October 11, 2020

WEEK 1

What is statistics?

→ Statistics is the art of learning from data. It is concerned with the collection of data, their subsequent description, and their analysis, which often leads to drawing of conclusions.

Descriptive statistics

- data collection
- organising data
- describing data

LEARNING OBJECTIVES

1. What is statistics?
   → descriptive statistics, inferential statistics
   → Distinguish b/w a sample and a population
2. Understand how data are collected
   → Identify variables and cases (observations) in a data set.
3. Types of data
   → classify data as categorical (qualitative) or numerical (quantitative) data.
   → Understand cross-sectional versus time-series data
   → Measurement scales.
4. Creating data sets; Downloading & manipulating data sets; working on subsets of data.
5. Framing Questions that can be answered from data.

Date
October 12, 2020

↳ Analysis → better understanding

MAJOR BRANCH OF STATISTICS

1. DESCRIPTION - describing data
The part of statistics concerned with the description and summarization of data is called Descriptive Statistics.

2. INFERENCE
The part of statistics concerned with the drawing of conclusions from data is called Inferential Statistics.

» To be able to draw a conclusion from the data, we must take into account the possibility of chance - intro to probability.

POPULATION & SAMPLE

→ Population : The total collection of all the elements that we are interested in is called a population

→ Sample : A subgroup of the population that will be studied in detail is called a sample.

PURPOSE OF STATISTICAL ANALYSIS

* If the purpose of the analysis is to examine & explore information for its own intrinsic interest only, the study is descriptive.

* If the info. is obtained from a sample of a population and the purpose of the study is to use that info to draw conclusions about the population, the study is inferential.

* A descriptive study may be performed either on a sample or on a population.

* When an inference is made about the population, based on information obtained from the sample, does the study become inferential.

SUMMARY

Statistics - (descriptive and inferential) main branches

Population & Sample

[Major Branch of Statistics]

Descriptive → (can be done on either sample or on population.)

Inferential → (can be done only on sample)

Date: October 10, 2020

Understanding Data

WHAT IS DATA ?

In order to learn something, we need to collect data.

* Data are the facts and figures collected, analysed and summarised for presentation and interpretation.

→ Statistics relies on data, information that is around us.

WHY DO WE COLLECT DATA ?

* Interested in the characteristics of some group or groups of people, places, things or events.
* EXAMPLE : to know about temperatures in a particular month in Chennai, India.
* EXAMPLE : To know about the marks obtained by students in their class 12.
* to know how many people like a new song / product / video collected through comments.

DATA COLLECTION

* Data available : published data
* Data not available : need to collect, generate data

We assume data is available and our objective is to do a statistical analysis of available data.

* Analyse → to look at or think about the different parts or details of something carefully in order to understand or explain it !
* Interpret → to explain or understand meaning of something

UNSTRUCTURED & STRUCTURED DATA

* For the information in a database to be useful, we must know the context of the numbers and text it holds.
* When they are scattered about with no structure, the information is of very little use.
* Hence, we need to organize data.

DATASET

* A structured collection of data → is Dataset !
* It is a collection of values - could be numbers, names, roll numbers.

VARIABLES & CASES

* Case (observation) : A unit from which data are collected.
  ↳ Rows represent cases
* Variable : → columns represents variables.
  - intuitive → A variable is that 'varies'
  - formally → A characteristic or attribute that varies across all units
* In our school data set :
  - Case : each student
  - Variable : names, marks obt., Board etc.

* Rows represent cases : for each case, same attributes is recorded.

* Columns represent variables : for each variables, same type of value for each case is recorded.

SUMMARY

We have organised data in a spreadsheet into a table.

→ Each variable must have its own column.
→ Each observation must have its own row.

Date October 12, 2020

CLASSIFICATION OF DATA

CATEGORICAL & NUMERICAL Variable

DATA

- Categorical (Qualitative)
- Numerical (Quantitative)
  - Discrete (countable)
  - Continuous (measurable)

* Categorical data
  → also called qualitative variables
  → Identify group membership { gender, Board, Blood Grp, Jersey no., mobile no. }

* Numerical data
  → Also called quantitative variables
  → Describe numerical properties of cases
  → Have measurement units { marks, height, weight, temperature, score }

* Measurement Units : scale that defines the meaning of numerical data, such as weights measured in kilograms, prices in rupees, heights in cms etc.

  · The data that makeup a numerical ~~value~~ variable in a data table must share a common unit.

CROSS-SECTIONAL & TIME SERIES DATA

* Time series - data recorded over time (on a particular variable, example, quantity)

* Timeplot - graph of a time series showing values in chronological order

* Cross sectional - data observed at the same time

SUMMARY

- classify data as categorical or numerical
- For numerical data, find out unit of measurement
- Check whether data is collected at a point of time (cross-sectional data) or over time (time-series data)

October 12, 2020

SCALES OF MEASUREMENT

* Data collection requires one of the following scales of measurement : nominal, ordinal, interval or ratio.
  (nominal, ordinal → CATEGORICAL; interval, ratio → ORDINAL)

NOMINAL SCALE OF MEASUREMENT

→ When the data for a variable consists of labels or names used to identify the characteristic of an observation, the scale of measurement is considered a nominal scale.

→ Examples : Name, Board, Gender, Blood Group etc.

* Sometimes nominal variables might be numerically coded.
  → For example, we might code Men as 1 & Women as 2 or Men as 3 & Women as 5. Both codes are valid.

→ There is no ordering in the variable.

→ NOMINAL : name categories without implying order.

ORDINAL SCALE OF MEASUREMENT

→ Data exhibits properties of nominal data and the order or rank of data is meaningful, the scale of measurement is considered an Ordinal scale.

→ Each customer who visits a restaurant provides a service rating of excellent, good or poor.

- The data obtained are the labels - excellent, good or poor - the data have the properties of nominal data.

- In addition the data can be ranked, or ordered, with respect to service quality.

→ ORDINAL : name categories that can be ordered

INTERVAL SCALE OF MEASUREMENT

→ If the data have all the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure, then the scale of measurement is interval scale.

→ Interval data are always numeric. Can find out difference between any 2 values.

→ Ratios of values have no meaning here because the value of zero is arbitrary.

→ INTERVAL : numerical values that can be added/subtracted (no absolute zero)

Example : TEMPERATURE

→ Suppose the response to a question on how hot the day is comfortable & uncomfortable, then the temperature be a variable is NOMINAL.

→ Suppose the answer to measuring the temperature of a liquid is cold, warm, hot - the variable is ORDINAL.

→ Example : Consider an AC room where temperature is set at 20°C and the temperature outside the room is 40°C. It is correct to say that the difference in temp is 20°C, but it is incorrect to say that the outdoors is twice as hot as indoors.

→ Temperature in degrees Fahrenheit or degrees centigrade is an interval variable. No absolute zero.

	Celsius	Fahrenheit
Freezing pt.	0	32
Boiling pt.	100	212

but in Kelvin scale there is absolute 0.

