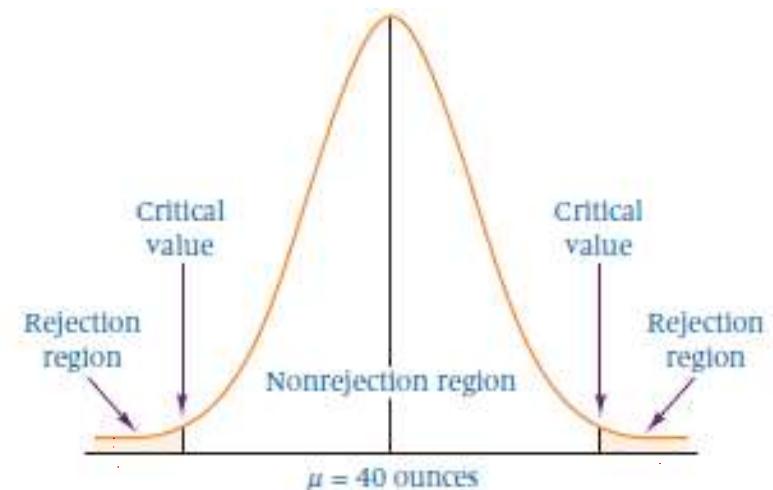
Example



- Consider the flour-packaging manufacturing example. The null hypothesis is that the average fill for the population of packages is 40 ounces.
- Suppose a sample of 100 such packages is randomly selected, and a sample mean of 50 ounces is obtained
- This sample mean may be so far from what is reasonable to expect for a population with a mean of 40 ounces that the decision is made to reject the null hypothesis.
- **This prompts the question: when is the sample mean so far away from the population mean that the null hypothesis is rejected?**

Critical values

- In each direction beyond the critical values lie the **rejection** regions.
- Any sample mean that falls in that region will lead the business researcher to reject the null hypothesis.
- Sample means that fall between the two critical values are close enough to the population mean that the business researcher will decide not to reject the null hypothesis. These means are in the **non-rejection** region.



$$\text{Test} = \begin{cases} \mu_1 < \mu_2 \Rightarrow \text{One - tailed test} \\ \mu_1 > \mu_2 \Rightarrow \text{One - tailed test} \\ \mu_1 \neq \mu_2 \Rightarrow \text{Two - tailed test} \end{cases}$$



Type I Errors

- **The null hypothesis is true, but the business researcher decides that it is not.**
- As an example, suppose the flour-packaging process actually is “in control” and is averaging 40 ounces of flour per package. Suppose also that a business researcher randomly selects 100 packages, weighs the contents of each, and computes a sample mean.
- It is possible, by chance, to randomly select 100 of the more extreme packages (mostly heavy weighted or mostly light weighted) resulting in a mean that falls in the rejection region.
- The decision is to reject the null hypothesis even though the population mean is actually 40 ounces. In this case, the business researcher has committed a Type I error.



Alpha

-
- Means that fall beyond the critical values will be considered so extreme that the business researcher chooses to reject the null hypothesis.
 - However, if the null hypothesis is true, any mean that falls in a rejection region will result in a decision that produces a Type I error.
 - The *probability of committing a Type I error* is called **alpha (α)** or **level of significance**.
 - **Alpha equals the area under the curve that is in the rejection region beyond the critical value(s).**



Type II error

- It is committed when a business researcher ***fails to reject a false null hypothesis.***
- In this case, the null hypothesis is false, but a decision is made to not reject it.
- Suppose in the case of the flour problem that the packaging process is actually producing a population mean of 41 ounces even though the null hypothesis is 40 ounces.
- A sample of 100 packages yields a sample mean of 40.2 ounces, which falls in the non-rejection region. The business decision maker decides not to reject the null hypothesis.
- A Type II error has been committed.

Beta



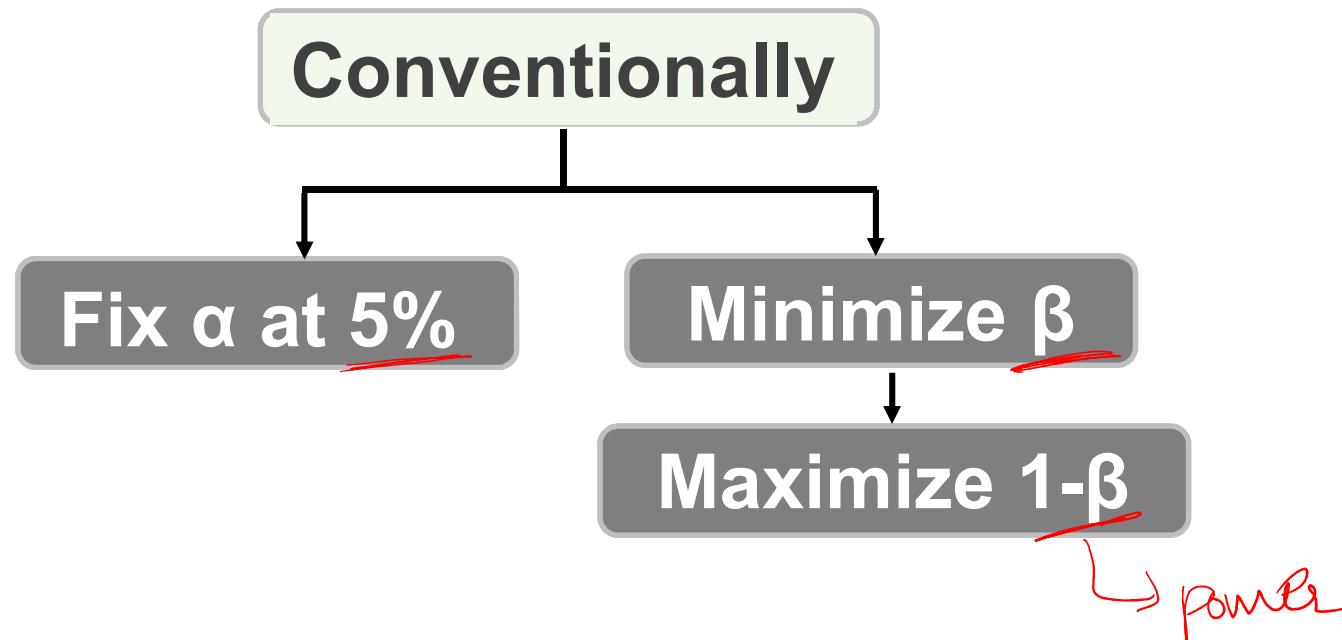
-
- The probability of committing a Type II error is **beta** ().
 - Beta occurs only when the null hypothesis is not true, the computation of beta varies with the many possible alternative parameters that might occur.
 - For example, in the flour-packaging problem, if the population mean is not 40 ounces, then what is it? It could be 41, 38, or 42 ounces. A value of beta is associated with each of these alternative means

Relation between alpha and beta

- Power, which is equal to $1 - \beta$, is the probability of a statistical test rejecting the null hypothesis when the null hypothesis is false.
- A business researcher cannot commit both a Type I error and a Type II error at the same time on the same hypothesis test.
- Generally, alpha and beta are inversely related.
- If alpha is reduced, then beta is increased, and vice versa.

		Reality	
		H_0 is true	H_0 is false
Decision	Reject H_0 , (conclude H_a)	Type I error	☺ Correct decision
	Fail to reject H_0	☺ Correct decision	Type II error

Decision on α –error and β - error





Steps involved in Testing of Hypothesis

Typically, statisticians and researchers present the hypothesis testing process in terms of an eight-step approach:

- Step 1. Establish a null and alternative hypothesis.
- Step 2. Determine the appropriate statistical test.
- Step 3. Set the value of alpha, the Type I error rate.
- Step 4. Establish the decision rule.
- Step 5. Gather sample data.
- Step 6. Analyze the data.
- Step 7. Reach a statistical conclusion.
- Step 8. Make a business decision.



Testing hypotheses about a population mean using the Z statistic



Z Test For A Single Mean

- Below formula can be used to test hypotheses about a single population mean when σ is known if the sample size is large ($n \geq 30$) for any population and for small samples ($n < 30$) if x is known to be normally distributed in the population.

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



Example

- A survey of CPAs across the United States found that the average net income for sole proprietor CPAs is \$74,914. Because this survey is now more than ten years old, an accounting researcher wants to test this figure by taking a random sample of 112 sole proprietor accountants in the United States to determine whether the net income figure changed. The researcher could use the eight steps of hypothesis testing to do so. Assume the population standard deviation of net incomes for sole proprietor CPAs is \$14,530.



Solution

- At step 1, the hypotheses must be established. Because the researcher is testing to determine whether the figure has changed, the alternative hypothesis is that the mean net income is not \$74,914.
- The null hypothesis is that the mean still equals \$74,914. These hypotheses follow.

$$H_0: \mu = \$74,914$$

$$H_1: \mu \neq \$74,914$$

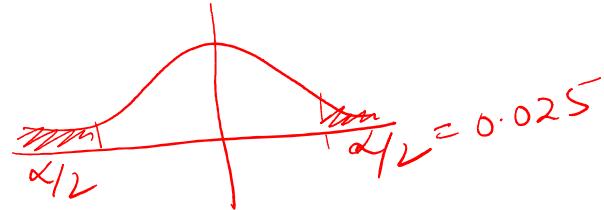


TEST:

- Step 2 is to determine the appropriate statistical test and sampling distribution. Because the population standard deviation is known (\$14,530) and the researcher is using the sample mean as the statistic, **the z test for a single mean is the appropriate test statistic.**

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

.....



- Step 3 is to specify the Type I error rate, or alpha, which is .05 in this problem.
- Step 4 is to state the decision rule. Because the test is two tailed and alpha is .05, there is 2 or .025 area in each of the tails of the distribution. Thus, the rejection region is in the two ends of the distribution with 2.5% of the area in each.
- There is a .4750 area between the mean and each of the critical values that separate the tails of the distribution (the rejection region) from the non-rejection region.
- By using this .4750 area and Z Table, the critical z value can be obtained.

$$z_{\alpha/2} = \pm 1.96 \quad \xrightarrow{\text{critical z value}}$$



Step 5 is to gather the data. Suppose the 112 CPAs who respond produce a sample mean of \$78,695. At step 6, the value of the test statistic is calculated by using $\bar{x} = \$78,695$, $n = 112$, $\sigma = \$14,530$, and a hypothesized $\mu = \$74,914$:

$$\text{observed } z \text{ value} = \frac{78,695 - 74,914}{\frac{14530}{\sqrt{112}}}$$

$$= 2.75$$



ACTION:

- Because this test statistic, $z = 2.75$, is greater than the critical value of z in the upper tail of the distribution, $\underline{z = +1.96}$,
- the statistical conclusion reached at step 7 of the hypothesis- testing process is to reject the null hypothesis.
- *The calculated test statistic* is often referred to as the **observed value**. Thus, the observed value of z for this problem is $\underline{2.75}$ and the critical value of z for this problem is 1.96.

Reject NULL hypothesis



Testing the Mean with a Finite Population

- Remember that if the sample size is less than 5% of the population, the finite correction factor does not significantly alter the solution.

FORMULA TO TEST
HYPOTHESES ABOUT
 μ WITH A FINITE
POPULATION (9.2)

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}}$$

- In the CPA net income example, suppose only 600 sole proprietor CPAs practice in the United States.
- Z= ?

$$Z = \frac{78,695 - 7491}{\sqrt{112} \sqrt{\frac{600-112}{600-1}}} = \frac{3781}{1239.2} = 3.05$$

Reject NULL hypothesis



Using the p-Value to Test Hypotheses

- Another way to reach a statistical conclusion in hypothesis testing problems is by using the **p-value**, sometimes referred to as **observed significance level**
- The p-value defines the smallest value of alpha for which the null hypothesis can be rejected.
- For example, if the p-value of a test is .038, the null hypothesis cannot be rejected at $\alpha = .01$ because .038 is the smallest value of alpha for which the null hypothesis can be rejected. However, the null hypothesis can be rejected for $\alpha = .05$.

Rejecting the Null Hypothesis Using p-Values



Range of p-Values	Rejection Range
$p\text{-value} > .10$	Cannot reject the null hypothesis for commonly accepted values of alpha
$.05 < p\text{-value} \leq .10$	Reject the null hypothesis for $\alpha = .10$
$.01 < p\text{-value} \leq .05$	Reject the null hypothesis for $\alpha = .05$
$.001 < p\text{-value} \leq .01$	Reject the null hypothesis for $\alpha = .01$
$.0001 < p\text{-value} \leq .001$	Reject the null hypothesis for $\alpha = .001$

Critical Value Method to Test Hypotheses

- The critical value method determines the **critical mean value** required for z to be in the rejection region and uses it to test the hypotheses.

$$\alpha = 0.05, \quad Z_c = \pm 1.96$$

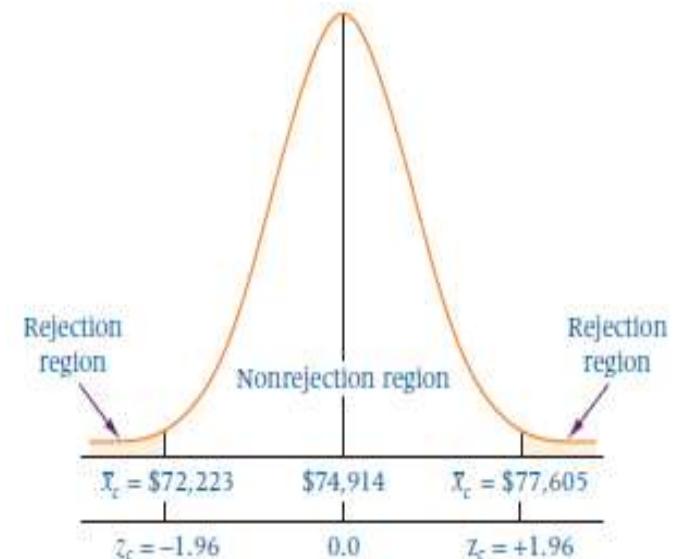
$$\pm 1.96 = \frac{\bar{x}_0 - 74914}{\frac{14530}{\sqrt{112}}}$$

$$\bar{x}_0 = 74914 \pm 2629$$

$$72,223 \leq \bar{x}_0 \leq 77,605$$

$$\text{Sample } \bar{x} = 78,695$$

Reject NULL hypothesis





Exercise

- In an attempt to determine why customer service is important to managers in the United Kingdom, researchers surveyed managing directors of manufacturing plants in Scotland. One of the reasons proposed was that customer service is a means of retaining customers. On a scale from 1 to 5, with 1 being low and 5 being high, the survey respondents rated this reason more highly than any of the others, with a mean response of **4.30**. Suppose U.S. researchers believe American manufacturing managers would not rate this reason as highly and conduct a hypothesis test to prove their theory. Alpha is set at **.05**. Data are gathered and the following results are obtained. Use these data and the eight steps of hypothesis testing to determine whether U.S. managers rate this reason significantly lower than the **4.30** mean ascertained in the United Kingdom. Assume from previous studies that the population standard deviation is **0.574**.

✓ 3 4 5 5 4 5 5 4 4 4 4
4 4 4 4 5 4 4 4 3 4 4
4 3 5 4 4 5 4 4 4 5
↑ sample data

$$\bar{x} \approx 4.156$$

Solution

① $H_0: \mu = 4.3$, $H_a: \mu < 4.3$

② σ is known z test & $n = 32$
one tail test

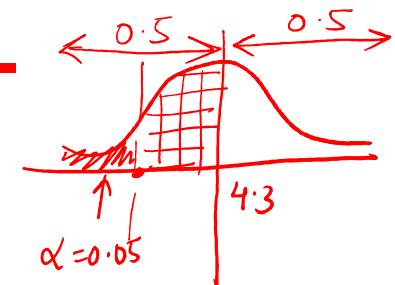
③ $\alpha = 0.05$

④ shaded area = $0.5 - 0.05 = 0.45$

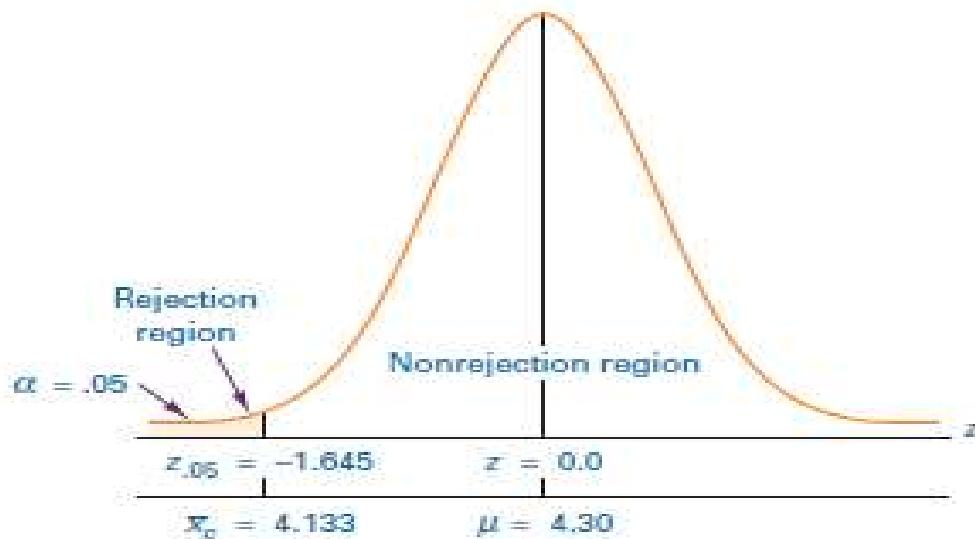
z value for 0.45 -1.645
 $z_c = -1.645$

⑤ observed value of z = $\frac{4.156 - 4.3}{0.574/\sqrt{32}} \approx -1.42$

$\therefore -1.42$ is not in rejection area
we will accept the
NULL hypothesis.
or fail to reject H_0 .



Solution



the probability of getting a z value at least this extreme when the null hypothesis is true is $.5000 - .4222 = .0778$. Hence, the null hypothesis cannot be rejected at $\alpha = .05$ because the smallest value of alpha for which the null hypothesis can be rejected is $.0778$. Had $\alpha = .10$, the decision would have been to reject the null hypothesis.

Using the critical value method: For what sample mean (or more extreme) value would the null hypothesis be rejected? This critical sample mean can be determined by using the critical z value associated with alpha, $z_{.05} = -1.645$.

$$z_c = \frac{\bar{x}_c - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$-1.645 = \frac{\bar{x}_c - 4.30}{\frac{.574}{\sqrt{32}}}$$

$$\bar{x}_c = 4.133$$

The decision rule is that a sample mean less than 4.133 would be necessary to reject the null hypothesis. Because the mean obtained from the sample data is 4.156, the researchers fail to reject the null hypothesis. The preceding diagram includes a scale with the critical sample mean and the rejection region for the critical value method.



Testing hypotheses about a population mean using the t Statistic



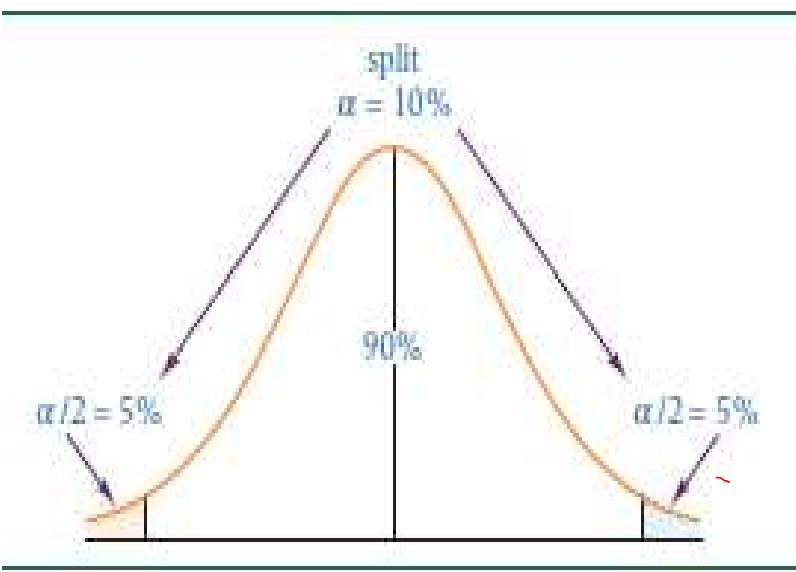
Introduction

- Very often when a business researcher is gathering data to test hypotheses about a single population mean, the value of the population standard deviation is unknown and the researcher must use the sample standard deviation as an estimate of it. In such cases, the z test cannot be used.
- Gosset developed the **t distribution**, which is used instead of the z distribution for doing inferential statistics on the population mean when the population standard deviation is unknown and the population is normally distributed. The formula for the t statistic is

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

Reading the t Distribution Table

- For example, if a 90% confidence interval is being computed the total area in the two tails is 10%. Thus, α is .10 and $\alpha/2$ is .05, as indicated. $n=25$



Degrees of Freedom	$t_{.10}$	$t_{.05}$	$t_{.025}$	$t_{.95}$	$t_{.005}$	$t_{.001}$
...
23						
24						
25						
...

A vertical arrow points from the row labeled '24' down to the value '1.711' in the $t_{.05}$ column.



Reading the t Distribution Table

- To find a value in the t distribution table requires knowing the degrees of freedom; each different value of degrees of freedom is associated with a different t distribution.
- The degrees of freedom for the t statistic presented in this section are computed by $n - 1$.
- The emphasis in the t table is on α , and each tail of the distribution contains of the area under the curve when confidence intervals are constructed.
- For confidence intervals, **the table t value is found in the column under the value of $\alpha/2$ and in the row of the degrees of freedom (df) value**

Example

The U.S. Farmers' Production Company builds large harvesters. For a harvester to be properly balanced when operating, a 25-pound plate is installed on its side. The machine that produces these plates is set to yield plates that average 25 pounds. The distribution of plates produced from the machine is normal. However, the shop supervisor is worried that the machine is out of adjustment and is producing plates that do not average 25 pounds. To test this concern, he randomly selects 20 of the plates produced the day before and weighs them. Table 9.1 shows the weights obtained, along with the computed sample mean and sample standard deviation.

$$\textcircled{1} \quad H_0: \mu = 25 \text{ pounds}$$

$$H_a: \mu \neq 25 \text{ pounds}$$

TABLE 9.1

Weights in Pounds of a Sample of 20 Plates

22.6	22.2	23.2	27.4	24.5
27.0	26.6	28.1	26.9	24.9
26.2	25.3	23.1	24.2	26.1
25.8	30.4	28.6	23.5	23.6
$\bar{x} = 25.51, s = 2.1933, n = 20$				

Solution



② $\because \sigma$ is not known, we will use the t-statistic, 2-tailed test

③ $\alpha = 0.05$

④ $\alpha = 0.05$ (\because 2 tail test)

$$\alpha/2 = 0.025$$

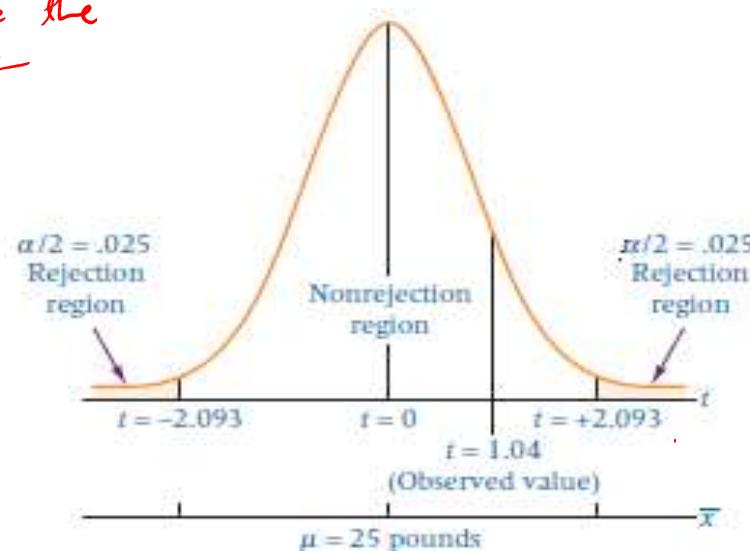
$$n = 20$$

$$df = 20 - 1 = 19$$

$$t_{\alpha/2, 19} = \pm 2.093$$

⑤ observed value of $t = \frac{25.51 - 25}{\sqrt{\frac{2.1933}{20}}} \approx 1.04$

Fail to reject / Accept Null hypo.



Exercise

(HW)



- Figures released by the U.S. Department of Agriculture show that the average size of farms has increased since 1940. In 1940, the mean size of a farm was 174 acres; by 1997, the average size was 471 acres. Between those years, the number of farms decreased but the amount of tillable land remained relatively constant, so now farms are bigger. This trend might be explained, in part, by the inability of small farms to compete with the prices and costs of large-scale operations and to produce a level of income necessary to support the farmers' desired standard of living. Suppose an agribusiness researcher believes the average size of farms has now increased from the 1997 mean figure of 471 acres. To test this notion, she randomly sampled 23 farms across the United States and ascertained the size of each farm from county records. The data she gathered follow. Use a 5% level of significance to test her hypothesis. Assume that number of acres per farm is normally distributed in the population.

445 489 474 505 553 477 454 463 466
557 502 449 438 500 466 477 557 433
545 511 590 561 560

Solution



① $H_0: \mu = 471$ $H_1: \mu \geq 471$

② $\because \sigma$ is not known t test

③ $\alpha = 0.05$

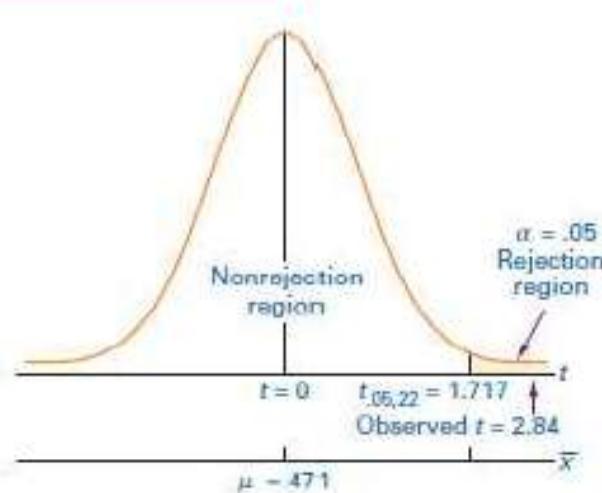
④ One tail test df = 22

$$t_{0.05, 22} = 1.717$$

⑤ Observed t value

$$\bar{x} = 498.78, s = 46.94$$

$$t = \frac{498.78 - 471}{46.94/\sqrt{23}} = 2.84$$



\therefore Observed value $>$ than the critical value

Reject Null hypothesis



BITS Pilani
Pilani Campus

Hypothesis Testing

Akanksha Bharadwaj
Asst. Professor, CS/IS Department





The Difference In Two Means Using The Z Statistic

- In some research designs, the sampling plan calls for selecting two independent samples, calculating the sample means and using the difference in the two sample means to estimate or test the **difference in the two population means**.
- The object might be to determine whether the two samples come from the same population or, if they come from different populations, to determine the amount of difference in the populations.
- This type of analysis can be used to determine, for example, whether the **effectiveness of two brands** of toothpaste differs or whether two brands of tires wear differently.

Central Limit Theorem

- The central limit theorem states that the difference in two sample means, $\bar{x}_1 - \bar{x}_2$, is normally distributed for large sample sizes (both n_1 and $n_2 \geq 30$) regardless of the shape of the populations.

$$\begin{aligned}\mu_{\bar{x}_1 - \bar{x}_2} &= \mu_1 - \mu_2 \\ \sigma_{\bar{x}_1 - \bar{x}_2} &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\end{aligned}$$

- These expressions lead to a z formula for the difference in two sample means.

***z* FORMULA FOR THE
DIFFERENCE IN TWO
SAMPLE MEANS
(INDEPENDENT SAMPLES
AND POPULATION
VARIANCES KNOWN) (10.1)**

where

μ_1 = the mean of population 1
 μ_2 = the mean of population 2
 n_1 = size of sample 1
 n_2 = size of sample 2

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$





Hypothesis Testing

- As a specific example, suppose we want to conduct a hypothesis test to determine whether the average annual wage for an advertising manager is different from the average annual wage of an auditing manager.
- Because we are testing to determine whether the means are different, it might seem logical that the null and alternative hypotheses would be

$$H_0: \mu_1 = \mu_2 \rightarrow H_0: \mu_1 - \mu_2 = 0$$

$$H_a: \mu_1 \neq \mu_2 \quad H_a: \mu_1 - \mu_2 \neq 0$$



Exercise

A sample of 87 professional working women showed that the average amount paid annually into a private pension fund per person was \$3352. The population standard deviation is \$1100. A sample of 76 professional working men showed that the average amount paid annually into a private pension fund per person was \$5727, with a population standard deviation of \$1700. A women's activist group wants to "prove" that women do not pay as much per year as men into private pension funds. If they use $\alpha = .001$ and these sample data, will they be able to reject a null hypothesis that women annually pay the same as or more than men into private pension funds? Use the eight-step hypothesis-testing process.

① $H_0: \mu_w - \mu_m = 0$

$$H_a/H_1: \mu_w - \mu_m < 0$$

② z-test

Solution

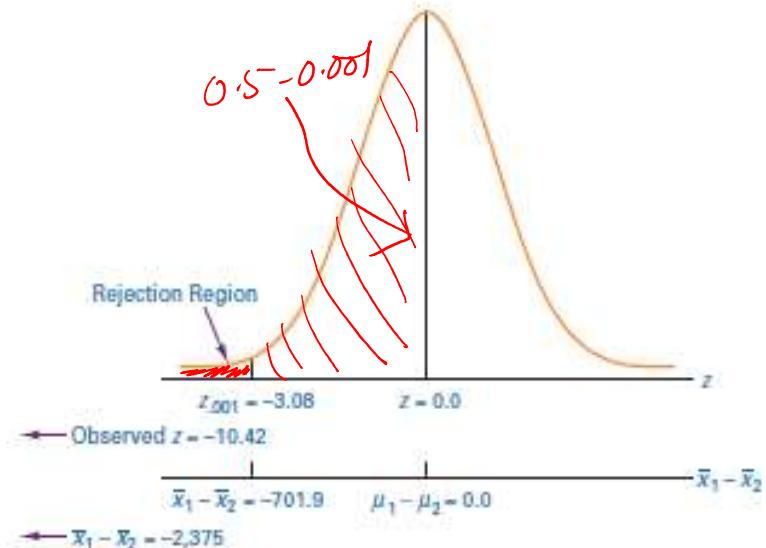
③ $\alpha = 0.001$, one-tailed test

④ critical value?
z-value for 0.499 is -3.08

⑤ observed value =?

$$Z = \frac{(3352 - 5727) - 0}{\sqrt{\frac{1100^2}{87} + \frac{1700^2}{76}}} =$$

$$= \frac{-2375}{227.9} = -10.42$$



Reject the NULL Hypo.

If this problem were worked by the critical value method, what critical value of the difference in the two means would have to be surpassed to reject the null hypothesis for a table z value of -3.08? The answer is

$$\begin{aligned}(\bar{X}_1 - \bar{X}_2)_c &= (\mu_1 - \mu_2) - Z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\&= 0 - 3.08(227.9) = -701.9\end{aligned}$$

The difference in sample means would need to be at least 701.9 to reject the null hypothesis. The actual sample difference in this problem was -2375 (3352 - 5727), which is considerably larger than the critical value of difference. Thus, with the critical value method also, the null hypothesis is rejected.

Confidence Intervals

- Sometimes being able to estimate the difference in the means of two populations is valuable.
- Algebraically, formula 10.1 can be manipulated to produce a formula for constructing confidence intervals for the difference in two population means.

CONFIDENCE INTERVAL TO
ESTIMATE $\mu_1 - \mu_2$ (10.2)

$$(\bar{x}_1 - \bar{x}_2) - z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$



BITS Pilani
Pilani Campus



SS ZG536, ADV STAT TECHNIQUES FOR ANALYTICS

Contact Session 7



BITS Pilani



Testing Hypotheses About A Proportion

Introduction

- To validly use this test, the sample size must be large enough such that $n*p \geq 5$ and $n*q \geq 5$.

z TEST OF A POPULATION PROPORTION (9.4)

where

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}}$$

\hat{p} = sample proportion

p = population proportion

$q = 1 - p$



Example

- A manufacturer believes exactly 8% of its products contain at least one minor flaw. Suppose a company researcher wants to test this belief. The null and alternative hypotheses are

$$\underline{H_0: p = .08}$$

$$\underline{H_a: p \neq .08}$$

- This test is two-tailed because the hypothesis being tested is whether the proportion of products with at least one minor flaw is .08. Alpha is selected to be .10

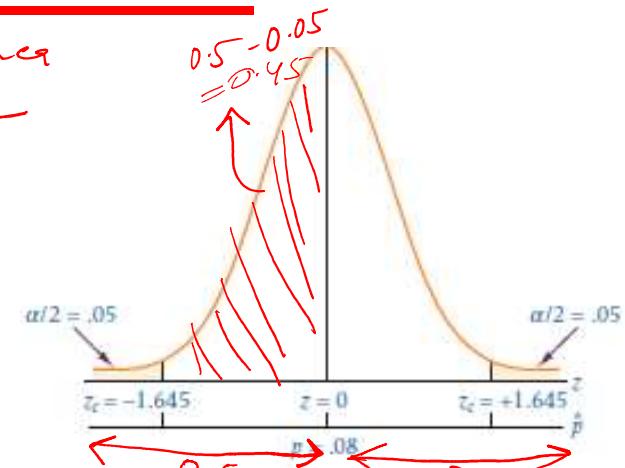
The business researcher randomly selects a sample of 200 products, inspects each item for flaws, and determines that 33 items have at least one minor flaw. Calculating the sample proportion gives:

$$\hat{p} = ? = \frac{33}{200} = 0.165$$

observed value of $Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.165 - 0.08}{\sqrt{\frac{(0.08)(0.92)}{200}}}$

$$= \frac{0.085}{0.019} = 4.43$$

$$Z_c \text{ for } 0.45 \text{ area} \\ = \pm 1.645$$



For the business researcher to reject the null hypothesis, the observed z value must be greater than 1.645 or less than -1.645

Reject null hypo.

Exercise

A survey of the morning beverage market shows that the primary breakfast beverage for 17% of Americans is milk. A milk producer in Wisconsin, where milk is plentiful, believes the figure is higher for Wisconsin. To test this idea, she contacts a random sample of 550 Wisconsin residents and asks which primary beverage they consumed for breakfast that day. Suppose 115 replied that milk was the primary beverage. Using a level of significance of .05, test the idea that the milk figure is higher for Wisconsin.

① $H_0: p = 0.17$
 $H_a: p > 0.17$

② z-test , one-tailed test

③ $\alpha = 0.05$

Solution

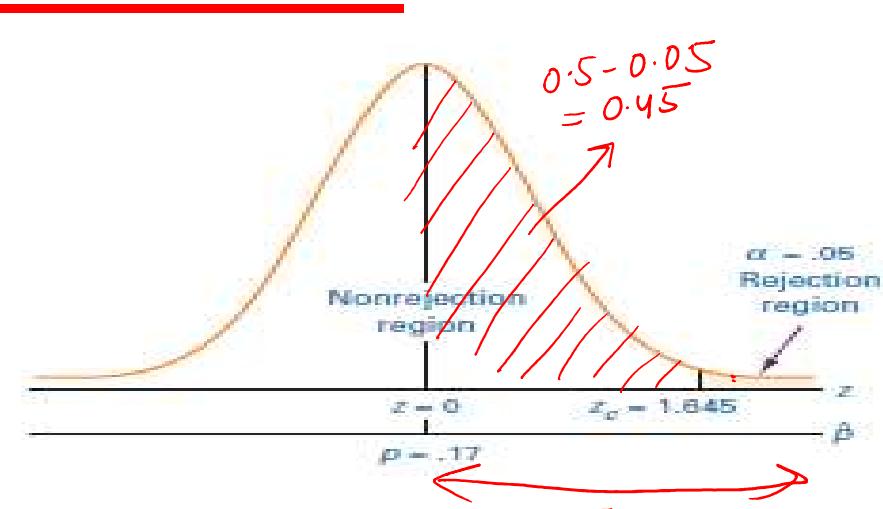
(4) z value for 0.45 area
 $Z_c = 1.645$

(5) observed value of $z = ?$
 $\hat{p} = \frac{115}{550} = 0.209$

$$z = \frac{0.209 - 0.17}{\sqrt{\frac{0.17 \times 0.83}{550}}} = \frac{0.039}{0.016}$$

$$= 2.44$$

Reject Null hypo.





Solving For Type II Errors

- A researcher reaches the statistical conclusion to fail to reject the null hypothesis
- If the null hypothesis is true, the researcher makes a correct decision.
- If the null hypothesis is false, then the result is a Type II error.



-
- Determining the probability of committing a Type II error is more complex than finding the probability of committing a Type I error.
 - The probability of committing a Type I error either is given in a problem or is stated by the researcher before proceeding with the study.
 - A Type II error, , varies with possible values of the alternative parameter

Example

- Suppose a researcher is conducting a statistical test on the following hypotheses.

$$H_0: \mu = 12 \text{ ounces}$$

$$H_a: \mu < 12 \text{ ounces}$$

- Often, when the null hypothesis is false, the value of the alternative mean is unknown, so the researcher will compute the probability of committing Type II errors for several possible values.
- Suppose that, in testing the preceding hypotheses, a sample of 60 cans of beverage yields a sample mean of 11.985 ounces. Assume that the population standard deviation is 0.10 ounces. From $\alpha=0.05$ and a one-tailed test, the table $z_{.05}$ value is -1.645. The observed z value from sample data is

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{11.985 - 12}{0.1 / \sqrt{60}} = -1.16$$

Accept NULL hypo.



- What is the probability of committing a Type II error in this problem if the population mean actually is 11.99?

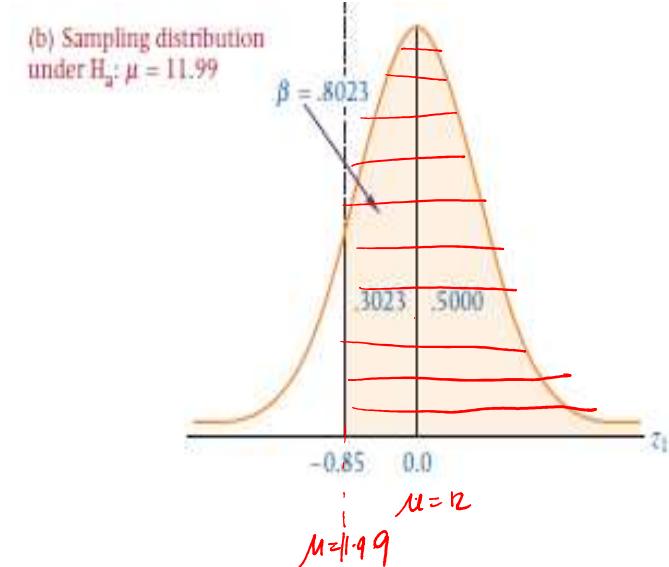
$$Z_c = -1.645$$
$$\bar{x}_c = -z_c \times \frac{\sigma}{\sqrt{n}} + \mu = -1.6425 \times \frac{0.1}{\sqrt{60}} + 12$$

$$\bar{x}_c = 11.979$$

If μ actually equals 11.99ounces, what is the probability of failing to reject $\mu=12$ ounces when 11.979 ounces is the critical value?

$$Z_1 = \frac{11.979 - 11.99}{0.10/\sqrt{60}} = -0.85$$

$$\text{prob. of type 2 error} = 0.5 + 0.3023 \\ = 0.8023$$





Exercise

Re-compute the probability of committing a Type II error for the soft drink example if the alternative mean is 11.96 ounces.

$$Z_c = -1.645 \quad \bar{x}_c = 11.979$$

$$\begin{aligned} Z_1 &= ? \\ &= \frac{11.979 - 11.96}{\frac{0.10}{\sqrt{60}}} = 1.47 \end{aligned}$$

Solution

The null hypothesized mean is still 12 ounces, the critical value is still 11.979 ounces, and $n = 60$.

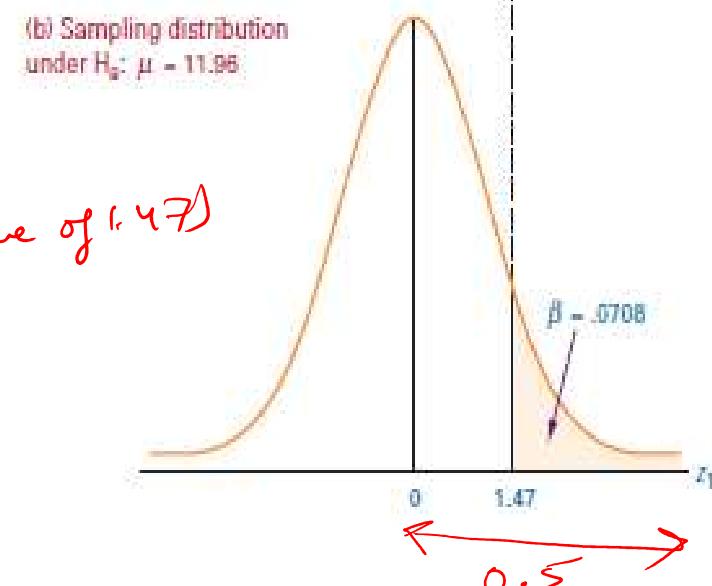
$$Z_1 = 1.47$$

area to the right of Z_1
is the rejection region

$$= 0.5 - (\text{area for } Z_1 \text{ value of } 1.47)$$

$$= 0.5 - 0.4292$$

$$\beta = 0.0708$$



Exercise (HW)

Suppose you are conducting a two-tailed hypothesis test of proportions. The null hypothesis is that the population proportion is .40. The alternative hypothesis is that the population proportion is not .40. A random sample of 250 produces a sample proportion of .44. With alpha of .05, the table z value for $\alpha/2$ is ± 1.96 . The observed z from the sample information is

$$z = \frac{|\hat{p} - p|}{\sqrt{\frac{p \cdot q}{n}}} = \frac{.44 - .40}{\sqrt{\frac{.031}{250}}} = 1.29$$

Thus the null hypothesis is not rejected. Either a correct decision is made or a Type II error is committed. Suppose the alternative population proportion really is .36. What is the probability of committing a Type II error?

Solution

Solve for the critical value of the proportion.

$$I_c = \frac{\hat{p}_c - p}{\sqrt{\frac{p \cdot q}{n}}}$$
$$\pm 1.96 = \frac{\hat{p}_c - .40}{\sqrt{\frac{(.40)(.60)}{250}}}$$
$$\hat{p}_c = .40 \pm .06$$

The critical values are .34 on the lower end and .46 on the upper end. The alternative population proportion is .36. The following diagram illustrates these results and the remainder of the solution to this problem.

Solving for the area between $p_0 = .34$ and $p_1 = .36$ yields

$$z_1 = \frac{.34 - .36}{\sqrt{\frac{(.36)(.64)}{250}}} = -0.66$$

The area associated with $z_1 = -0.66$ is .2454.

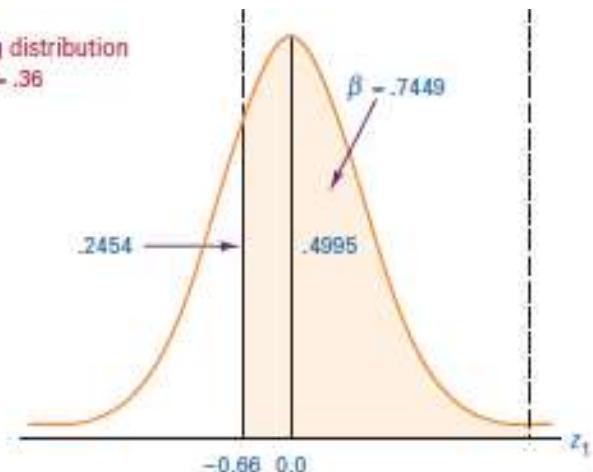
The area between .36 and .46 of the sampling distribution under $H_0: p = .36$ (graph (b)) can be solved for by using the following z value.

$$z = \frac{.46 - .36}{\sqrt{\frac{(.36)(.64)}{250}}} = 3.29$$

The area from Table A.5 associated with $z = 3.29$ is .4995. Combining this value with the .2454 obtained from the left side of the distribution in graph (b) yields the total probability of committing a Type II error:

$$.2454 + .4995 = .7449$$

(b) Sampling distribution under $H_0: p = .36$





BITS Pilani



Statistical Inferences For Two Related Populations



Introduction

- In this section, a method is presented to analyze **dependent samples** or related samples.
- Some researchers refer to this test as the **matched-pairs** test. Others call it the ***t test for related measures*** or the ***correlated t test***.
- **Example:** Sometimes as an experimental control mechanism, the same person or object is measured both before and after a treatment. Certainly, the after measurement is not independent of the before measurement because the measurements are taken on the same person or object in both case



Hypothesis Testing

- The approach to analyzing two related samples is different from the techniques used to analyze independent samples.
- The matched-pairs test for related samples requires that the **two samples** be the **same size** and that the individual related scores be matched.

**t FORMULA TO TEST THE
DIFFERENCE IN TWO
DEPENDENT POPULATIONS
(10.5)**

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

$$df = n - 1$$

where

n = number of pairs

d = sample difference in pairs

D = mean population difference

s_d = standard deviation of sample difference

\bar{d} = mean sample difference

-
- This t test for dependent measures uses the sample difference, d , between individual matched sample values as the basic measurement of analysis instead of individual sample values.
 - Analysis of the d values effectively converts the problem from a two-sample problem to a single sample of differences, which is an adaptation of the single-sample means formula.

FORMULAS FOR \bar{d} AND s_d (10.6 AND 10.7)

$$\bar{d} = \frac{\sum d}{n}$$
$$s_d = \sqrt{\frac{\sum (d - \bar{d})^2}{n-1}} = \sqrt{\frac{g d^2 - \frac{(\sum d)^2}{n}}{n-1}}$$

Example

- Suppose a stock market investor is interested in determining whether there is a significant difference in the P/E (price to earnings) ratio for companies from one year to the next. In an effort to study this question, the investor randomly samples nine companies from the *Handbook of Common Stocks* and records the P/E ratios for each of these companies at the end of year 1 and at the end of year 2. The data are shown in Table 10.5. These data are related data because each P/E value for year 1 has a corresponding year 2 measurement on the same company. Because no prior information indicates whether P/E ratios have gone up or down, the hypothesis tested is two tailed. Assume $\alpha=.01$. Assume that differences in P/E ratios are normally distributed in the population.

Company	Year 1 P/E Ratio	Year 2 P/E Ratio
1	8.9	12.7
2	38.1	45.4
3	43.0	10.0
4	34.0	27.2
5	34.5	22.8
6	15.2	24.1
7	20.3	32.3
8	19.9	40.1
9	61.9	106.5

Solution

① $H_0: D = 0$
 $H_1: D \neq 0$

② t -test, 2 tailed test

③ $\alpha = 0.01$
 $\alpha/2 = 0.005$ $n = 9$
 $\therefore df = 9 - 1 = 8$

④ $t_{0.005, 8} = \pm 3.355$

Company	Year 1 P/E	Year 2 P/E	d
1	8.9	12.7	-3.8
2	38.1	45.4	-7.3
3	43.0	10.0	33.0
4	34.0	27.2	6.8
5	34.5	22.8	11.7
6	15.2	24.1	-8.9
7	20.3	32.3	-12.0
8	19.9	40.1	-20.2
9	61.9	106.5	-44.6

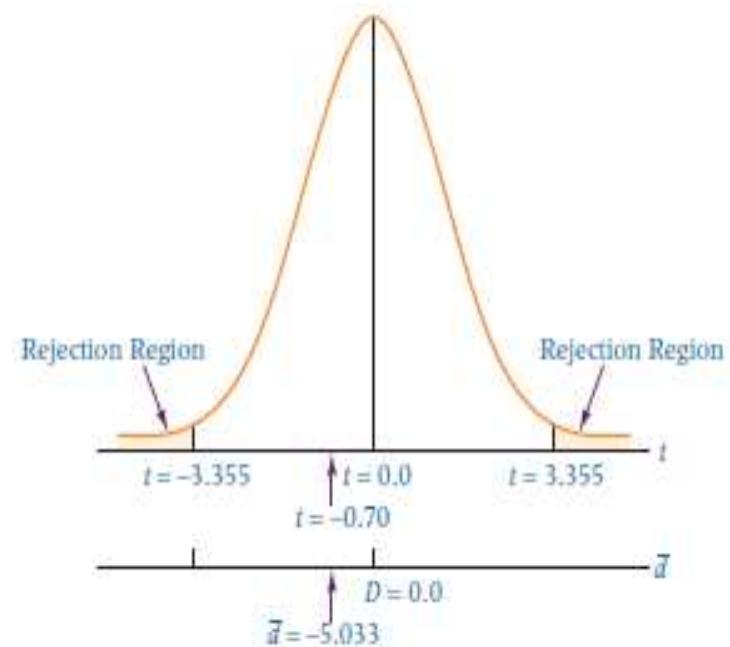
$\bar{d} = -5.033, s_d = 21.599, n = 9$

Observed $t = \frac{-5.033 - 0}{\frac{21.599}{\sqrt{9}}} = -0.70$

③ observed value of t

$$t = \frac{-5.033 - 0}{\frac{21.599}{\sqrt{9}}} = -0.70$$

fail to reject NULL hypo.





Exercise (HW)

- Let us use this hypothetical study in which consumers are asked to rate a company both before and after viewing a video on the company twice a day for a week. Use an alpha of .05 to test to determine whether there is a significant increase in the ratings of the company after the one-week video treatment. Assume that differences in ratings are normally distributed in the population.

Individual	Before	After
1	32	39
2	11	15
3	21	35
4	17	13
5	30	41
6	38	39
7	14	22

$$\textcircled{1} \quad H_0: D = 0 \\ H_a: D < 0$$

Solution

Step 1

$$H_0: D = 0$$

$$H_a: D < 0$$

Step 2

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

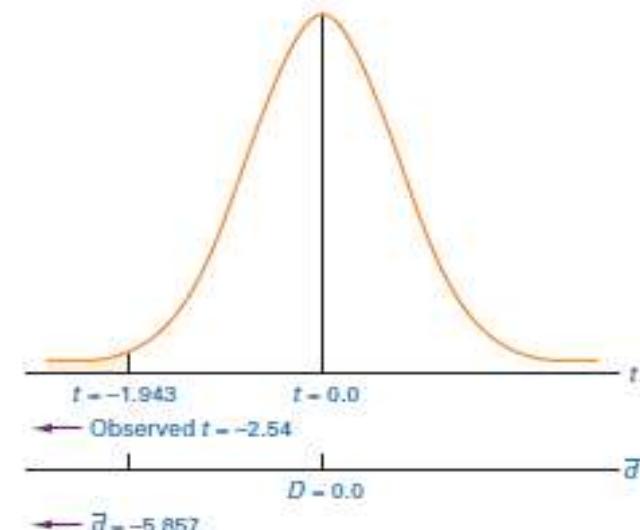
Step 3

$$\alpha = 0.05$$

Step 4

$$df = n-1 = 7-1 = 6 \quad [\text{one tail test}]$$

$$t_{0.05, 6} = -1.943$$





Step 5 $\bar{d} = -5.857$ $s_d = 6.0945$

Step 6 $t = \frac{-5.857 - 0}{\frac{6.0945}{\sqrt{7}}} = -2.54$

Individual	Before	After	d
1	32	39	-7
2	11	15	-4
3	21	35	-14
4	17	13	+4
5	30	41	-11
6	38	39	-1
7	14	22	-8
$\bar{d} = -5.857$		$s_d = 6.0945$	

Step 7 Action:- $\because -2.547$ is less than the critical value of -1.943

REJECT the NULL hypothesis.



Confidence Intervals

- Sometimes a researcher is interested in estimating the mean difference in two populations for related samples.
- A confidence interval for D , the mean population difference of two related samples, can be constructed by algebraically rearranging formula (10.5), which was used to test hypotheses about D .

CONFIDENCE INTERVAL
FORMULA TO ESTIMATE
THE DIFFERENCE IN
RELATED POPULATIONS,
 D (10.8)

$$\bar{d} - t \frac{s_d}{\sqrt{n}} \leq D \leq \bar{d} + t \frac{s_d}{\sqrt{n}}$$
$$df = n - 1$$



Example

- The sale of new houses apparently fluctuates seasonally. Superimposed on the seasonality are economic and business cycles that also influence the sale of new houses. In certain parts of the country, new-house sales increase in the spring and early summer and drop off in the fall. Suppose a national real estate association wants to estimate the average difference in the number of new-house sales per company in Indianapolis between 2008 and 2009. To do so, the association randomly selects 18 real estate firms in the Indianapolis area and obtains their new-house sales figures for May 2008 and May 2009. The numbers of sales per company are shown in Table 10.7. Using these data, the association's analyst estimates the average difference in the number of sales per real estate company in Indianapolis for May 2008 and May 2009 and constructs a 99% confidence interval. The analyst assumes that differences in sales are normally distributed in the population.

Realtor	May 2008	May 2009
1	8	11
2	19	30
3	5	6
4	9	13
5	3	5
6	0	4
7	13	15
8	11	17
9	9	12
10	5	12
11	8	6
12	2	5
13	11	10
14	14	22
15	7	8
16	12	15
17	6	12
18	10	10

Solution

99% confidence interval

$$\alpha = 1 - 0.99 = 0.01$$

$$\alpha/2 = 0.005$$

$$n = 18 \Rightarrow df = 17$$

$$\bar{d} - \frac{t \times s_d}{\sqrt{n}} \leq D \leq \bar{d} + \frac{t \times s_d}{\sqrt{n}}$$

$$t_{0.005, 17} = 2.898$$

$$-5.625 \leq D \leq -1.153$$

Realtor	May 2008	May 2009	d
1	8	11	-3
2	19	30	-11
3	5	6	-1
4	9	13	-4
5	3	5	-2
6	0	4	-4
7	13	15	-2
8	11	17	-6
9	9	12	-3
10	5	12	-7
11	8	6	+2
12	2	5	-3
13	11	10	+1
14	14	22	-8
15	7	8	-1
16	12	15	-3
17	6	12	-6
18	10	10	0

$\bar{d} = -3.389$ and $s_d = 3.274$



BITS Pilani



Statistical Inferences About Two Population Proportions, $P_1 - P_2$



Introduction

- Sometimes a researcher wishes to make inferences about the difference in two population proportions.
- This type of analysis has many applications in business, such as comparing the market share of a product for two different markets, studying the difference in the proportion of female customers in two different geographic regions, or comparing the proportion of defective products from one period to another

Z-test for difference between proportions

Z-test



Difference between proportion of two population ($P_1 - P_2$)

Assumptions

Assume that the samples are drawn from normal population

The sample size should be more than or equal to 30

Subjects should be selected randomly

Two groups should be independent of each other

Z formula

**z FORMULA FOR THE
DIFFERENCE IN TWO
POPULATION
PROPORTIONS (10.9)**

where

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 \cdot q_1}{n_1} + \frac{p_2 \cdot q_2}{n_2}}}$$

\hat{p}_1 = proportion from sample 1

\hat{p}_2 = proportion from sample 2

n_1 = size of sample 1

n_2 = size of sample 2

p_1 = proportion from population 1

p_2 = proportion from population 2

$q_1 = 1 - p_1$

$q_2 = 1 - p_2$



Hypothesis Testing

- The sample proportions are combined by using a weighted average to produce \bar{p} , which, in conjunction with \bar{q} and the sample sizes, produces a point estimate of the standard deviation of the difference in sample proportions.

**FORMULA TO TEST THE
DIFFERENCE IN
POPULATION
PROPORTIONS (10.10)**

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{(\bar{p} \cdot \bar{q}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$ and $\bar{q} = 1 - \bar{p}$



Example

- Do consumers and CEOs have different perceptions of ethics in business? A group of researchers attempted to determine whether there was a difference in the proportion of consumers and the proportion of CEOs who believe that fear of getting caught or losing one's job is a strong influence of ethical behavior. In their study, they found that 57% of consumers said that fear of getting caught or losing one's job was a strong influence on ethical behavior, but only 50% of CEOs felt the same way. Suppose these data were determined from a sample of 755 consumers and 616 CEOs. Does this result provide enough evidence to declare that a significantly higher proportion of consumers than of CEOs believe fear of getting caught or losing one's job is a strong influence on ethical behaviour?

$$\begin{array}{c} \text{consumers} \leftarrow \\ \textcircled{1} \quad H_0: p_1 - p_2 = 0 \rightarrow \text{CEOs} \\ H_a: p_1 - p_2 > 0 \end{array}$$



Solution

② z-test , one-tailed test

③ $\alpha = 0.1$

$$Z_c = 1.28 \text{ (for area } 0.4)$$

$$n_1 = 755$$

$$\hat{p}_1 = 0.57$$

$$n_2 = 616$$

$$\hat{p}_2 = 0.50$$

$$\bar{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{755 \times 0.57 + 616 \times 0.50}{755 + 616}$$

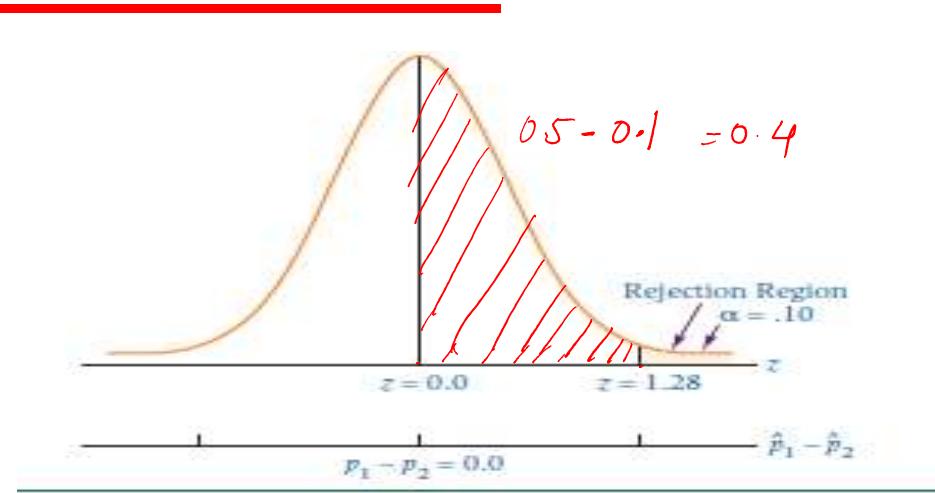
$$= 0.539$$

$$\bar{q} = 1 - \bar{p} = 0.461$$

observed value of Z =

$$\frac{(0.57 - 0.50) - 0}{\sqrt{0.539 \times 0.461} \left(\frac{1}{755} + \frac{1}{616} \right)} = 2.586$$

Reject NULL
hypo.





Exercise (HW)

$$\bar{P} = \frac{24 + 39}{100 + 95} = \frac{63}{195} = 0.323$$

A study of female entrepreneurs was conducted to determine their definition of success. The women were offered optional choices such as happiness/self-fulfillment, sales/profit, and achievement/challenge. The women were divided into groups according to the gross sales of their businesses. A significantly higher proportion of female entrepreneurs in the \$100,000 to \$500,000 category than in the less than \$100,000 category seemed to rate sales/profit as a definition of success. Suppose you decide to test this result by taking a survey of your own and identify female entrepreneurs by gross sales. You interview 100 female entrepreneurs with gross sales of less than \$100,000, and 24 of them define sales/profit as success. You then interview 95 female entrepreneurs with gross sales of \$100,000 to \$500,000, and 39 cite sales/profit as a definition of success. Use this information to test to determine whether there is a significant difference in the proportions of the two groups that define success as sales/profit. Use alpha = .01.

$$z_{\text{c for area}} = \frac{(0.5 - 0.323)}{\sqrt{0.323 \cdot 0.677 / 195}} = 2.495$$



Solution

Step 1 $H_0: p_1 - p_2 = 0$

$H_a: p_1 - p_2 \neq 0$

Step 2 z-test for difference in proportions.

Step 3 & 4 $\alpha = 0.01$ $Z_{\alpha/2} = 0.005$

critical value $\leftarrow Z_{0.005} = \pm \underline{\underline{2.575}}$

[for area $0.5 - 0.005 = 0.495$]

Step 5 $n_1 = 100$

$x_1 = 24$

$\hat{p}_1 = \frac{24}{100} = 0.24$

$n_2 = 95$

$x_2 = 39$

$\hat{p}_2 = \frac{39}{95} = 0.41$

here,
 $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{24 + 39}{100 + 95}$

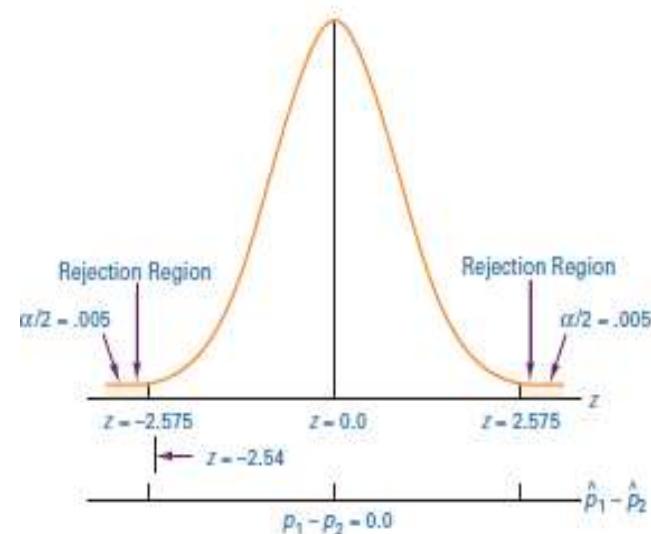
$\bar{p} = \frac{63}{195} = 0.323$

Step 6

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}\hat{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{(0.24 - 0.41) - (0)}{\sqrt{(0.323)(0.677) \left(\frac{1}{100} + \frac{1}{95} \right)}} = -\underline{\underline{2.54}}$$

Step 7

\therefore observed value is in the non-rejection region
NULL hypothesis is not rejected.





Confidence Intervals

- Sometimes in business research the investigator wants to estimate the difference in two population proportions.
- For example, what is the difference, if any, in the population proportions of workers in the Midwest who favor union membership and workers in the South who favor union membership?

CONFIDENCE INTERVAL TO
ESTIMATE $p_1 - p_2$ (10.11)

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \leq p_1 - p_2 \leq (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$



Example

Suppose that in an attempt to target its clientele, managers of a supermarket chain want to determine the difference between the proportion of morning shoppers who are men and the proportion of after-5 P.M. shoppers who are men. Over a period of two weeks, the chain's researchers conduct a systematic random sample survey of 400 morning shoppers, which reveals that 352 are women and 48 are men. During this same period, a systematic random sample of 480 after-5 P.M. shoppers reveals that 293 are women and 187 are men. Construct a 98% confidence interval to estimate the difference in the population proportions of men.

Morning Shoppers	After-5 P.M. Shoppers
$n_1 = 400$	$n_2 = 480$
$x_1 = 48$ men	$x_2 = 187$ men
$\hat{p}_1 = .12$	$\hat{p}_2 = .39$
$\hat{q}_1 = .88$	$\hat{q}_2 = .61$

for 98%. CI



Solution

$$\alpha = ? \\ = 1 - 0.98 = 0.02$$

$$\alpha/2 = 0.01$$

$$\text{value of } z \text{ for area } 0.5 - 0.01 \\ = 0.49$$

$$z = 2.33$$

$$(0.12 - 0.39) - 2.33 \sqrt{\frac{(0.12)(0.88)}{480} + \frac{0.39 \times 0.61}{480}} \leq (P_1 - P_2) \leq (0.12 - 0.39) + 2.33 \sqrt{\frac{0.12 \times 0.88 + 0.39 \times 0.61}{480}}$$

$$0.334 \geq (P_1 - P_2) \geq -0.206$$



BITS Pilani
Pilani Campus

ANOVA

Akanksha Bharadwaj
Asst. Professor, CS/IS Department





Need for ANOVA

- In the machine operator example, **is it possible to analyze the four samples by using a t test** for the difference in two sample means?
- These four samples would require ${}^4C_2 = 6$ individual t tests to accomplish the analysis of two groups at a time.
- Recall that if $\alpha = .05$ for a particular test, there is a 5% chance of rejecting a null hypothesis that is true (i.e., committing a Type I error).
- If enough tests are done, eventually one or more null hypotheses will be falsely rejected by chance.
- Hence, $\alpha = .05$ is valid only for one t test. In this problem, with six t tests, the error rate compounds, so when the analyst is finished with the problem there is a much greater than .05 chance of committing a Type I error.



Analysis of Variance

- When there are more than two groups to be compared, it is not correct to compare the groups in pairs, as this type of comparison will not take the within variability into consideration
- The Analysis procedure used in such comparisons is known as ANALYSIS OF VARIANCE



Example

- As an example of a completely randomized design, suppose a researcher decides to analyze the effects of the machine operator on the valve opening measurements of valves produced in a manufacturing plant, like those shown in Table below.

6.26	6.19	6.33	6.26	6.50
6.19	6.44	6.22	6.54	6.23
6.29	6.40	6.23	6.29	6.58
6.27	6.38	6.58	6.31	6.34
6.21	6.19	6.36	6.56	

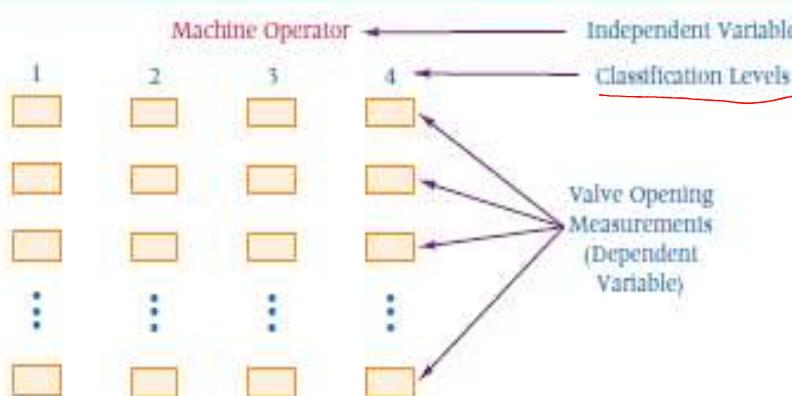
$$\bar{x} = 6.34 \text{ Total Sum of Squares Deviation} = SST = \sum(x_i - \bar{x})^2 = .3915$$

- The independent variable in this design is machine operator.

Example continued

independent variable

- Suppose further that four different operators operate the machines. These four machine operators are the levels of treatment, or classification, of the independent variable.
- The dependent variable is the opening measurement of the valve.
- Figure below shows the structure of this completely randomized design.
- Table below contains the valve opening measurements for valves produced under each operator.



Valve Openings by Operator

1	2	3	4
6.33	6.26	6.44	6.29
6.26	6.36	6.38	6.23
6.31	6.23	6.58	6.19
6.29	6.27	6.54	6.21
6.40	6.19	6.56	
	6.50	6.34	
	6.19	6.58	
		6.22	

Machine operators

One Way ANOVA

ANOVA

Testing equality of k group means against not equal

- Samples are drawn from normal population
- The population variances should be equal
- The sample size should be less than 30 (i.e., $n < 30$)
- Groups should be independent
- Subjects should be allocated randomly to both groups
- However even if sample size more than 30 (i.e., $n > 30$) ANOVA should be continue to apply, because of central limit theorem it approaches normal.



Hypothesis in ANOVA

- In general, if k samples are being analyzed, the following hypotheses are being tested in a one-way ANOVA.

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

$H_1:$ At least one of the means is different from the others.

- The null hypothesis states that the population means for all treatment levels are equal.
- Because of the way the alternative hypothesis is stated, if even one of the population means is different from the others, the null hypothesis is rejected.



Testing hypotheses

Testing these hypotheses by using one-way ANOVA is accomplished by partitioning the total variance of the data into the following two variances.

1. The variance resulting from the treatment (columns)



2. The error variance, or that portion of the total variance unexplained by the treatment



Total Sum of Squares of Variation

The error variation can be viewed at this point as variation due to individual differences within treatment groups.

$$\sum_{i=1}^{n_i} \sum_{j=1}^C \underline{(x_{ij} - \bar{x})^2} = \sum_{j=1}^C n_j (\bar{x}_j - \bar{x})^2 + \sum_{i=1}^{n_i} \sum_{j=1}^C \underline{(x_{ij} - \bar{x}_j)^2}$$

where

SST = total sum of squares

SSC = sum of squares column (treatment)

SSE = sum of squares error

i = particular member of a treatment level

j = a treatment level

C = number of treatment levels

n_j = number of observations in a given treatment level

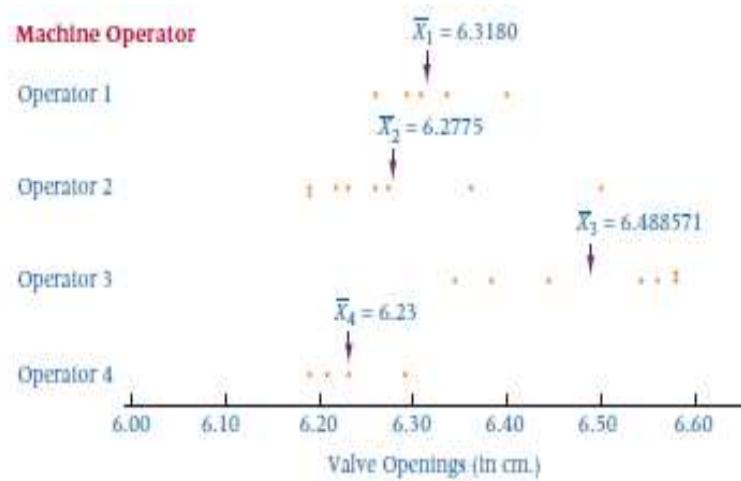
\bar{x} = grand mean

\bar{x}_j = mean of a treatment group or level

x_{ij} = individual value

Example

- Figure below displays the data from the machine operator example in terms of treatment level.
- Note the variation of values (x) *within* each treatment level. Now examine the variation between levels 1 through 4 (the difference in the machine operators).





Assumptions

- Analysis of variance is used to determine statistically whether the variance between the treatment level means is greater than the variances within levels (error variance).
- Several important assumptions underlie analysis of variance:
 1. Observations are drawn from normally distributed populations.
 2. Observations represent random samples from the populations.
 3. Variances of the populations are equal.

These assumptions are similar to those for using the *t* test for independent samples



Formula

FORMULAS FOR COMPUTING A ONE-WAY ANOVA

$$SSC = \sum_{j=1}^C n_j (\bar{x}_j - \bar{x})^2$$

$$SSE = \sum_{i=1}^n \sum_{j=1}^C (x_{ij} - \bar{x}_j)^2$$

$$SST = \sum_{i=1}^n \sum_{j=1}^C (x_{ij} - \bar{x})^2$$

$$df_C = C - 1$$

$$df_E = N - C$$

$$df_T = N - 1$$

$$MSC = \frac{SSC}{df_C}$$

$$MSE = \frac{SSE}{df_E}$$

$$F = \frac{MSC}{MSE}$$

where

i = a particular member of a treatment level

j = a treatment level

C = number of treatment levels

n_j = number of observations in a given treatment level

\bar{x} = grand mean

\bar{x}_j = column mean

x_{ij} = individual value



- SST is the total sum of squares and is a measure of all variation in the dependent variable.
- As shown previously, SST contains both SSC and SSE and can be partitioned into SSC and SSE.
- MSC, MSE, and MST are the mean squares of column, error, and total respectively.
- Mean square is an average and is computed by dividing the sum of squares by the degrees of freedom.
- Finally, the F value is determined by dividing the treatment variance (MSC) by the error variance (MSE).
- As discussed earlier, the F is a ratio of two variances.
- In the ANOVA situation, the **F value** is a *ratio of the treatment variance to the error variance*.