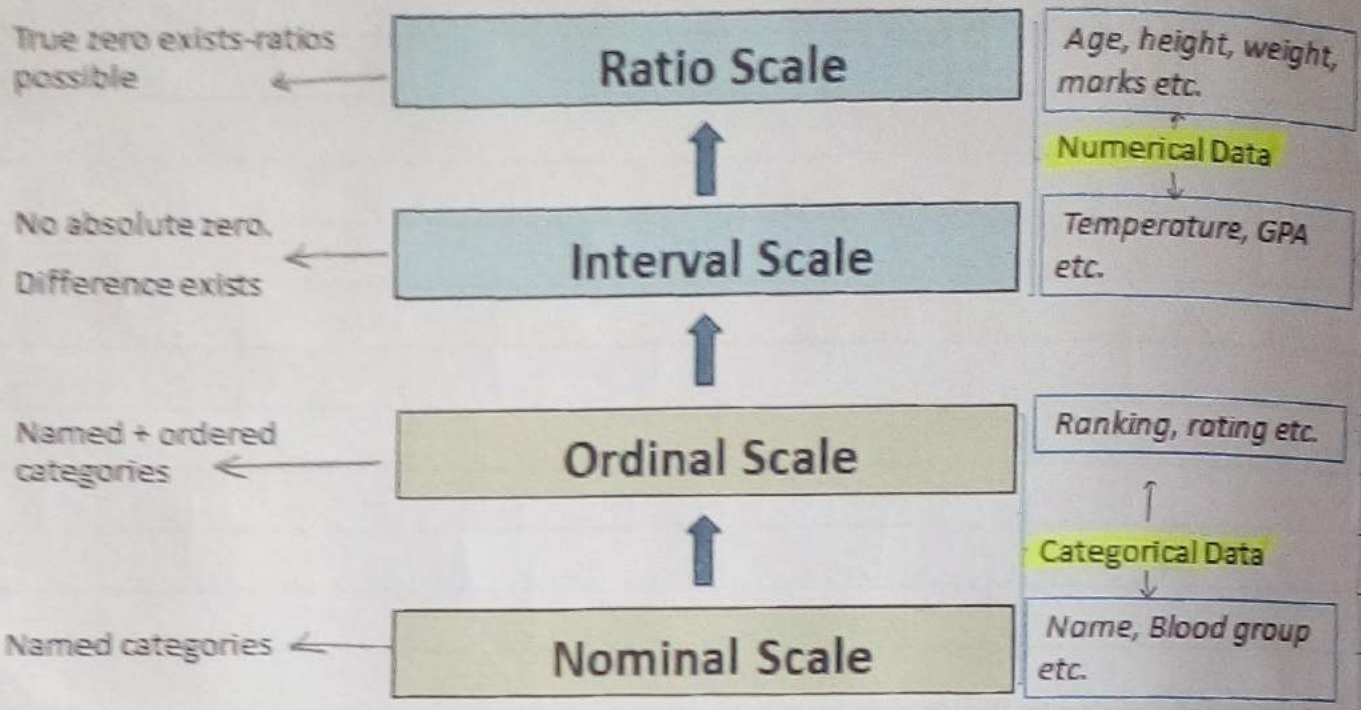
RATIO SCALE OF MEASUREMENT

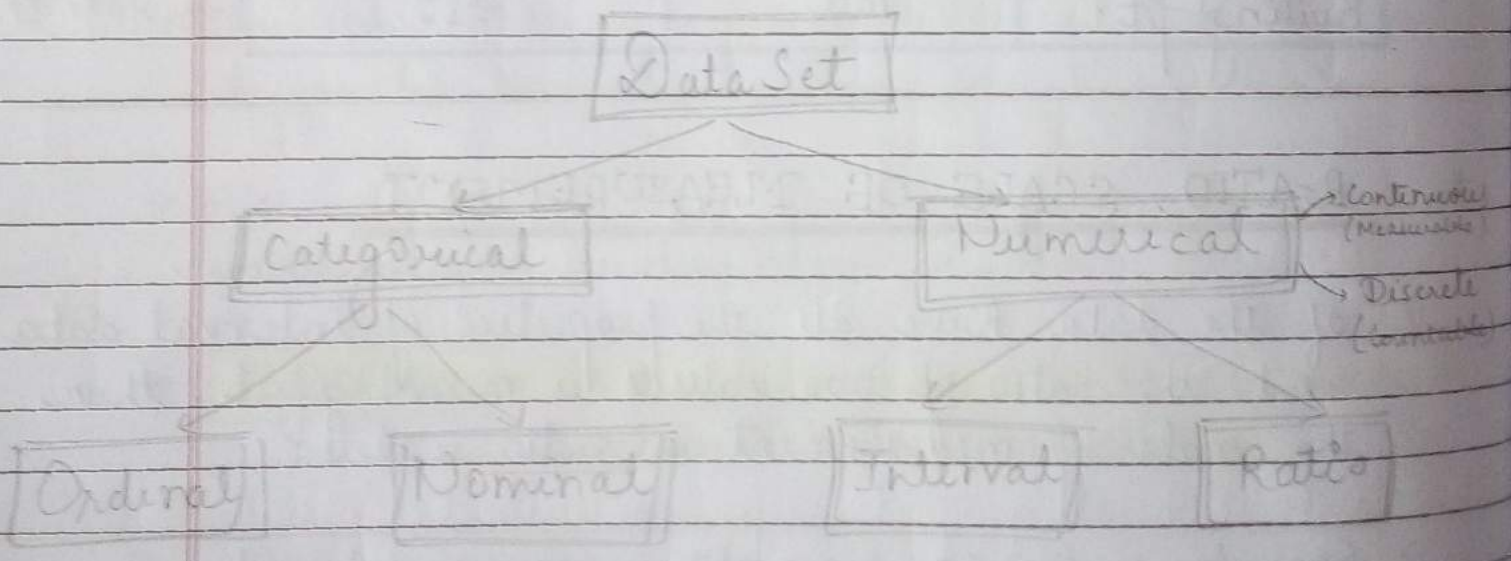
→ If the data have all the properties of interval data and the ratio of two values is meaningful, then the scale of measurement is ratio scale.

→ Example : height, weight, age, marks etc.

→ RATIO : numerical values that can be added, subtracted, multiplied or divided (makes ratio comparisons possible)

SUMMARY


True zero exists-ratios possible
Ratio Scale
Age, height, weight, marks etc.
Numerical Data
No absolute zero.
Difference exists
Interval Scale
Temperature, GPA etc.
Named + ordered categories
Ordinal Scale
Ranking, rating etc.
Categorical Data
Named categories
Nominal Scale
Name, Blood group etc.


Data Set
Categorical
Numerical
Continuous
Discrete
Ordinal
Nominal
Interval
Ratio

Date
October 15, 2020

classmate
Date
Page

WEEK 2

Describing Categorical Data

FREQUENCY DISTRIBUTIONS

A frequency distribution of qualitative data is a listing of the distinct values and their frequencies. (count)

Each row of a frequency table lists a category along with the number of cases in this category. COUNT

CONSTRUCT A FREQUENCY DISTRIBUTION

The steps to construct a frequency distribution:

STEP 1: List the distinct values of the observations in the data set in the first column of a table.

STEP 2: For each observation, place a tally mark in the second column of the table in the row of the appropriate distinct value.

STEP 3: Count the tallies for each distinct value and record the totals in the third column of the table.

FREQUENCY TABLE IN A GOOGLESHEET

STEP 1: Select / Highlight the cells having data you want to visualize.

STEP 2: In the Formatting bar click on the Data option.

STEP 3: In the Data option go to Pivot Table option and create a new sheet

STEP 4: After creating Pivot Table, go in Pivot Table Editor and in that first add row & then values.

RELATIVE FREQUENCY

The ratio of the frequency to the total no. of observation is called relative frequency.

→ The steps to construct a relative frequency distribution

Step 1: Obtain a frequency distribution of the data

Step 2: Divide each frequency by the total no. of observation.

Why relative frequency ?

→ for comparing two data sets.

→ because relative frequencies always fall b/w 0 and 1, they provide a standard for comparison.

Date October 7, 2020

CHARTS OF CATEGORICAL DATA

→ The two most common displays of a categorical variable are a bar chart and a pie chart.

→ Both describe a categorical variable by displaying its frequency table.

★ PIE CHARTS

A pie chart is a circle divided into pieces proportional to the relative frequencies of the qualitative data.

The steps to construct a pie-chart -:

Step 1: Obtain a relative frequency distribution of data.

Step 2: Divide a circle into pieces proportional to the relative frequencies.

Step 3: Label the slices with the distinct values and their relative frequencies.

EXAMPLE

1. A, A, B, C, A, D, A, B, D, C, A, B, C, D, A

Category	Tally mark	Frequency	R.F.	Degrees
A	𝍸 |	6	0.4	144
B	|||	3	0.2	72
C	|||	3	0.2	72
D	|||	3	0.2	72
Total		15	1	360°

Pie Chart in a google sheet

STEP 1 : Select / Highlight the cells having data you want to visualize.

STEP 2 : Click the insert chart option in Google Sheets toolbar.

STEP 3 : Change the visualization type in chart Editor

STEP 4 : Sleed in Chart Editor, Chart Type to Pie Chart.

SECTIONAL SUMMARY

- A pie chart is used to show the proportions of the a categorical variable
- A pie chart is a good way to show that one category makes up more than half of the total.

★ BAR CHART

A bar chart displays the distinct values of the qualitative data on a horizontal axis and the relative frequencies (or frequencies or percents) of those values on a vertical axis.

The frequency / relative frequency of each distinct value is represented by a vertical bard whose height is equal to the frequency / relative frequency of that value. The bars should be positioned so that they do not touch each other.

The steps to construct a bar chart -

Step 1: Obtain a frequency / relative frequency distribution of the data.

Step 2: Draw a horizontal axis on which to place the bars and a vertical axis on which to display the frequencies / relative frequencies.

Step 3: For each distinct value, construct a vertical bar whose height equals the frequency / relative frequency of that value.

Step 4: Label the bars with the distinct values, the horizontal axis with the name of the variable, and the vertical axis with "Frequency" / "Relative Frequency"

Bar Chart In a Google Sheet

STEP 1: Select / Highlight the cells having data you want to visualize.

STEP 2: Click the Insert Chart option in Google Sheets Toolbar.

STEP 3: Change the visualization type in Chart Editor.

STEP 4: Select in Chart Editor, Chart Type to Bar Chart

★ PARETO CHARTS

When the categories in a bar chart are sorted by frequency, the bar chart is sometimes called a Pareto Chart.

Pareto charts are popular in quality control to identify problems in a business process.

▶ If the categorical variable is ORDINAL, then the bar chart must preserve the ordering.

SECTIONAL SUMMARY

- A bar chart is used to show the frequencies / relative frequencies of a categorical variable.
- If ordinal, the order of categories is preserved.
- The bars can be oriented either horizontally or vertically.
- A Pareto chart is a bar chart where the categories are sorted by frequency.

October 17, 2020

BEST PRACTICES

KNOW YOUR PURPOSE

→ Have a purpose for every table or graph you create
  · choose the table / graph to serve the purpose

→ Pie Charts are best to use when you are trying to compare parts of a whole.

→ Bar graphs are used to compare things between different groups.

LABEL YOUR DATA

→ Label your chart to show the categories and indicate whether some have been combined or omitted.

→ Name the bars in a bar chart

→ Name the slices in a pie chart

→ If you have omitted some of the cases, make sure the label of the plot defines the collection that is summarized.

MANY CATEGORIES

- A bar chart or pie chart with too many categories might conceal the more important categories.
- In some cases, grouping other categories together might be done.

October 17, 2020

MISLEADING GRAPHS

classmate
Date
Page

THE AREA PRINCIPLE

→ Displays of data must obey a fundamental rule called the area principle.

→ The area principle says that the area occupied by a part of the graph should correspond to the amount of data it represents.

→ Violations of the area principle are a common way to mislead with statistics.

MISLEADING GRAPHS

Lec 2.3-2.4 Pg-18

(1) Violating Area Principle

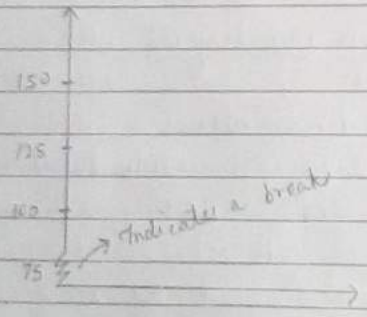
→ Decorated Graphics : charts decorated to attract attention often violate the area principle

(2) Truncated Graphs - Pg 23

→ Another common violation is when the baseline of the bar chart is not at zero.

Left graph exaggerates the no. coming from the south and North. Graph on right shows same data with the baseline at zero.

INDICATING A Y-AXIS BREAK



ROUND-OFF ERRORS - Pg 26

→ Important to check for round off errors.

→ When the table entries are percentages or proportions, the total may sum to a value slightly different from 100% or 1. This might result in a pie chart where the total does not add up.

SECTIONAL SUMMARY

- Know your purpose & choose the table/graph appropriately
- Labels your charts
- Handle multiple categories appropriately
- Respect Area Principle
  - Avoid Overly Decorated graphs
  - Avoid Truncated Graphs (use special symbols to indicate vertical axis has been modified)
  - Check for round off errors.

October 18, 2020

SUMMARIZING CATEGORICAL DATA

→ Graphical summaries of categorical data : bar chart and pie chart.

→ Need for a compact measure.

→ Numbers that are used to describe data sets are called DESCRIPTIVE MEASURES.

→ Descriptive measures that indicate where the center or most typical value of a data set lies are called MEASURES OF CENTRAL TENDENCY.

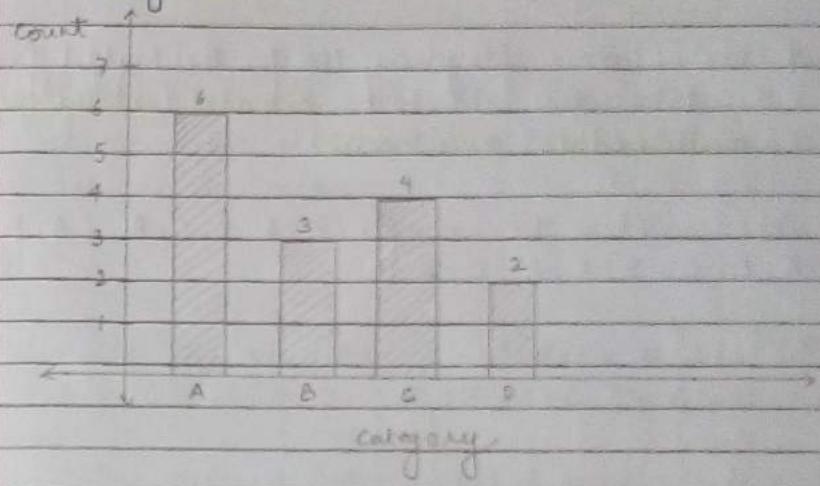
MODE

The mode of a categorical variable is the most common category, the category with the highest frequency.