Machine operator Example

Machine Operator	1	2	3	4
6.33	6.26	6.44	6.29	
6.26	6.36	6.38	6.23	
6.31	6.23	6.58	6.19	
6.29	6.27	6.54	6.21	
6.40	6.19	6.56		
	6.50	6.34		
	6.19	6.58		
	6.22			

Treatment levels

$$SSC = \sum_{j=1}^n n_j (\bar{x}_j - \bar{x})^2$$

$$SSE = \sum_{i=1}^n \sum_{j=1}^c (x_{ij} - \bar{x}_j)^2$$

$$SST = \sum_{i=1}^n \sum_{j=1}^c (x_{ij} - \bar{x})^2$$

$$\begin{array}{lll}
 T_j: & T_1 = 31.59 & T_2 = 50.22 & T_3 = 45.42 & T_4 = 24.92 & T = 152.15 \\
 n_j: & n_1 = 5 & n_2 = 8 & n_3 = 7 & n_4 = 4 & N = 24 \\
 \bar{x}_j: & \bar{x}_1 = 6.318 & \bar{x}_2 = 6.2775 & \bar{x}_3 = 6.488571 & \bar{x}_4 = 6.230 & \bar{x} = 6.339583
 \end{array}$$



$$\begin{aligned}SSC = \sum_{j=1}^C n_j (\bar{x}_j - \bar{\bar{x}})^2 &= [5(6.318 - 6.339583)^2 + 8(6.2775 - 6.339583)^2 \\&\quad + 7(6.488571 - 6.339583)^2 + 4(6.230 - 6.339583)^2] \\&= 0.00233 + 0.03083 + 0.15538 + 0.04803 \\&= \underline{\underline{0.23658}}\end{aligned}$$

$$\begin{aligned}SSE = \sum_{i=1}^n \sum_{j=1}^C (x_{ij} - \bar{x}_j)^2 &= [(6.33 - 6.318)^2 + (6.26 - 6.318)^2 + (6.31 - 6.318)^2 \\&\quad + (6.29 - 6.318)^2 + (6.40 - 6.318)^2 + (6.26 - 6.2775)^2 \\&\quad + (6.36 - 6.2775)^2 + \dots + (6.19 - 6.230)^2 + (6.21 - 6.230)^2] \\&= \underline{\underline{0.15492}}\end{aligned}$$

$$\begin{aligned}df_C &= C - 1 = 4 - 1 = 3 & df_T &= N - 1 = 24 - 1 \\df_E &= N - C = 24 - 4 = 20 & &= 23\end{aligned}$$



$$\underline{SST} = \sum_{i=1}^n \sum_{j=1}^C (x_{ij} - \bar{x})^2 = [(6.33 - 6.339583)^2 + (6.26 - 6.339583)^2 + (6.31 - 6.339583)^2 + \dots + (6.19 - 6.339583)^2 + (6.21 - 6.339583)^2] = 0.39150$$

$$df_C = C - 1 = 4 - 1 = 3$$

$$df_E = N - C = 24 - 4 = 20$$

$$df_T = N - 1 = 24 - 1 = 23$$

$$\underline{MSC} = \frac{\underline{SSC}}{df_C} = \frac{.23658}{3} = .078860$$

$$\underline{MSE} = \frac{\underline{SSE}}{df_E} = \frac{.15492}{20} = .007746$$

$$\boxed{F} = \frac{.078860}{.007746} = \underline{10.18}$$

From these computations, an analysis of variance chart can be constructed

Source of Variance	df	SS	MS	F
Between	3	0.23658	0.078860	10.18
Error	20	0.15492	0.007746	
Total	23	0.39150		



$$\alpha = 0.05$$

- Associated with every F value in the table are two unique df values: degrees of freedom in the numerator (df_C) and degrees of freedom in the denominator (df_E).
- For the machine operator example, $df_C = 3$ and $df_E = 20$, $F_{.05,3,20}$ is **3.10**. This value is the **critical** value of the F test.
- Analysis of variance tests are always one-tailed tests with the rejection region in the upper tail.
- The decision rule is to **reject** the null hypothesis if the observed F value is greater than the critical F value



Comparison of F and t Values

- Analysis of variance can be used to test hypotheses about the difference in two means.
- Analysis of data from two samples by both a *t* test and an ANOVA shows that the observed
- *F* value equals the observed *t* value squared.
$$F = t^2 \text{ for } df_C = 1$$
- The ***t* test of independent samples actually is a special case of one-way ANOVA** when there are only two treatment levels ($df_C = 1$).
- The *t* test is computationally simpler than ANOVA for two groups.
- However, some statistical computer software packages do not contain a *t* test.
- In these cases, the researcher can perform a one-way ANOVA and then either take the square root of the *F* value to obtain the value of *t* or use the generated probability with the *p*-value method to reach conclusions.



Exercise (HW)

A company has three manufacturing plants, and company officials want to determine whether there is a difference in the average age of workers at the three locations. The following data are the ages of five randomly selected workers at each plant. Perform a one-way ANOVA to determine whether there is a significant difference in the mean ages of the workers at the three plants. Use $\alpha = .01$ and note that the sample sizes are equal.

Plant (Employee Ages)	1	2	3
29	32	25	
27	33	24	
30	31	24	
27	34	25	
28	30	26	

$$\begin{array}{ll} SSC & df_c = 3 - 1 = 2 \\ SSE & df_{FE} = 15 - 3 = 12 \\ SST & df_T = 15 - 1 = 14 \\ \textcircled{F} & \text{critical value of } F_{0.01, 2, 12} = 6.93 \end{array}$$



Solution

HYPOTHESIZE:

STEP 1. The hypotheses follow.

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

H_a : At least one of the means is different from the others.

TEST:

STEP 2. The appropriate test statistic is the F test calculated from ANOVA.

STEP 3. The value of α is .01.

STEP 4. The degrees of freedom for this problem are $3 - 1 = 2$ for the numerator and $15 - 3 = 12$ for the denominator. The critical F value is $F_{.01, 2, 12} = 6.93$.

Because ANOVAs are always one tailed with the rejection region in the upper tail, the decision rule is to reject the null hypothesis if the observed value of F is greater than 6.93.



$$T_f: \quad T_1 = 141 \quad T_2 = 160 \quad T_3 = 124 \quad T = 425$$

$$n_f: \quad n_1 = 5 \quad n_2 = 5 \quad n_3 = 5 \quad N = 15$$

$$\bar{X}_f: \quad \bar{X}_1 = 28.2 \quad \bar{X}_2 = 32.0 \quad \bar{X}_3 = 24.8 \quad \bar{X} = 28.33$$

$$SSC = 5(28.2 - 28.33)^2 + 5(32.0 - 28.33)^2 + 5(24.8 - 28.33)^2 = 129.73$$

$$SSE = (29 - 28.2)^2 + (27 - 28.2)^2 + \dots + (25 - 24.8)^2 + (26 - 24.8)^2 = 19.60$$

$$SST = (29 - 28.33)^2 + (27 - 28.33)^2 + \dots + (25 - 28.33)^2 \\ + (26 - 28.33)^2 = 149.33$$

$$df_C = 3 - 1 = 2$$

$$df_E = 15 - 3 = 12$$

$$df_T = 15 - 1 = 14$$



Source of Variance	SS	df	MS	F
Between	129.73	2	64.87	39.80
Error	19.60	12	1.63	
Total	149.33	14		

ACTION:

STEP 7. The decision is to reject the null hypothesis because the observed F value of 39.80 is greater than the critical table F value of 6.93.



References

-
- Probability and Statistics for Engineering and Sciences, 8th Edition, Jay L Devore, Cengage Learning
 - Applied Business Statistics by Ken Black



BITS Pilani
Pilani Campus

Linear Regression

Akanksha Bharadwaj
Asst. Professor, CS/IS Department



SS ZG536, ADV STAT TECHNIQUES FOR ANALYTICS

Contact Session 8

Covariance

- Variables may change in relation to each other
- Covariance measures how much the movement in one variable predicts the movement in a corresponding variable
- “Covariance” indicates the **direction** of the linear relationship between variables.

Variance Vs Covariance

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.

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

Covariance

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

- E[x] is the expected value or mean of a sample 'x', then cov(x,y) can be represented in the following way:

$$\begin{aligned}
 \text{cov}(x, y) &= E[(x - \mu_x)(y - \mu_y)] \\
 &= E[xy] - E[x]E[y] \\
 &= E[xy] - \mu_x\mu_y \\
 \text{And } \mu_x \text{ & } \mu_y &= E[x] \text{ & } E[y] \text{ respectively.}
 \end{aligned}$$

Sampled variance

- 's²' or sampled variance is basically the covariance of a variable with itself

$$s^2 = \text{cov}(x, x) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

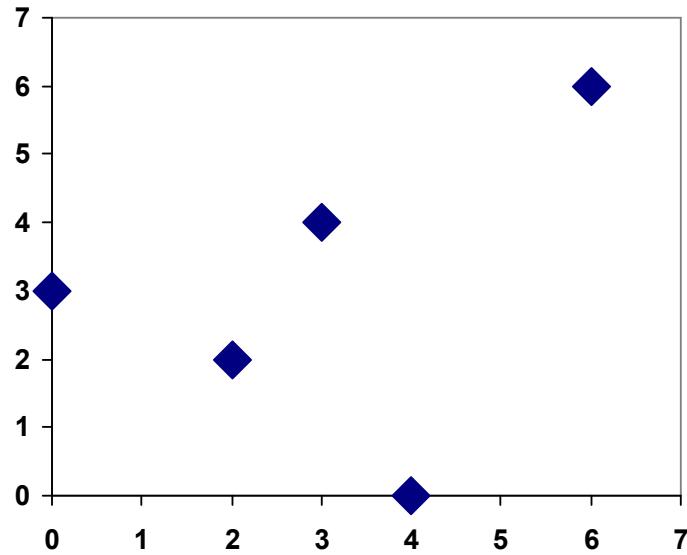
$$= \frac{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}{n-1} \quad \text{--- (A)}$$

for 2 variables:

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

- In the above formula, the numerator of the equation(A) is called the **sum of squared deviations**.
- In equation(B) with two variables x and y, it is called the **sum of cross products**.
- In the above formula, n is the number of samples in the data set. The value (n-1) indicates the degrees of freedom.

Example Covariance



x | y | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$

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
$\bar{x} = 3$		$\bar{y} = 3$		$\sum = 7$

$n = 5$

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

What does this number tell us?

If both variables tend to increase or decrease together, the coefficient is positive.

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.
 - 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

Correlation

- It is a measure of ***the degree of relatedness of variables***. It can help a business researcher determine, for example, whether the stocks of two airlines rise and fall in any related manner.
- For a sample of pairs of data, correlation analysis can yield a numerical value that represents the degree of relatedness of the two stock prices over time.
- it is obtained by dividing the covariance of the two variables by the product of their standard deviations

Mathematical representation

$$\rho_{x,y} = \text{corr}(x,y) = \frac{\text{cov}(x,y)}{s_x s_y} = \frac{E[(x-\mu_x)(y-\mu_y)]}{s_x s_y}$$
$$= \frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x \sigma_y}$$

* $\sigma_x = s_x$ = standard deviation of 'x'

$\sigma_y = s_y$ = standard deviation of 'y'

Questions a Pearson correlation answers

- Is there a statistically significant relationship between age and height?
 - Is there a relationship between temperature and ice cream sales?
 - Is there a relationship among job satisfaction, productivity, and income?
 - Which two variable have the strongest co-relation between age, height, weight, size of family and family income?
-

Assumptions

- For the Pearson r correlation, both variables should be **normally distributed**.
- There should be **no significant outliers**.
- Each variable should be **continuous**
- The two variables have a **linear relationship**.
- The observations are **paired observations**. That is, for every observation of the independent variable, there must be a corresponding observation of the dependent variable.

Pearson product-moment correlation coefficient

- Researchers virtually always deal with sample data, this section introduces a widely used sample **coefficient of correlation**, r .
 - The term r is a *measure of the linear correlation of two variables*. It is a number that ranges from -1 to 0 to +1, representing the strength of the relationship between the variables.
 - An r value of +1 denotes a perfect positive relationship between two sets of numbers.
 - An r value of -1 denotes a perfect negative correlation, which indicates an inverse relationship between two variables: as one variable gets larger, the other gets smaller.
 - An r value of 0 means no linear relationship is present between the two variables.
-

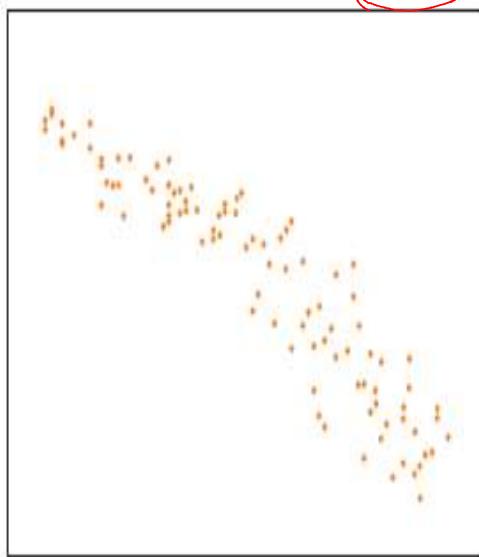
Formula

PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENT (12.1)

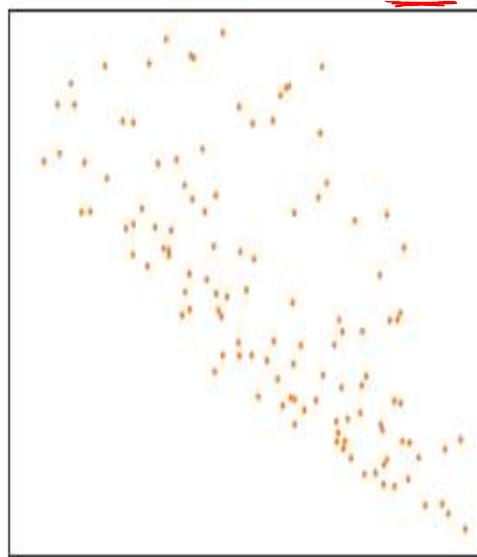
$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n} \right] \left[\sum y^2 - \frac{(\sum y)^2}{n} \right]}}$$

- The closer it is to +1 or -1, the more closely are the two variables are related.
- The positive sign signifies the direction of the correlation i.e. if one of the variables increases, the other variable is also supposed to increase.

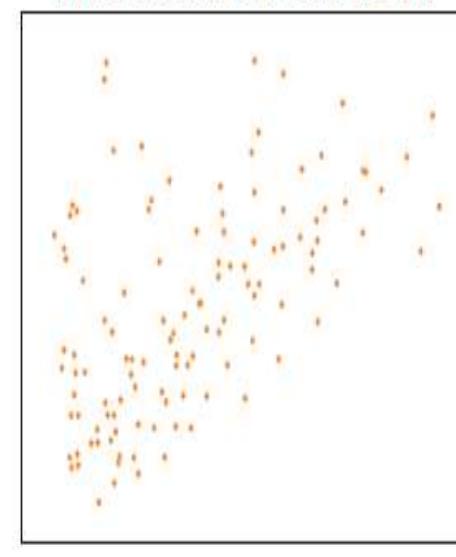
(a) Strong Negative Correlation ($r = -.933$)



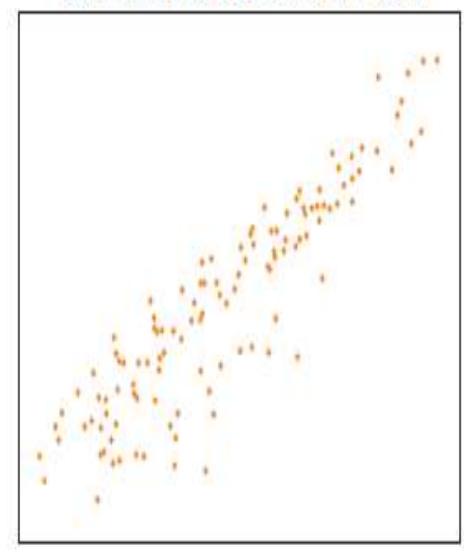
(b) Moderate Negative Correlation ($r = -.674$)



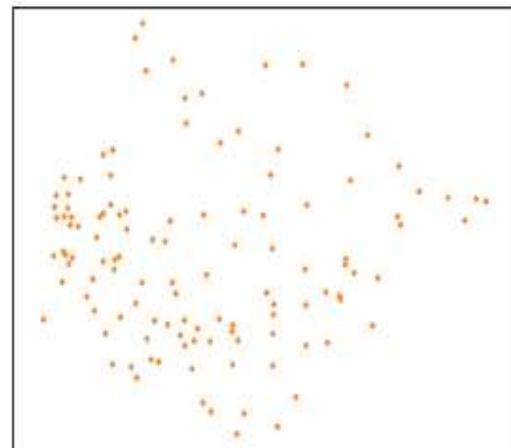
(c) Moderate Positive Correlation ($r = .518$)



(d) Strong Positive Correlation ($r = .909$)



(e) Virtually No Correlation ($r = -.004$)



Correlation does not have units but Covariance always has units

- As we see from the formula of covariance, it assumes the units from the product of the units of the two variables.
- On the other hand, correlation is dimensionless. It is a unit-free measure of the relationship between variables.
- This is because we divide the value of covariance by the product of standard deviations which have the same units.
- The value of covariance is affected by the change in scale of the variables.
- However, on doing the same, the value of correlation is not influenced by the change in scale of the values.

Advantages of the Correlation Coefficient

The Correlation Coefficient has several advantages over covariance for determining strengths of relationships:

- While correlation coefficients lie between -1 and +1, covariance can take any value between $-\infty$ and $+\infty$.
- Because of its numerical limitations, correlation is more useful for determining **how strong** the relationship is between the two variables.
- Correlation isn't affected by changes in the center (i.e. mean) or scale of the variables

Exercise

What is the measure of correlation between the interest rate of federal funds and the commodities futures index? With data such as those shown in Table 12.1, which represent the values for interest rates of federal funds and commodities futures indexes for a sample of 12 days.

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

$$\text{Var} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

TABLE 12.1		
	Data for the Economics Example	
Day	Interest Rate	Futures Index
1	7.43	221
2	7.48	222
3	8.00	226
4	7.75	225
5	7.60	224
6	7.63	223
7	7.68	223
8	7.67	226
9	7.59	226
10	8.07	235
11	8.03	233
12	8.00	241

Day	Interest x	Futures Index y	Futures Index		
			x ²	y ²	xy
1	7.43	221	55.205	48,841	1,642.03
2	7.48	222	55.950	49,284	1,660.56
3	8.00	226	64.000	51,076	1,808.00
4	7.75	225	60.063	50,625	1,743.75
5	7.60	224	57.760	50,176	1,702.40
6	7.63	223	58.217	49,729	1,701.49
7	7.68	223	58.982	49,729	1,712.64
8	7.67	226	58.829	51,076	1,733.42
9	7.59	226	57.608	51,076	1,715.34
10	8.07	235	65.125	55,225	1,896.45
11	8.03	233	64.481	54,289	1,870.99
12	8.00	241	64.000	58,081	1,928.00
$\Sigma x = 92.93$		$\Sigma y = 2,725$	$\Sigma x^2 = 720.220$	$\Sigma y^2 = 619,207$	$\Sigma xy = 21,115.07$

$$r = \frac{(21,115.07) - \frac{(92.93)(2725)}{12}}{\sqrt{\left[(720.22) - \frac{(92.93)^2}{12}\right] \left[(619,207) - \frac{(2725)^2}{12}\right]}} = .815$$

strong + relationship

Correlation Vs Covariance

- In simple words, both the terms measure the relationship and the dependency between two variables.
 - “Covariance” indicates the direction of the linear relationship between variables.
 - “Correlation” on the other hand measures both the strength and direction of the linear relationship between two variables.
 - Correlation is a function of the covariance.
 - **What sets them apart is the fact that correlation values are standardized whereas, covariance values are not.**
-

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!**

Regression Analysis

- **Regression analysis** is *the process of constructing a mathematical model or function that can be used to predict or determine one variable by another variable or other variables.*
- The most elementary regression model is called **simple regression** or **bivariate regression** involving two variables in which one variable is predicted by another variable.

Regression Analysis

- Regression analysis is a way of mathematically sorting out which of those variables does indeed have an impact.

It answers the questions:

- Which factors matter most?
- Which can we ignore?
- How do those factors interact with each other?
- And, perhaps most importantly, how certain are we about all of these factors?

In regression analysis, those **factors** are called **variables**.

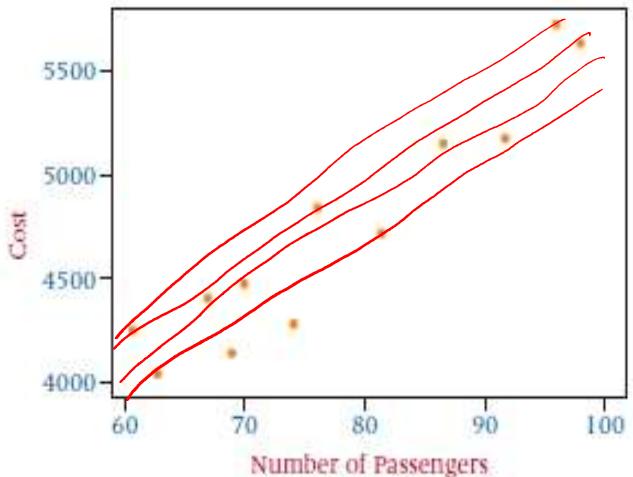
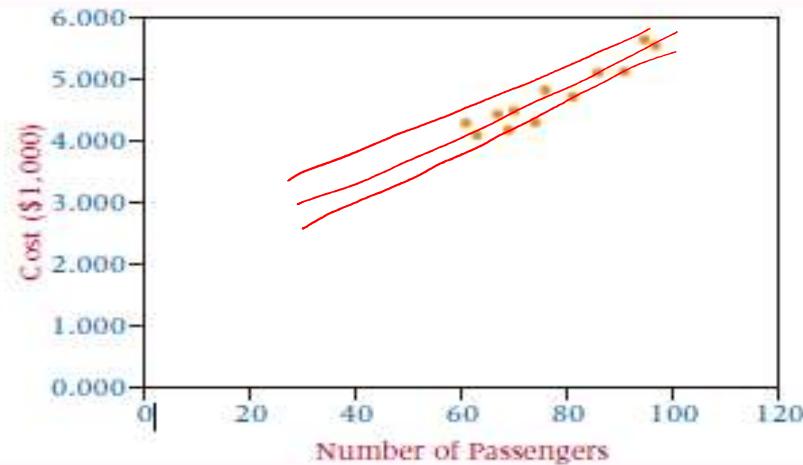
Dependent and independent variable

- In simple regression, *the variable to be predicted* is called the **dependent variable** and is designated as y .
- The *predictor* is called the **independent variable**, or *explanatory variable*, and is designated as x .
- In simple regression analysis, only a straight-line relationship between two variables is examined.
- Nonlinear relationships and regression models with more than one independent variable can be explored by using multiple regression models

Scatter Plot

- Usually, the first step in simple regression analysis is to construct a **scatter plot**
- Graphing the data in this way yields preliminary information about the shape and spread of the data

Scatter Plot



- Try to imagine a line passing through the points. Is a linear fit possible? Would a curve fit the data better?



Simple Linear Regression

Equation of regression line

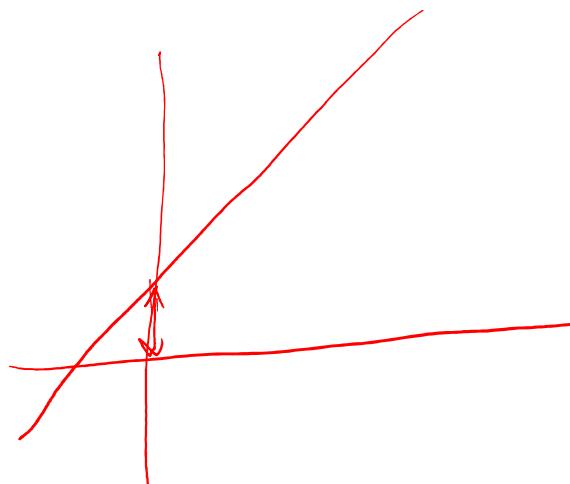
- In math courses, the slope-intercept form of the equation of a line often takes the form

$$y = mx + b$$

where

m = slope of the line

b = y intercept of the line



Equation of regression line

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

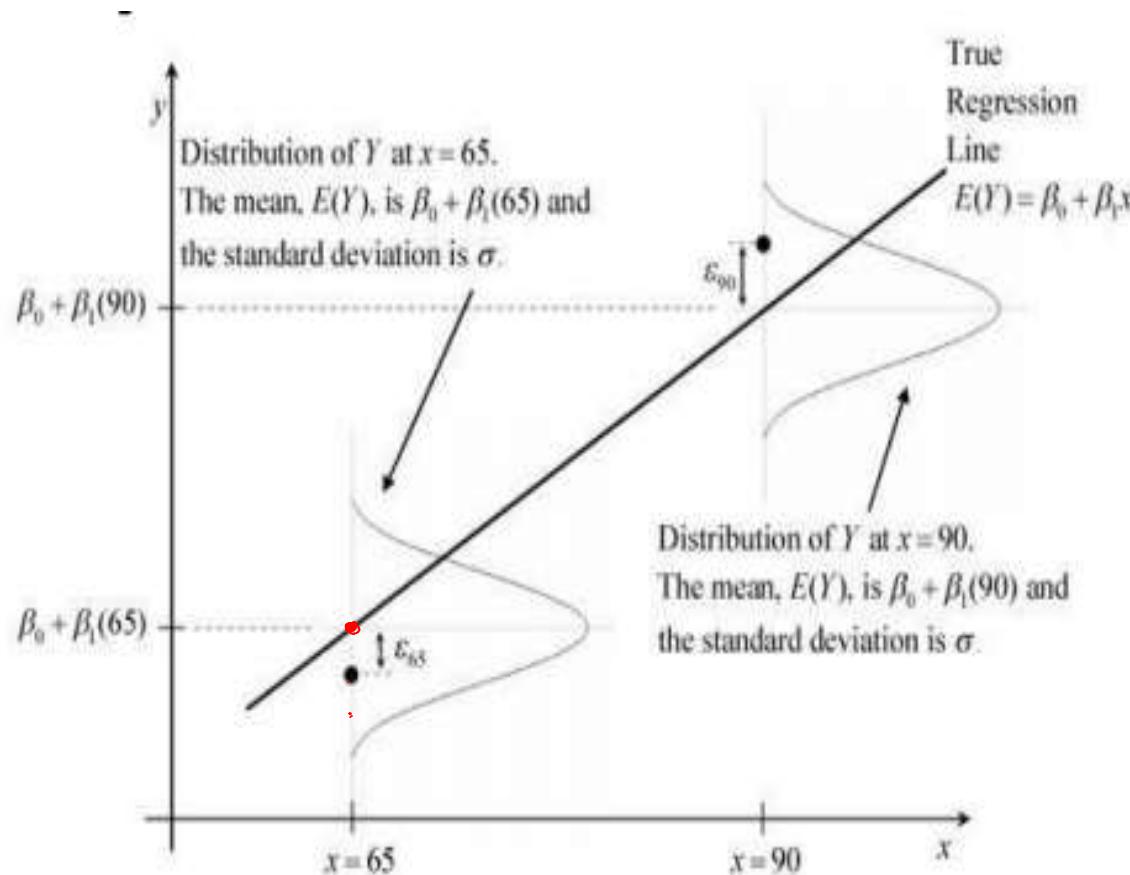
↓ ↓ ↓ ↓
 dependent variable independent variable population slope error of prediction
 population y intercept

- Unless the points being fitted by the regression equation are in perfect alignment, the regression line will miss at least some of the points.
- In the preceding equation, ϵ represents the error of the regression line in fitting these points. *sum of error \approx zero*
- If a point is on the regression line the value of error will be zero

Assumptions about the Error

- *Summation of all the error terms $E(\varepsilon_i)$ almost equals to 0.*
 - $\sigma(\varepsilon_i) = \sigma_\varepsilon$ where σ_ε is unknown.
 - The errors are independent, that is, the error in the i th observation is independent of the error observed in the j th observation.
 - The ε_i are normally distributed (with mean 0 and standard deviation σ_ε).
-

Assumptions



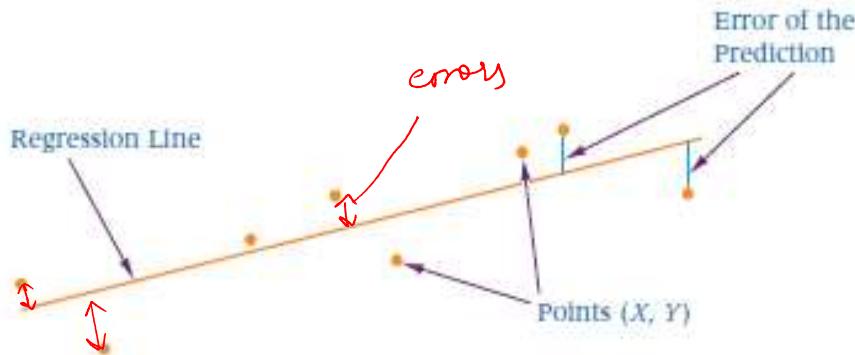
Estimated Regression model

$$\hat{y} = b_0 + b_1 x$$

dependent variable sample intercept sample slope

- Sample Regression line provides estimate of population regression line
- To determine the equation of the regression line for a sample of data, the researcher must determine the values for b_0 and b_1 .

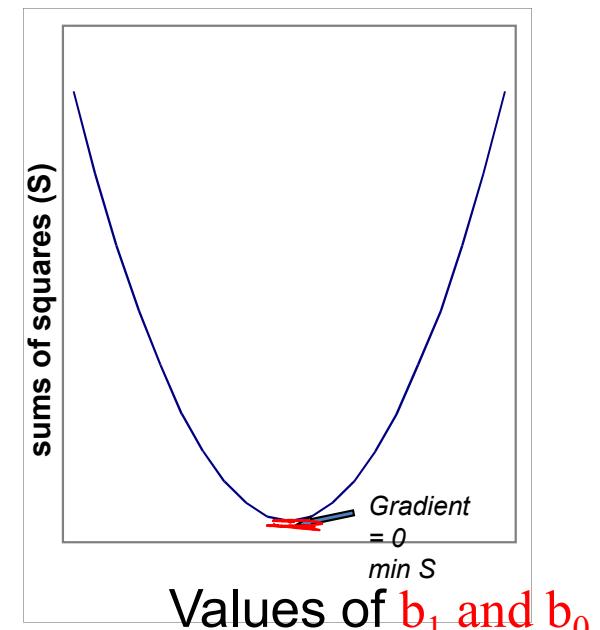
Least Squares regression line



- Observe that the line does not actually pass through any of the points. The vertical distance from each point to the line is the **error** of the prediction.
- In theory, an infinite number of lines could be constructed to pass through these points in some manner.
- The **least squares regression line** is the regression line that results in the smallest sum of errors squared.

Minimising sums of squares

- Need to minimise $\sum(y - \hat{y})^2$ → predicted value
- $\hat{y} = b_1x + b_0$
- so need to minimise: $\sum(y - b_1x + b_0)^2$
- If we plot the sums of squares for all different values of b_1 and b_0 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.



Slope of the regression line

- Formula 12.2 is an equation for computing the value of the sample slope.

**SLOPE OF THE REGRESSION
LINE (12.2)**

$$\textcircled{b_1} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$\underline{SS_{xy}} = \sum(x - \bar{x})(y - \bar{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$$\underline{SS_{xx}} = \sum(x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

**ALTERNATIVE FORMULA
FOR SLOPE (12.3)**

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

Y intercept of the regression line

Y INTERCEPT OF THE
REGRESSION LINE (12.4)

$$b_0 = \bar{y} - b_1 \bar{x} = \frac{\sum y}{n} - b_1 \frac{(\sum x)}{n}$$

Exercise

What is the Best-fit Line for this data?

TABLE 12.3

Airline Cost Data

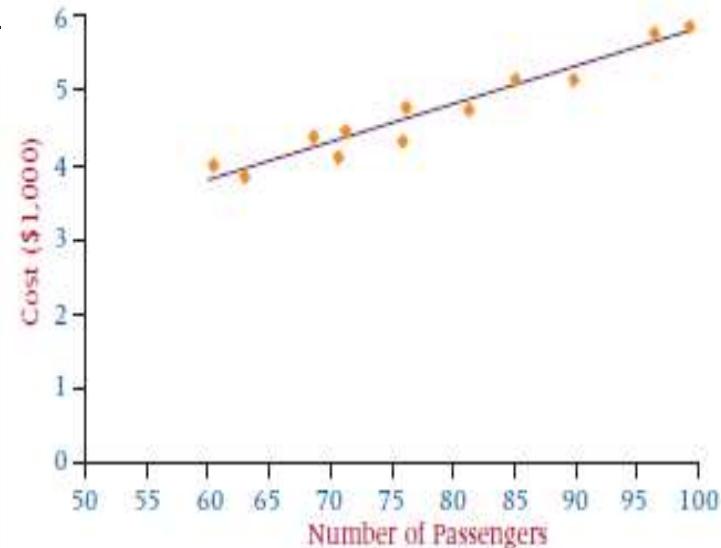
Number of Passengers	Cost (\$1,000)
61	4.280
63	4.080
67	4.420
69	4.170
70	4.480
74	4.300
76	4.820
81	4.700
86	5.110
91	5.130
95	5.640
97	5.560

Solution

$$b_0 = \frac{\sum y}{n} - b_1 \frac{\sum x}{n} = 1.569$$

$$\hat{y} = 1.569 + 0.0407x$$

Number of Passengers	Cost (\$1,000)		
x	y	x^2	xy
61	4.280	3,721	261.080
63	4.080	3,969	257.040
67	4.420	4,489	296.140
69	4.170	4,761	287.730
70	4.480	4,900	313.600
74	4.300	5,476	318.200
76	4.820	5,776	366.320
81	4.700	6,561	380.700
86	5.110	7,396	439.460
91	5.130	8,281	466.830
95	5.640	9,025	535.800
97	5.560	9,409	539.320
$\Sigma x = 930$	$\Sigma y = 56.690$	$\Sigma x^2 = 73,764$	$\Sigma xy = 4462.220$



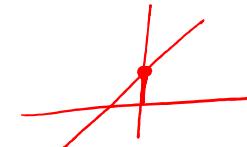
$$S_{xy} = \frac{\sum xy - (\sum x)(\sum y)}{n} = \frac{4462.220 - (930)(56.690)}{12} = 68.745$$

$$S_{xx} = \frac{\sum x^2 - (\sum x)^2}{n} = \frac{73,764 - (930)^2}{12} = 1689$$

$$b_1 = \frac{S_{xy}}{S_{xx}} = 0.0407$$

Interpretation of slope and intercept

- b_1 estimated change in the average value of y as a result of unit change in x
- b_0 estimated average value of y when the value of x is 0



Exercise (HW)

- A specialist in hospital administration stated that the number of FTEs (full-time employees) in a hospital can be estimated by counting the number of beds in the hospital (a common measure of hospital size). A healthcare business researcher decided to develop a regression model in an attempt to predict the number of FTEs of a hospital by the number of beds. She surveyed 12 hospitals and obtained the following data. The data are presented in sequence, according to the number of beds.

$$\begin{aligned}\sum x &= 592 \\ \sum y &= 1692\end{aligned}$$

Number of Beds	FTEs	Number of Beds	FTEs
23	69	50	138
29	95	54	178
29	102	64	156
35	118	66	184
42	126	76	176
46	125	78	225

$$\sum x^2 = 33044$$

$$\sum xy = 92,038$$

SS_{xy} ? SS_{xx} ?

Solution

Hospital	Number of Beds <i>x</i>	FTEs <i>y</i>	<i>x</i> ²	<i>xy</i>
1	23	69	529	1,587
2	29	95	841	2,755
3	29	102	841	2,958
4	35	118	1,225	4,130
5	42	126	1,764	5,292
6	46	125	2,116	5,750
7	50	138	2,500	6,900
8	54	178	2,916	9,612
9	64	156	4,096	9,984
10	66	184	4,356	12,144
11	76	176	5,776	13,376
12	<u>78</u>	<u>225</u>	<u>6,084</u>	<u>17,550</u>
	$\Sigma x = 592$	$\Sigma y = 1,692$	$\Sigma x^2 = 33,044$	$\Sigma xy = 92,038$

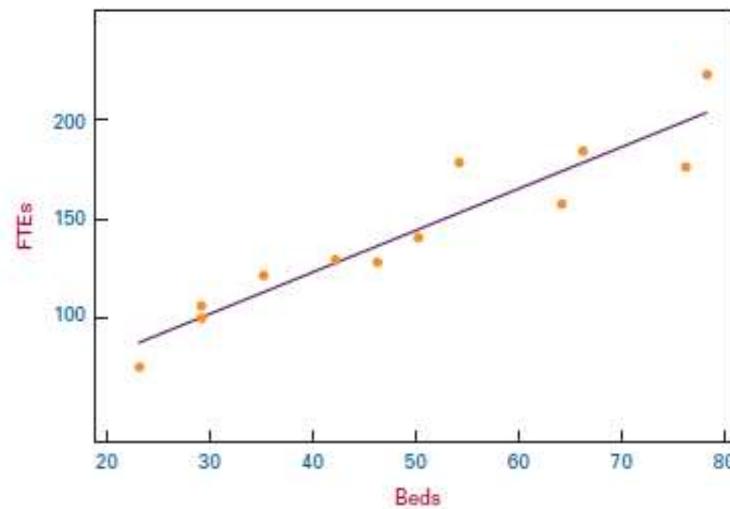
Using these values, the researcher solved for the sample slope (b_1) and the sample y -intercept (b_0).

$$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 92,038 - \frac{(592)(1692)}{12} = 8566$$

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 33,044 - \frac{(592)^2}{12} = 3838.667$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{8566}{3838.667} = 2.232$$

$$b_0 = \frac{\sum y}{n} - b_1 \frac{\sum x}{12} = \frac{1692}{12} - (2.232) \frac{592}{12} = 30.888$$



The least squares equation of the regression line is

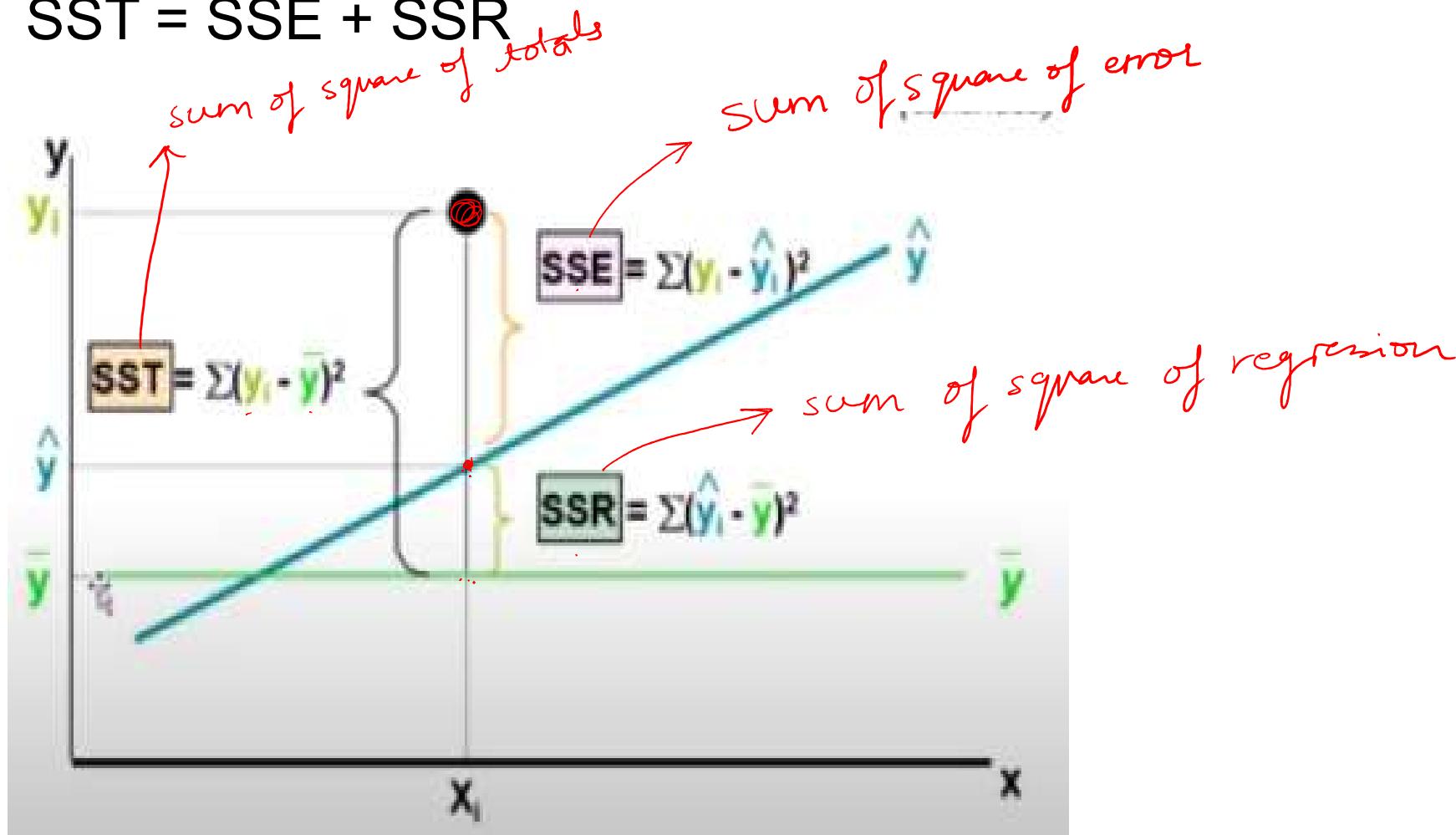
$$\hat{y} = 30.888 + 2.232x$$

Coefficient of determination

- The coefficient of determination is *the proportion of variability of the dependent variable (y) accounted for or explained by the independent variable (x)*.
- The coefficient of determination ranges from 0 to 1.
- An r^2 of zero means that the predictor accounts for none of the variability of the dependent variable and that there is no regression prediction of y by x.
- An r^2 of 1 means perfect prediction of y by x and that 100% of the variability of y is accounted for by x.

Sum of Square of Totals

$$SST = SSE + SSR$$



- The dependent variable, y , being predicted in a regression model has a variation that is measured by the sum of squares of y (SS_{yy}):

$$SST \leftarrow \underline{SS_{yy}} = \sum(y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n}$$

- This variation can be broken into two additive variations: the explained variation, measured by the sum of squares of regression (**SSR**), and the unexplained variation, measured by the sum of squares of error (**SSE**). This relationship can be expressed in equation form as
- Coefficient of determination, $r^2 = \frac{SSR}{SST}$
- SST also called as SS_{yy}

$$SS_{yy} = SSR + SSE$$

If each term in the equation is divided by SS_{yy} , the resulting equation is

$$1 = \frac{SSR}{SS_{yy}} + \frac{SSE}{SS_{yy}}$$

The term r^2 is the proportion of the y variability that is explained by the regression model and represented here as

$$r^2 = \frac{SSR}{SS_{yy}}$$

Substituting this equation into the preceding relationship gives

$$1 = r^2 + \frac{SSE}{SS_{yy}}$$

Solving for r^2 yields formula 12.5.

COEFFICIENT OF DETERMINATION (12.5)

$$r^2 = 1 - \frac{SSE}{SS_{yy}} = 1 - \frac{SSE}{\sum y^2 - \frac{(\sum y)^2}{n}}$$

Note: $0 \leq r^2 \leq 1$

Significance of R-squared

In Graph 1:

All the points lie on the line
and the R² value is a perfect 1

In Graph 2:

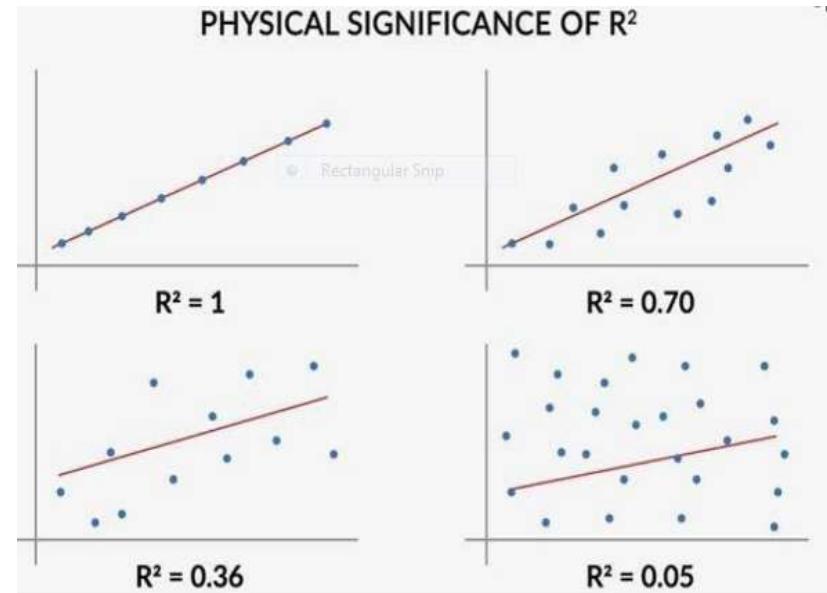
Some points deviate from
the line and the error is represented
by the lower R² value of 0.70

In Graph 3:

The deviation further
increases and the R² value further
goes down to 0.36

In Graph 4:

The deviation is further higher with a very low R² value of 0.05





BITS Pilani
Pilani Campus

Regression

Akanksha Bharadwaj
Asst. Professor, CS/IS

Significance of R-squared

In Graph 1:

All the points lie on the line
and the R² value is a perfect 1

In Graph 2:

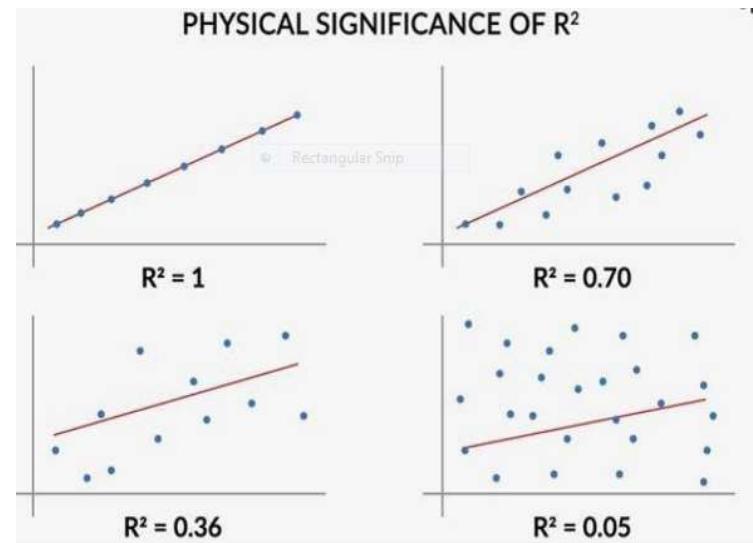
Some points deviate from
the line and the error is represented
by the lower R² value of 0.70

In Graph 3:

The deviation further
increases and the R² value further
goes down to 0.36

In Graph 4:

The deviation is further higher with a very low R² value of 0.05





Testing the slope of the regression line

- For example, the slope of the regression line for the airline cost data is .0407. This value is obviously not zero.
 - The problem is that this slope is obtained from a sample of 12 data points; and if another sample was taken, it is likely that a different slope would be obtained.
 - For this reason, the population slope is statistically tested using the sample slope.
 - The question is: If all the pairs of data points for the population were available, would the slope of that regression line be different from zero?
-

Hypothesis testing

- Here the sample slope, b_1 , is used as evidence to test whether the population slope is different from zero. The hypotheses for this test follow.

$$\left\{ \begin{array}{l} H_0: \beta_1 = 0 \\ H_a: \beta_1 \neq 0 \end{array} \right.$$

- Note that this test is two tailed. The null hypothesis can be rejected if the slope is either negative or positive.
- A negative slope indicates an inverse relationship between x and y.



Hypothesis Testing

- To determine whether there is a significant positive relationship between two variables, the hypotheses would be one tailed, or

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 > 0$$

or inverse

- To test for a significant negative relationship between two variables, the hypotheses also would be one tailed, or

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 < 0$$

- In each case, testing the null hypothesis involves a t test of the slope.

t test of the slope

t TEST OF SLOPE

where

$$t = \frac{b_1 - \beta_1}{s_b}$$

$$s_b = \frac{s_t}{\sqrt{SS_{xx}}}$$

$$s_t = \sqrt{\frac{SSE}{n - 2}}$$

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

β_1 = the hypothesized slope

df = $n - 2$

t test for airline example

The test of the slope of the regression line for the airline cost regression model for $\alpha = .05$ follows. The regression line derived for the data is

$$\hat{y} = 1.57 + .0407x \quad \text{slope}$$

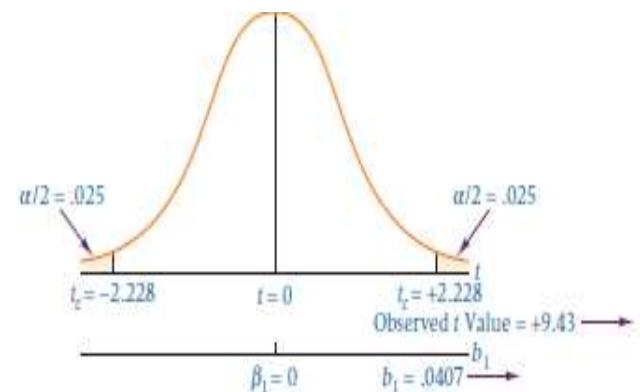
The sample slope is $.0407 = b_1$. The value of s_e is $.1773$, $\sum x = 930$, $\sum x^2 = 73,764$, and $n = 12$. The hypotheses are

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

The $df = n - 2 = 12 - 2 = 10$. As this test is two tailed, $\alpha/2 = .025$. The table t value is $t_{.025, 10} = \pm 2.228$. The observed t value for this sample slope is

$$t = \frac{.0407 - 0}{.1773 \sqrt{\frac{73,764 - (930)^2}{12}}} = 9.43$$



The null hypothesis that the population slope is zero is rejected.



Accessing the model fit

-
- 1. t statistic:** Used to determine the p-value and hence, helps in determining whether the coefficient is significant or not

 - 2. R-squared:** it tells the extent of the fit, i.e. how well the straight line describes the variance in the data. Its value ranges from 0 to 1, with the value 1 being the best fit and the value 0 showcasing the worst.



References

- Probability and Statistics for Engineering and Sciences, 8th Edition, Jay L Devore, Cengage Learning
- Applied Business Statistics by Ken Black
- <https://towardsdatascience.com/let-us-understand-the-correlation-matrix-and-covariance-matrix-d42e6b643c22>
- <https://towardsdatascience.com/overfitting-vs-underfitting-a-complete-example-d05dd7e19765>
- <https://towardsdatascience.com/multicollinearity-why-is-it-a-problem-398b010b77ac>
- <https://pdfs.semanticscholar.org/d1ee/9331a2fe0fb9c8ad27fbf378e3d4cb20163.pdf>
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3412358/>



“Correlation Is Not Causation”

- Whenever you work with regression analysis or any other analysis that tries to explain the impact of one factor on another, you need to remember the important adage:
- Correlation is not causation.
- This is critical and here's why: It's easy to say that there is a correlation between rain and monthly sales a product.
- Regression might show that they are indeed related.
- But it's an entirely different thing to say that rain *caused* the sales. Unless you're selling raincoats.



BITS Pilani
Pilani Campus



SS ZG536, ADV STAT TECHNIQUES FOR ANALYTICS

Contact Session 9



Moving from SLR to MLR

The new aspects to consider when moving from simple to multiple linear regression are:

- **Overfitting**
 - As you keep adding the variables, the model may become far too complex
 - It may end up memorising the training data and will fail to generalize
 - A model is generally said to overfit when the training accuracy is high while the test accuracy is very low
- **Multicollinearity**
 - Associations between predictor variables
- **Feature selection**
 - Selecting the optimal set from a pool of given features, many of which might be redundant becomes an important task



Multicollinearity

- It refers to the phenomenon of having related predictor variables in the input dataset.
- In simple terms, in a model which has been built using several independent variables, some of these variables might be interrelated.
- You drop some of these related independent variables as a way of dealing with multicollinearity.



Identifying Multicollinearity

- **Looking at pairwise correlations:** Looking at the correlation between different pairs of independent variables
- **Variance Inflation Factor (VIF):** The VIF assesses how much the variance of an estimated regression coefficient increases if your predictors are correlated.
- If there is no correlation the VIF will be 1. So the larger the number the more correlated the two variables are.



Variance Inflation Factor (VIF)

The VIF is given by:

$$VIF_j = \frac{1}{1 - R_i^2}$$

where R_j^2 is the R^2 -value obtained by regressing the j^{th} predictor on the remaining predictors.

- A VIF of 1 means that there is no correlation among the j^{th} predictor and the remaining predictor variables
- The general rule of thumb is that VIFs exceeding 4 warrant further investigation,
- while VIFs exceeding 10 are signs of serious multicollinearity requiring correction.



Fixing Multicollinearity

- 1. Feature Engineer:** If you can find a way to aggregate or combine the two features and turn it into one variable

- 2. Drop One:** It is common to drop one of the variables that are too highly correlated with another.



Equation for multiple linear regression

- Extending this notion to multiple regression gives the general equation for the probabilistic multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k + \epsilon$$

where

y = the value of the dependent variable

β_0 = the regression constant

β_1 = the partial regression coefficient for independent variable 1

β_2 = the partial regression coefficient for independent variable 2

β_3 = the partial regression coefficient for independent variable 3

β_k = the partial regression coefficient for independent variable k

k = the number of independent variables



-
- In virtually all research, these values are estimated by using sample information.

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_kx_k$$

where

\hat{y} = the predicted value of y

b_0 = the estimate of the regression constant

b_1 = the estimate of regression coefficient 1

b_2 = the estimate of regression coefficient 2

b_3 = the estimate of regression coefficient 3

b_k = the estimate of regression coefficient k

k = the number of independent variables



Multiple regression Steps

1. Generate the list of possible dependent and independent variable
 2. Collect data on variables
 3. Check the relationship between each independent variable and the dependent variable using scatterplots and correlation
 4. Check the relationship between independent variables using scatterplots and correlation – *multicollinearity*
 5. Use the non redundant independent variables in the analysis to find the best fitting model
 6. Use best fitting model to make predictions about the dependent variable
-

Some Problems with R-squared



- **Problem 1:** Every time you add a predictor to a model, the R-squared increases, even if due to chance alone. It never decreases. Consequently, a model with more terms may appear to have a better fit simply because it has more terms.
- **Problem 2:** If a model has too many predictors and higher order polynomials, it begins to model the random noise in the data. This condition is known as overfitting the model and it produces misleadingly high R-squared values and a lessened ability to make predictions.



Adjusted R-squared

- The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model.
- The adjusted R-squared increases only if the new term improves the model more than would be expected by chance.
- It decreases when a predictor improves the model by less than expected by chance.
- Use the adjusted R-square to compare models with different numbers of predictors

Adjusted R²

- Ranges from 0 to 1 with values closer to 1 indicating a stronger relationship
- Adjusted R^2 is the value of R^2 which has been penalized for the number of variables added to the model
- Therefore Adjusted R^2 is always smaller than R^2

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$$R^2 = \frac{SS_{regression}}{SS_{total}}$$

$$\underline{R^2_{adj}} = R^2(1 - R^2) \left(\frac{p}{n - p - 1} \right)$$

p = number of predictor variables
(regressors, not including intercept)

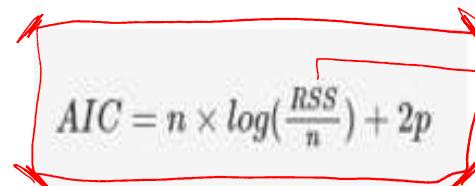
n = sample size

Akaike Information Criterion

- AIC considers both the fit of the model and the number of parameters used. More parameters result in a penalty
- Allows us to balance over- and under-fitting in our modelled relationships
 - We want a model that is as simple as possible, but no simpler
 - A reasonable amount of explanatory power is traded off against model size
 - AIC measures the balance of this for us

$$AIC = n \times \log\left(\frac{RSS}{n}\right) + 2p$$

Residual sum of squares



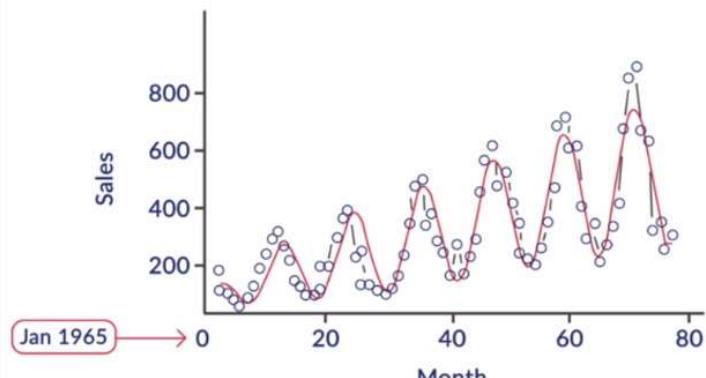
- Here, n is the sample size meaning the number of rows you'd have in the dataset and p is the number of predictor variables.

Non-linear Regression



In the first example, notice that the data points oscillate and follow a sine or cosine type of function.

Sales Figure

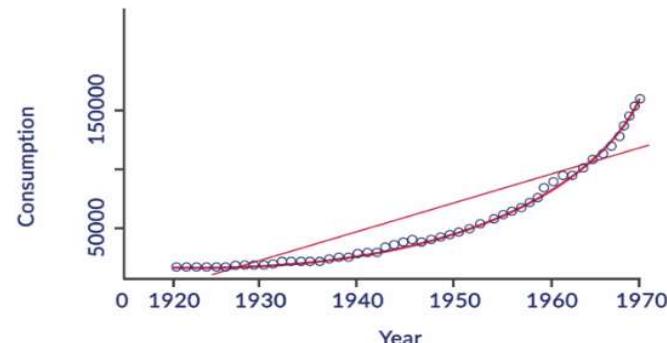


$$S(m) = 30.99 * \sqrt{m} * \cos(0.53 * m) + 5.83 * m + 76.61$$

Sales Figure

In the second example of electricity consumption, the data points gradually increase non-linearly, indicative of a polynomial or an exponential function:

Total Electric Consumption in USA

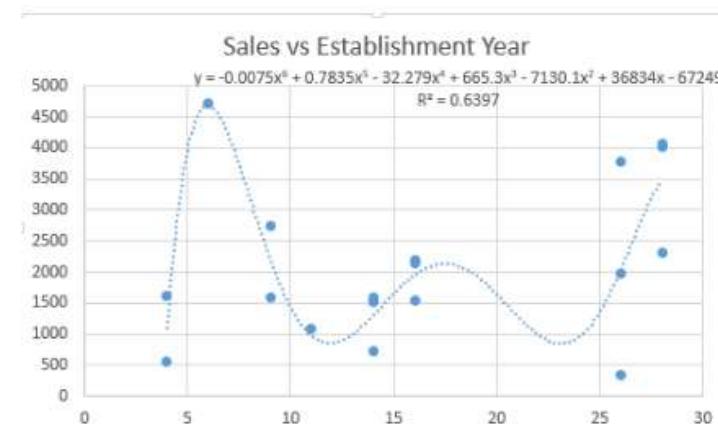


$$C(Y) = -112637300000 + 175430200.0 * Y - 91078.44 * Y^2 + 15.762219 * Y^3$$

Total Electric Consumption in USA

Non-linear Regression

- Polynomial regression is another form of regression in which the maximum power of the independent variable is more than one.
- In this regression technique, the best fit line is not a straight line instead it is in the form of a curve.





What is bias?

- Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.
- Model with high bias pays very little attention to the training data and oversimplifies the model.
- It always leads to high error on training and test data.



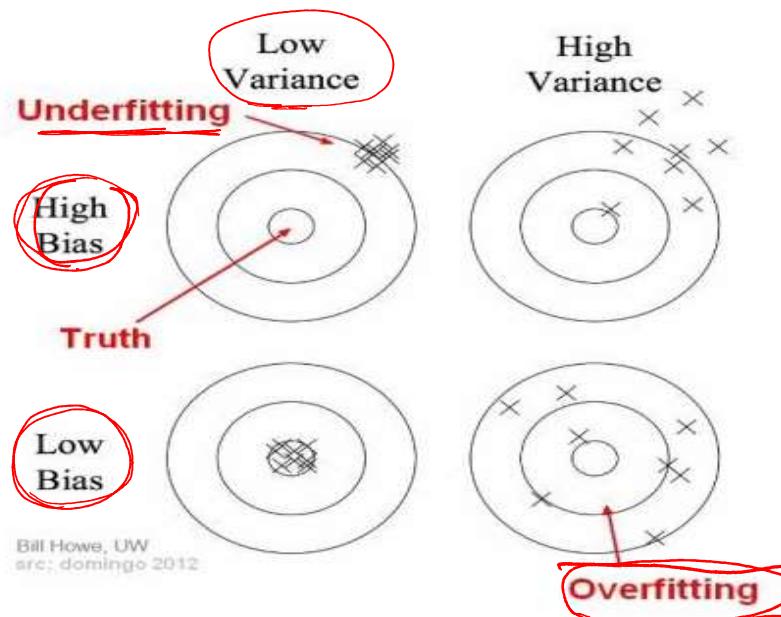
What is variance?

- Variance is the variability of model prediction for a given data point or a value which tells us spread of our data.
- Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before.
- As a result, such models perform very well on training data but has high error rates on test data.

Bias and variance using bulls-eye diagram



- Center of the target is a model that perfectly predicts correct values.
- As we move away from the bulls-eye our predictions get worse and worse.



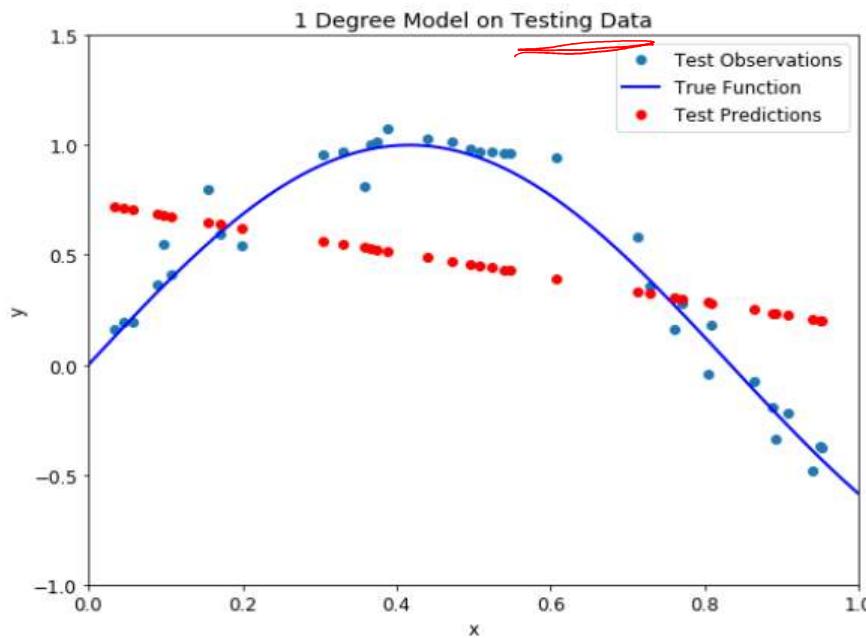
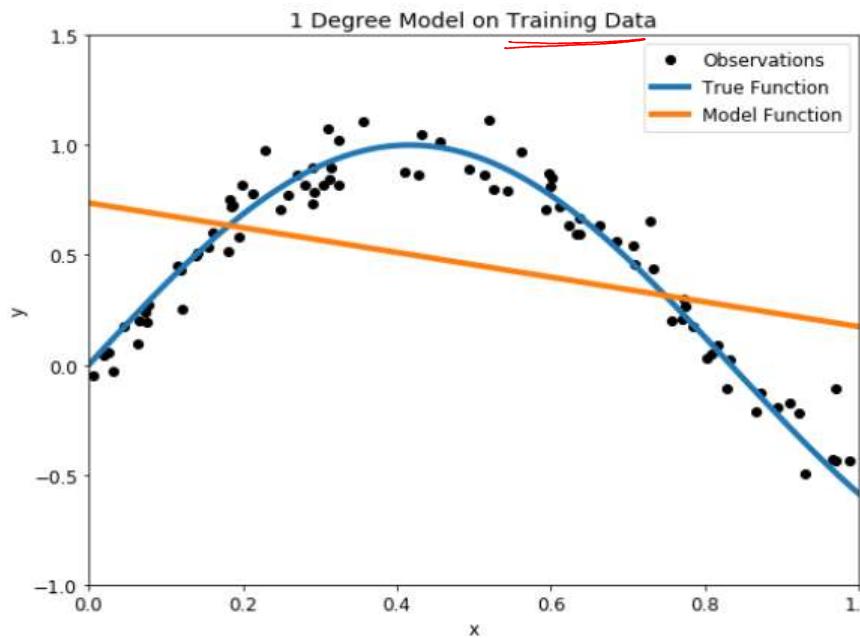


Overfitting

- A model learns relationships between the inputs, called **features**, and outputs, called **labels**, from a training dataset.
- During training the model is given both the features and the labels and learns how to map the former to the latter.
- A trained model is evaluated on a testing set, where we only give it the features and it makes predictions.
- We compare the predictions with the known labels for the testing set to calculate accuracy.
- Overfitting happens because your model is trying too hard to capture the noise in your training dataset

Overfitting vs. Underfitting

- The problem of Overfitting vs Underfitting finally appears when we talk about the polynomial degree.
- The degree represents how much flexibility is in the model, with a higher power allowing the model freedom to hit as many data points as possible.

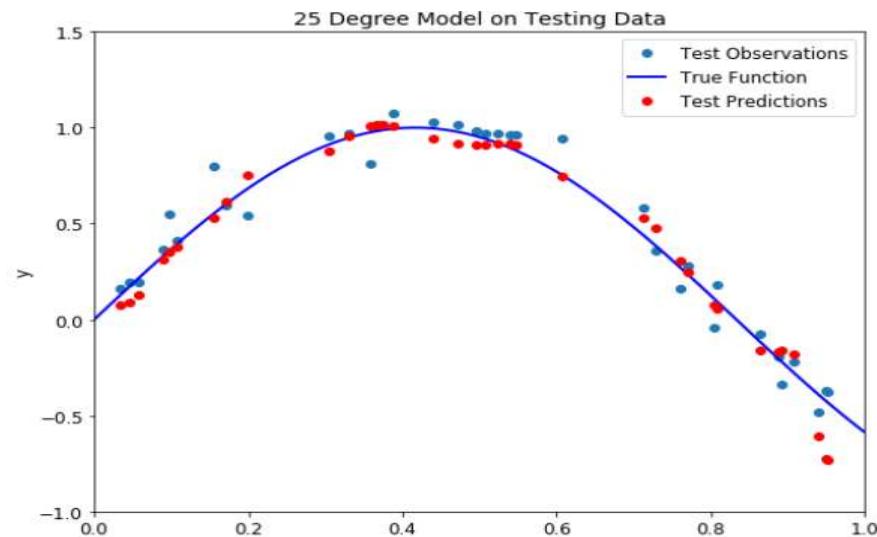
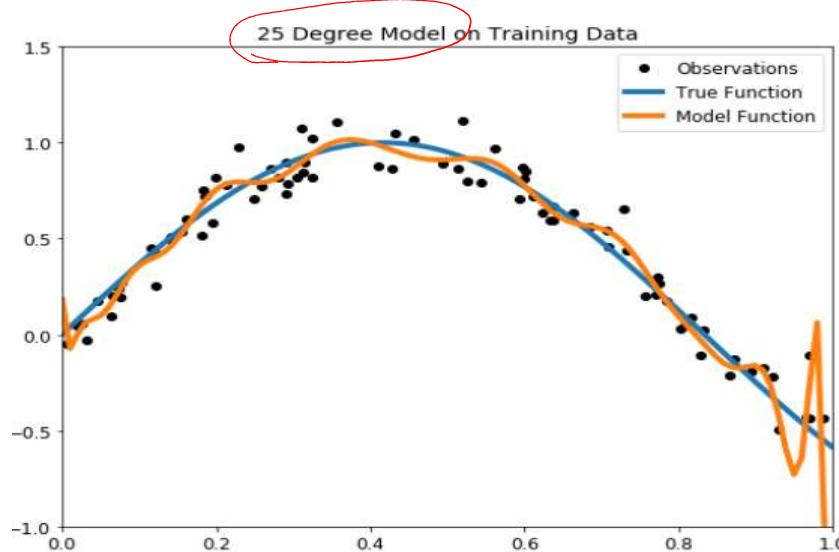




Underfitting Example

- Our model passes straight through the training set with no regard for the data! This is because an underfit model has low variance and high bias.
- Variance refers to how much the model is dependent on the training data.
- For the case of a 1 degree polynomial, the model depends very little on the training data because it barely pays any attention to the points!
- Instead, the model has high bias, which means it makes a strong assumption about the data.
- For this example, the assumption is that the data is linear, which is evidently quite wrong. When the model makes test predictions, the bias leads it to make inaccurate estimates.
- The model failed to learn the relationship between x and y because of this bias, a clear example of **underfitting**.

Overfitting Example



- This is a model with a high variance, because it will change significantly depending on the training data.
- The predictions on the test set are better than the one degree model, but the twenty five degree model still does not learn the relationship because it essentially memorizes the training data and the noise.



Solution

- Our problem is that we want a model that does not “memorize” the training data, but learns the actual relationship!
- How can we find a balanced model with the right polynomial degree?
- If we choose the model with the best score on the training set, we will just select the overfitting model but this cannot generalize well to testing data.
- Fortunately, there is a well-established data science technique for developing the optimal model: **validation**.



Validation

- We need some sort of pre-test to use for model optimization and evaluate. This pre-test is known as a validation set.
- A basic approach would be to use a validation set in addition to the training and testing set.
- This presents a few problems though: we could just end up overfitting to the validation set and we would have less training data.



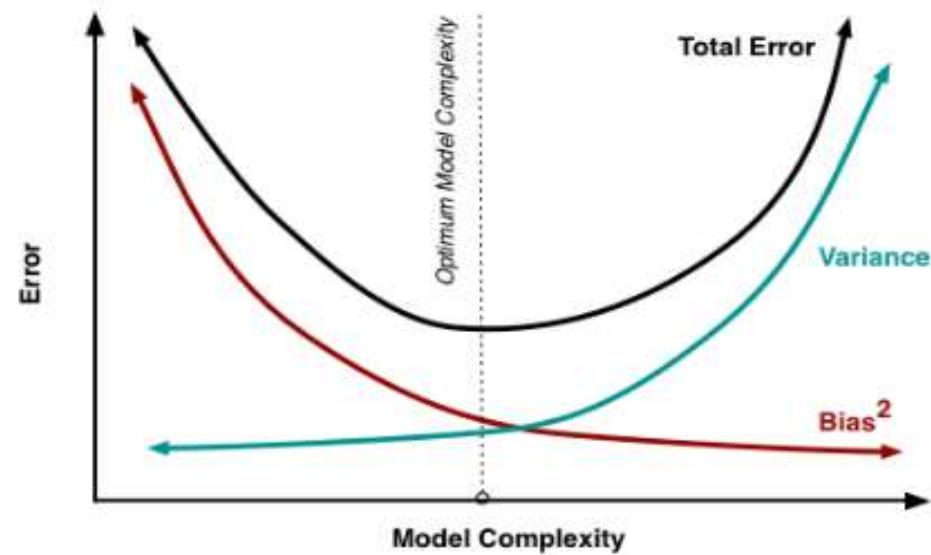
k-fold cross-validation

- A smarter implementation of the validation concept is k-fold cross-validation
- Let's use five folds as an example. We perform a series of train and evaluate cycles where each time we train on 4 of the folds and test on the 5th, called the hold-out set.
- We repeat this cycle 5 times, each time using a different fold for evaluation.
- At the end, we average the scores for each of the folds to determine the overall performance of a given model.
- This allows us to optimize the model before deployment without having to use additional data.

Bias Variance Tradeoff

- To build a good model, we need to find a good balance between bias and variance such that it minimizes the total error.
- An optimal balance of bias and variance would never overfit or underfit the model.

$$\text{Total Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$





Regularization

- This technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.
- A simple relation for linear regression looks like this.

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

- The fitting procedure involves a loss function, known as residual sum of squares or RSS. The coefficients are chosen, such that they minimize this loss function.

$$\text{RSS} = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2.$$

Regularization



- If there is noise in the training data, then the estimated coefficients won't generalize well to the future data.
- This is where regularization comes in and shrinks or regularizes these learned estimates towards zero.



Ridge Regression

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^p \beta_j^2$$

- *RSS is modified by adding the shrinkage quantity.*
- Now, the coefficients are estimated by minimizing this function.
- The increase in flexibility of a model is represented by increase in its coefficients, and if we want to minimize the above function, then these coefficients need to be small.
- This is how the Ridge regression technique prevents coefficients from rising too high.
- Also, notice that we shrink the estimated association of each variable with the response, except the intercept β_0 . This intercept is a measure of the mean value of the response when $x_{i1} = x_{i2} = \dots = x_{ip} = 0$.



Lasso/L1 Regression

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^p |\beta_j|.$$

LASSO (Least Absolute Shrinkage Selector Operator), is quite similar to ridge



What does Regularization achieve?

- Regularization, significantly reduces the variance of the model, without substantial increase in its bias.
- So the tuning parameter λ , used in the regularization techniques described above, controls the impact on bias and variance.
- As the value of λ rises, it reduces the value of coefficients and thus reducing the variance.
- Till a point, this increase in λ is beneficial as it is only reducing the variance(hence avoiding overfitting), without loosing any important properties in the data.
- But after certain value, the model starts loosing important properties, giving rise to bias in the model and thus underfitting.
- Therefore, the value of λ should be carefully selected.



Logistic Regression

- Logistic Regression is used when the dependent variable(target) is categorical (binary in nature).
- For example,
 - To predict whether an email is spam (1) or (0)
 - Whether the tumor is malignant (1) or not (0)



Example: Logistic Regression

- *E.g.* When we have to predict if a student passes or fails in an exam when the number of hours spent studying is given as a feature, the response variable has two values, pass and fail.
- This type of a problem is referred to as **Binomial Logistic Regression**, where the response variable has two values 0 and 1 or pass and fail or true and false.
- Multinomial Logistic Regression deals with situations where the response variable can have three or more possible values.



Logistic Regression

Univariate logistic regression equation : $\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x$

The left-hand side of this equation is what is called log odds.

Basically, the odds of having an event ($P/1-P$).