The mode labels
- the longest bar in the bar charts
- the widest slice in a pie chart
- In a Pareto chart, the mode is the first category shown.

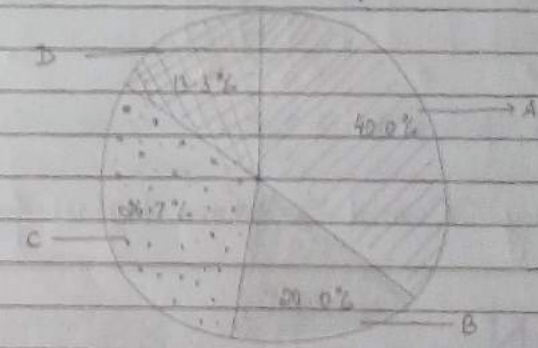
Example : Lets consider the example A, A, B, C, A, D, A, B, C, C, A, B, C, D, A

→ The longest bar in the bar chart



The most common category is "A"

→ The widest slice in a pie chart

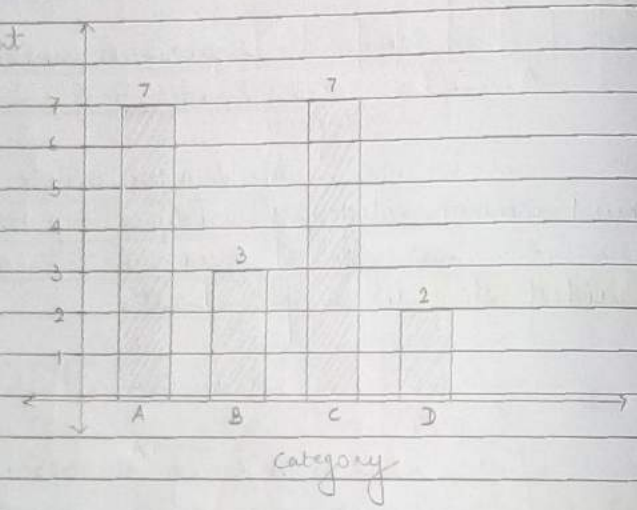


The most common category is "A".

BIMODAL & MULTIMODAL DATA

→ If two or more categories tie for the highest frequency, the data are said to be bimodal (in the case of 2) or multimodal (more than 2).

→ Lets consider the example A, A, B, C, A, C, A, B, C, C, A, C, C, D, A, A, C, D, B.



→ Both category "A" and "C" have highest frequency.

MEDIAN

→ Ordinal data offer another summary, the median, that is not available unless the data can be put into order.

* The MEDIAN of an ordinal variable is the category of the middle observation of the sorted values.

→ If there are an even no. of observations, choose the category on either side of the middle of the sorted list as the median.

Example :

(1) Consider the grades of 15 students which is listed as A, B, B, C, A, D, B, B, A, C, B, B, C, D, A

Odd median = $(n+1) \div 2$ (n = no. of observation)

→ The ordered data is A, A, A, A, B, B, B, B, B, B, C, C, C, D, D
→ The median grade is the category "B", i.e. category associated with the 8th observation.

Even median = $\left(\frac{n}{2}\right)$ and $\left(\frac{n}{2}+1\right)$ (n = no. of observation)

(2) Consider the grade of 14 students which is listed as A, B, B, C, A, D, B, B, A, C, B, B, C, D.

→ The ordered data is A, A, A, B, B, B, B, B, C, C, C, D, D
→ The median grade is the category associated with the 7 or 8 observation which is "B".

SUMMARY

- The mode of a categorical variable is the most common category.
- The median of an ordinal variable is the category of the middle observation of the sorted values.

Summary Of Week 2

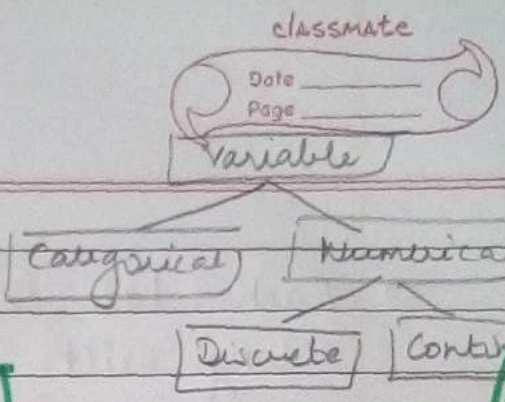
1. Tabulate Data : frequency and relative frequency

2. Charts of Categorical Data
   - Pie Charts
   - Bar Charts and Pareto Charts

3. Best Practices & Misleading Graphs
   - Label your Data
   - Dealing with multiple categories
   - Area Principle
   - Misleading Graphs
     - Decorated graphs
     - Truncated graphs (Missing Baseline)
     - Round off errors

4. Descriptive Measures
   - Mode (most frequency)
   - Median for ordinal data

October 29, 2020



WEEK 3

Describing Numerical Data

ORGANISING NUMERICAL DATA

- A discrete variable usually involves a count of something whereas a continuous variable usually involves a measurement of something.

- First, group the observations into classes (also known as categories or bins) and then treat the classes as the distinct values of qualitative data.

- Once we group the quantitative data into classes, we can construct frequency and relative frequency distribu-tions of the data in exactly the same way as we did for categorical data.

ORGANIZING DISCRETE DATA (SINGLE VALUE)

- If the data set contains only a relatively small number of distinct, or different, values, it is convinient to represent it in a frequency table

→ Each class represents a distinct value (single value) along with its frequency of occurrence.

EXAMPLE : Suppose the dataset reports the no. of people in a household. The following data is the response from 15 individuals.

→ 2, 1, 3, 4, 5, 2, 3, 3, 3, 4, 4, 1, 2, 3, 4

→ The distinct values the variable, no. of people in each household, takes is 1, 2, 3, 4, 5

→ The frequency distribution table is

Value	Tally Mark	Frequency	Relative Frequency
1	II	2	0.133
2	III	3	0.2
3	IIII	5	0.333
4	IIII	4	0.266
5	I	1	0.066
TOTAL		15	1

ORGANIZING CONTINUOUS DATA

Organize the data into a number of classes to make the data understandable. However, there are few guidelines that need to be followed. They are:

1. Number of classes → The appropriate no. is a subjective choice, the rule of thumb is to have between 5 and 20 classes.
2. Each Observation should belong to some class and no observation should belong to more than one class.
3. It is common, although not essential, to choose class intervals of equal length.

SOME NEW TERMS

1. Lower class limit : The smallest value that could go in a class.
2. Upper class limit : The largest value that could go in a class.
3. Class Width : The difference between the lower limit of a class & the lower limit of the next higher class.
4. Class Mark : The average of the two class limits of a class.
5. A class interval contains its left-end but not its right-end boundary point.

EXAMPLE : The marks obtained by 50 students in a particular course.

→ 68, 79, 38, 68, 35, 70, 61, 47, 58, 66, 60, 45, 61, 60, 59, 45, 39, 80, 59, 62, 49, 76, 54, 60, 53, 55, 62, 58, 67, 55, 86, 56, 63, 64, 67, 50, 51, 78, 56, 62, 57, 69, 58, 52, 42, 66, 42, 36, 58

Class Interval	Tally Mark	Frequency	Relative Frequency
30 - 40	III	3	0.06
40 - 50	IIII I	6	0.12
50 - 60	IIII IIII IIII III	18	0.36
60 - 70	IIII IIII IIII II	17	0.34
70 - 80	IIII	4	0.08
80 - 90	II	2	0.04
TOTAL		50	1

SECTION SUMMARY

- Frequency table for discrete single value data.
- Frequency table for continuous data using class intervals

GRAPHICAL SUMMARIES

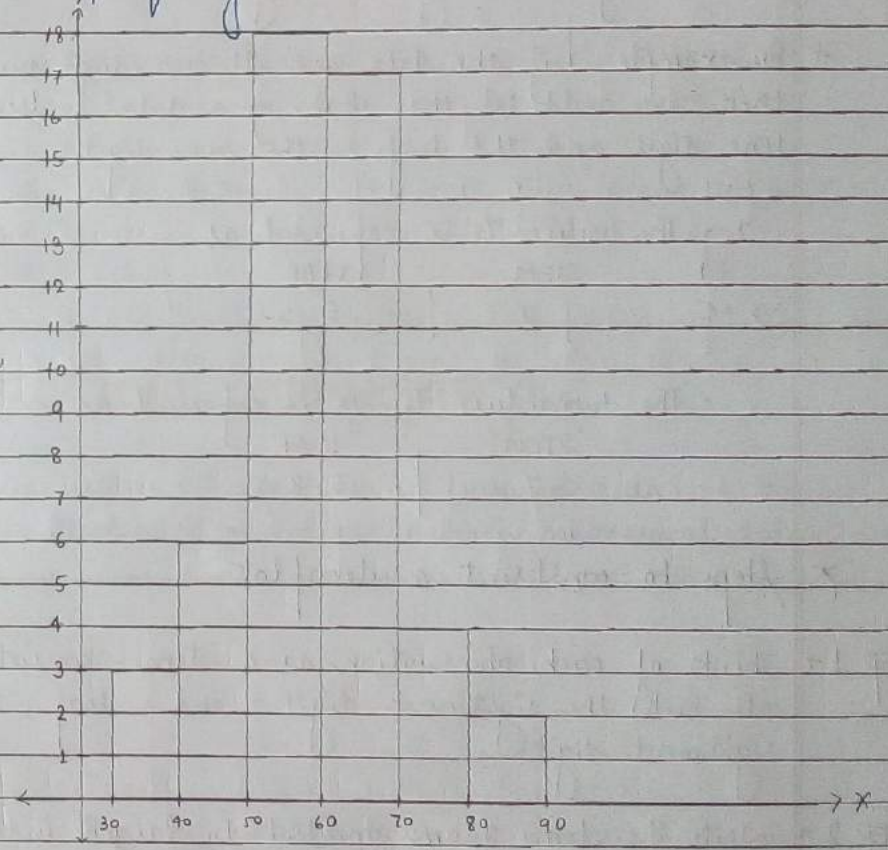
Steps to construct a histogram

STEP 1 → Obtain a frequency (relative frequency) distribution of the data.

STEP 2 → Draw a horizontal axis on which to place the classes and a vertical axis on which to display the frequencies.

STEP 3 → For each class, construct a vertical bar whose height equals the frequency of that class.

STEP 4 → Label the bars with the classes, the horizontal axis with the name of the variable, and the vertical axis with "frequency".



HISTOGRAM

STEM AND LEAF DIAGRAM

In a stem-and-leaf diagram (or stemplot), each observation is separated into two parts, namely, a stem - consisting of all but the rightmost digit - and a leaf, the rightmost digit.

→ For example, if the data are all two-digit numbers, then we could let the stem of a data value be the tens digit and the leaf be the ones digit.

- The value 75 is expressed as

STEM	LEAF
7	5

- The two values 75, 78 is expressed as

STEM	LEAF
7	5, 8

→ Steps to construct a stemplot

STEP 1→ Think of each observation as a stem - consisting of all but the rightmost digit - and a leaf, the rightmost digit.

STEP 2→ Write the stems from smallest to largest in a vertical column to the left of a vertical rule.

STEP 3→ Write each leaf to the right of the vertical rule in the row that contains the appropriate stem.

STEP 4→ Arrange the leaves in each row in ascending order.

EXAMPLE: The following are the ages, to the nearest year, of 11 patients admitted in a certain hospital:
15, 22, 29, 36, 31, 23, 45, 10, 25, 28, 48.

Draw a stem-and-leaf plot for this data set.

1	0, 5
2	2, 3, 5, 8, 9
3	1, 6
4	5, 8

SECTION SUMMARY

- construct a histogram for grouped data.
- construct a stemplot to describe numerical data.

October 30, 2020

NUMERICAL SUMMARIES

DESCRIPTIVE MEASURES

- The objective is to develop measures that can be used to summarize a data set.
- These descriptive measures are quantities whose values are determined by the data.
- Most commonly used descriptive measures can be categorized as

→ Measures of Central Tendency : These are measures that indicate the most typical value or centre of a data set.

→ Measures of Dispersion : These measures indicate the variability or spread of a dataset.

MEASURES OF CENTRAL TENDENCY

1. MEAN

The most commonly used measure of central tendency is the mean.

Definition : The mean of a data set is the sum of the observations divided by the number of observations

→ The mean is usually refered to as 'Average'.

→ Arithmetic average ; divide the sum of the values by the number of values (another typical value)

→ For DISCRETE observations :

- Sample mean : $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$
- Population mean : $\mu = \frac{x_1 + x_2 + \dots + x_N}{N}$

Example : The marks obtained by ten students in an exam is
68, 79, 38, 68, 35, 70, 61, 47, 58, 66

→ The sample mean is

$$\frac{68 + 79 + 38 + 68 + 35 + 70 + 61 + 47 + 58 + 66}{10} = \frac{590}{10} = 59$$

★ MEAN FOR GROUPED DATA (DISCRETE)

→ The following data is the response from 15 individuals
2, 1, 3, 4, 5, 2, 3, 3, 3, 4, 4, 1, 2, 3, 4.

→ $$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \dots + f_n x_n}{n}$$

where, f_i is frequency
x_i is value
n is total observations.

Value (x_i)	Tally Mark	Frequency (f_i)	$f_i x_i$
1	II	2	2
2	III	3	6
3	IIII	5	15
4	IIII	4	16
5	I	1	5
TOTAL		15	44

$$\text{Mean} = \frac{\sum f_i x_i}{\sum f_i} = \frac{44}{15} = 2.93$$

★ MEAN FOR GROUPED DATA (CONTINUOUS)

$$\bar{x} = \frac{f_1 m_1 + f_2 m_2 + \ldots + f_n m_n}{n}$$

where m is MID POINT of class interval.

CLASS INTERVAL	TALLY MARK	FREQUENCY (f_i)	MID POINT (m_i)	$f_i m_i$
30-40	III	3	35	105
40-50	IIII I	6	45	270
50-60	IIII IIII IIII III	18	55	990
60-70	IIII IIII IIII II	17	65	1105
70-80	IIII	4	75	300
80-90	II	2	85	170
TOTAL		50		2940

→ AVERAGE $= \frac{2940}{50} = 58.8$

→ 58.8 is an approximate and not exact value of the mean.

★ ADDING A CONSTANT

- Let $y_i = x_i + c$ where c is a constant then $\bar{y} = \bar{x} + c$
- Example : Recall the marks of students
  68, 79, 38, 68, 35, 70, 61, 47, 58, 66

→ Suppose the teacher has decided to add 5 marks to each student.

→ Then the data becomes
73, 84, 43, 73, 40, 75, 66, 52, 63, 71

→ The mean of the new data set is $\frac{640}{10} = 64 = 59 + 5$
$\bar{y} = \bar{x} + c$

★ MULTIPLYING A CONSTANT

- Let $y_i = x_i c$ where c is a constant then $\bar{y} = \bar{x}c$
- Example : Recall the marks of of students
  68, 79, 38, 68, 35, 70, 61, 47, 58, 66

→ Suppose the teacher has decided to scale down each mark by 40%, in other words each mark is multiplied by 0.4.

→ Then the data becomes
27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4

→ The mean of the new data set is $\frac{236}{10} = 23.6 = 59 \times 0.4$
$\bar{y} = \bar{x}c$

SECTION SUMMARY

1. Mean or average is a measure of central tendency
2. Compute sample mean for
   - → ungrouped data
   - → grouped discrete data
   - → grouped continuous data
3. Manipulating Data
   - → Adding a constant to each data pt.
   - → Multiplying each data point with a constant

Date October 31, 2020

NUMERICAL SUMMARIES

Measures of Central Tendency

2. MEDIAN

→ Another frequently used measure of center is the median.

→ Essentially, the median of a data set is the number that divides the bottom 50% of the data from the top 50%.

Definition : The median of a data set is the middle value in its ordered list.

Steps to obtain median

Arrange the data in increasing order. Let n be the total no. of observations in the dataset.

1. If the no. of observations is odd, then the median is the observation exactly in the middle of the ordered list, i.e. $(n+1)/2$ observation.
2. If the no. of observation is even, then the median is the mean of the two middle observations in the ordered list, i.e, mean of $n/2$ and $n/2 + 1$ observation.

EXAMPLES : (1) 2, 12, 5, 6, 7, 3

- Arrange data in increasing order → 2, 3, 5, 6, 7, 12
- $n = 7$ (odd) $\therefore$ median $= \left(\frac{n+1}{2}\right)^{th}$ observation $= \frac{8}{2} = 4^{th}$ observation $= "6"$

(2) 2, 105, 5, 7, 6, 7, 3

- Arrange in ↑ order : 2, 3, 5, 6, 7, 7, 105
- n = 7 (odd)

∴ Median = $\left(\frac{n+1}{2}\right)^{th}$ observation = $\left(\frac{7+1}{2}\right) = \frac{8}{2} = 4^{th}$ obs. = "6"

(3) 2, 105, 5, 7, 6, 3

- Arrange in ↑ order : 2, 3, 5, 6, 7, 105
- n = 6 (even)

∴ Median is avg. of $\left(\frac{n}{2}\right)^{th}$ & $\left(\frac{n}{2}\right)+1^{th}$ observation

$= \frac{5+6}{2} = 5.5$ is median.

Difference Between Mean & Median

EXAMPLE : (1) 2, 12, 5, 6, 7, 3, 7

Sample mean = $\frac{2+3+6+5+7+7+12}{7} = 6$

Sample median = 6

(2) 2, 117, 5, 7, 6, 7, 3

Sample mean = $\frac{2+3+5+6+7+7+117}{7} = 21$

Sample median = 6

Note → The sample mean is sensitive to outliers, whereas the sample median is not sensitive to outliers.

★ ADDING A CONSTANT

- Let $y_i = x_i + c$, where c is the constant, then

new median = old median + c

EXAMPLE, Recall the marks of students

68, 79, 38, 68, 35, 70, 61, 47, 58, 66

→ Arranging in ascending order, 35, 38, 47, 58, 61, 66, 68, 68, 70, 79

→ The median for the data is avg. of (n/2) & (n/2 + 1) observation which is (61+66)/2 = 127/2 = 63.5

→ Suppose the teacher has decided to add 5 marks to each student.

→ Then data in ascending order is

40, 43, 52, 63, 66, 71, 73, 73, 75, 84

→ The median of the new data set is $\frac{66+71}{2} = \frac{137}{2} = 68.5$

→ 68.5 = 63.5 + 5

new median = old median + constant

★ MULTIPLYING A CONSTANT

- Let $y_i = x_i c$ where c is a constant then

new median = old median × c

EXAMPLE, Recall the marks of students

68, 79, 38, 68, 35, 70, 61, 47, 58, 66

We already know median for this data is 63.5.

→ Suppose the teacher has decided to scale down each mark by 40%, in other words each mark is multiplied by 0.4.

→ Then the data becomes
27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4
The ascending order is
14, 15.2, 18.8, 23.2, 24.4, 26.4, 27.2, 28, 31.6

The median of the new data set is $(24.4 + 26.4) \div 2$
$= 50.8 \div 2 = 25.4$

→ Note, $25.4 = 0.4 \times 63.5$
new median = constant × old median

3. MODE

Another measure of central tendency is the sample mode

Definition : The mode of the data set is its most frequently occurring value.

Steps To Obtain Mode

- If no value occurs more than once, then the data set has no mode.
- Else, the value that occurs with the greatest frequency is a mode of the data set.

★ ADDING A CONSTANT

- Let $y_i = x_i + c$ where c is a constant then
new mode = old mode + c

EXAMPLE, Recall the marks of students
68, 79, 38, 68, 35, 70, 61, 47, 58, 66
The mode for this data is 68.

→ Suppose the teacher has decided to add 5 marks to each student.
Then the data in ascending order is
40, 43, 52, 63, 66, 71, 73, 73, 75, 84

→ The mode of the new data set is 73

→ $73 = 68 + 5$
$\text{new}_{\text{mode}} = \text{old}_{\text{mode}} + \text{constant}$

★ MULTIPLYING A CONSTANT

- Let $y_i = x_i c$, where c is a constant then

new mode = old mode × c

SUMMARY

→ Measure of Central Tendency : Mean, Median, Mode
→ Impact of adding a constant or multiplying with constant on the measures.

Date: October 31, 2020

Measures Of Dispersion

Why Do We Need a Measure of Dispersion?

→ Consider the two data sets given below :
- Dataset 1 → 3, 3, 3, 3, 3
- Dataset 2 → 1, 2, 3, 4, 5

→ The measures of central tendency for both datasets are:

	Dataset 1	Dataset 2
MEAN	3	3
MEDIAN	3	3
MODE	3	not available

→ The mean median are same for both datasets. However the datasets are not same. They are different.

→ To describe that difference quantitatively, we use a descriptive measure, that indicates the amount of variation, or spread, in a data set.

→ Such descriptive measures are refered to as
- measures of dispersion, or
- measures of variation, or
- measures of spread

→ In this course, we will be discussing about the following measures of dispersion.

1. Range
2. Variance
3. Standard Deviation
4. Interquartile range

1. RANGE

Definition : The range of a dataset is the difference b/w its largest and smallest values.