Multivariate logistic regression equation :

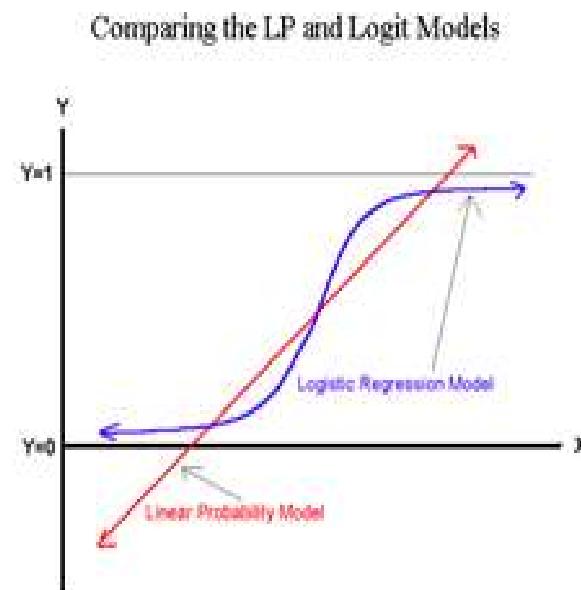
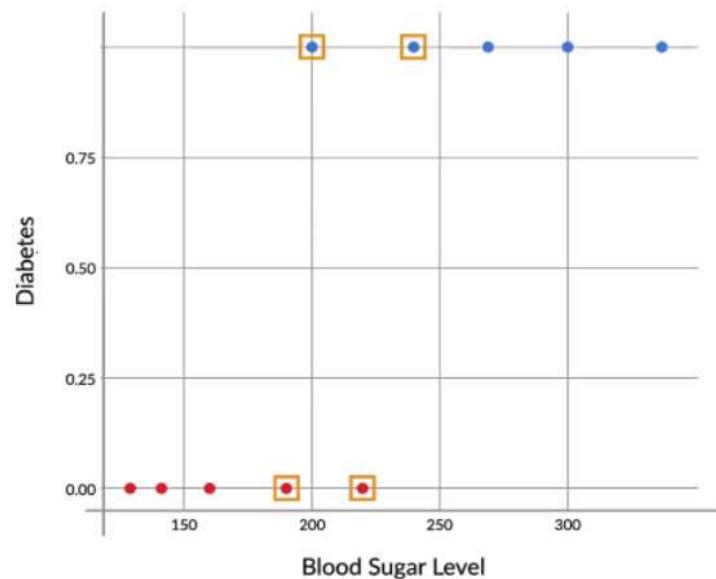
$$P = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\dots)}}$$

Where P denotes the probability of the event we are trying to predict with multiple independent variables.

Logistic regression example

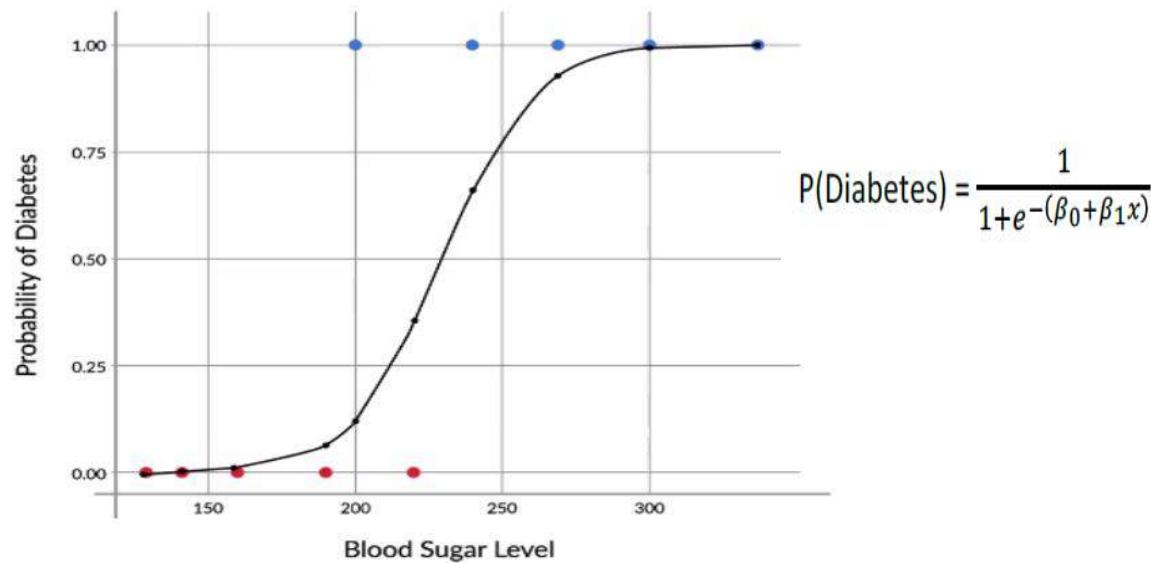


- Consider a case where we have blood sugar level of 10 patients along-with the status of them being diabetic.



Logistic regression example

One such curve which can model the probability of diabetes very well, is the **sigmoid curve**.





References

- Probability and Statistics for Engineering and Sciences, 8th Edition, Jay L Devore, Cengage Learning
- Applied Business Statistics by Ken Black
- <https://towardsdatascience.com/let-us-understand-the-correlation-matrix-and-covariance-matrix-d42e6b643c22>
- <https://towardsdatascience.com/overfitting-vs-underfitting-a-complete-example-d05dd7e19765>
- <https://towardsdatascience.com/multicollinearity-why-is-it-a-problem-398b010b77ac>
- <https://pdfs.semanticscholar.org/d1ee/9331a2fe0fb9c8ad27fbf378e3d4cb20163.pdf>
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3412358/>



BITS Pilani
Pilani Campus

Forecasting

Akanksha Bharadwaj
Asst. Professor, CS/IS



BITS Pilani
Pilani Campus



SS ZG536, ADV STAT TECHNIQUES FOR ANALYTICS

Contact Session 10



Definition of Forecasting

- Forecasting is the process of making predictions of the future based on past and present data and most commonly by analysis of trends.
- A common place example might be estimation of some variable of interest at some specified future date.



Forecasting is required in many situations

- Deciding whether to build another power generation plant in the next five years requires forecasts of future demand
 - Scheduling staff in a call center next week requires forecasts of call volumes
 - Stocking an inventory requires forecasts of stock requirements.
-



Forecasting

-
- Some things are easier to forecast than others. The time of the sunrise tomorrow morning can be forecast precisely.
 - On the other hand, tomorrow's lotto numbers cannot be forecast with any accuracy.



The predictability of an event or a quantity depends on several factors including:

- how well we understand the factors that contribute to it;
- how much data is available;
- whether the forecasts can affect the thing we are trying to forecast.



Example

- For example, forecasts of electricity demand can be highly accurate because all three conditions are usually satisfied.
 - We have a good idea of the contributing factors: electricity demand is driven largely by temperatures, with smaller effects for calendar variation such as holidays, and economic conditions.
-



-
- Provided there is a sufficient history of data on electricity demand and weather conditions,
 - and we have the skills to develop a good model linking electricity demand and the key driver variables, the forecasts can be remarkably accurate.



Example

- On the other hand, when forecasting currency exchange rates, only one of the conditions is satisfied: there is plenty of available data.
 - However, we have a limited understanding of the factors that affect exchange rates, and forecasts of the exchange rate have a direct effect on the rates themselves.
-



-
- forecasting whether the exchange rate will rise or fall tomorrow is about as predictable as forecasting whether a tossed coin will come down as a head or a tail.
 - In both situations, you will be correct about 50% of the time, whatever you forecast.
 - In situations like this, forecasters need to be aware of their own limitations, and not claim more than is possible.

FORECASTING



- Virtually all areas of business, including production, sales, employment, transportation, distribution, and inventory, produce and maintain time-series data.
- Table provides an example of time-series data released by the Office of Market Finance, U.S. Department of the Treasury.
- The table contains the bond yield rates of three-month Treasury Bills for a 17-year period

TABLE 15.1	
Bond Yields of Three-Month Treasury Bills	
Year	Average Yield
1	14.03%
2	10.69
3	8.63
4	9.58
5	7.48
6	5.98
7	5.82
8	6.69
9	8.12
10	7.51
11	5.42
12	3.45
13	3.02
14	4.29
15	5.51
16	5.02
17	5.07



Principles of Forecasting

- There are many types of forecasting models.
- They differ in their degree of complexity, the amount of data they use, and the way they generate the forecast.
- However, some features are common to all forecasting models. They include the following:
- Forecasts are rarely perfect
- Forecasts are more accurate for groups or families of items rather than for individual items.
- Forecasts are more accurate for shorter than longer time horizons.



Time Series

- A time series is a collection of observations of well-defined data items obtained through repeated measurements over time.
- For example, measuring the value of retail sales each month of the year would comprise a time series.
- This is because sales revenue is well defined, and consistently measured at equally spaced intervals.
- Data collected irregularly or only once are not time series.



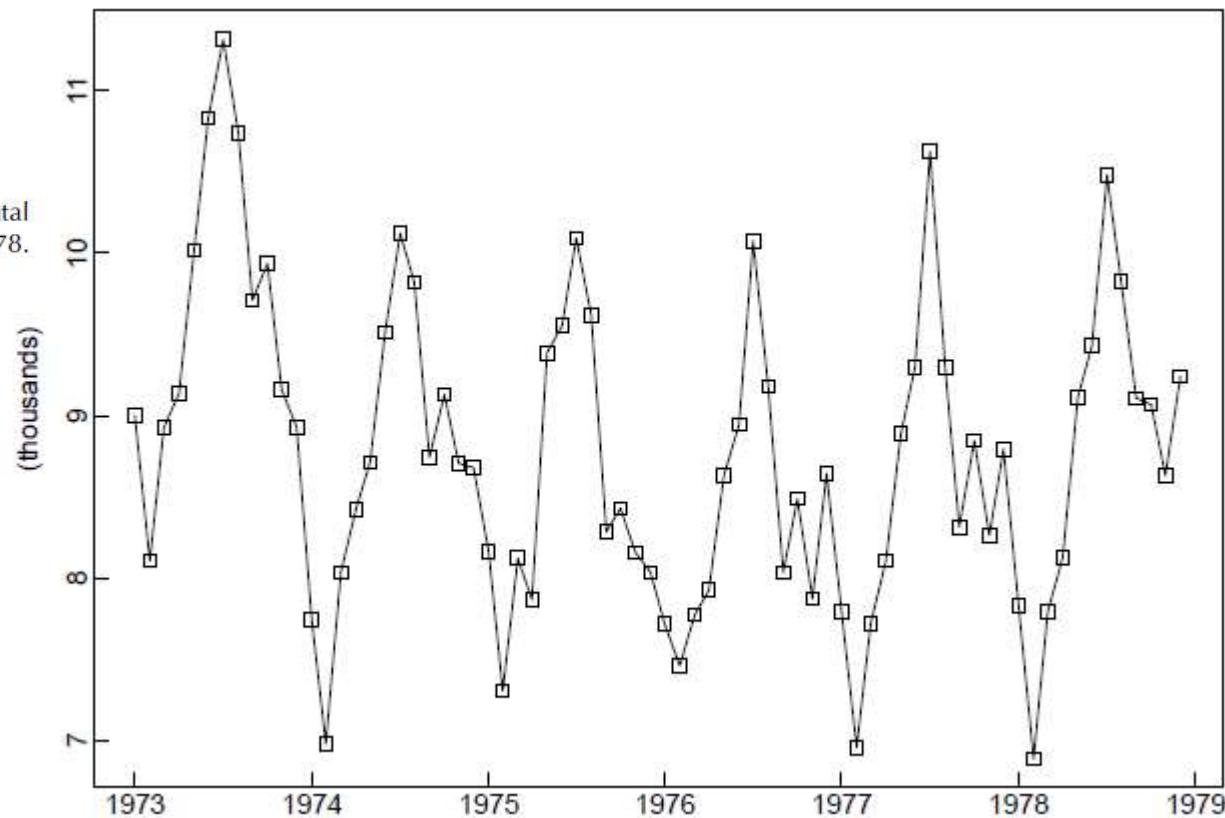
Time Series

- A **time series** is a set of observations x_t , each one being recorded at a specific time t .
- A discrete-time time series is one in which the set T_0 of times at which observations are made is a discrete set, as is the case, for example, when observations are made at fixed time intervals.
- Continuous time series are obtained when observations are recorded continuously over some time interval, e.g., when $T_0 = [0, 1]$.

Example



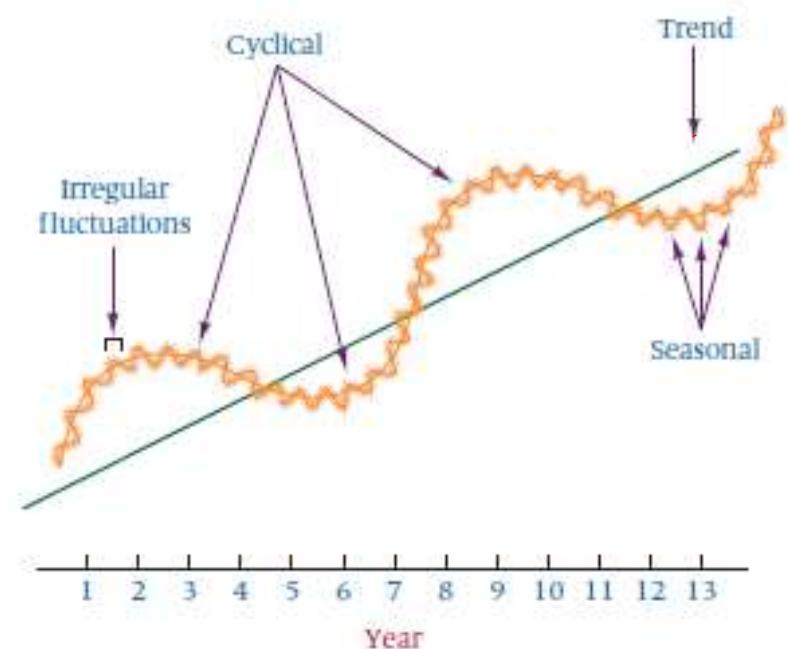
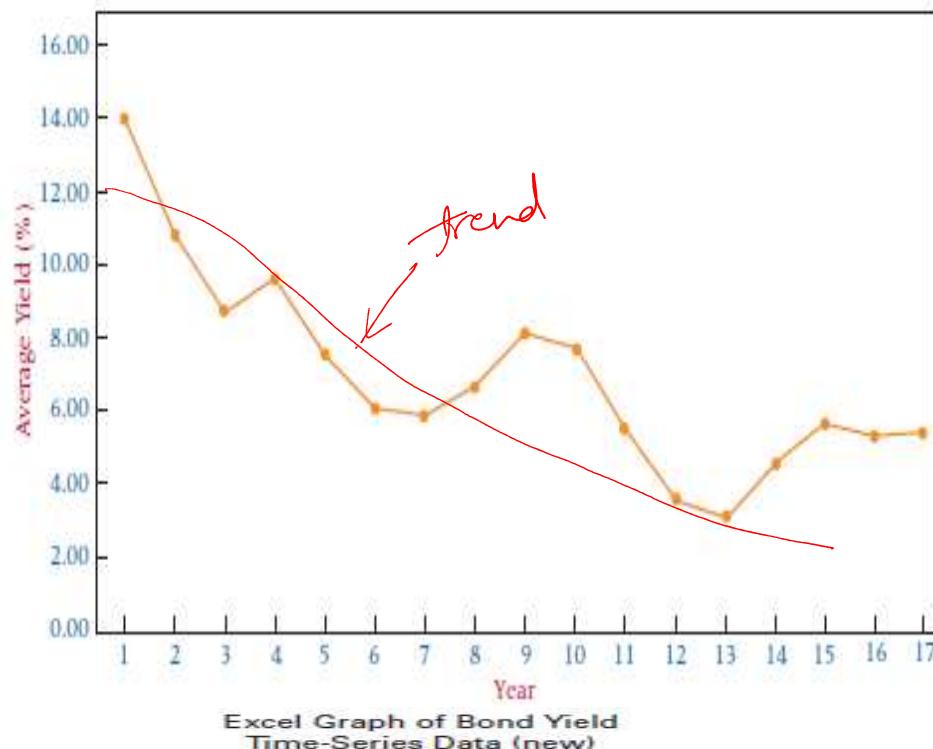
The monthly accidental deaths data, 1973–1978.





Why use time series data?

- To develop forecasting models
 - What will the rate of inflation be next year?
- To estimate dynamic causal effects
 - If the Fed increases the Federal Funds rate now, what will be the effect on the rates of inflation and unemployment in 3 months? in 12 months?



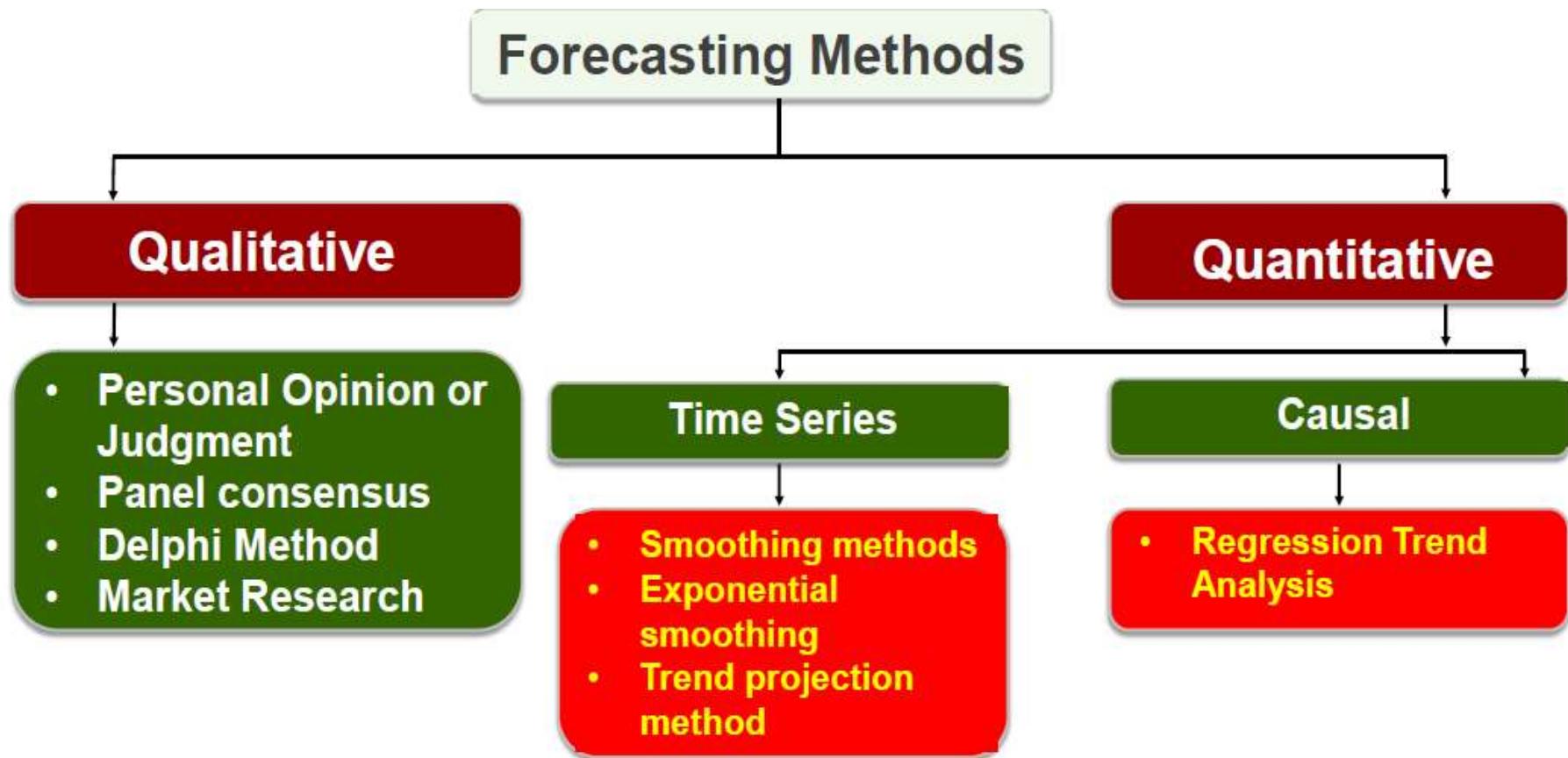


Time-Series Components

- **Trend:** *The long-term general direction of data*
 - **Cycles** are *patterns of highs and lows through which data move over time periods usually of more than a year.*
 - **Seasonal effects**, on the other hand, are *shorter cycles, which usually occur in time periods of less than one year.*
 - Often seasonal effects are measured by the month, but they may occur by quarter, or may be measured in as small a time frame as a week.
 - **Irregular fluctuations** are *rapid changes or “bleeps” in the data, which occur in even shorter time frames than seasonal effects.*
 - Irregular fluctuations can happen as often as day to day. They are subject to momentary change and are often unexplained.
-



Types of Forecasting methods





Qualitative methods

Personal Opinion

- Individuals forecasts future based on their own judgment or opinion without any formal model

Panel consensus

- Panel of individuals are encouraged to share information, opinions, and assumptions, if any, to predict future value of some variable under study



Qualitative method

Type	Characteristics	Strengths	Weaknesses
Executive opinion	A group of managers meet & come up with a forecast	Good for strategic or new-product forecasting	One person's opinion can dominate the forecast
Market research	Uses surveys & interviews to identify customer preferences	Good determinant of customer preferences	It can be difficult to develop a good questionnaire
Delphi method	Seeks to develop a consensus among a group of experts	Excellent for forecasting long-term product demand,	Time consuming to develop



Quantitative forecasting

Time Series Models:

- Assumes information needed to generate a forecast is contained in a time series of data
- Assumes the future will follow same patterns as the past

Causal Models or Associative Models

- Explores cause-and-effect relationships
- Uses leading indicators to predict the future
- Eg. House sales and appliance sales

The Measurement of Forecasting Error



How does a decision maker know which forecasting technique is doing the best job in predicting the future?

- One way is to compare forecast values with actual values and determine the amount of **forecasting error**
- An examination of individual errors gives some insight into the accuracy of the forecasts.
- However, this process can be tedious, especially for large data sets

ERROR OF AN
INDIVIDUAL FORECAST

where

$$e_t = \underline{x_t} - \underline{F_t}$$

e_t = the error of the forecast
 x_t = the actual value
 F_t = the forecast value

Mean Absolute Deviation (MAD)



- The **mean absolute deviation (MAD)** is *the mean, or average, of the absolute values of the errors*.

MEAN ABSOLUTE
DEVIATION

$$\text{MAD} = \frac{\sum |e_i|}{\text{Number of Forecasts}}$$

Actual value = 10

$$P_1 = 9$$

$$P_2 = 12$$

$$e_1 = 10 - 9 = 1$$

$$e_2 = 10 - 12 = -2$$



MAD example

Example

TABLE 15.2

Nonfarm Partnership
Tax Returns

Year	Actual	Forecast	Error
1	1,402	—	—
2	1,458	1,402	56.0
3	1,553	1,441.2	111.8
4	1,613	1,519.5	93.5
5	1,676	1,585.0	91.0
6	1,755	1,648.7	106.3
7	1,807	1,723.1	83.9
8	1,824	1,781.8	42.2
9	1,826	1,811.3	14.7
10	1,780	1,821.6	-41.6
11	1,759	1,792.5	-33.5

The mean absolute error can be computed for the forecast errors in Table 15.2 as follows.

$$MAD = \frac{|56.0| + |111.8| + |93.5| + |91.0| + |106.3| + |83.9| + |42.2| + |14.7| + |-41.6| + |-33.5|}{10} = 67.45$$



Mean Square Error (MSE)

- The MSE is *computed by squaring each error (thus creating a positive number) and averaging the squared errors.*

MEAN SQUARE ERROR

$$\text{MSE} = \frac{\sum e_i^2}{\text{Number of Forecasts}}$$



MSE example

Example

TABLE 15.2

Nonfarm Partnership
Tax Returns

Year	Actual	Forecast	Error
1	1,402	—	—
2	1,458	1,402	56.0
3	1,553	1,441.2	111.8
4	1,613	1,519.5	93.5
5	1,676	1,585.0	91.0
6	1,755	1,648.7	106.3
7	1,807	1,723.1	83.9
8	1,824	1,781.8	42.2
9	1,826	1,811.3	14.7
10	1,780	1,821.6	-41.6
11	1,759	1,792.5	-33.5

The mean square error can be computed for the errors shown in Table 15.2 as follows.

$$MSE = \frac{(56.0)^2 + (111.8)^2 + (93.5)^2 + (91.0)^2 + (106.3)^2 + (83.9)^2 + (42.2)^2 + (14.7)^2 + (-41.6)^2 + (-33.5)^2}{10} = 5,584.7$$



Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE)

$$= \frac{\sum \left| \frac{x_i - F_i}{x_i} \right|}{\text{Number of forecasts}} \times 100$$

X is actual value and F is forecasted value



When to prefer MAD

- MAD is a better measure of error than MSE if forecast error does not have a symmetric distribution.
- It gives us the average difference between F and D, disregarding the accuracy of F.



When to prefer MAPE

- MAPE (mean absolute percentage error): good measure of forecast error when the underlying forecast has significant **seasonality** and demand varies considerably from one period to the next.



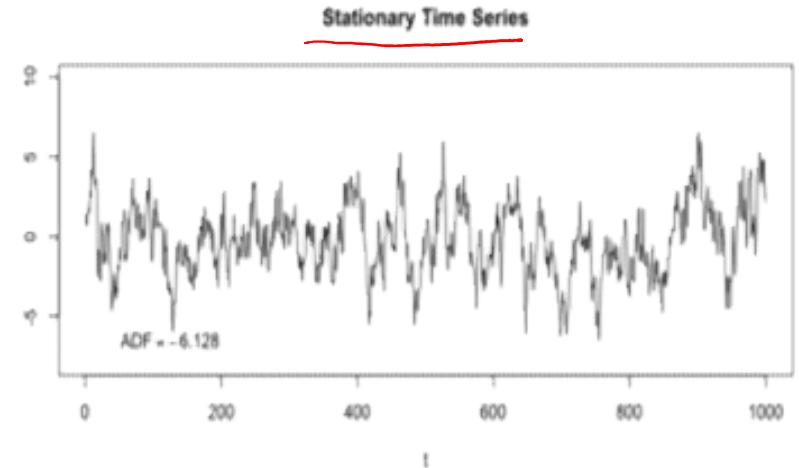
When to prefer MSE

- MSE (mean squared error): MSE is related to the **variance** of the forecast error.
- We estimate the random component of demand has a mean of 0 and a **variance** of MSE.
- The MSE penalizes large errors much more significantly than small errors because all errors are squared.
- Because of this, it is a good idea to use the MSE to compare forecasting methods if **the cost of a large error is much larger than the gains from very accurate forecasts**.
- MSE is appropriate when forecast error has a distribution that is symmetric about zero.

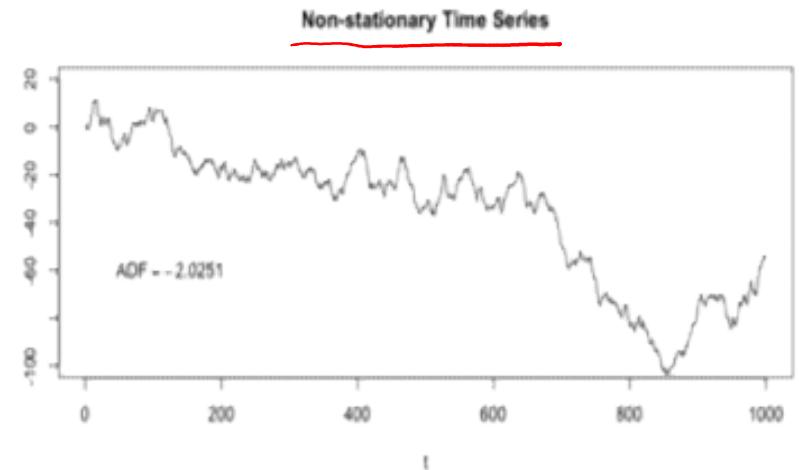
Stationary time series



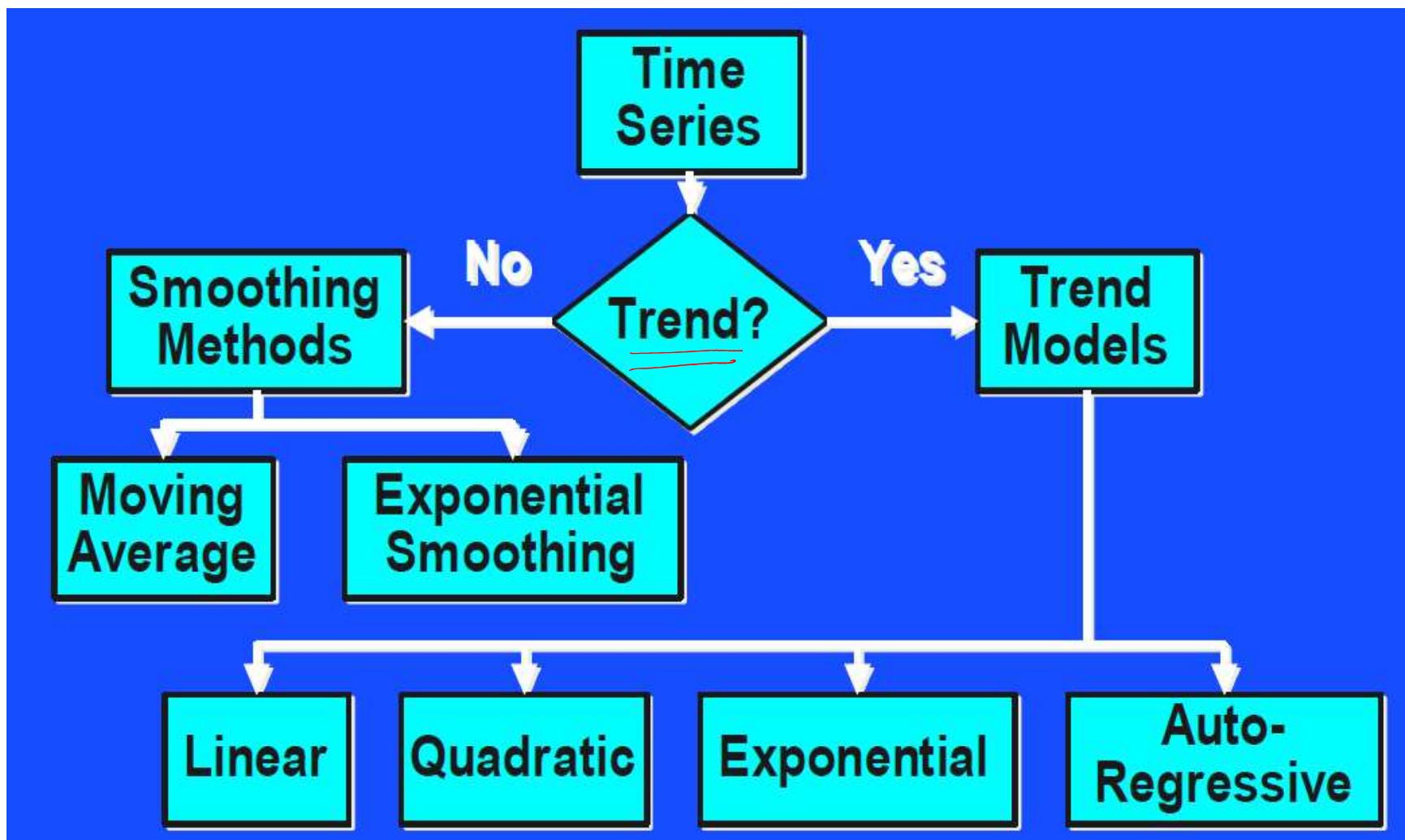
- *Time-series data that contain no trend, cyclical, or seasonal effects are said to be stationary.*



- Techniques used to forecast stationary data analyze only the irregular fluctuation effects.



Analysis





SMOOTHING TECHNIQUES

- Several techniques are available to forecast time-series data that are **stationary** or that include no significant trend, cyclical, or seasonal effects.
- These techniques are often referred to as **smoothing techniques** because they *produce forecasts based on “smoothing out” the irregular fluctuation effects in the time-series data.*
 - (1) naive forecasting models,
 - (2) averaging models, and
 - (3) exponential smoothing



Naïve Forecasting Models

- They are *simple models in which it is assumed that the more recent time periods of data represent the best predictions or forecasts for future outcomes.*
- Naïve models do not take into account data trend, cyclical effects, or seasonality.
- For this reason, naive models seem to work better with data that are reported on a daily or weekly basis or in situations that show no trend or seasonality.
- The simplest of the naïve forecasting methods is the model in which the forecast for a given time period is the value for the previous time period.

$$F_t = \underline{x_{t-1}}$$

where

F_t = the forecast value for time period t

x_{t-1} = the value for time period $t - 1$

Example

- Table is representing the total reported domestic rail, truck, and air shipments of bell peppers in the United States for a given year reported by the U.S. Department of Agriculture.
- Prediction for Jan next year can be taken as 412 using naïve model
- Use 336 as the prediction for Jan next year

Month	Shipments (millions of pounds)
January	336
February	308
March	582
April	771
May	935
June	808
July	663
August	380
September	333
October	412
November	458
December	412



Averaging Models

- **Averaging models** are computed by *averaging data from several time periods and using the average as the forecast for the next time period.*

Moving Averages

- Suppose we were to attempt to forecast the heating oil cost for October of year 3 by using MA as the forecasting method.
- It would seem to make sense to use the 12 months prior to October of year 3 (i.e. October of year 2 through September of year 3) to average for the new forecast.
- Suppose in September of year 3 the cost of heating oil is 53.3 cents

$$= \frac{56.7 + 57.2 + 58.0 + 58.2 + 58.3 + 57.7 + 56.7 + 56.8 + 55.5 + 53.8 + 52.8 + 53.3}{12} = 56.25$$

12

**Cost of Residential Heating Oil
(cents per gallon)**

Time Frame	Cost of Heating Oil
January (year 1)	66.1
February	66.1
March	66.4
April	64.3
May	63.2
June	61.6
July	59.3
August	58.1
September	58.9
October	60.9
November	60.7
December	59.4
January (year 2)	61.3
February	63.3
March	62.1
April	59.8
May	58.4
June	57.6
July	55.7
August	55.1
September	55.7
October	56.7
November	57.2
December	58.0
January (year 3)	58.2
February	58.3
March	57.7
April	56.7
May	56.8
June	55.5
July	53.8
August	52.8

Sep

53.3



-
- The Moving Average basically filters out rapid fluctuations .i.e. high frequency noise. Thus, it acts as a low-pass filter.

Disadvantages:

- It is difficult to choose the optimal length of time for which to compute the moving average, and
- moving averages do not usually adjust for such time-series effects as trend, cycles, or seasonality.
- To determine the more optimal lengths for which to compute the moving averages, we would need to forecast with several different average lengths and compare the errors produced by them.



Example

- Shown here are shipments (in millions of dollars) for electric lighting and wiring equipment over a 12-month period. Use these data to compute a 4-month moving average for all available months.

4-Month Moving Forecast				Month	Shipments
Month	Shipments	Average	Error		
January	1056	—	—	January	1056
February	1345	—	—	February	1345
March	1381	—	—	March	1381
April	1191	—	—	April	1191
May	1259	1243.25	15.75	May	1259
June	1361	1294.00	67.00	June	1361
July	1110	1298.00	-188.00	July	1110
August	1334	1230.25	103.75	August	1334
September	1416	1266.00	150.00	September	1416
October	1282	1305.25	-23.25	October	1282
November	1341	1285.50	55.50	November	1341
December	1382	1343.25	38.75	December	1382



Weighted Moving Averages

- A forecaster may want to place more weight on certain periods of time than on others. For example, a forecaster might believe that the previous month's value is three times as important in forecasting as other months.
- *A moving average in which some time periods are weighted differently than others is called a **weighted moving average***

As an example, suppose a 3-month weighted average is computed by weighting last month's value by 3, the value for the previous month by 2, and the value for the month before that by 1. This weighted average is computed as

$$\bar{x}_{\text{weighted}} = \frac{3(M_{t-1}) + 2(M_{t-2}) + 1(M_{t-3})}{6}$$

Example

- Compute a 4-month weighted moving average for the electric lighting and wiring data, using weights of 4 for last month's value, 2 for the previous month's value, and 1 for each of the values from the 2 months prior to that

Month	Shipments	4-Month Weighted Moving Average		Month	Shipments
		Forecast	Error		
January	1056	—	—	January	1056
February	1345	—	—	February	1345
March	1381	—	—	March	1381
April	1191	—	—	April	1191
May	1259	1240.9	18.1	May	1259
June	1361	1268.0	93.0	June	1361
July	1110	1316.8	-206.8	July	1110
August	1334	1201.5	132.5	August	1334
September	1416	1272.0	144.0	September	1416
October	1282	1350.4	-68.4	October	1282
November	1341	1300.5	40.5	November	1341
December	1382	1334.8	47.2	December	1382



Exponential Smoothing

- Another forecasting technique, **exponential smoothing**, is *used to weight data from previous time periods with exponentially decreasing importance in the forecast.*
- Exponential smoothing is accomplished by multiplying the actual value for the present time period, X_t , by a value between 0 and 1

EXPONENTIAL SMOOTHING

$$\underline{F_{t+1} = \alpha X_t + (1 - \alpha) \cdot F_t}$$

where

F_{t+1} = the forecast for the next time period ($t + 1$)

F_t = the forecast for the present time period (t)

X_t = the actual value for the present time period

α = a value between 0 and 1 referred to as the exponential smoothing constant.



Example

- The U.S. Census Bureau reports the total units of new privately owned housing started over a 16-year recent period in the United States are given here. Use exponential smoothing to forecast the values for each ensuing time period. Work the problem using $\alpha = .2, .5$, and $.8$.