→ The range of a dataset is given by the formula

$$\text{Range} = \text{Max} - \text{Min}$$

where max and min denote the maximum & minimum observations, respectively.

	Dataset 1	Dataset 2
	3, 3, 3, 3, 3	1, 2, 3, 4, 5
→ Max	3	5
Min	3	1
Range	0	4

∴ Range is sensitive to outliers. For example, consider two datasets as given below,

	Dataset 1	Dataset 2
	1, 2, 3, 4, 5	1, 2, 3, 4, 15
Max	5	15
Min	1	1
Range	4	14

→ Though the two datasets differ only in one datapoint, we can see that the contributed to the value of range significantly. This happens because the range takes into consideration only the Min & Max of the dataset.

2. VARIANCE

- In contrast to the range, the variance takes into account all the observations.
- One way of measuring the variability of a dataset is to consider the deviations of the data values from a central value.

POPULATION VARIANCE & SAMPLE VARIANCE

Recall when we refer to a dataset from a population, we assume the dataset has N observations, whereas, when refer to a dataset from a sample, we assume the datbset has n observations.

- The variance is computed using the following formulae-

→ Population Variance : $\sigma^2 = \frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \dots + (x_N-\mu)^2}{N}$

→ Sample Variance : $s^2 = \frac{(x_1-\bar{x})^2 + (x_1-\bar{x})^2 + \dots + (x_n-\bar{x})^2}{n-1}$

- The numerator is the sum of squared deviations of every observation from its mean.
- The denominator for computing population variance is N, the total no. of observations.
- The denominator for computing sample variance is (n-1).

EXAMPLE : Recall marks of students obtained by 10 students in an exam is 68, 79, 38, 68, 35, 70, 61, 47, 58, 66

→ The mean was computed to be 59.

→ The deviations of each data pt. from its mean is given in the table below :

	Data	Deviation $(x_i - \bar{x})$	Squared Deviations $(x_i - \bar{x})^2$
1	68	9	81
2	79	20	400
3	38	-21	441
4	68	9	81
5	35	-24	576
6	70	11	121
7	61	2	4
8	47	-12	144
9	58	-1	1
10	66	-7	49
TOTAL	590	0	1898

1. Population Variance = $\frac{1898}{10}$ = 189.8

2. Sample Variance = $\frac{1898}{9}$ = 210.88

★ ADDING A CONSTANT

→ Let $y_i = x_i + c$, where c is constant then,

new variance = old variance

EXAMPLE, Recall the marks of students
68, 79, 38, 68, 35, 70, 61, 47, 58, 66
Sample Variance = 210.88

→ Suppose the teacher has decided to add 5 marks to each student.
Then the data is 73, 84, 43, 73, 40, 75, 66, 52, 63, 71

→ The variance of new dataset is $\frac{1898}{9}$ = 210.88

→ In general, adding a constant does not change variability of a dataset, and hence it is the same.

★ MULTIPLYING A CONSTANT

• Let $y_i = x_i c$, where c is a constant then

new variance = c^2 x old variance

3. STANDARD DEVIATION

Another very useful measure of dispersion is the standard deviation.

Definition : The quantity

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}}$$

which is the square root of sample variance is the sample standard deviation.

UNITS OF STANDARD DEVIATION

→ The sample variance is expressed in units of square units if original variable. For example, instead of marks if the data were weights of 10 students measured in kg. Then the unit of variance would be kg^2.

→ The sample standard deviation is measured in the same units as the original data. That is, for instance, if the data are in kg, then the units of std. deviation are also in kg.

★ ADDING A CONSTANT

• Let $y_i = x_i + c$, where c is a constant then,

new variance = old variance

★ MULTIPLYING A CONSTANT

- Let $y_i = x_i c$ where c is a constant then,

  new variance = c^2 x old variance

SECTION SUMMARY

- Measures of Dispersion
  1. Range
  2. Variance
  3. Standard Deviation
- Impact of Adding a constant or multiplying with a constant on the measures.

Date October 31, 2020

PERCENTILES

→ The sample 100p percentile is that data value having the property that at least 100p percent of the data are less than or equal to it and at least $\{100(1-p)$ percent of the data values are greater than or equal to it.

→ If two data values satisfy this condition, then the sample 100p percentile is the arithmetic avg of these values.

→ Median is the 50^{th} percentile.

Computing Percentile

- To find the sample 100p percentile of a data set of size n.

1. Arrange the data in increasing order.
2. If np is not an integer, determine the smallest integer greater than np. The data value in that position is the sample 100p percentile.
3. If np is an integer, then the avg. of the values in positions np & np + 1 is the sample 100p percentile.

Example : Let n = 10

→ Arrange data in ascending order

35, 38, 47, 58, 61, 66, 68, 68, 70, 79

p	np	Percentile Value
0.1	1	36.5
0.25	2.5	47
0.5	5	63.5
0.75	7.5	68
1	10	79

QUARTILES

Definition : The sample 25th percentile is called the first quartile. The sample 50th percentile is called the median or second quartile. The sample 75th percentile is called the third quartile.

In other words,

the quartiles break up a data set into four parts with about 25 percent of the data values being less than the first (lower) quartile, about 25 percent being b/w the first & second quartiles, about 25 percent being b/w the second & third (upper) quartiles, and about 25 percent being larger than the third quartile.

The Five Number Summary

→ Minimum
→ Q_1 : First Quartile or Lower Quartile
→ Q_2 : Second Quartile or Median
→ Q_3 : Third Quartile or Upper Quartile
→ Maximum

4. THE INTERQUARTILE RANGE

Definition : The interquartile range, IQR, is the difference between the first and third quartiles; that is

$$IQR = Q_3 - Q_1$$

→ IQR for the example

- First Quartile, $Q_1 = 49.75$
- Third Quartile, $Q_3 = 68$
- $IQR = Q_3 - Q_1 = 18.25$

SECTION SUMMARY

- Definition of Percentiles
- How to compute percentiles
- Definition of quartile
- Five no. summary
- Inter quartile range as a measure of dispersion

SUMMARY OF WEEK 3

1. Frequency Tables
   - Frequency tables for discrete data
   - Frequency table for continuous data

2. Graphical Summaries
   - Histograms
   - Stem-and-leaf plot

3. Numerical Summaries
   - Measures of central Tendency
     → Mean, Median, Mode
   - Measures of Dispersion
     → Range, Variance, Std. Deviation
   - Percentiles
     → Interquartile range as a measure of dispersion

Date
June 2, 2021

classmate
Date
Page

WEEK 4

Association Between 2 Categorical Variables

LEARNING OBJECTIVES (for Week 4)

(1) Use of two way contingency tables to understand association between two categorical variables.

(2) Understand association between numerical variable through scatter plots ; compute and interpret correlation

(3) Understand relationship between a categorical and numerical variable

---

INTRODUCTION

- To understand the association between two categorical variables.
- Learn how to construct two-way contingency table
- Learn concept of relative row/column frequencies and how to use them to determine whether there is an association between the categorical variables.

EXAMPLE 1 (Nominal Variables)

Gender v/s Use of Smartphone

- A market research firm is interested in finding out whether ownership of a smartphone is associated with gender of a student. In other words, they want to find out whether more females owns a smartphone while compared to males, or whether owning a smartphone is independent of gender.

- To answer this question, a group of 100 college going children were surveyed about whether they owned a smartphone or not.

- The categorical variables in this example are:
  - → Gender : Male, Female (Nominal)
  - → Own a smartphone : Yes, No (Nominal)

Summarize Data :

→ We have the following summary statistics

(1) There are 44 female & 56 male students
(2) 76 students owned a smartphone, 24 did not own.
(3) 34 female students owned a smartphone, 42 male students owned a smartphone

→ The data given in the example can be organised using a two way table, referred to as contingency table.

	Own a smartphone		
Gender	No	Yes	Row Total
Female	10	34	44
Male	14	42	56
Col. Total	24	76	100

EXAMPLE 2 (Nominal & Ordinal Variable)

Income Vs Use of Smartphone

→ The categorical variables in this example are
- Income : Low, Medium, High (Ordinal)
- Own a smartphone : Yes, No (Nominal)

Contingency Table (summarizing data)

→ We have the foll. summary statistics :

(1) There are 20 high income, 66 medium income and 14 low income participants.

(2) 62 participants owned a smartphone, 38 did not own.

(3) 18 high income participants, 39 medium incomes participants and 5 low income participants owned a smartphone.

The contingency table corresponding to the data is given below:

Income Level	Own a smartphone		
	NO	YES	Row Total
HIGH	2	18	20
MEDIUM	27	39	66
LOW	9	5	14
Column Total	38	62	100

SECTION SUMMARY

- Organize bivariate categorical data into a two-way table : contingency table
- If data is ordinal, maintain order of the variable in the table.

June 3, 2021

RELATIVE FREQUENCY

ROW RELATIVE FREQUENCIES

→ What proportion of total participants own a smartphone ?

→ What proportion of female participants own a smartphone ?

Gender	Own a smartphone		
	NO	YES	Row Total
FEMALE	10	34	44
MALE	14	42	56
Column Total	24	76	100

* Row Relative Frequency : Divide each cell frequency in a row by its row total

Example 1 : GENDER Vs OWN A SMARTPHONE

Gender	Own a smartphone		
	NO	YES	Row Total
FEMALE	10/44	34/44	44
MALE	14/56	42/56	56
Column Total	24/100	76/100	100

Gender	Own a Smartphone		
	NO	YES	Row Total
FEMALE	22.73%	77.27%	44
MALE	25.00%	75.00%	56
Column Total	24.00%	76.00%	100

COLUMN RELATIVE FREQUENCIES

→ What proportion of total participants are females?

→ What proportion of smartphone owners are females?

* Column Relative Frequency : Divide each cell frequency in a column by its column total.

EXAMPLE : GENDER V/S OWN A SMARTPHONE

GENDER	OWN A SMARTPHONE		
	NO	YES	Row Total
FEMALE	10/24 = 41.67%	34/76 = 44.74%	44/100 = 44%
MALE	14/24 = 58.33%	42/76 = 55.26%	56/100 = 56%
Column Total	24	76	100

ASSOCIATION BETWEEN TWO VARIABLES

- What do we mean by stating two variables are associated?
  Knowing information about one variable provides information about the other variable.

→ To determine if two categorical variables are associated, we use the notion of relative row frequencies and relative column frequencies.

(1)→ If the row relative frequencies (the column relative frequencies) are the <u>same</u> for all rows (columns), then we say that two variables are <u>not associated</u> with each other.

(2)→ If the row relative frequencies (the column relative frequencies) are different for some rows (columns) then we say that the two variable are <u>associated</u> with each other.