Year	Total Units (1000)
1	1193
2	1014
3	1200
4	1288
5	1457
6	1354
7	1477
8	1474
9	1617
10	1641
11	1569
12	1603
13	1705
14	1848
15	1956
16	2068

Solution

Year	Total Units (1000)	$\alpha = .2$		$\alpha = .5$		$\alpha = .8$	
		F	e	F	e	F	e
1	1193	—	—	—	—	—	—
2	1014	1193.0	-179.0	1193.0	-179.0	1193.0	-179.0
3	1200	1157.2	42.8	1103.5	96.5	1049.8	150.2
4	1288	1165.8	122.2	1151.8	136.2	1170.0	118.0
5	1457	1190.2	266.8	1219.9	237.1	1264.4	192.6
6	1354	1243.6	110.4	1338.4	15.6	1418.5	-64.5
7	1477	1265.7	211.3	1346.2	130.8	1366.9	110.1
8	1474	1307.9	166.1	1411.6	62.4	1455.0	19.0
9	1617	1341.1	275.9	1442.8	174.2	1470.2	146.8
10	1641	1396.3	244.7	1529.9	111.1	1587.6	53.4
11	1569	1445.2	123.8	1585.5	-16.5	1630.3	-61.3
12	1603	1470.0	133.0	1577.2	25.8	1581.3	21.7
13	1705	1496.6	208.4	1590.1	114.9	1598.7	106.3
14	1843	1538.3	309.7	1647.6	200.4	1683.7	164.3
15	1956	1600.2	355.8	1747.8	208.2	1815.1	140.9
16	2068	1671.4	396.6	1851.9	216.1	1927.8	140.2
		$\alpha = .2$	$\alpha = .5$	$\alpha = .8$			
	MAD:	209.8	128.3	111.2			
	MSE:	53,110.5	21,628.6	15,245.4			



Exercise (HW)

- Following are time-series data for eight different periods. Use exponential smoothing to forecast the values for periods 3 through 8. Use the value for the first period as the forecast for the second period. Compute forecasts using two different values of alpha, $\alpha = 0.1$ and $\alpha = 0.8$. Compute the errors for each forecast and compare the errors produced by using the two different exponential smoothing constants

$$F_3 = 0.1 \times (228) + 0.9 \times 211 \\ \approx 212.7$$
$$F_4 = 0.1 \times (236) + 0.9 \times 212.7$$

Time Period	Value	Time Period	Value
1	211	5	242
2	228	6	227
3	236	7	217
4	241	8	203



Solution

Period	Value	$F(\alpha=.1)$	Error	$F(\alpha=.8)$	Error	Difference
1	211					
2	228	211				
3	236	213	23	225	11	12
4	241	215	26	234	7	19
5	242	218	24	240	2	22
6	227	220	7	242	-15	22
7	217	221	-4	230	-13	9
8	203	220	-17	220	-17	0

Using alpha of .1 produced forecasting errors that were larger than those using alpha = .8 for the first three forecasts. For the next two forecasts (periods 6 and 7), the forecasts using alpha = .1 produced smaller errors. Each exponential smoothing model produced the same amount of error in forecasting the value for period 8. There is no strong argument in favor of either model.



SEASONAL EFFECTS

- **Seasonal effects** are *patterns of data behavior that occur in periods of time of less than one year.*
- How can we separate out the seasonal effects?
- One of the main techniques for isolating the effects of seasonality is **decomposition**



Time series decomposition

Objective is to estimate the overall time series as a combination of long term trend and seasonality

- Additive Model
- Multiplicative Model
- The trend and seasonality can be decomposed using smoothing and regression methods
- Exponential smoothing is suitable with constant variance and no seasonality. Recommended for short-term forecast.
- Another method for stationarizing the time series is by doing transformation .e.g. Differencing



Procedure for decomposition

- Assess the trend component by smoothing or curve fitting (regression with time)
- Assess/account for seasonality i.e. deseasonalize the data
- For **Additive Adjustment**: Look at all periods of a given type (eg. First Qtr periods where data is quarterly ,or all August period where the data is monthly) & compute an average deviation of the actual values from the smooth or fitted values in those periods. The average can then be added to the trend to adjust for seasonality.



Procedure for decomposition

- For **multiplicative adjustment**: Instead of calculating the average deviation, compute an average ratio, also called seasonal indices, of the actual values to the smooth or fitted values in those periods. The indices are then used as multiplier to adjust for seasonality.
- Forecast by projecting trend component in to the future and then adding or multiplying the seasonal component, as the case may be according to your chosen model



Example

- Let us consider the following quarterly sales data for J C Penney Sales company for the last 24 quarters
- Forecast sales for the 25th Quarter ?

Period	JC Penney Sales	Period	JC Penney Sales
1	4452	13	7339
2	4507	14	7104
3	5537	15	7639
4	8157	16	9661
5	6481	17	7528
6	6420	18	7207
7	7208	19	7538
8	9509	20	9573
9	6755	21	7522
10	6483	22	7211
11	7129	23	7729
12	9072	24	9542



Solution

Step-1: Fit a trend line. We can have a Linear or a polynomial fit using Regression with t

$$\hat{y}_t = 5903.2174 + 118.75261t$$

$$\hat{y}_t = 6354.9514 + 118.75261t - 9.4274932(t - 12.5)^2$$

The forecast value of trend component for the 25th Qtr will be

$$\begin{aligned}\hat{y}_{25} &= 5903.2174 + 118.75261(25) \\ &= 8872\end{aligned}$$

Step-2: Compute additive/multiplicative seasonal adjustment factor for the first quarter

Period, t	y_t	\hat{y}_t	$y_t - \hat{y}_t$	$\frac{y_t}{\hat{y}_t}$
1	4452	6022	-1570	.7393
5	6481	6497	-16	.9975
9	6755	6972	-217	.9685
13	7339	7447	-108	.9855
17	7528	7922	-394	.9503
21	7522	8397	-875	.8958
	<u>-3180</u>		<u>5.5369</u>	

Then note that the average $y_t - \hat{y}_t$ is

$$\frac{-3180}{6} = -530$$

and the average y_t/\hat{y}_t is

$$\frac{5.5369}{6} = .9228$$

Step-3: So, forecast for the 25th Qtr period is obtained by adjusting the trend fit for seasonality as either

$$\hat{y}_{25} = \underline{8872} + (-530) = \underline{\underline{8342}}$$

(making use of an "additive" seasonality adjustment) or as

$$\hat{y}_{25} = \underline{8872} (.9228) = \underline{\underline{8187}}$$

(making use of a "multiplicative" seasonality adjustment).



Exercise (HW)

- For the same example, what will be the sales forecast for the 28th Qtr period? (Hint: calculate seasonality factor for the 4th Qtr .i.e. 4, 8, 12,16, 20, 24)



Time series data raises new technical issues

- Time lags
- Correlation over time (serial correlation, a.k.a. autocorrelation)
- Forecasting models built on regression methods like autoregressive (AR) models
- Conditions under which dynamic effects can be estimated, and how to estimate them
- Calculation of standard errors when the errors are serially correlated



Stationarity

- A common assumption in many time series techniques is that the data are stationary.
- A stationary process has the property that the mean, variance and autocorrelation structure do not change over time.
- If past effects accumulate and the values increase toward infinity, then stationarity is not met.



Independent and identically distributed (iid) noise

- simplest model for a time series is one in which there is no trend or seasonal component and in which the observations are simply independent and identically distributed (iid) random variables with zero mean.
- We refer to such a sequence of random variables X_1, X_2, \dots as iid noise
$$\{X_t\} \sim \text{IID}(0, \sigma^2)$$
- Indicates that the random variables X_t are independent and identically distributed, each with mean 0 and variance σ^2 .
- Although iid noise is a rather uninteresting process for forecasting, it plays an important role as a building block for more complicated time series models.



BITS Pilani
Pilani Campus

Forecasting

Akanksha Bharadwaj
Asst. Professor, CS/IS



BITS Pilani
Pilani Campus



SS ZG536, ADV STAT TECHNIQUES FOR ANALYTICS Contact Session 11



Python example: Minimum Daily Temperatures Dataset

- The dataset describes the minimum daily temperatures over 10 years (1981-1990) in the city Melbourne, Australia.
- The units are in degrees Celsius and there are 3,650 observations. The source of the data is credited as the Australian Bureau of Meteorology.

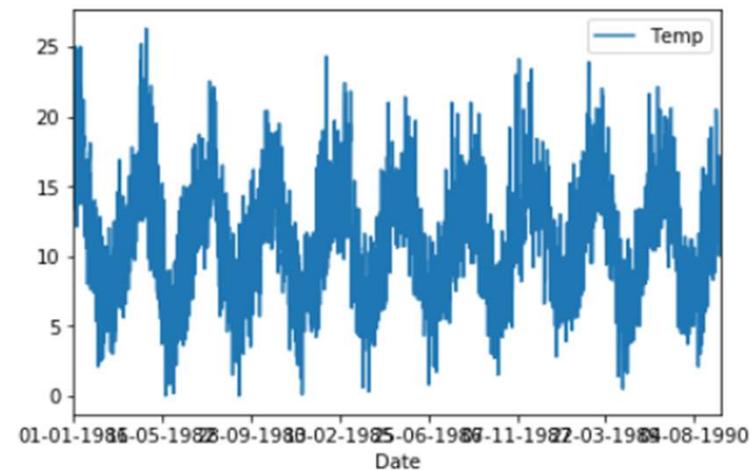


Load the dataset as a Pandas Series.

- Running the example prints the first 5 rows from the loaded dataset.
- A line plot of the dataset is then created.

```
from pandas import read_csv  
from matplotlib import pyplot  
series = read_csv('C:\Akanksha\Data\daily-minimum-  
temperatures.csv', header=0, index_col=0)  
print(series.head())  
series.plot()  
pyplot.show()
```

Date	Temp
01-01-1981	20.7
02-01-1981	17.9
03-01-1981	18.8
04-01-1981	14.6
05-01-1981	15.8



Quick Check for Autocorrelation



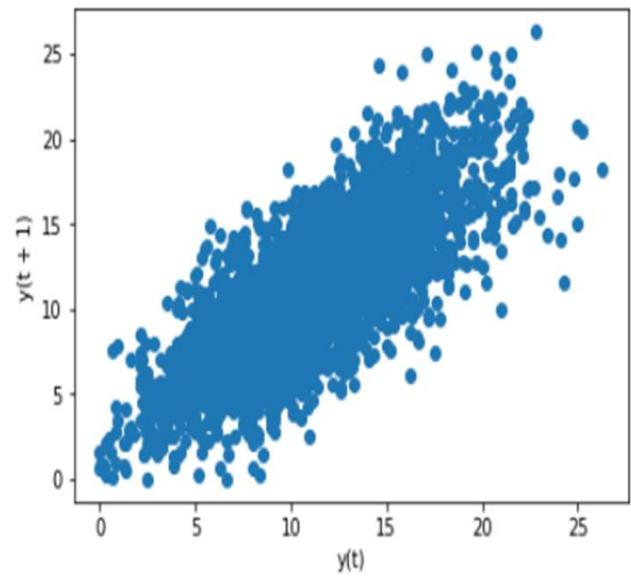
- We can plot the observation at the previous time step (t) with the observation at the next time step ($t+1$) as a scatter plot.
 - This could be done manually by first creating a lag version of the time series dataset and using a built-in scatter plot function in the Pandas library.
 - Easy method: Pandas provides a built-in plot to do exactly this, called the `lag_plot()` function.
-



- Running the example plots the temperature data (t) on the x-axis against the temperature on the previous day ($t+1$) on the y-axis.

```
from pandas import read_csv  
from matplotlib import pyplot  
from pandas.plotting import lag_plot  
series = read_csv('C:\Akanksha\Data\daily-minimum-  
temperatures.csv', header=0, index_col=0)  
lag_plot(series)  
pyplot.show()
```

- We can see a large ball of observations along a diagonal line of the plot. It clearly shows a relationship or **some correlation**.





Pearson correlation coefficient

- We can use a statistical test like the Pearson correlation coefficient.
- This produces a number to summarize how correlated two variables are between -1 (negatively correlated) and +1 (positively correlated) with small values close to zero indicating low correlation and high values above 0.5 or below -0.5 showing high correlation.
- Correlation can be calculated easily using the corr() function on the DataFrame of the lagged dataset.



-
- It shows a strong positive correlation (0.77) between the observation and the lag=1 value.
 - This is good but tedious if we want to check a large number of lag variables in our time series.

```
from pandas import read_csv
from pandas import DataFrame
from pandas import concat
from matplotlib import pyplot
series = read_csv('C:\Akanksha\Data\daily-minimum-temperatures.csv', header=0, index_col=0)
values = DataFrame(series.values)
dataframe = concat([values.shift(1), values], axis=1)
dataframe.columns = ['t-1', 't+1']
result = dataframe.corr()
print(result)
```

	t-1	t+1
t-1	1.00000	0.77487
t+1	0.77487	1.00000



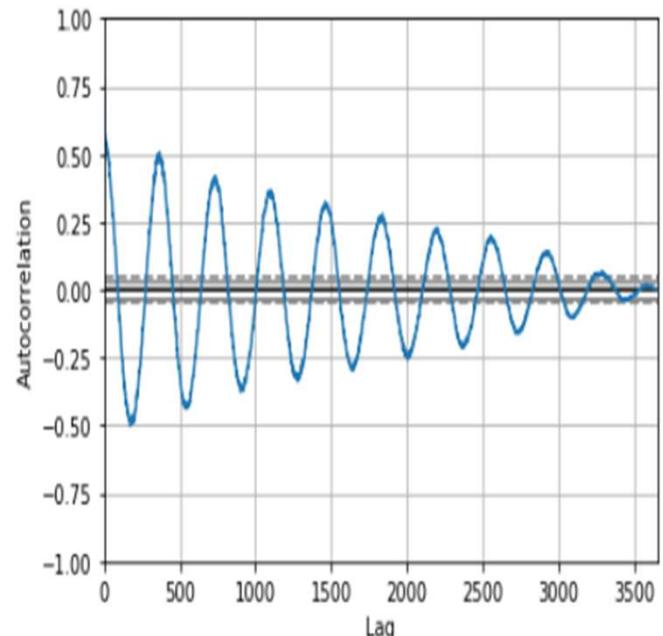
Autocorrelation Plots

- We can plot the correlation coefficient for each lag variable.
- We could manually calculate the correlation values for each lag variable and plot the result.
- But, Pandas provides a built-in plot called the `autocorrelation_plot()` function.
- The plot provides the lag number along the x-axis and the correlation coefficient value between -1 and 1 on the y-axis.
- The plot also includes solid and dashed lines that indicate the 95% and 99% confidence interval for the correlation values.
- Correlation values above these lines are more significant than those below the line, providing a threshold or cutoff for selecting more relevant lag values.



- Figure shows the swing in positive and negative correlation as the temperature values change cross summer and winter seasons each previous year.

```
from pandas import read_csv
from matplotlib import pyplot
from pandas.plotting import autocorrelation_plot
series = read_csv('C:\Akanksha\Data\daily-minimum-
    temperatures.csv', header=0, index_col=0)
autocorrelation_plot(series)
pyplot.show()
```





Autoregression Model

- An autoregression model is a linear regression model that uses lagged variables as input variables.
- We could calculate the linear regression model manually using the LinearRegression class in scikit-learn and manually specify the lag input variables to use.
- Alternately, the statsmodels library provides an autoregression model that automatically selects an appropriate lag value using statistical tests and trains a linear regression model. It is provided in the AR class.
- We can use this model by first creating the model AR() and then calling fit() to train it on our dataset. This returns an AR Result object.
- Once fit, we can use the model to make a prediction by calling the predict() function for a number of observations in the future.



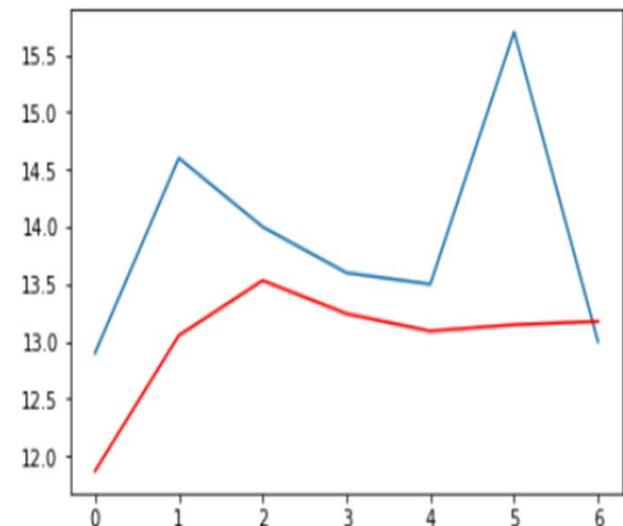
- Running the example first prints the chosen optimal lag and the list of coefficients in the trained linear regression model.
- We can see that a 29-lag model was chosen and trained. This is interesting given how close this lag is to the average number of days in a month.
- The 7 day forecast is then printed and the mean squared error of the forecast is summarized

```
Lag: 29
Coefficients: [ 5.57543506e-01  5.88595221e-01 -9.08257090e-02  4.82615092e-02
 4.00650265e-02  3.93020055e-02  2.59463738e-02  4.46675960e-02
 1.27681498e-02  3.74362239e-02 -8.11700276e-04  4.79081949e-03
 1.84731397e-02  2.68908418e-02  5.75906178e-04  2.48096415e-02
 7.40316579e-03  9.91622149e-03  3.41599123e-02 -9.11961877e-03
 2.42127561e-02  1.87870751e-02  1.21841870e-02 -1.85534575e-02
 -1.77162867e-03  1.67319894e-02  1.97615668e-02  9.83245087e-03
 6.22710723e-03 -1.37732255e-03]
predicted=11.871275, expected=12.900000
predicted=13.053794, expected=14.600000
predicted=13.532591, expected=14.000000
predicted=13.243126, expected=13.600000
predicted=13.091438, expected=13.500000
predicted=13.146989, expected=15.700000
predicted=13.176153, expected=13.000000
Test MSE: 1.502
```

Predictions from fixed AR model



- A plot of the expected (blue) vs the predicted values (red) is made.
 - The forecast does look pretty good (about 1 degree Celsius out each day), with big deviation on day 5.
 - The statsmodels API does not make it easy to update the model as new observations become available.
 - One way would be to re-train the AR model each day as new observations become available, and that may be a valid approach, if not computationally expensive.
-





Moving Averages model

- The moving average (MA) method models the next step in the sequence as a linear function of the residual errors from a mean process at prior time steps.
- A moving average model is different from calculating the moving average of the time series.
- The notation for the model involves specifying the order of the model q as a parameter to the MA function, e.g. $\text{MA}(q)$. For example, $\text{MA}(1)$ is a first-order moving average model.
- The method is suitable for univariate time series without trend and seasonal components.



Moving Average (MA) Models

The notation $\text{MA}(q)$ refers to the moving average model of order q :

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

- where μ is the mean of the series, the $\theta_1, \dots, \theta_q$ are the parameters of the model and the $\varepsilon_t, \varepsilon_{t-1}, \dots, \varepsilon_{t-q}$ are white noise error terms. The value of q is called the order of the MA model.
- Thus, a moving-average model is conceptually a linear regression of the current value of the series against current and previous (observed) white noise error terms or random shocks.
- The random shocks at each point are assumed to be mutually independent and to come from the same distribution, typically a normal distribution, with location at zero and constant scale.
- The distinction in this model is that these random shocks are propagated to future values of the time series. Fitting the MA estimates is more complicated than with AR models because the error terms are not observable.



Moving Average Models

-
- The MA model should not be confused with Moving Average Smoothing as they are not the same thing.
 - A **moving average** term in a time series model is a past error (multiplied by a coefficient).

Let r.v. $w_t \sim \text{iid}(0, \sigma^2)$, meaning that the w_t are identically, independently distributed, having mean 0 and the same variance.

The **1st order moving average** model, denoted by MA(1) is:

$$x_t = \mu + w_t + \theta_1 w_{t-1}$$

The **2nd order moving average** model, denoted by MA(2) is:

$$x_t = \mu + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2}$$

The **qth order moving average** model, denoted by MA(q) is:

$$x_t = \mu + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \dots + \theta_q w_{t-q}$$

Theoretical Properties of a Time Series with an MA(1) Model



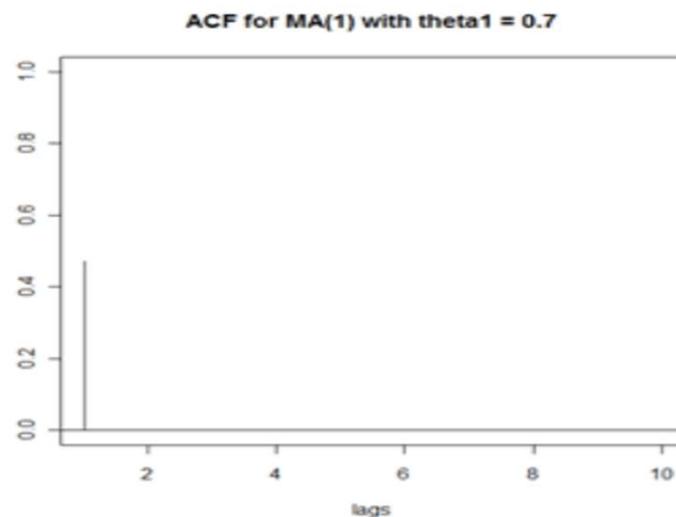
- Mean is $E(x_t) = \mu$
- Variance is $Var(x_t) = \sigma_w^2(1 + \theta_1^2)$
- Autocorrelation function (ACF) is:

$$\rho_1 = \frac{\theta_1}{1 + \theta_1^2}, \text{ and } \rho_h = 0 \text{ for } h \geq 2$$

Example

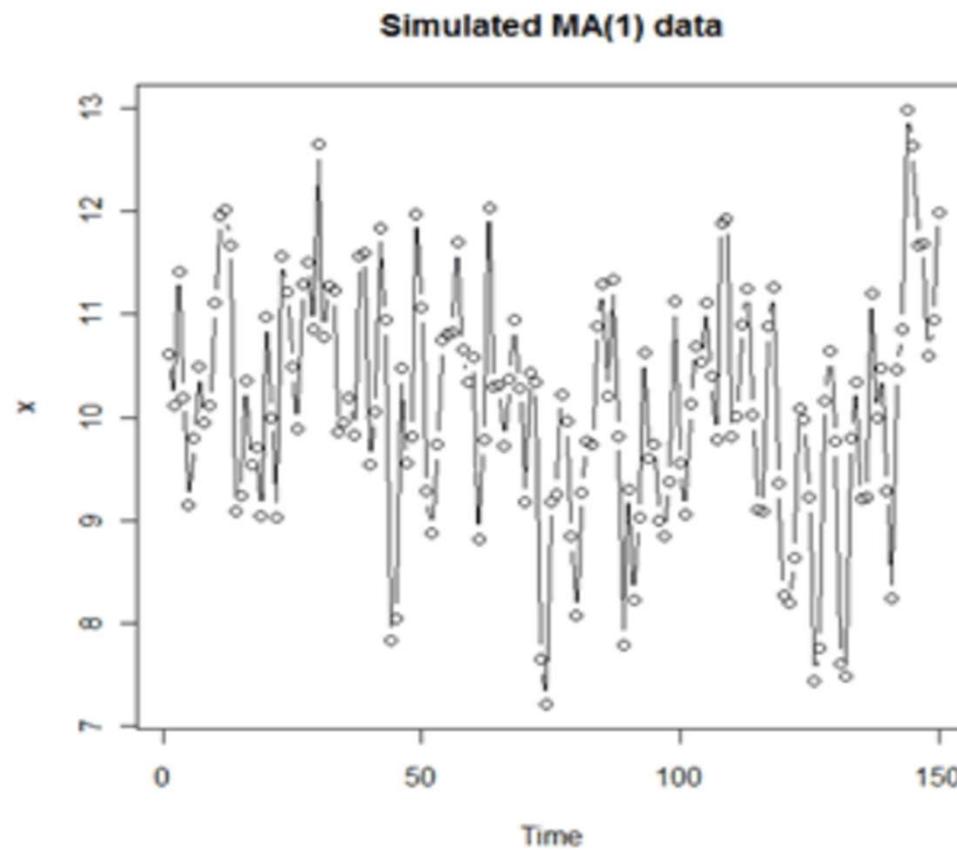
Suppose that an MA(1) model is $x_t = 10 + w_t + .7w_{t-1}$, where $w_t \stackrel{iid}{\sim} N(0, 1)$. Thus the coefficient $\theta_1 = 0.7$. The theoretical ACF is given by:

$$\rho_1 = \frac{0.7}{1 + 0.7^2} = 0.4698, \text{ and } \rho_h = 0 \text{ for all lags } h \geq 2$$

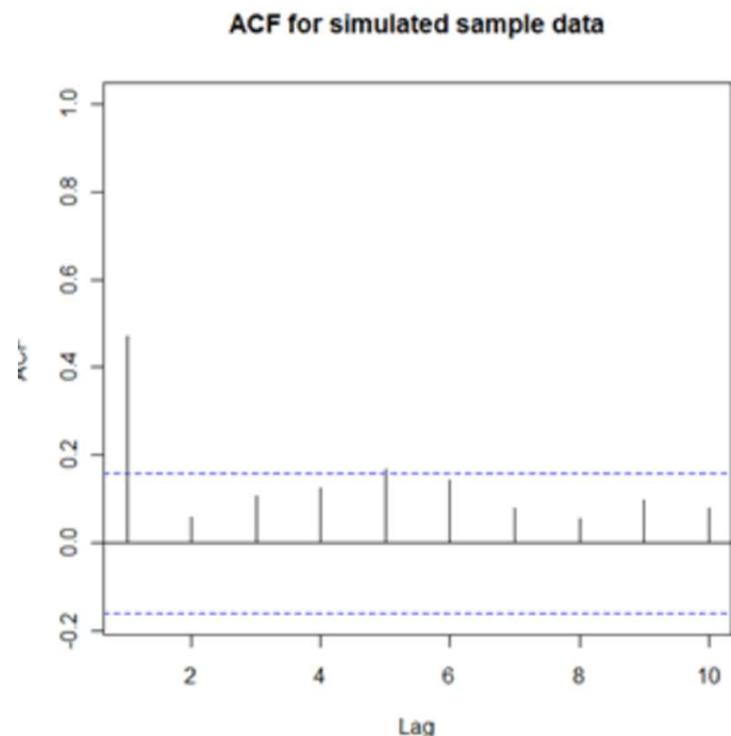


- That the *only nonzero value in the theoretical ACF is for lag 1*. All other autocorrelations are 0. Thus a sample ACF with a significant autocorrelation only at lag 1 is an indicator of a possible MA(1) model.

Using R, we simulated $n = 100$ sample values using the model $x_t = 10 + w_t + .7w_{t-1}$ where $w_t \stackrel{iid}{\sim} N(0, 1)$. For this simulation, a time series plot of the sample data follows. We can't tell much from this plot.



- The sample ACF does not match the theoretical pattern of the underlying MA(1), which is that all autocorrelations for lags past 1 will be 0.
- A different sample would have a slightly different sample ACF shown below, but would likely have the same broad features.



Theoretical Properties of a Time Series with an MA(2) Model



Theoretical Properties of a Time Series with an MA(2) Model

For the MA(2) model, theoretical properties are the following:

- Mean is $E(x_t) = \mu$
- Variance is $Var(x_t) = \sigma_w^2(1 + \theta_1^2 + \theta_2^2)$
- Autocorrelation function (ACF) is:

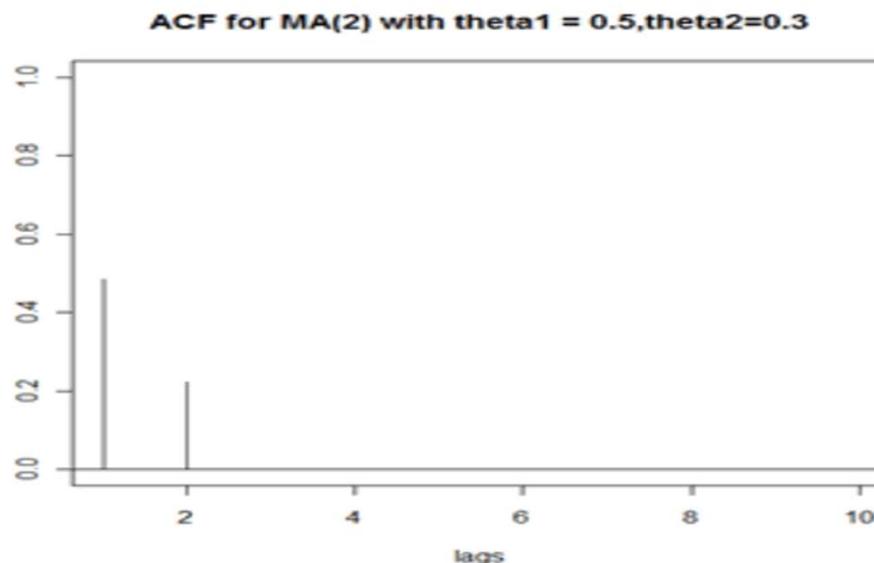
$$\rho_1 = \frac{\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2}, \rho_2 = \frac{\theta_2}{1 + \theta_1^2 + \theta_2^2}, \text{ and } \rho_h = 0 \text{ for } h \geq 3$$

Example

Consider the MA(2) model $x_t = 10 + w_t + .5w_{t-1} + .3w_{t-2}$, where $w_t \stackrel{iid}{\sim} N(0, 1)$. The coefficients are $\theta_1 = 0.5$ and $\theta_2 = 0.3$. Because this is an MA(2), the theoretical ACF will have nonzero values only at lags 1 and 2.

Values of the two nonzero autocorrelations are:

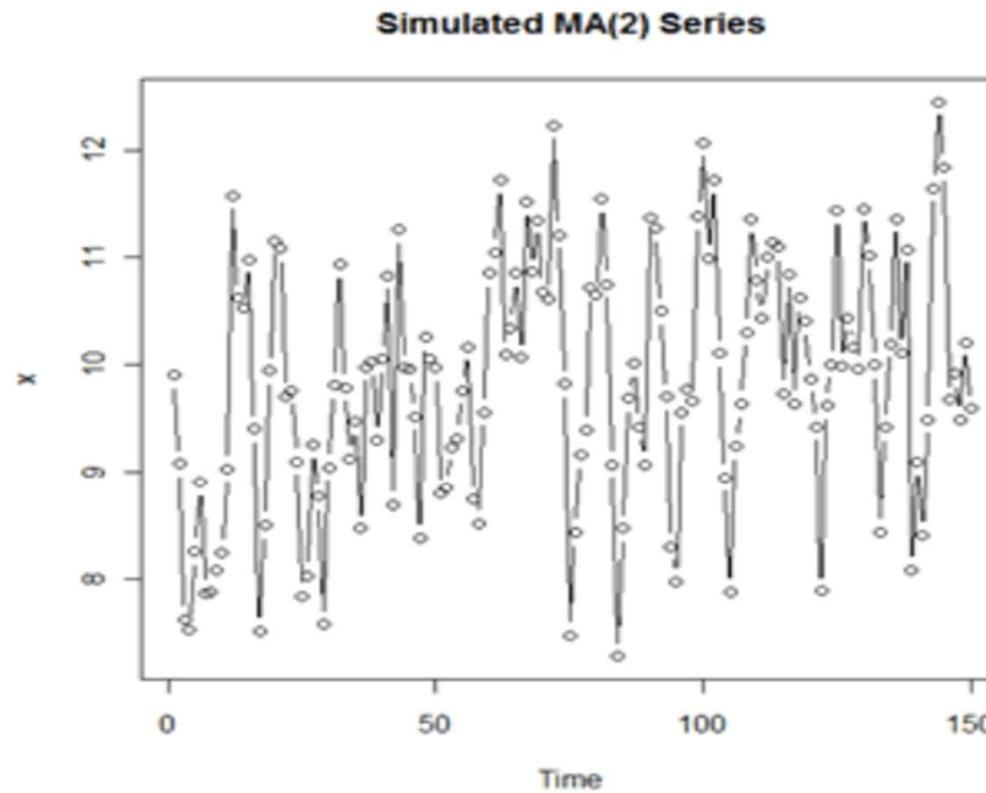
$$\rho_1 = \frac{0.5 + 0.5 \times 0.3}{1 + 0.5^2 + 0.3^2} = 0.4851 \text{ and } \rho_2 = \frac{0.3}{1 + 0.5^2 + 0.3^2} = 0.2239$$



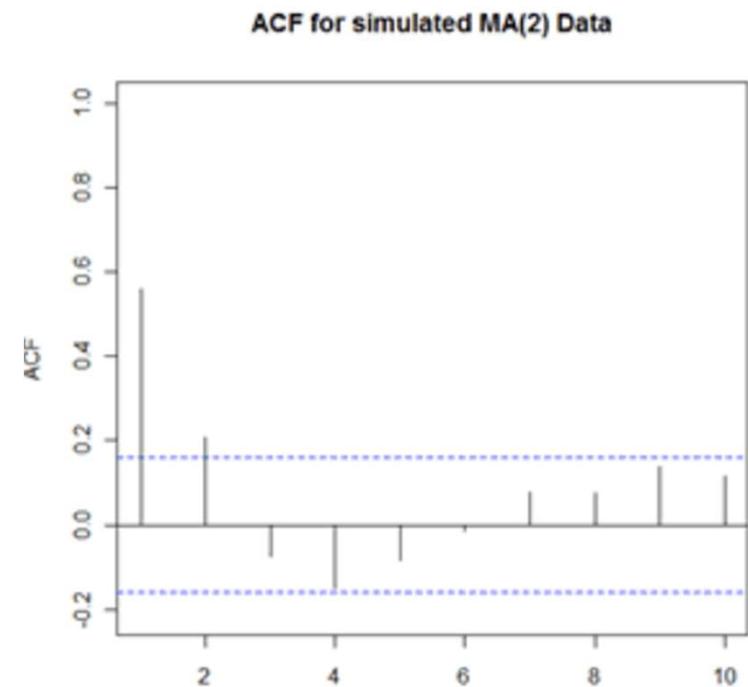
The only nonzero values in the theoretical ACF are for lags 1 and 2. Autocorrelations for higher lags are 0. So, a sample ACF with significant autocorrelations at lags 1 and 2, but non-significant autocorrelations for higher lags indicates a possible MA(2) model.

We simulated $n = 150$ sample values for the model

$x_t = 10 + w_t + .5w_{t-1} + .3w_{t-2}$, where $w_t \stackrel{iid}{\sim} N(0, 1)$. The time series plot of the data follows. A with the time series plot for the MA(1) sample data, you can't tell much from it.



- The pattern is typical for situations where an MA(2) model may be useful. There are two statistically significant “spikes” at lags 1 and 2 followed by non-significant values for other lags.
- Note that due to sampling error, the sample ACF did not match the theoretical pattern exactly.



ACF for General MA(q) Models



A property of MA(q) models in general is that there are nonzero autocorrelations for the first q lags and autocorrelations = 0 for all lags > q.

Non-uniqueness of connection between values of θ_1 and ρ_1 in MA(1) Model.

In the MA(1) model, for any value of θ_1 , the reciprocal $1/\theta_1$ gives the same value for:

$$\rho_1 = \frac{\theta_1}{1 + \theta_1^2}$$

As an example, use +0.5 for θ_1 , and then use $1/(0.5) = 2$ for θ_1 . You'll get $\rho_1 = 0.4$ in both instances.

To satisfy a theoretical restriction called **invertibility**, we restrict MA(1) models to have values with absolute value less than 1. In the example just given, $\theta_1 = 0.5$ will be an allowable parameter value, whereas $\theta_1 = 1/0.5 = 2$ will not.



Python Code

- We can use the ARMA class to create an MA model and setting a zeroth-order AR model. We must specify the order of the MA model in the order argument.