EXAMPLE 1

Gender Vs Smartphone Ownership

Gender	Own a smartphone		
	NO	YES	Row Total
FEMALE	22.73%	77.27%	44
MALE	25.00%	75.00%	56
Column Total	24.00%	76.00%	100

' Row Frequency '

→ Now there we can see that there is not much difference in the relative frequencies of row 1, 2, 3.
So in accordance to point (1), the two variables are not associated.

i.e, Gender & Smartphone Ownership are not associated.

EXAMPLE 2

Income Vs Smartphone Ownership

INCOME	OWNERSHIP		Row Total
(Row Percentage)	No	Yes	
High	10%	90%	20
Medium	40.91%	59.09%	66
Low	64.29%	35.71%	14
Column Total	38.00%	62.00%	100

→ Now, there we can see that the relative frequencies of all the three or four rows are very much different from each other. Hence, from point (2), we know that the two variables are associated.

i.e, Income & Smartphone Ownership are associated.

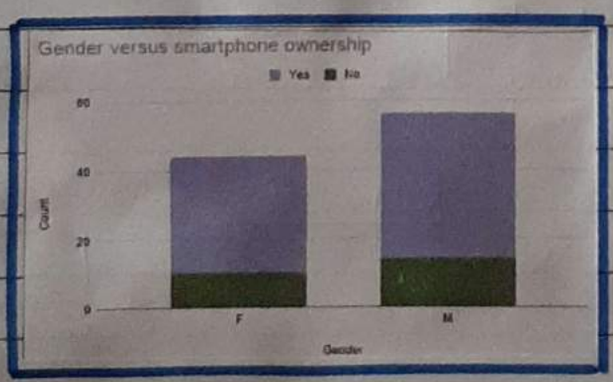
STACKED BAR CHART

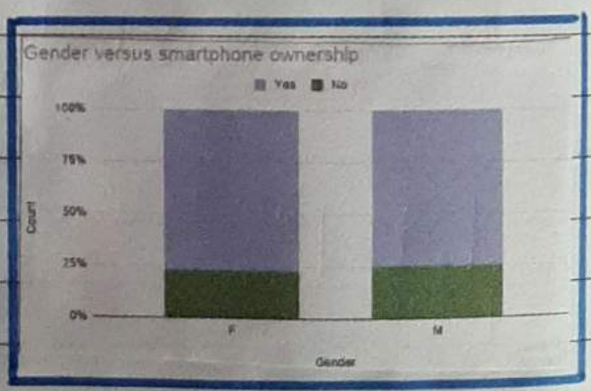
- Recall, a bar chart summarized the data for a categorical variable. It presented a graphical summary of the categorical variable under consideration with the length of the bars representing the frequency of occurence of a particular category.

- A STACKED BAR CHART represents the counts for a particular category. In addition, each bar is further broken down into smaller segments, with each segment representing the frequency of that particular category within the segment. A stacked bar chart is also referred to as a segmented bar chart.

EXAMPLE 1

	Own a smartphone		
Gender	No	Yes	Row total
Female	22.73%	77.27%	44
Male	25.00%	75.00%	56
Column total	24.00%	76.00%	100

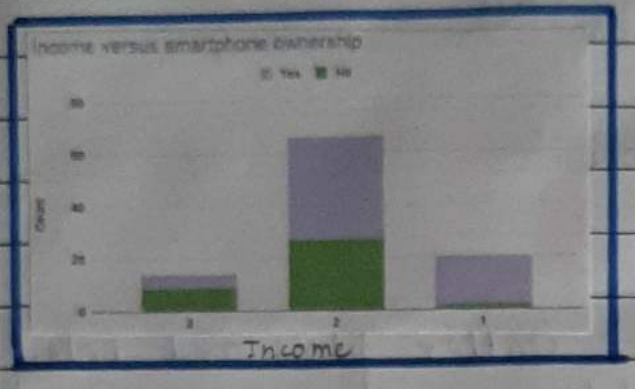


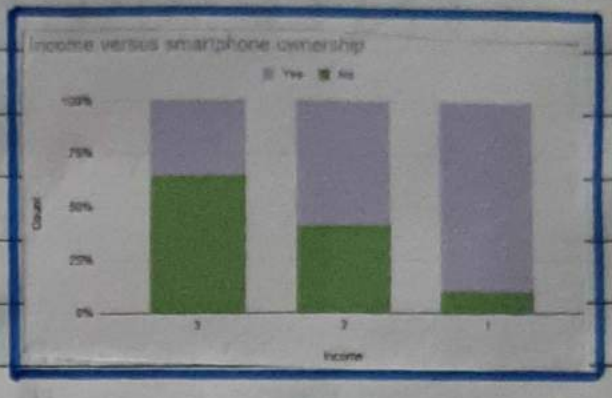


A 100% stacked bar chart is useful to part-to-whole relationships

EXAMPLE 2

	Own a smartphone		
Income level	No	Yes	Row Total
High	10.00%	90.00%	20
Medium	40.91%	59.09%	66
Low	64.29%	35.71%	14
Column Total	38.00%	62.00%	100





Frequency, Stacked Bar Chart

100% stacked bar Chart

Date
June 4, 2021

classmate
Date
Page

Association Between Two Numerical Var.

INTRODUCTION

- To understand the association between two numerical variables.
- Learn how to construct scatter plots and interpret association in scatter plots
- Summarize association with a line
- Correlation matrix

SCATTER PLOT

- We use a scatterplot to look for association between numerical variables

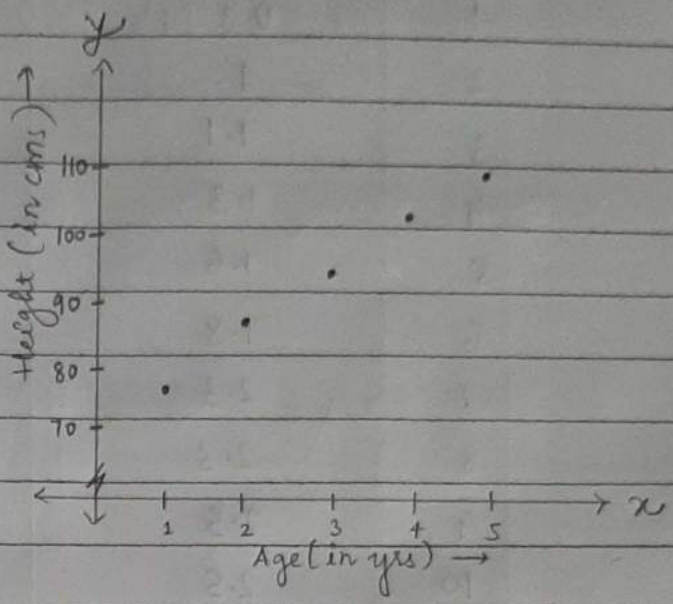
Definition: A scatter plot is a graph that displays pairs of values as points on a two-dimensional plane.

→ To decide which variable to put on the x-axis and which to put on y-axis, display the variable you would like to explain along the y-axis (referred as

response variable) and the variable which explains on x-axis (referred as explanatory variable).

EXAMPLE 1

AGE (in yrs)	HEIGHT (cms)
1	75
2	85
3	94
4	101
5	108



EXAMPLE 2

A real estate agent collected the prices of different size of homes. He wanted to see what was the relationship between the price of a home and size of home. In particular, he wanted to know if the prices of homes increased linearly with the size or in any other way?

To answer this question, he collected data on 15 homes. The data he recorded was

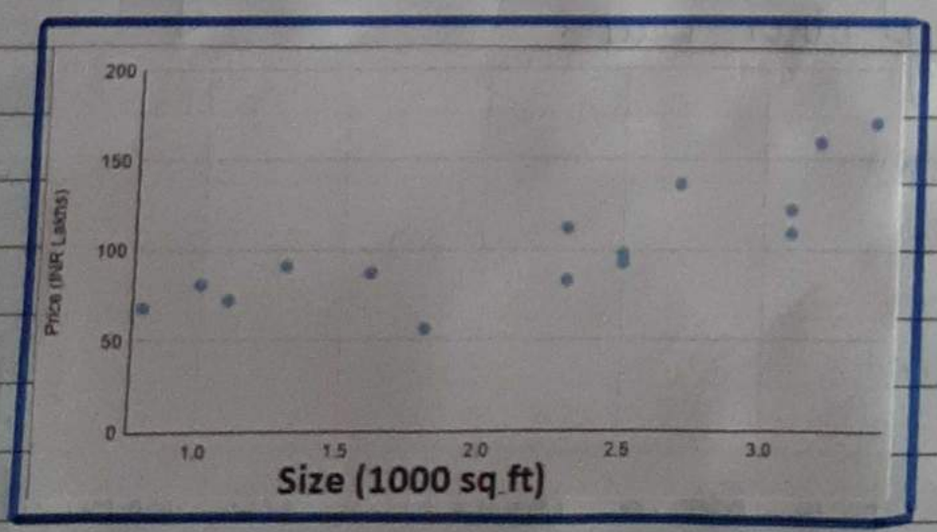
(1) Size of a home measured in 1000 of square feet

(2) Price of a home measured in lakh of rupees.

HOUSING DATA

S.no	Size (1000 sq ft)	Price (INR Lakhs)
1	0.8	68
2	1	81
3	1.1	72
4	1.3	91
5	1.6	87
6	1.8	56
7	2.3	83
8	2.3	112
9	2.5	93
10	2.5	98
11	2.7	136
12	3.1	109
13	3.1	122
14	3.2	159
15	3.4	170

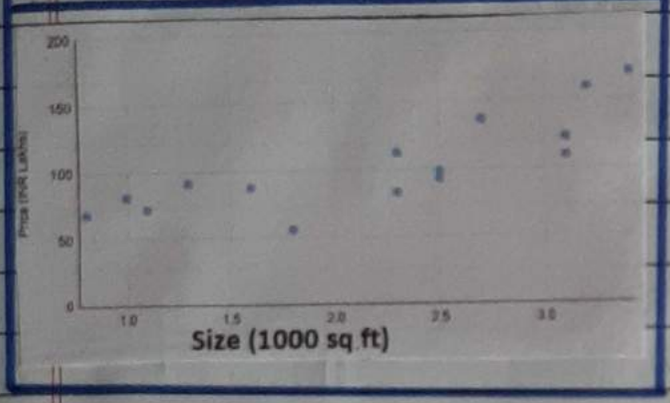
SCATTER PLOT



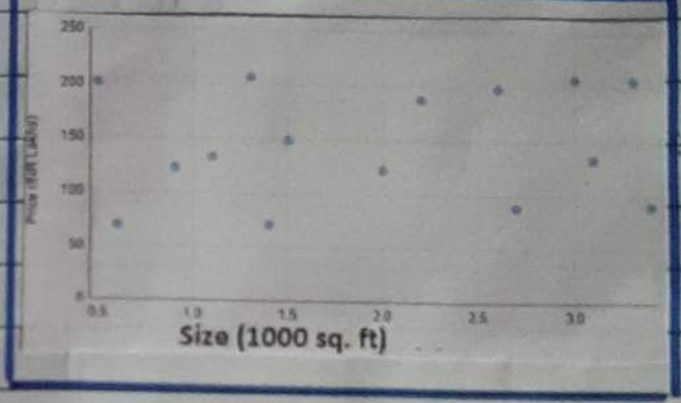
VISUAL TEST FOR ASSOCIATION

- Do we see a pattern in the scatter plot?
- In other words, if I know about the x-value, can I use it to say something about the y-value or guess y-value?