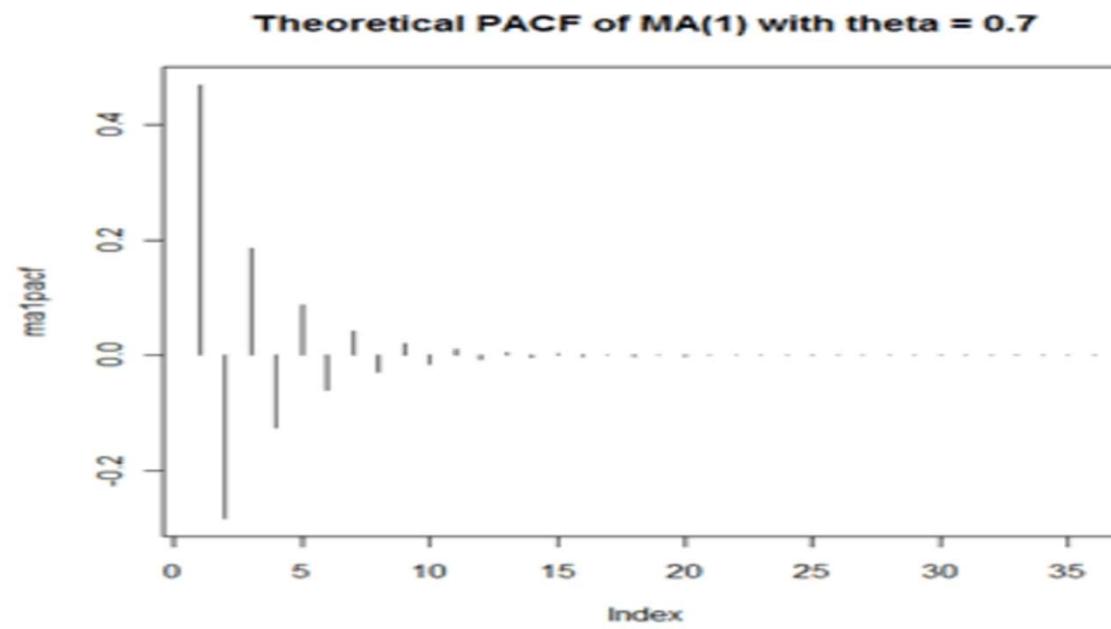
```
# MA example
from statsmodels.tsa.arima_model import ARMA
from random import random
# contrived dataset
data = [x + random() for x in range(1, 100)]
# fit model
model = ARMA(data, order=(0, 1))
model_fit = model.fit(disp=False)
# make prediction
yhat = model_fit.predict(len(data), len(data))
print(yhat)
```

```
In [1]: # MA example
from statsmodels.tsa.arima_model import ARMA
from random import random
# contrived dataset
data = [x + random() for x in range(1, 100)]
# fit model
model = ARMA(data, order=(0, 1))
model_fit = model.fit(disp=False)
# make prediction
yhat = model_fit.predict(len(data), len(data))
print(yhat)
```

[75.00274978]

Important conclusion for MA(q) Process

- For an MA model, the theoretical PACF does not shut off, but instead tapers toward 0 in some manner.
- A clearer pattern for an MA model is in the ACF. The ACF will have non-zero autocorrelations only at lags involved in the model.





Box-Jenkins models

ARIMA models, also called Box-Jenkins models, are models that may possibly include autoregressive terms, moving average terms, and differencing operations. Various abbreviations are used:

- When a model only involves autoregressive terms it may be referred to as an AR model. When a model only involves moving average terms, it may be referred to as an MA model.
- When no differencing is involved, the abbreviation ARMA may be used.



Autoregressive Moving Average (ARMA)

- The Autoregressive Moving Average (ARMA) method models the next step in the sequence as a linear function of the observations and residual errors at prior time steps.
- It combines both Autoregression (AR) and Moving Average (MA) models.
- The notation for the model involves specifying the order for the AR(p) and MA(q) models as parameters to an ARMA function, e.g. ARMA(p, q). An ARIMA model can be used to develop AR or MA models.
- The method is suitable for univariate time series without trend and seasonal components.



ARMA

ARMA model is simply the merger between AR(p) and MA(q) models:

- AR(p) models try to explain the momentum and mean reversion effects often observed in trading markets (market participant effects).
- MA(q) models try to capture the shock effects observed in the white noise terms. These shock effects could be thought of as unexpected events affecting the observation process e.g. Surprise earnings, wars, attacks, etc.

ARMA Models



-
- Autoregressive and Moving Average Models can be combined to form ARMA models
 - ARMA(p,q) time series model has the following form:

$$X_t = w_t + \sum_{i=1}^p \phi_i X_{t-i} + \sum_{j=1}^q \theta_j w_{t-j}$$

where $\phi_p, \theta_q \neq 0$ and $w_t \sim \text{white noise}$



ARMA(1,1)

ARMA(1,1) model is:

$$x(t) = a^*x(t-1) + b^*e(t-1) + e(t)$$

$e(t)$ is white noise with $E[e(t)] = 0$



Identifying the Model

- Three items should be considered to determine a first guess at an ARIMA model: a time series plot of the data, the ACF and the PACF.
- Time series plot will show if there is a trend, if so, detrend it by taking first difference of the consecutive time series values, transforming the series, smoothing or by subtracting a regression estimate of the trend
- For an ARMA model, ACF and PACF gives only a guess to select the p , q values.

	$AR(p)$	$MA(q)$	$ARMA(p, q)$
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

Estimating & diagnosing a possible model



- Estimate the model using software such as R, SAS, Minitab. etc.
- Once the model has been estimated, do the following for diagnosis:
 - Look at the significance of the coefficients. Compare p values to the significance level or calculate a t-statistic, $t = \text{estimated coeff.}/\text{std. error of coeff}$ & compare with $t_{\alpha, df}$. When n is large, coeff./std. error of coeff can be compared to 1.96.
 - Look at ACF of the residuals. For a good model, ACF of the residual series should be non-significant.
 - Look at time series plot of the residuals for randomness
 - Look at Box-Pierce tests for possible residual autocorrelation at various lags.
- If something looks wrong, revise your guess of the model and do the steps again.



What if more than one model looks OK !!

- Possibly choose the model with the fewest parameter
- Examine standard error of forecast values such as MSE, MAPE and pick the model with the lowest standard errors
- Compare models using criterion such as AIC and BIC.



Example: choosing from competing models

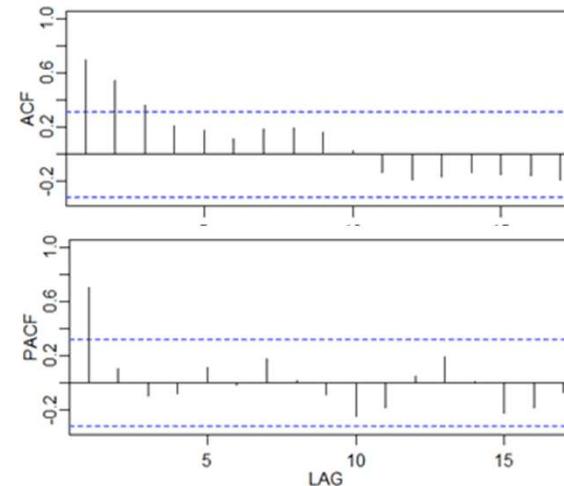
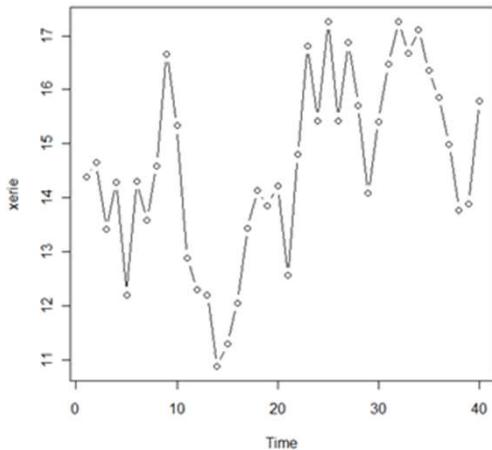
- Let an AR process gives the following data

Order of AR process, p	1	2	3	4	5
coeff	0.5970	0.7111 -0.1912	0.7136 -0.2003 0.0128	0.7137 -0.2016 0.0176 -0.0066	0.7141 -0.2027 0.0302 -0.0515 0.629
AIC	5751.32	5679.274	5680.945	5682.857	5676.917
SSE	2072.832	1997.007	1996.678	1996.590	1988.660

- Which AR process order will you choose?
- AR(1) is the best as it is parsimonious given not much difference in AIC.

Example

- N=40 consecutive annual measurements of the lake Erie in a particular month are given
- Model identification: A time series plot of the data is obtained, the ACF and the PACF are drawn

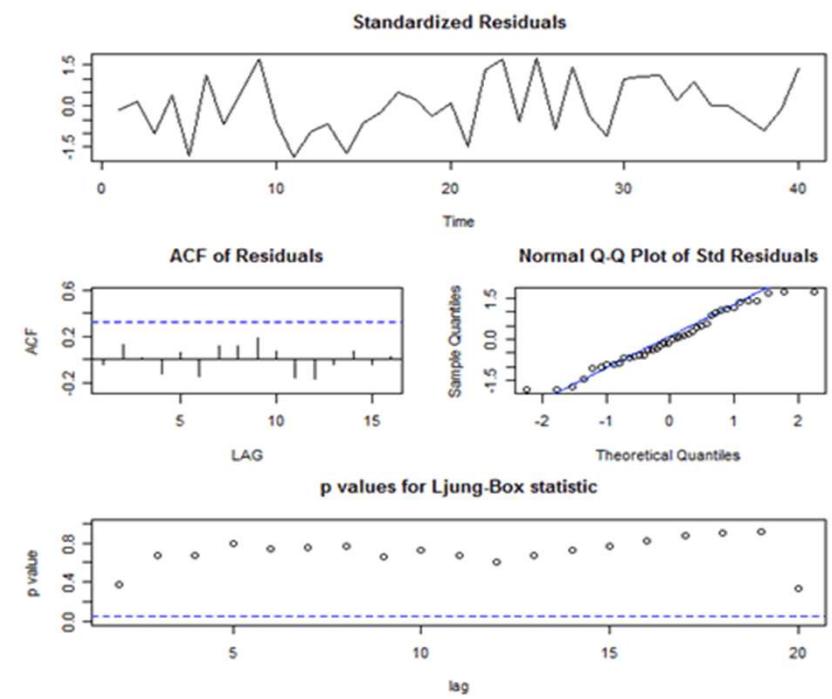


- The plot shows almost no trend. The ACF tapers off & the PACF shows a single spike. **AR(1)** _____ model is indicated.

Example



- N=40 consecutive annual measurements of the lake Erie in a particular month are given
- Model estimation: Using the software, the AR(1) model was estimated to be $x_t = 4.522 + 0.6909 x_{t-1}$
- Model diagnosis: We check the z value = 6.315 for the AR coeff. which is statistically significant and do the residual diagnostics
- The time series plot of the residuals no trend, no change in variance – Good!
The ACF shows no significant autocorrelations – Good!
Q-Q plot shows residuals are normally distributed - Good!
- Thus the estimated model can be used for forecasting





So how do we decide the values of p and q ?

- To fit data to an ARMA model, we use the Akaike Information Criterion (AIC) across a subset of values for p,q to find the model with minimum AIC and then apply the Ljung-Box test to determine if a good fit has been achieved, for particular values of p,q .
- If the p-value of the test is greater than the required significance, we can conclude that the residuals are independent and white noise.



Python code

```
# ARMA example
from statsmodels.tsa.arima_model import ARMA
from random import random
# contrived dataset
data = [random() for x in range(1, 100)]
# fit model
model = ARMA(data, order=(2, 1))
model_fit = model.fit(disp=False)
# make prediction
yhat = model_fit.predict(len(data), len(data))
print(yhat)
```

```
: # ARMA example
from statsmodels.tsa.arima_model import ARMA
from random import random
# contrived dataset
data = [random() for x in range(1, 100)]
# fit model
model = ARMA(data, order=(2, 1))
model_fit = model.fit(disp=False)
# make prediction
yhat = model_fit.predict(len(data), len(data))
print(yhat)

[0.48230949]
```

Autoregressive Integrated Moving Average (ARIMA)



- The Autoregressive Integrated Moving Average (ARIMA) method models the next step in the sequence as a linear function of the differenced observations and residual errors at prior time steps.
- It combines both Autoregression (AR) and Moving Average (MA) models as well as a differencing pre-processing step of the sequence to make the sequence stationary, called integration (I).
- The notation for the model involves specifying the order for the AR(p), I(d), and MA(q) models as parameters to an ARIMA function, e.g. ARIMA(p, d, q). An ARIMA model can also be used to develop AR, MA, and ARMA models.
- The method is suitable for univariate time series with trend and without seasonal components

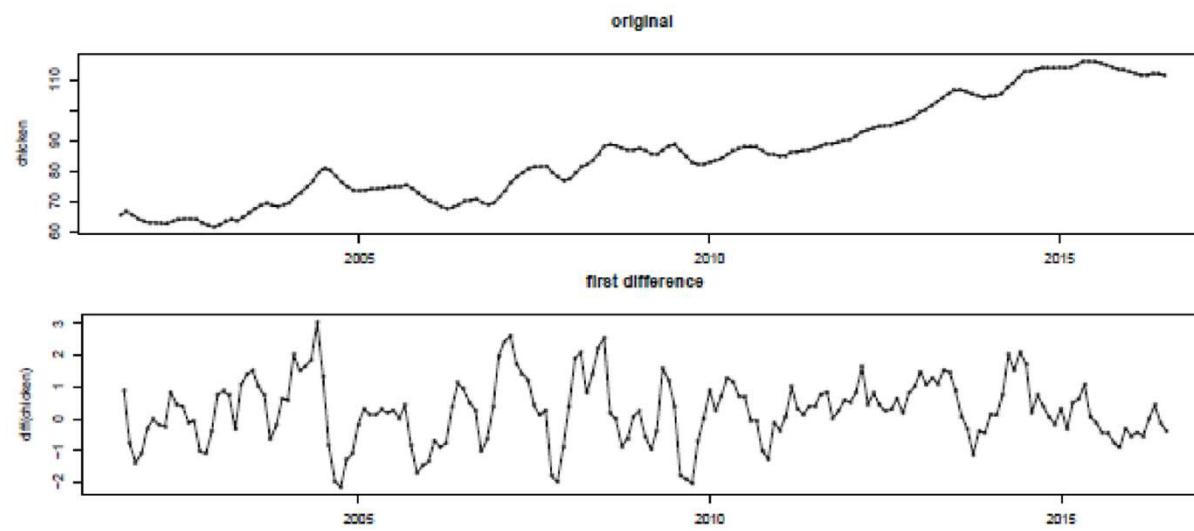


ARIMA Models

- ARIMA stands for Autoregressive, Integrated, Moving Average
- AR and MA Models work on stationary time series and I is a preprocessing procedure to “stationarize” the time series, if required, through differencing operation.
- ARIMA Model is specified by 3 parameters ARIMA(p,d,q)
- Eg. A MA(2) Model will be specified as ARIMA(0,0,2)
- ARIMA Model can be configured to perform the function of an ARMA Model, and even a simple AR, I or MA Model.

ARIMA Models

- Real life data sets are non-stationary.
- During the model identification stage, if the time series plot indicates non-stationarity, we stationarize the process
- Eg. If the process has a trend then we can stationarize by differencing and the ARMA process so obtained will be analysed using the tools already discussed.





Python code

```
# ARIMA example
from statsmodels.tsa.arima_model import ARIMA
from random import random
# contrived dataset
data = [x + random() for x in range(1, 100)]
# fit model
model = ARIMA(data, order=(1, 1, 1))
model_fit = model.fit(disp=False)
# make prediction
yhat = model_fit.predict(len(data), len(data), typ='levels')
print(yhat)
```

```
# ARIMA example
from statsmodels.tsa.arima_model import ARIMA
from random import random
# contrived dataset
data = [x + random() for x in range(1, 100)]
# fit model
model = ARIMA(data, order=(1, 1, 1))
model_fit = model.fit(disp=False)
# make prediction
yhat = model_fit.predict(len(data), len(data), typ='levels')
print(yhat)
```

[100.51557447]



References

- Introduction to Time Series and Forecasting by Peter J. Brockwell, Richard A. Davis
- Applied Business Statistics by Ken Black
- Introduction-to-time-series-forecasting by Jason Brownlee
- https://www.math-stat.unibe.ch/e237483/e237655/e243381/e281679/files281691/Chap12_ger.pdf
- <https://faculty.chicagobooth.edu/jeffrey.russell/teaching/bstats/timeseries.pdf>



BITS Pilani
Pilani Campus

Multivariate Analytics

Akanksha Bharadwaj
Asst. professor, CS/IS Department



SS ZG536, ADV STAT TECHNIQUES FOR ANALYTICS

Contact Session 12

Introduction

- In science and in real life, we are often interested in two (or more) random variables at the same time.
- For example, we might measure the IQ and birthweight of children, or the level of air pollution and rate of respiratory illness in cities.



BITS Pilani

Pilani Campus



Joint Distribution

Joint PMF

- Suppose X and Y are two discrete random variables and that X takes values $\{x_1, x_2, \dots, x_n\}$ and Y takes values $\{y_1, y_2, \dots, y_m\}$.
- The ordered pair (X, Y) take values in the product $\{(x_1, y_1), (x_1, y_2), \dots, (x_n, y_m)\}$.
- If X and Y are discrete, this distribution can be described with a **joint probability mass function**.
- The joint probability mass function (joint pmf) of X and Y is the function $p(x_i, y_j)$ giving the probability of the joint outcome $X = x_i, Y = y_j$.

Joint probability table

$X \setminus Y$	y_1	y_2	...	y_j	...	y_m
x_1	$p(x_1, y_1)$	$p(x_1, y_2)$...	$p(x_1, y_j)$...	$p(x_1, y_m)$
x_2	$p(x_2, y_1)$	$p(x_2, y_2)$...	$p(x_2, y_j)$...	$p(x_2, y_m)$
...
...
x_i	$p(x_i, y_1)$	$p(x_i, y_2)$...	$p(x_i, y_j)$...	$p(x_i, y_m)$
...
x_n	$p(x_n, y_1)$	$p(x_n, y_2)$...	$p(x_n, y_j)$...	$p(x_n, y_m)$

Properties of joint probability mass function

A joint probability mass function must satisfy two properties:

1. $0 \leq p(x_i, y_j) \leq 1$ ✓
2. The total probability is 1. We can express this as

$$\sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) = 1$$

Example

- Two balls are selected at random from a bag containing three green, two blue and four red balls.

- (Handwritten note: green and blue are written above the circles, and X and Y are circled in red)*
- If X and Y are respectively the numbers of green and blue balls included among the two balls drawn from the bag, find the probabilities associated with all possible pairs of value of X and Y .

Solution



- Here the possible pairs are $(0, 0), (0, 1), (1, 0), (1, 1), (0, 2), (2, 0)$.
- To obtain the probability associated with $\underline{(1, 0)}$, we see that we are dealing with the event of getting one of the three green balls, no blue ball and hence, one of the red ball is the number of ways in which we get this event

$${}^3C_1 \times {}^2C_0 \times {}^4C_1 = 12$$

- total number of ways in which two ball are drawn out of nine

$${}^9C_2 = 36$$

- probability of the event associated with $(1, 0)$ is $\frac{12}{36} = \frac{1}{3}$

Similarly,

		X		
		0	1	2
Y	0	1/6	1/3	1/12
	1	2/9	1/6	0
2	1/36	0	0	

Marginal Probability?

$$P(X=0) = \frac{1}{6} + \frac{2}{9} + \frac{1}{36}$$

$$P(X=1) = \frac{1}{3} + \frac{1}{6}$$

$$P(X=2) = \frac{1}{12}$$

$$P(Y=0) = \frac{1}{6} + \frac{1}{3} + \frac{1}{2}$$

$$P(Y=1) = \frac{2}{9} + \frac{1}{6}$$

$$P(Y=2) = \frac{1}{36}$$

Joint PDF

- If X takes values in $[a, b]$ and Y takes values in $[c, d]$ then the pair (X, Y) takes values in the product $[a, b] \times [c, d]$.
- If X and Y are continuous, this distribution can be described with a **joint probability density function**
- The joint probability density function (joint pdf) of X and Y is a function $f(x, y)$ giving the probability density at (x, y) .
- That is, the probability that (X, Y) is in a small rectangle of width dx and height dy around (x, y) is $f(x, y) dx dy$.

Properties of joint probability distribution function

A joint probability density function must satisfy two properties:

1. $0 \leq f(x, y)$.
2. The total probability is 1. We now express this as

$$\int_c^d \int_a^b f(x, y) dx dy = 1$$

Exercise

- A bank operates both a drive-up facility and a walk-up window. On a randomly selected day, let X = the proportion of time that the drive-up facility is in use (at least one customer is being served or waiting to be served) and Y = the proportion of time that the walk-up window is in use. Then the set of possible values for (X, Y) is the rectangle D $\{(x, y): 0 \leq x \leq 1, 0 \leq y \leq 1\}$. Suppose the joint pdf of (X, Y) is given by

$$\underline{f(x, y)} = \begin{cases} \frac{6}{5}(x + y^2) & \underline{0 \leq x \leq 1, 0 \leq y \leq 1} \\ 0 & \text{otherwise} \end{cases}$$

- Verify that this is a legitimate pdf
- The probability that neither facility is busy more than one-quarter of the time

Solution

$$\begin{aligned}
 (1) \quad & \int_0^1 \int_0^1 \frac{6}{5} (x+y^2) dx dy = \int_0^1 \int_0^1 \frac{6}{5} x dx dy + \int_0^1 \int_0^1 \frac{6}{5} y^2 dx dy \\
 &= \int_0^1 \frac{6}{5} x dx [y]_0^1 + \int_0^1 \frac{6}{5} y^2 dy [x]_0^1 = \int_0^1 \frac{6}{5} x dx + \int_0^1 \frac{6}{5} y^2 dy \\
 &= \frac{6}{5} \left[\frac{x^2}{2} \right]_0^1 + \frac{6}{5} \left[\frac{y^3}{3} \right]_0^1 = \frac{6}{10} + \frac{6}{15} = 1 \\
 &\therefore \text{It is legitimate.}
 \end{aligned}$$

$$\begin{aligned}
 (2) \quad & P(0 \leq x \leq \frac{1}{4}, 0 \leq y \leq \frac{1}{4}) \\
 &= \frac{6}{5} \int_0^{\frac{1}{4}} \int_0^{\frac{1}{4}} x dx dy + \frac{6}{5} \int_0^{\frac{1}{4}} \int_0^{\frac{1}{4}} y^2 dx dy = \frac{6}{5} \int_0^{\frac{1}{4}} \left[\frac{x^2}{2} \right]_0^{\frac{1}{4}} dy + \frac{6}{5} \int_0^{\frac{1}{4}} \left[\frac{y^3}{3} \right]_0^{\frac{1}{4}} dx \\
 &= \frac{6}{5} \times \frac{1}{32} \int_0^{\frac{1}{4}} dy + \frac{6}{5} \times \frac{1}{\frac{64}{3}} \int_0^{\frac{1}{4}} dz = \frac{6}{5} \times \frac{1}{32} \times \frac{1}{4} + \frac{6}{5} \times \frac{1}{\frac{64}{3}} \times \frac{1}{4} \\
 &= \frac{7}{640}
 \end{aligned}$$

Exercise

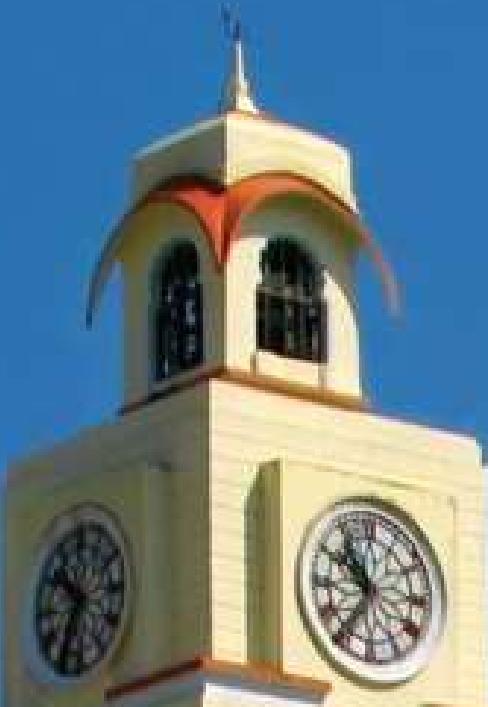
Suppose the random variables X and Y have the joint density function defined by

$$f(x, y) = \begin{cases} c(2x + y) & 2 < x < 6, \quad 0 < y < 5 \\ 0 & \text{otherwise} \end{cases}$$

Find value of c.

Solution

$$\begin{aligned}
 I &= \int_2^6 \int_0^5 C(2x+y) dy dx = \int_2^6 C \left[2xy + \frac{y^2}{2} \right]_0^5 dx \\
 I &= \int_2^6 C \left(10x + \frac{25}{2} \right) dx = C \left(\frac{10 \times \frac{x^2}{2} + \frac{25}{2}x}{2} \right)_2^6 \\
 I &= C \left[\frac{10 \times 6^2}{2} + \frac{25}{2} \times 6 - \frac{10 \times 2^2}{2} - \frac{25 \times 2}{2} \right] \\
 I &= C [180 + 75 - 20 - 25] \\
 \therefore C &= 1/210
 \end{aligned}$$



BITS Pilani
Pilani Campus

Multivariate Analytics

Akanksha Bharadwaj
Asst. professor, CS/IS Department



SS ZG536, ADV STAT TECHNIQUES FOR ANALYTICS

Contact Session 12

Introduction

- In science and in real life, we are often interested in two (or more) random variables at the same time.
- For example, we might measure the IQ and birthweight of children, or the level of air pollution and rate of respiratory illness in cities.



BITS Pilani

Pilani Campus



Joint Distribution

Joint PMF

- Suppose X and Y are two discrete random variables and that X takes values $\{x_1, x_2, \dots, x_n\}$ and Y takes values $\{y_1, y_2, \dots, y_m\}$.
- The ordered pair (X, Y) take values in the product $\{(x_1, y_1), (x_1, y_2), \dots, (x_n, y_m)\}$.
- If X and Y are discrete, this distribution can be described with a **joint probability mass function**.
- The joint probability mass function (joint pmf) of X and Y is the function $p(x_i, y_j)$ giving the probability of the joint outcome $X = x_i, Y = y_j$.

Joint probability table

$X \setminus Y$	y_1	y_2	...	y_j	...	y_m
x_1	$p(x_1, y_1)$	$p(x_1, y_2)$...	$p(x_1, y_j)$...	$p(x_1, y_m)$
x_2	$p(x_2, y_1)$	$p(x_2, y_2)$...	$p(x_2, y_j)$...	$p(x_2, y_m)$
...
...
x_i	$p(x_i, y_1)$	$p(x_i, y_2)$...	$p(x_i, y_j)$...	$p(x_i, y_m)$
...
x_n	$p(x_n, y_1)$	$p(x_n, y_2)$...	$p(x_n, y_j)$...	$p(x_n, y_m)$

Properties of joint probability mass function

A joint probability mass function must satisfy two properties:

1. $0 \leq p(x_i, y_j) \leq 1$ ✓
2. The total probability is 1. We can express this as

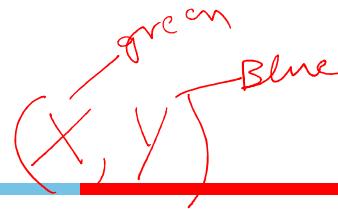
$$\sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) = 1$$

Example

- Two balls are selected at random from a bag containing three green, two blue and four red balls.

- (Handwritten note: green and blue are written above the circles, and X and Y are circled in red)*
- If X and Y are respectively the numbers of green and blue balls included among the two balls drawn from the bag, find the probabilities associated with all possible pairs of value of X and Y .

Solution



- Here the possible pairs are $(0, 0), (0, 1), (1, 0), (1, 1), (0, 2), (2, 0)$.
- To obtain the probability associated with $\underline{(1, 0)}$, we see that we are dealing with the event of getting one of the three green balls, no blue ball and hence, one of the red ball is the number of ways in which we get this event

$${}^3C_1 \times {}^2C_0 \times {}^4C_1 = 12$$

- total number of ways in which two ball are drawn out of nine

$${}^9C_2 = 36$$

- probability of the event associated with $(1, 0)$ is $\frac{12}{36} = \frac{1}{3}$

Similarly,

		X		
		0	1	2
Y	0	1/6	1/3	1/12
	1	2/9	1/6	0
2	1/36	0	0	

Marginal Probability?

$$P(X=0) = \frac{1}{6} + \frac{2}{9} + \frac{1}{36}$$

$$P(X=1) = \frac{1}{3} + \frac{1}{6}$$

$$P(X=2) = \frac{1}{12}$$

$$P(Y=0) = \frac{1}{6} + \frac{1}{3} + \frac{1}{2}$$

$$P(Y=1) = \frac{2}{9} + \frac{1}{6}$$

$$P(Y=2) = \frac{1}{36}$$

Joint PDF

- If X takes values in $[a, b]$ and Y takes values in $[c, d]$ then the pair (X, Y) takes values in the product $[a, b] \times [c, d]$.
- If X and Y are continuous, this distribution can be described with a **joint probability density function**
- The joint probability density function (joint pdf) of X and Y is a function $f(x, y)$ giving the probability density at (x, y) .
- That is, the probability that (X, Y) is in a small rectangle of width dx and height dy around (x, y) is $f(x, y) dx dy$.

Properties of joint probability distribution function

A joint probability density function must satisfy two properties:

1. $0 \leq f(x, y)$.
2. The total probability is 1. We now express this as

$$\int_c^d \int_a^b f(x, y) dx dy = 1$$

Exercise

- A bank operates both a drive-up facility and a walk-up window. On a randomly selected day, let X = the proportion of time that the drive-up facility is in use (at least one customer is being served or waiting to be served) and Y = the proportion of time that the walk-up window is in use. Then the set of possible values for (X, Y) is the rectangle D $\{(x, y): 0 \leq x \leq 1, 0 \leq y \leq 1\}$. Suppose the joint pdf of (X, Y) is given by

$$\underline{f(x, y)} = \begin{cases} \frac{6}{5}(x + y^2) & \underline{0 \leq x \leq 1, 0 \leq y \leq 1} \\ 0 & \text{otherwise} \end{cases}$$

- Verify that this is a legitimate pdf
- The probability that neither facility is busy more than one-quarter of the time

Solution

$$\begin{aligned}
 (1) \quad & \int_0^1 \int_0^1 \frac{6}{5} (x+y^2) dx dy = \int_0^1 \int_0^1 \frac{6}{5} x dx dy + \int_0^1 \int_0^1 \frac{6}{5} y^2 dx dy \\
 &= \int_0^1 \frac{6}{5} x dx [y]_0^1 + \int_0^1 \frac{6}{5} y^2 dy [x]_0^1 = \int_0^1 \frac{6}{5} x dx + \int_0^1 \frac{6}{5} y^2 dy \\
 &= \frac{6}{5} \left[\frac{x^2}{2} \right]_0^1 + \frac{6}{5} \left[\frac{y^3}{3} \right]_0^1 = \frac{6}{10} + \frac{6}{15} = 1 \\
 &\therefore \text{It is legitimate.}
 \end{aligned}$$

$$\begin{aligned}
 (2) \quad & P(0 \leq x \leq \frac{1}{4}, 0 \leq y \leq \frac{1}{4}) \\
 &= \frac{6}{5} \int_0^{\frac{1}{4}} \int_0^{\frac{1}{4}} x dx dy + \frac{6}{5} \int_0^{\frac{1}{4}} \int_0^{\frac{1}{4}} y^2 dx dy = \frac{6}{5} \int_0^{\frac{1}{4}} \left[\frac{x^2}{2} \right]_0^{\frac{1}{4}} dy + \frac{6}{5} \int_0^{\frac{1}{4}} \left[\frac{y^3}{3} \right]_0^{\frac{1}{4}} dx \\
 &= \frac{6}{5} \times \frac{1}{32} \int_0^{\frac{1}{4}} dy + \frac{6}{5} \times \frac{1}{\frac{64}{3}} \int_0^{\frac{1}{4}} dz = \frac{6}{5} \times \frac{1}{32} \times \frac{1}{4} + \frac{6}{5} \times \frac{1}{\frac{64}{3}} \times \frac{1}{4} \\
 &= \frac{7}{640}
 \end{aligned}$$

Exercise

Suppose the random variables X and Y have the joint density function defined by

$$f(x, y) = \begin{cases} c(2x + y) & 2 < x < 6, \quad 0 < y < 5 \\ 0 & \text{otherwise} \end{cases}$$

Find value of c.

Solution

$$\begin{aligned}
 I &= \int_2^6 \int_0^5 C(2x+y) dy dx = \int_2^6 C \left[2xy + \frac{y^2}{2} \right]_0^5 dx \\
 I &= \int_2^6 C \left(10x + \frac{25}{2} \right) dx = C \left(\frac{10 \times \frac{x^2}{2} + \frac{25}{2}x}{2} \right)_2^6 \\
 I &= C \left[\frac{10 \times 6 \times 6}{2} + \frac{25}{2} \times 6 - \frac{10 \times 2 \times 2}{2} - \frac{25 \times 2}{2} \right] \\
 I &= C [180 + 75 - 20 - 25] \\
 \therefore C &= 1/210
 \end{aligned}$$

Joint cumulative distribution function

- Suppose X and Y are jointly-distributed random variables.
- We will use the notation ' $X \leq x, Y \leq y$ ' to mean the event ' $X \leq x$ and $Y \leq y$ '.
- The joint cumulative distribution function (joint cdf) is defined as $F(x, y) = P(X \leq x, Y \leq y)$

Joint cumulative distribution function

- If X and Y are continuous random variables with joint density $f(x, y)$ over the range $[a, b] \times [c, d]$ then the joint cdf is given by the double integral

$$F(x, y) = \int_c^y \int_a^x f(u, v) du dv.$$

- If X and Y are discrete random variables with joint pmf $p(x_i, y_j)$ then the joint cdf is give by the double sum

$$\underline{F(x, y)} = \sum_{x_i \leq x} \sum_{y_j \leq y} p(x_i, y_j).$$

Properties of the joint cdf

The joint cdf $F(x, y)$ of X and Y must satisfy several properties:

1. $F(x, y)$ is non-decreasing: i.e. if x or y increase then $F(x, y)$ must stay constant or increase.

2. $F(x, y) = 0$ at the lower-left of the joint range.

If the lower left is $(-\infty, -\infty)$ then this means $\lim_{(x,y) \rightarrow (-\infty, -\infty)} F(x, y) = 0$.

3. $F(x, y) = 1$ at the upper-right of the joint range.

If the upper-right is (∞, ∞) then this means $\lim_{(x,y) \rightarrow (\infty, \infty)} F(x, y) = 1$.

Exercise

A nut company markets cans of deluxe mixed nuts containing almonds, cashews, and peanuts. Suppose the net weight of each can is exactly 1 lb, but the weight contribution of each type of nut is random. Because the three weights sum to 1, a joint probability model for any two gives all necessary information about the weight of the third type. Let $X = \text{the weight of almonds in a selected can}$ and $Y = \text{the weight of cashews}$. Then the region of positive density is $D = \{(x, y): 0 \leq x \leq 1, 0 \leq y \leq 1, x + y \leq 1\}$, the shaded region pictured in Figure 5.2.

Now let the joint pdf for (X, Y) be

$$f(x, y) = \begin{cases} 24xy & 0 \leq x \leq 1, 0 \leq y \leq 1, x + y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

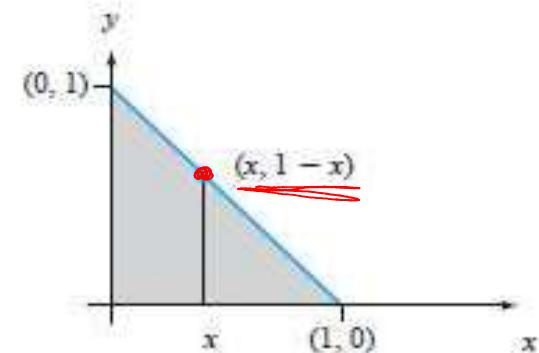


Figure 5.2 Region of positive density for Example

Solution

$P(X+Y=1)$

$$\begin{aligned} \int_0^1 \left[\int_0^{1-x} 24xy \, dy \right] dx &= \int_0^1 24x \left[\frac{y^2}{2} \right]_0^{1-x} dx \\ &= \int_0^1 12x(1-x)^2 dx = 1. \end{aligned}$$

To compute the probability that the two types of nuts together make up at most 50% of the can, let $A = \{(x, y): 0 \leq x \leq 1, 0 \leq y \leq 1, \text{ and } x + y \leq .5\}$, as shown in Figure 5.3. Then

$$P((X, Y) \in A) = \int_A \int f(x, y) \, dx \, dy = \int_0^.5 \int_0^{.5-x} 24xy \, dy \, dx = .0625$$

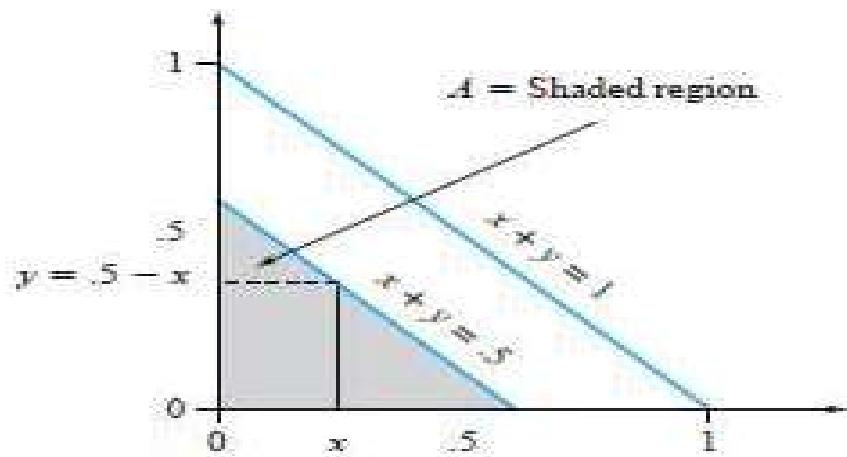


Figure 5.3 Computing $P((X, Y) \in A)$ for Example 5.5

Independence

- Events A and B are independent if $P(A \cap B) = P(A)P(B).$

- Jointly-distributed random variables X and Y are independent if their joint cdf is the product of the marginal cdf's

$$\underbrace{F(X, Y)}_{\text{joint CDF}} = \underbrace{F_X(x)}_{\text{marginal CDF}} \underbrace{F_Y(y)}_{\text{marginal CDF}}.$$

- For discrete variables this is equivalent to the joint pmf being the product of the marginal pmf's

$$p(x_i, y_j) = \underbrace{p_X(x_i)}_{\text{marginal pmf}} \underbrace{p_Y(y_j)}_{\text{marginal pmf}}.$$

- For continuous variables this is equivalent to the joint pdf being the product of the marginal pdf's

$$f(x, y) = f_X(x)f_Y(y).$$

Exercise

- Consider two random variables X and Y with joint PMF given in Table
- Find $P(X \leq 2, Y \leq 4)$.
- Find the marginal PMFs of X and Y.
- Find $P(Y=2|X=1)$.
- Are X and Y independent?