DATASET 1 (Scatter Plot)



DATASET 2 (Scatter Plot)



→ Here we can see that, Dataset 1 follows some kind of pattern but there is no pattern being followed in Dataset 2.

DESCRIBING ASSOCIATION

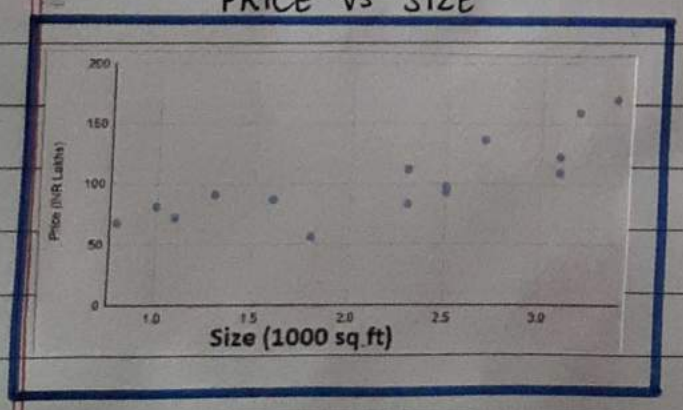
When describing association between variables in a scatter plot, there are four key questions that need to be answered.

1. DIRECTION : Does the pattern trend up, down, or both?
2. CURVATURE : Does the pattern appear to be linear or does it curve?
3. VARIATION : Are the points tightly clustered along the pattern?
4. OUTLIERS : Did you find something unexpected?

1. DIRECTION

Does the pattern trend up, down or both?

PRICE VS SIZE



UP

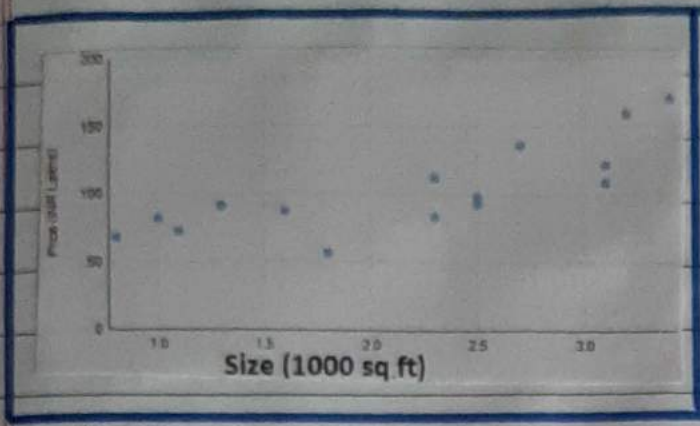
PRICE VS AGE OF CAR

DOWN

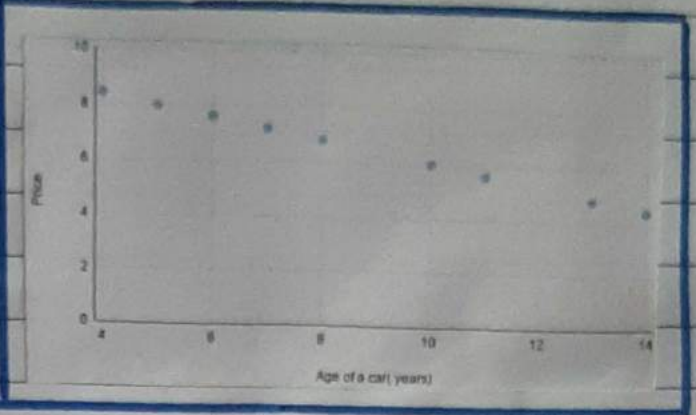
2. CURVATURE

Does the pattern appear to be linear or does it curve?

LINEAR

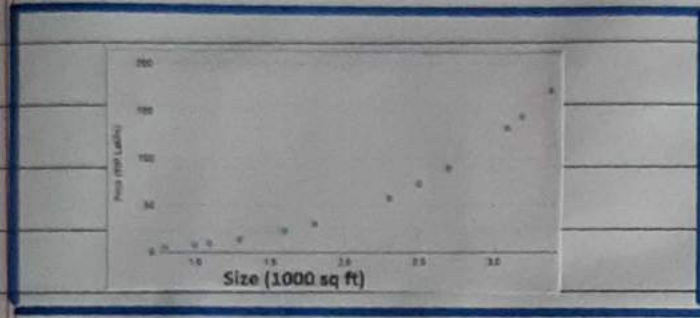


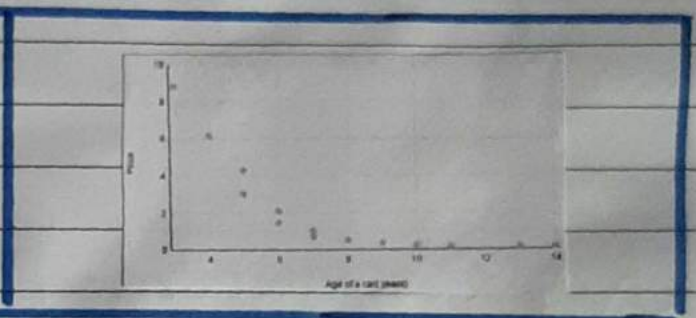
PRICE VS SIZE



PRICE VS AGE OF CAR

CURVE

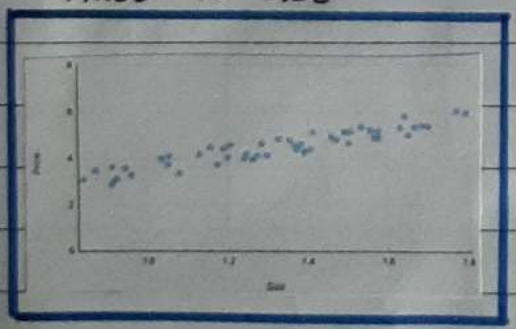




3. VARIATION

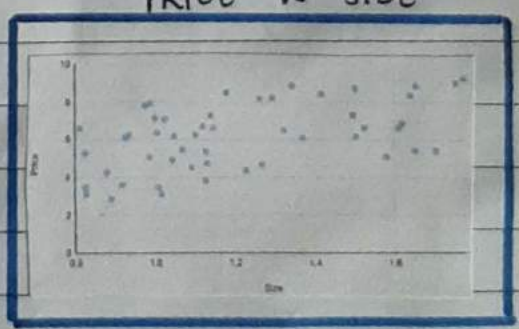
Are the points tightly clustered along the pattern?

PRICE VS SIZE



TIGHTLY CLUSTERED

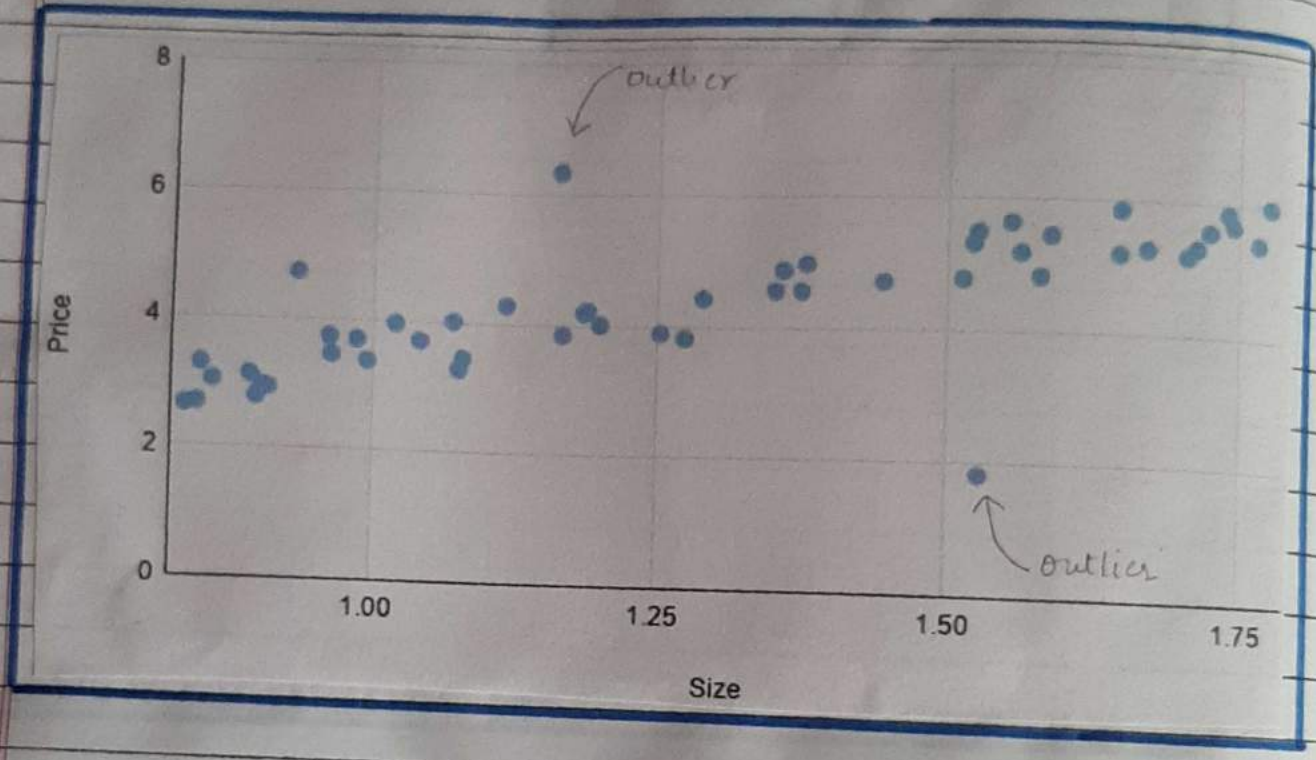
PRICE VS SIZE



VARIABLE

4. OUTLIERS

Did you find something unexpected ?



June 5, 2021

MEASURES OF ASSOCIATION

How do we measure the strength of association between two variables?

1. Covariance
2. Correlation

1. COVARIANCE

Covariance quantifies the strength of the linear association between two numerical variables.

EXAMPLE 1

Recall, the association between age and height of a person.