	$Y = 2$	$Y = 4$	$Y = 5$
$X = 1$	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{1}{24}$
$X = 2$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{8}$
$X = 3$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{12}$

$$P(Y|X) = \frac{P(Y \cap X)}{P(X)}$$

Solution

To find $P(X \leq 2, Y \leq 4)$, we can write

$$\begin{aligned} P(X \leq 2, Y \leq 4) &= P_{XY}(1, 2) + P_{XY}(1, 4) + P_{XY}(2, 2) + P_{XY}(2, 4) \\ &= \frac{1}{12} + \frac{1}{24} + \frac{1}{6} + \frac{1}{12} = \frac{3}{8}. \end{aligned}$$

$$\begin{aligned} P(Y = 2|X = 1) &= \frac{P(X = 1, Y = 2)}{P(X = 1)} \\ &= \frac{P_{XY}(1, 2)}{P_X(1)} \\ &= \frac{\frac{1}{12}}{\frac{1}{6}} = \frac{1}{2}. \end{aligned}$$

$$P(X = 2, Y = 2) = \frac{1}{6} \neq P(X = 2)P(Y = 2) = \frac{3}{16}.$$

Thus, we conclude that X and Y are not independent.

$$P_X(x) = \begin{cases} \frac{1}{6} & x = 1 \\ \frac{3}{8} & x = 2 \\ \frac{11}{24} & x = 3 \\ 0 & \text{otherwise} \end{cases}$$

$$P_Y(y) = \begin{cases} \frac{1}{2} & y = 2 \\ \frac{1}{4} & y = 4 \\ \frac{1}{4} & y = 5 \\ 0 & \text{otherwise} \end{cases}$$

Expected Value

Let X and Y be jointly distributed rv's with pmf $p(x, y)$ or pdf $f(x, y)$ according to whether the variables are discrete or continuous. Then the expected value of a function $h(X, Y)$, denoted by $E[h(X, Y)]$ or $\mu_{h(X, Y)}$, is given by

$$E[h(X, Y)] = \begin{cases} \sum_x \sum_y h(x, y) \cdot p(x, y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) \cdot f(x, y) \, dx \, dy & \text{if } X \text{ and } Y \text{ are continuous} \end{cases}$$

Example

- Five friends have purchased tickets to a certain concert. If the tickets are for seats 1–5 in a particular row and the tickets are randomly distributed among the five, what is the expected number of seats separating any particular two of the five? Let X and Y denote the seat numbers of the first and second individuals, respectively. Possible (X, Y) pairs are $\{(1, 2), (1, 3), \dots, (5, 4)\}$, and the joint pmf of (X, Y) is

$$p(x, y) = \begin{cases} \frac{1}{20} & x = 1, \dots, 5; y = 1, \dots, 5; x \neq y \\ 0 & \text{otherwise} \end{cases}$$

The number of seats separating the two individuals is $h(X, Y) = |X - Y| - 1$. The accompanying table gives $h(x, y)$ for each possible (x, y) pair.

Solution

$$|X-Y| - 1$$

$x \neq y$

		x				
		1	2	3	4	5
y	1	—	0	1	2	3
	2	0	—	0	1	2
	3	1	0	—	0	1
	4	2	1	0	—	0
	5	3	2	1	0	—

$$\sum_{(x,y)} h(x,y) \cdot P(x,y)$$

$$\frac{1}{20} (1 + 2 + 3 + 1 + 2 + 1 + 1 + 2 + 1 + 3 + 2 + 1) = 1$$

Exercise (HW)

- Suppose that 2 batteries are randomly chosen without replacement from the following group of 12 batteries:
 - 3 new
 - 4 used (working)
 - 5 defective
- Let X denote the number of new batteries chosen. Let Y denote the number of used batteries chosen.
- Find $f_{XY}(x, y)$

Solution

- Though X can take on values 0, 1, and 2, and Y can take on values 0, 1, and 2, when we consider them jointly, $X + Y \leq 2$. So, not all combinations of (X, Y) are possible.

There are 6 possible cases...

CASE: no new, no used (so all defective)

$$f_{XY}(0, 0) = \frac{\binom{5}{2}}{\binom{12}{2}} = 10/66$$

CASE: no new, 1 used

$$f_{XY}(0, 1) = \frac{\binom{4}{1} \binom{5}{1}}{\binom{12}{2}} = 20/66$$

CASE: no new, 2 used

$$f_{XY}(0, 2) = \frac{\binom{4}{2}}{\binom{12}{2}} = 6/66$$

CASE: 1 new, no used

$$f_{XY}(1, 0) = \frac{\binom{3}{1} \binom{5}{1}}{\binom{12}{2}} = 15/66$$

CASE: 2 new, no used

$$f_{XY}(2, 0) = \frac{\binom{3}{2}}{\binom{12}{2}} = 3/66$$

CASE: 1 new, 1 used

$$f_{XY}(1, 1) = \frac{\binom{3}{1} \binom{4}{1}}{\binom{12}{2}} = 12/66$$

x= number of *new* chosen

	0	1	2
0	10/66	15/66	3/66
1	20/66	12/66	
2	6/66		

There are 6 possible (X, Y) pairs.

And, $\sum_x \sum_y f_{XY}(x, y) = 1$.

Exercise

Compute $E(X)$ and $E(Y)$.

Compute $E(XY)$.

$$\begin{aligned}
 E(X) &= 0 \times 0.2 \\
 &\quad + \\
 &0 \times 0.1 \\
 &\quad + \\
 &1 \times 0.0 \\
 &\quad + \\
 &1 \times 0.2 \\
 &\quad + \\
 &2 \times (0.5) \\
 &= 1.2
 \end{aligned}$$

x	y	$P(X = x, Y = y)$
0	1	0.2
0	2	0.1
1	1	0.0
1	2	0.2
2	1	0.3
2	2	0.2

Solution

$$E(Y) = 1 \times (0.2 + 0.0 + 0.3) + \\ 2 \times (0.1 + 0.2 + 0.2) = 1.5$$

$$E(XY) = 0 \times 1 \times 0.2 + 0 \times 2 \times 0.1 + 1 \times 1 \times 0.0 \\ + 1 \times 2 \times 0.2 + 2 \times 1 \times 0.3 + 2 \times 2 \times 0.2 \\ = 1.8$$

Covariance

- When two random variables X and Y are not independent, it is frequently of interest to assess how strongly they are related to one another.

The covariance between two rv's X and Y is

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

$$= \begin{cases} \sum_x \sum_y (x - \mu_X)(y - \mu_Y) p(x, y) & X, Y \text{ discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y) dx dy & X, Y \text{ continuous} \end{cases}$$

Example

		<i>y</i>	0	100	200
		0	.20	.10	.20
		100	.05	.15	.30
<i>x</i>		200			

<i>x</i>	100	250
<i>p_x(x)</i>	.5	.5

<i>y</i>	0	100	200
<i>p_y(y)</i>	.25	.25	.5

from which $\mu_x = \sum x p_x(x) = 175$ and $\mu_y = 125$. Therefore,

$$\begin{aligned}
 \mu_y &= 0 \times 0.25 + 100 \times 0.25 \\
 &\quad + 200 \times 0.5 \\
 &= 125
 \end{aligned}$$

Covariance???

$$\begin{aligned} \text{Cov}(X, Y) &= \sum_{(x,y)} (x - 175)(y - 125) p(x, y) \\ &= (100 - 175)(0 - 125)(0.2) + \dots + \dots \\ &\quad + (250 - 175)(200 - 125)(0.3) \end{aligned}$$

The following shortcut formula for $\text{Cov}(X, Y)$ simplifies the computations.

$$\text{Cov}(X, Y) = E(XY) - \mu_X \cdot \mu_Y$$

Covariance

$\text{Cov}(X, Y) = E(XY) - E(X) \cdot E(Y)$ where

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$

$$E(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy$$

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x, y) dx dy$$

$$V(X) = E(X^2) - [E(X)]^2$$

$$V(Y) = E(Y^2) - [E(Y)]^2$$

Correlation

The correlation coefficient of X and Y , denoted by $\text{Corr}(X, Y)$, $\rho_{X,Y}$, or just ρ , is defined by

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

- If X and Y are independent, then correlation is 0, but if correlation is 0 that does not imply independence.

Exercise

A large insurance agency services a number of customers who have purchased both a homeowner's policy and an automobile policy from the agency. For each type of policy, a deductible amount must be specified. For an automobile policy, the choices are \$100 and \$250, whereas for a homeowner's policy, the choices are 0, \$100, and \$200. Suppose an individual with both types of policy is selected at random from the agency's files. Let X the deductible amount on the auto policy and Y the deductible amount on the homeowner's policy

Suppose the joint pmf is given by

		y		
		0	100	200
x	100	.20	.10	.20
	250	.05	.15	.30

Find

- (i) Marginal probabilities of X and Y.
- (ii) $P(Y \geq 100)$

$$\begin{aligned} P_x(100) &= 0.2 + 0.1 + 0.2 \\ &= 0.5 \end{aligned}$$

$$\begin{aligned} P_x(250) &= 0.05 + 0.15 + 0.30 \\ &= 0.5 \end{aligned}$$

Solution

$$P_X(x) = \begin{cases} 0.5 & x=100, 250 \\ 0 & \text{otherwise} \end{cases}$$

$$P_Y(y) = \begin{cases} 0.25 & y=0, 100 \\ 0.5 & y=200 \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned}
 P(Y \geq 100) &= P(100, 100) + P(250, 100) + P(100, 200) \\
 &\quad + P(250, 200) \\
 &= 0.75
 \end{aligned}$$

Exercise (HW)

- When a certain method is used to collect a fixed volume of rock samples in a region, there are four resulting rock types. Let X_1 , X_2 , and X_3 denote the proportion by volume of rock types 1, 2, and 3 in a randomly selected sample (the proportion of rock type 4 is $1 - X_1 - X_2 - X_3$, so a variable X_4 would be redundant). If the joint pdf of X_1 , X_2 , X_3 is

$$f(x_1, x_2, x_3) = \begin{cases} kx_1x_2(1 - x_3) & 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1, 0 \leq x_3 \leq 1, x_1 + x_2 + x_3 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

then k is determined by

- The probability that rocks of types 1 and 2 together account for at most 50% of the sample is

Solution

$$1 = \int_{-a}^a \int_{-a}^a \int_{-a}^a f(x_1, x_2, x_3) dx_3 dx_2 dx_1$$

$$= \int_0^1 \left\{ \int_0^{1-x_1} \left[\int_0^{1-x_1-x_2} kx_1 x_2 (1-x_3) dx_3 \right] dx_2 \right\} dx_1$$

This iterated integral has value $k/144$, so $k = 144$.

$$P(X_1 + X_2 \leq .5) = \iiint \begin{cases} 0 \leq x_j \leq 1 \text{ for } j=1, 2, 3 \\ x_1 + x_2 + x_3 \leq 1, x_1 + x_2 \leq .5 \end{cases} f(x_1, x_2, x_3) dx_3 dx_2 dx_1$$

$$= \int_0^.5 \left\{ \int_0^{.5-x_1} \left[\int_0^{1-x_1-x_2} 144x_1 x_2 (1-x_3) dx_3 \right] dx_2 \right\} dx_1$$

$$= .6066$$





Multivariate normal distribution

Univariate Normal Distribution

- The normal distribution , also known as the Gaussian distribution, is so called because its based on the Gaussian function .
- This distribution is defined by two parameters: the mean μ , which is the expected value of the distribution, and the standard deviation σ , which corresponds to the expected deviation from the mean.

$$p(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

- Given this mean and variance we can calculate the probability density function (pdf) of the normal distribution with the normalised Gaussian function. For a value x the density is:

Multivariate normal distribution

- The multivariate normal distribution is a multidimensional generalisation of the one-dimensional normal distribution .
- “a random vector is said to be r-variate normally distributed if every linear combination of its r components has a univariate normal distribution”.
- A bivariate normal distribution is made up of two independent random variables. The two variables in a bivariate normal are both are normally distributed, and they have a normal distribution when both are added together.
- **Visually, the bivariate normal distribution is a three-dimensional bell curve.**

PDF of the Bivariate Normal Distribution

- The bivariate normal distribution can be defined as the probability density function (PDF) of two variables X and Y that are linear functions of the same independent normal random variables:

$$P(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{z}{2(1-\rho^2)}\right],$$

- Where σ is standard deviation
- ρ correlation of x_1 and x_2 .

$$\rho \equiv \text{cor}(x_1, x_2) = \frac{\langle x_1 x_2 \rangle - \langle x_1 \rangle \langle x_2 \rangle}{\sigma_1 \sigma_2}$$

$$z \equiv \frac{(x_1 - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1 \sigma_2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2},$$

Probability for Bivariate

Probabilities still relate to the area under the pdf:

$$P([a_x \leq X \leq b_x] \text{ and } [a_y \leq Y \leq b_y]) = \int_{a_x}^{b_x} \int_{a_y}^{b_y} \underline{\underline{f(x, y)dydx}} \quad (6)$$

where $\int \int f(x, y)dydx$ denotes the multiple integral of the pdf $f(x, y)$.

Defining $\mathbf{z} = (x, y)$, we can still define the cdf:

$$\begin{aligned} \underline{\underline{F(\mathbf{z})}} &= P(X \leq x \text{ and } Y \leq y) \\ &= \int_{-\infty}^x \int_{-\infty}^y f(u, v)dvdu \end{aligned} \quad (7)$$

Conditional distribution

The **conditional distribution** of a variable Y given $X = x$ is

$$f_{Y|X}(y|X = x) = \frac{f_{XY}(x, y)}{\underline{f_X(x)}}$$

where

- $f_{XY}(x, y)$ is the **joint pdf** of X and Y
- $f_X(x)$ is the **marginal pdf** of X

In the bivariate normal case, we have that

$$Y|X \sim N(\mu_*, \sigma_*^2)$$

where $\underline{\mu_*} = \mu_y + \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x)$ and $\sigma_*^2 = \underline{\sigma_y^2(1 - \rho^2)}$

Statistical Independence

Two variables X and Y are statistically independent if

$$\underline{f_{XY}(x,y)} = \underline{f_X(x)f_Y(y)} \quad (10)$$

where $f_{XY}(x,y)$ is joint pdf, and $f_X(x)$ and $f_Y(y)$ are marginals pdfs.

Note that if X and Y are independent, then

$$f_{Y|X}(y|X=x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{\cancel{f_X(x)} f_Y(y)}{\cancel{f_X(x)}} = \underline{\underline{f_Y(y)}} \quad (11)$$

so conditioning on $X = x$ does not change the distribution of Y .

Example

Let X be the height of the father, Y the height of the son, in a sample of father-son pairs. Assume X and Y bivariate normal, as found by Karl Pearson around 1900. Assume $E(X) = 68$ (inches), $E(Y) = 69$, $\sigma_X = \sigma_Y = 2$, $\rho = .5$. (We expect ρ to be positive because on the average, the taller the father, the taller the son.)

mean?

variance?

Solution

Given $X = 80$ (6 feet 8 inches), Y is normal with mean

$$\mu_y + \frac{P\sigma_y}{\sigma_x} (x - \mu_x) = 69 + \frac{0.5 \times 2}{2} (80 - 68)$$

$$= 75$$

which is 6 feet 3 inches. The variance of Y given $X = 80$ is

$$\sigma_y^2 (1 - P^2) = (2)^2 (1 - (0.5)^2)$$

$$= 4 \left(1 - \frac{1}{4}\right) = 3$$

The standard multivariate normal distribution

- The adjective "standard" is used to indicate that the mean of the distribution is equal to zero
- Its covariance matrix is equal to the identity matrix.

Standard MV-N random vectors are characterized as follows.

Definition Let x be a $K \times 1$ continuous random vector. Let its support be the set of K -dimensional real vectors:

$$R_X = \mathbb{R}^K$$

We say that x has a standard multivariate normal distribution if its joint probability density function is

$$f_X(x) = (2\pi)^{-K/2} \exp\left(-\frac{1}{2}x^\top x\right)$$

transpose

Expected value

- The expected value of a standard MV-N random vector X is $E[X]=0$

Proof

- All the components of X are standard normal random variables and a standard normal random variable has mean 0.

Covariance matrix

- Since the components of X are all standard normal random variables, their variances are all equal to 1, i.e.,

$$\text{Var}[X_1] = \dots = \text{Var}[X_K] = 1$$

Furthermore, since the components of X are mutually independent and independence implies zero-covariance, all the covariances are equal to 0, i.e.,

$$\text{Cov}[X_i, X_j] = 0 \quad \forall i, j$$

Therefore,

$$\begin{aligned} \text{Var}[X] &= \begin{bmatrix} \text{Var}[X_1] & \text{Cov}[X_1, X_2] & \dots & \text{Cov}[X_1, X_K] \\ \text{Cov}[X_1, X_2] & \text{Var}[X_2] & \dots & \text{Cov}[X_2, X_K] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[X_1, X_K] & \text{Cov}[X_2, X_K] & \dots & \text{Var}[X_K] \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = I \end{aligned}$$

The multivariate normal distribution in general

Definition Let x be a $K \times 1$ continuous random vector. Let its support be the set of K -dimensional real vectors:

$$R_X = \mathbb{R}^K$$

Let μ be a $K \times 1$ vector and V a $K \times K$ symmetric and positive definite matrix. We say that x has a **multivariate normal distribution** with mean μ and covariance V if its joint probability density function is

$$f_X(x) = (2\pi)^{-K/2} |\det(V)|^{-1/2} \exp\left(-\frac{1}{2}(x - \mu)^\top V^{-1}(x - \mu)\right)$$

We indicate that x has a multivariate normal distribution with mean μ and covariance V by

$$X \sim N(\mu, V)$$

The K random variables X_1, \dots, X_K constituting the vector x are said to be **jointly normal**.

Multivariate Normal probabilities



Probabilities still relate to the area under the pdf:

$$P(a_j \leq X_j \leq b_j \forall j) = \int_{a_1}^{b_1} \cdots \int_{a_p}^{b_p} f(\mathbf{x}) dx_p \cdots dx_1 \quad (13)$$

where $\int \cdots \int f(\mathbf{x}) dx_p \cdots dx_1$ denotes the multiple integral $f(\mathbf{x})$.

We can still define the cdf of $\mathbf{x} = (x_1, \dots, x_p)'$

$$\begin{aligned} F(\mathbf{x}) &= P(X_j \leq x_j \forall j) \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_p} f(\mathbf{u}) du_p \cdots du_1 \end{aligned} \quad (14)$$

Multivariate conditional distribution

Given variables $\mathbf{x} = (x_1, \dots, x_p)'$ and $\mathbf{y} = (y_1, \dots, y_q)'$, we have

$$f_{Y|X}(\mathbf{y}|X = \mathbf{x}) = \frac{f_{XY}(\mathbf{x}, \mathbf{y})}{f_X(\mathbf{x})} \quad (15)$$

where

- $f_{Y|X}(\mathbf{y}|X = \mathbf{x})$ is the conditional distribution of \mathbf{y} given \mathbf{x}
- $f_{XY}(\mathbf{x}, \mathbf{y})$ is the joint pdf of \mathbf{x} and \mathbf{y}
- $f_X(\mathbf{x})$ is the marginal pdf of \mathbf{x}

Conditional Normal Multivariate

Suppose that $\mathbf{z} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

- $\mathbf{z} = (\mathbf{x}', \mathbf{y}')' = (x_1, \dots, x_p, y_1, \dots, y_q)'$

- $\boldsymbol{\mu} = (\boldsymbol{\mu}_x', \boldsymbol{\mu}_y')' = (\mu_{1x}, \dots, \mu_{px}, \mu_{1y}, \dots, \mu_{qy})'$

Note: $\boldsymbol{\mu}_x$ is mean vector of \mathbf{x} , and $\boldsymbol{\mu}_y$ is mean vector of \mathbf{y}

- $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}'_{xy} & \boldsymbol{\Sigma}_{yy} \end{pmatrix}$ where $(\boldsymbol{\Sigma}_{xx})_{p \times p}$, $(\boldsymbol{\Sigma}_{yy})_{q \times q}$, and $(\boldsymbol{\Sigma}_{xy})_{p \times q}$,

Note: $\boldsymbol{\Sigma}_{xx}$ is covariance matrix of \mathbf{x} , $\boldsymbol{\Sigma}_{yy}$ is covariance matrix of \mathbf{y} , and $\boldsymbol{\Sigma}_{xy}$ is covariance matrix of \mathbf{x} and \mathbf{y}

In the multivariate normal case, we have that

$$\mathbf{y} | \mathbf{x} \sim \mathbf{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \quad (16)$$

where $\boldsymbol{\mu}_* = \boldsymbol{\mu}_y + \boldsymbol{\Sigma}'_{xy} \boldsymbol{\Sigma}_{xx}^{-1} (\mathbf{x} - \boldsymbol{\mu}_x)$ and $\boldsymbol{\Sigma}_* = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}'_{xy} \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy}$

Statistical Independence

Using Equation (16), we have that

$$\mathbf{y}|\mathbf{x} \sim N(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \equiv N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_{yy}) \quad (17)$$

if and only if $\boldsymbol{\Sigma}_{xy} = \mathbf{0}_{p \times q}$ (a matrix of zeros).

Note that $\boldsymbol{\Sigma}_{xy} = \mathbf{0}_{p \times q}$ implies that the p elements of \mathbf{x} are uncorrelated with the q elements of \mathbf{y} .

- For multivariate normal variables: uncorrelated \rightarrow independent
- For non-normal variables: uncorrelated $\not\rightarrow$ independent

Exercise

A statistics class takes two exams X (Exam 1) and Y (Exam 2) where the scores follow a bivariate normal distribution with parameters:

- $\mu_x = 70$ and $\mu_y = 60$ are the marginal means
- $\sigma_x = 10$ and $\sigma_y = 15$ are the marginal standard deviations
- $\rho = 0.6$ is the correlation coefficient

Suppose we select a student at random. What is the probability that...

- the student scores over 75 on Exam 2?
- the student scores over 75 on Exam 2, given that the student scored $X = 80$ on Exam 1?
- the sum of his/her Exam 1 and Exam 2 scores is over 150?
- the student did better on Exam 1 than Exam 2?
- $P(5X - 4Y > 150)$?

Solution: a

Solution

$$P(Y > 75) = P\left(Z > \frac{75 - 60}{15}\right)$$

$$\begin{aligned} & P(Z > 1) \\ &= 1 - P(Z < 1) = 1 - \phi(1) \\ &= 1 - 0.8413 \end{aligned}$$

where $\Phi(x) = \int_{-\infty}^x f(z)dz$ with $f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ denoting the standard normal pdf

Solution: b

Note that $(Y|X = 80) \sim N(\mu_*, \sigma_*^2)$ where

$$\begin{aligned}\mu_* &= \mu_y + \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x) \\ &= 60 + 0.6 \times \frac{15}{10} (80 - 70) = 69 \\ \sigma_*^2 &= \sigma_y^2 (1 - \rho^2) = \frac{15^2}{10} (1 - 0.6^2) = 144\end{aligned}$$

If a student scored $X = 80$ on Exam 1, the probability that the student scores over 75 on Exam 2 is

$$\begin{aligned}P(Y > 75 | X = 80) &= P\left(Z > \frac{75 - 69}{12}\right) \\ &= P(Z > 0.5) \\ &= 1 - P(Z < 0.5) \\ &= 1 - 0.6914\end{aligned}$$

Solution: c

Note that $(X + Y) \sim N(\mu_*, \sigma_*^2)$ where

✓ $\mu_* = \mu_X + \mu_Y = 70 + 60 = 130$

✓ $\sigma_*^2 = \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y = 10^2 + 15^2 + 2(0.6)(10)(15) = 505$

The probability that the sum of Exam 1 and Exam 2 is above 150 is

$$\begin{aligned} P(X + Y > 150) &= P\left(Z > \frac{150 - 130}{\sqrt{505}}\right) \\ &= P(Z > 0.8899883) \\ &= 1 - \Phi(0.8899883) \\ &= 1 - 0.8132639 \\ &= 0.1867361 \end{aligned}$$

Solution: d

Note that $(X - Y) \sim N(\mu_*, \sigma_*^2)$ where

- ✓ $\mu_* = \mu_X - \mu_Y = 70 - 60 = 10$
- ✓ $\sigma_*^2 = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y = 10^2 + 15^2 - 2(0.6)(10)(15) = 145$

The probability that the student did better on Exam 1 than Exam 2 is

$$\begin{aligned} P(X > Y) &= P(X - Y > 0) \\ &= P\left(Z > \frac{0 - 10}{\sqrt{145}}\right) \\ &= P(Z > -0.8304548) \\ &= 1 - \Phi(-0.8304548) \\ &= 1 - 0.2031408 \\ &= 0.7968592 \end{aligned}$$



Solution: e

Note that $(5X - 4Y) \sim N(\mu_*, \sigma_*^2)$ where

$$\mu_* = 5\mu_X - 4\mu_Y = 5(70) - 4(60) = 110$$

$$\sigma_*^2 = 5^2\sigma_X^2 + (-4)^2\sigma_Y^2 + 2(5)(-4)\rho\sigma_X\sigma_Y =$$

$$25(10^2) + 16(15^2) - 2(20)(0.6)(10)(15) = 2500$$

Thus, the needed probability can be obtained using

$$\begin{aligned} P(5X - 4Y > 150) &= P\left(Z > \frac{150 - 110}{\sqrt{2500}}\right) \\ &= P(Z > 0.8) \\ &= 1 - \Phi(0.8) \\ &= 1 - 0.7881446 \\ &= 0.2118554 \end{aligned}$$



BITS Pilani
Pilani Campus



Principal Component Analysis



Introduction

- Increasing the number of features does not always improve accuracy.
- Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.
- The idea of PCA is to reduce the number of variables of a data set, while preserving as much information as possible.



When should I use PCA?

- Do you want to reduce the number of variables, but aren't able to identify variables to completely remove from consideration?
- Do you want to ensure your variables are independent of one another?
- Are you comfortable making your independent variables less interpretable?

If you answered “yes” to all three questions, then PCA is a good method to use. If you answered “no” to question 3, you **should not** use PCA.

Terms



-
- **Variance** : It is a measure of the variability or it simply measures how spread the data set is.
 - **Covariance** : It is a measure of the extent to which corresponding elements from two sets of ordered data move in the same direction.

$$var(x) = \frac{\sum(x_i - \bar{x})^2}{N}$$

$$cov(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}$$



Terms

Sparseness

- Data lacks denseness, and its high percentage of the variable's cells do not contain actual data.
- Fundamentally full of “empty” or “N/A” values.

c1	c2	c3	c4	c5
0	0	0	5	0
0	0	0	0	0
0	0	1	0	0
0	0	0	0	0
3	0	0	0	0
0	0	0	0	0

Example of a problem where PCA is required



- There are 100 students in a class with m different features like grade, age, height, weight, hair color, and others.
- Most of the features may not be relevant that describe the student. Therefore, it is vital to find the critical features that characterize a student.

Features to Ignore Vs Features to Keep



Ignore

- Collinear features or linearly dependent features. e.g., leg size and height.
- Noisy features that are constant. e.g., the thickness of hair
- Constant features. e.g., Number of teeth.

Keep

- Non-collinear features or low covariance.
- Features that change a lot, high variance. e.g., grade. *hobbies*



What does Principal Component Analysis (PCA) do?

- PCA finds a new set of dimensions (or a set of basis of views) such that all the dimensions are orthogonal (and hence **linearly independent**)
- ranked according to the variance of data along them. It means more important principle axis occurs first. (**more important = more variance/more spread out data**)

Steps in PCA

height in cm 170, 190
weight in kg 60, 70



STEP 1: STANDARDIZATION

- The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis.
- Mathematically, this can be done by subtracting the mean and dividing by the standard deviation for each value of each variable.

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$



Steps in PCA

STEP 2: COVARIANCE MATRIX COMPUTATION

- The aim of this step is to understand how the variables of the input data set are varying from the mean with respect to each other, or in other words, to see if there is any relationship between them.
- For example, for a 3-dimensional data set with 3 variables x , y , and z , the covariance matrix is a 3×3 matrix of this form:

$$\begin{bmatrix} \textcircled{Cov}(x, x) & Cov(x, y) & Cov(x, z) \\ Cov(y, x) & \textcircled{Cov}(y, y) & Cov(y, z) \\ Cov(z, x) & Cov(z, y) & \textcircled{Cov}(z, z) \end{bmatrix}$$



Steps in PCA

STEP 3: COMPUTE THE EIGENVECTORS AND EIGENVALUES OF THE COVARIANCE MATRIX TO IDENTIFY THE PRINCIPAL COMPONENTS

- Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the *principal components* of the data.
- If we rank the eigenvalues in descending order, we get $\lambda_1 > \lambda_2$, which means that the eigenvector that corresponds to the first principal component (PC1) is v_1 and the one that corresponds to the second principal component (PC2) is v_2 .



Steps in PCA

STEP 4: FEATURE VECTOR

- The feature vector is simply a matrix that has as columns the eigenvectors of the components that we decide to keep.
- This makes it the first step towards dimensionality reduction, because if we choose to keep only p eigenvectors (components) out of n , the final data set will have only p dimensions.



Steps in PCA

LAST STEP: RECAST THE DATA ALONG THE PRINCIPAL COMPONENTS AXES

- In this step, which is the last one, the aim is to use the feature vector formed using the eigenvectors of the covariance matrix, to reorient the data from the original axes to the ones represented by the principal components
- This can be done by multiplying the transpose of the original data set by the transpose of the feature vector.

$$\text{FinalDataSet} = \text{FeatureVector}^T * \text{StandardizedOriginalDataSet}^T$$



Goal of PCA

- Find linearly independent dimensions (or basis of views) which can losslessly represent the data points.
- Those newly found dimensions should allow us to predict/reconstruct the original dimensions.
- The reconstruction/projection error should be minimized.



Exercise

Let our data matrix \mathbf{X} be the score of three students :

Student	Math	English	Art
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

0 to 100

same
scale

Solution



- Step 1: Take the whole dataset consisting of $d+1$ dimensions and ignore the labels such that our new dataset becomes d dimensional.
- Step 2: Compute the mean of every dimension of the whole dataset.

$$A = \begin{bmatrix} M & E & A \\ 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix}$$

Matrix A

$$\bar{A} = [\frac{90+90+66+60+30}{5} \quad 60 \quad 60]$$

Mean of Matrix A



- Step3: Compute the covariance matrix of the whole dataset (sometimes also called as the variance-covariance matrix)

$$cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (\underline{X_i} - \bar{x})(\underline{Y_i} - \bar{y})$$

	<i>Math</i>	<i>English</i>	<i>Art</i>
<i>Math</i>	504	360	180
<i>English</i>	360	360	0
<i>Art</i>	180	0	720

Covariance Matrix of A



-
- a) The covariance between math and English is positive (360), and the covariance between math and art is positive (180). This means the scores tend to covary in a positive way. As scores on math go up, scores on art and English also tend to go up; and vice versa.
 - b) The covariance between English and art, however, is zero. This means there tends to be no predictable relationship between the movement of English and art scores.



* CAT PCA

- Step 4: Compute Eigenvalues and corresponding Eigenvalues
- The eigenvalues of A are roots of the characteristic equation

$$\det(A - \lambda I) = 0$$

identity matrix

$$\det \left(\begin{pmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right)$$

- Simplifying the matrix first, we can calculate the determinant later,

$$\begin{pmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{pmatrix} - \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}$$

$$\begin{pmatrix} 504 - \lambda & 360 & 180 \\ 360 & 360 - \lambda & 0 \\ 180 & 0 & 720 - \lambda \end{pmatrix}$$

- Now that we have our simplified matrix, we can find the determinant of the same :

$$\det \begin{pmatrix} 504 - \lambda & 360 & 180 \\ 360 & 360 - \lambda & 0 \\ 180 & 0 & 720 - \lambda \end{pmatrix}$$

$\lambda^3 + 1584\lambda^2 - 641520\lambda + 25660800$



-
- After solving this equation for the value of λ , we get the following value

$$\lambda \approx 44.81966\dots, \lambda \approx 629.11039\dots, \lambda \approx 910.06995\dots$$

- So, after solving for *eigenvectors* we would get the following solution for the corresponding *eigenvalues*

$$\begin{pmatrix} -3.75100\dots \\ 4.28441\dots \\ 1 \end{pmatrix}, \begin{pmatrix} -0.50494\dots \\ -0.67548\dots \\ 1 \end{pmatrix}, \begin{pmatrix} 1.05594\dots \\ 0.69108\dots \\ 1 \end{pmatrix}$$



-
- Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a $d \times k$ dimensional matrix W.
 - So, after sorting the eigenvalues in decreasing order, we have

$$\begin{matrix} \checkmark & \left(\begin{array}{c} 910.06995 \\ 629.11039 \\ 44.81966 \end{array} \right) \end{matrix}$$

-
- For our simple example, where we are reducing a 3-dimensional feature space to a 2-dimensional feature subspace, we are combining the two eigenvectors with the highest eigenvalues to construct our $d \times k$ dimensional eigenvector matrix W .
 - So, *eigenvectors* corresponding to two maximum eigenvalues are :

$$W = \begin{bmatrix} 1.05594 & -0.50494 \\ 0.69108 & -0.67548 \\ 1 & 1 \end{bmatrix}$$

=====



Eigen values & Eigen Vectors

$$A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$$

$$|A - \lambda I| = 0$$

$$\left| \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

$$\left| \begin{bmatrix} -\lambda & 1 \\ -2 & -3-\lambda \end{bmatrix} \right| = 0$$

$$\lambda^2 + 3\lambda + 2 = 0$$

$$\lambda_1 = -1, \quad \lambda_2 = -2$$

eigen values



v_1 - Eigen vector for λ_1 ,

$$Av_1 = \lambda_1 v_1$$

$$(A - \lambda_1 I) v_1 = 0$$

$$\left(\begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} - \lambda_1 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) v_1 = 0$$

$$\lambda_1 = -1$$

$$\begin{bmatrix} -\lambda_1 & 1 \\ -2 & -3 - \lambda_1 \end{bmatrix} v_1 = 0$$

$$\begin{bmatrix} +1 & 1 \\ -2 & -2 \end{bmatrix} v_1 = 0$$

$$v_{11} + v_{12} = 0$$

$$-2v_{11} - 2v_{12} = 0$$

$$v_{11} = -v_{12}$$

$$\begin{bmatrix} 1 & 1 \\ -2 & -2 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = 0$$

$$v_1 = k_1 \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$\lambda_2 = -2$$



$$A.v_2 = \lambda_2 v_2$$

$$(A - \lambda_2 I) v_2 = 0$$

$$\begin{bmatrix} -\lambda_2 & 1 \\ -2 & -3 - \lambda_2 \end{bmatrix} v_2 = \begin{bmatrix} 2 & 1 \\ -2 & -1 \end{bmatrix} \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = 0$$

$$2v_{21} = -v_{22}$$

$$-2v_{21} = v_{22}$$

$$v_2 = k_2 \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$



Self Study

-
- <https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html>
 - http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf
 - <https://textbooks.math.gatech.edu/ila/eigenvectors.html>
 - <https://towardsdatascience.com/eigenvalues-and-eigenvectors-all-you-need-to-know-df92780c591f>



References

- Probability and Statistics for engineering and sciences
- A Tutorial on Principal Component Analysis by Jonathon Shlens*
Google Research
- <https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202>
- Book: <http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/ADAfaEPoV.pdf>
- Numerical: <https://towardsdatascience.com/the-mathematics-behind-principal-component-analysis-fff2d7f4b643>,
<https://www.itl.nist.gov/div898/handbook/pmc/section5/pmc552.htm>
- Bivariate:
<http://personal.kenyon.edu/hartlaub/MellonProject/Bivariate2.html>
- <https://www.statlect.com/probability-distributions/multivariate-normal-distribution>
- Material by Prof. Nathaniel E. Helwig
- <https://medium.com/towards-artificial-intelligence/principal-component-analysis-pca-with-python-examples-tutorial-67a917bae9aa>