Age (in yrs) x	Height (in cms) y	Deviation of x $(x_i - \bar{x})$	Dev. of y $(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	75	-2	-17.6	35.2
2	85	-1	-7.6	7.6
3	94	0	1.4	0
4	101	1	8.4	8.4
5	108	2	15.4	30.8

$\bar{x} = 3$ $\quad \bar{y} = 92.6$

~~Pop variance, $\sigma^2 = \frac{82}{5} = 16.4$~~

~~Sam. variance, $s^2 = \frac{82}{4} = 20.5$~~

EXAMPLE 2 Variables : Age of a car & price of a car

Age (in yrs) x	Price (INR lakhs) Y	Dev. of x $(x_i - \bar{x})$	Dev of y $(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	6	-2	2	-4
2	5	-1	1	-1
3	4	0	0	0
4	3	1	-1	-1
5	2	2	-2	-4
$\bar{x} = 3$	$\bar{y} = 4$	~~$\frac{10}{5}$~~	~~$\frac{10}{4}$~~	~~2.5~~

Key Observation :

- When large (small) values of x tend to be associated with large (small) values of y - the signs of the deviations, $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will also be tend to be same.

- When large (small) values of x tend to be associated with small (large) values of y - the signs of deviations, $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will also tend to be different.

X	Y	Signs of Dev.
Large	Large	Same
Small	Small	Same
Large	Small	Diff.
Small	Large	Diff.

Definition: Let x_i denote the i^{th} observation of variable x, and y_i denote the i^{th} observation of variable y. Let (x_i, y_i) be the i^{th} paired observation of a population (sample) dataset having $N(n)$ observations. The Covariance between the variables x and y is given by

→ Population Covariance : $$Cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{N}$$

→ Sample Covariance : $$Cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Units of Covariance:

→ The size of covariance, however, is difficult to interpret because the covariance has units.

→ The units of the covariance are those of x-variable times those of the y-variable.

Example 1: Population variance $= \frac{-17.6 + (-7.6) + 1.4 + 8.4 + 15.4}{5}$

$= 82 \div 5 = 16.4$

Sample variance $= \frac{82}{4} = 20.5$

Example 2: Population variance $= \frac{-4 + (-1) + 0 + (-1) + (-4)}{5} = \frac{-10}{5}$

$= -2$

Sample Variance $= \frac{-10}{4} = -2.5$

2. CORRELATION

- A more easily interpretted measure of linear association between two numerical variables is correlation.

- It is derived from covariance.

- To find the correlation between two numerical variables x and y divide the covariance between x and y by the product of the std. deviations of x and y. The pearson correlation coefficient, r, between x & y is given by

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{cov(x,y)}{s_x s_y}$$

NOTE :
- The units of the standard deviations cancel out the units of covariance.
- It can be shown that the correlation measure always lies between -1 and $+1$.

EXAMPLE 1

Age x	Height Y	sq. Devn of x $(x_i - \bar{x})^2$	sq. Devn of y $(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	75	4	309.76	35.2
2	85	1	57.76	7.6
3	94	0	1.96	0
4	101	1	70.56	8.4
5	108	4	237.16	30.8
		10	677.2	82

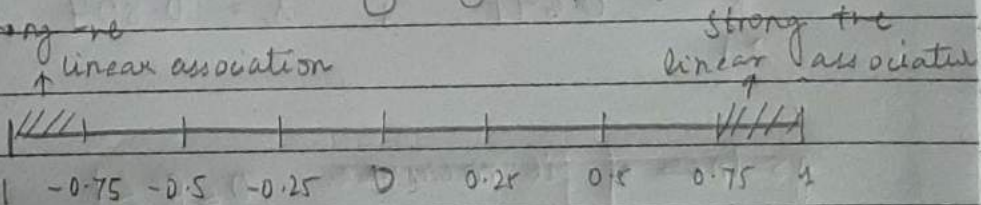
→ $S_x = 1.58$, $S_y = 13.01$

→ $r = \frac{82}{\sqrt{10 \times 677.2}}$ or $\frac{20.5}{1.58 \times 13.01} = 0.9964$

Covariance → if $cov(x,y) \rightarrow +ve$
It means both are going up

if $cov(x,y) \rightarrow -ve$
one is going up & other down.

Correlation :



FITTING A LINE

Learning Objectives

→ Summarize the linear association between two variables using the equation of a line.

→ Understand the significance of R^2

SUMMARIZING THE ASSOCIATION WITH A LINE

- The strength of linear association between the variables was measured using the measures of covariance & correlation.
- The linear association can be described using the equation of a line.

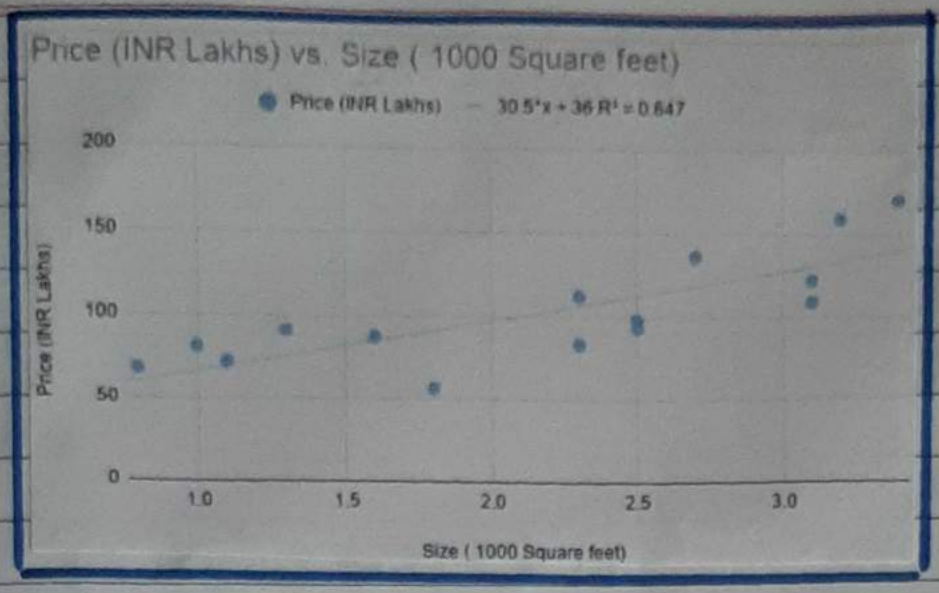
EXAMPLE 1 SIZE Versus PRICE OF HOMES : Equation

$y = mx + c$

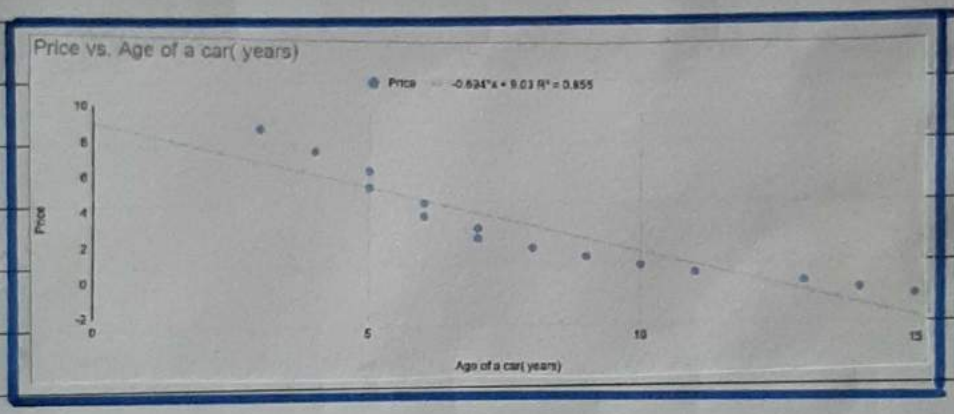
Equation of the line : Price = 30.5 X Size + 36

$R^2 = 0.647$; $r = 0.804$

goodness of fit measure $0 \leq R^2 \leq 1$



EXAMPLE 2 AGE VS PRICE OF CARS : Equation



Equation of Line : Price = $-0.694 \times$ Age $+ 9.03$

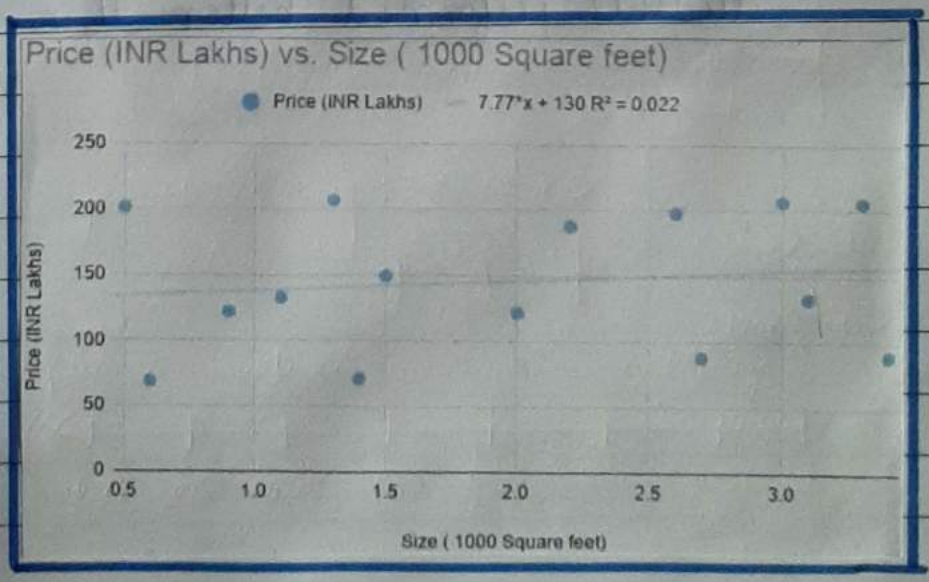
$R^2 = 0.855$; $r = -0.9247$

EXAMPLE 3 :

Eqn of Line :

Price = $7.77 \times$ Size $+ 130$

$R^2 = 0.022$; $r = 0.149$



June 6, 2021

Association Between Categorical & Numerical Variables

INTRODUCTION

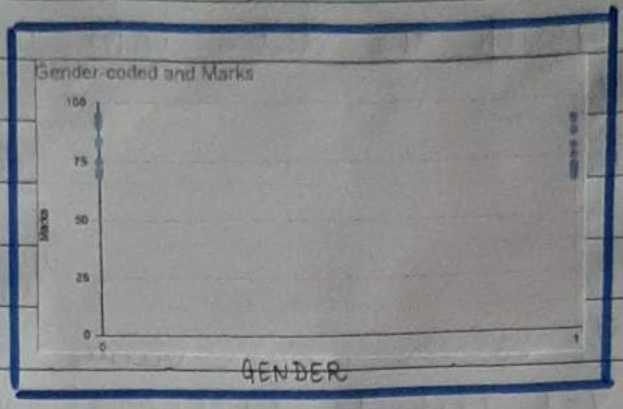
- Understand the association between a categorical variable and numerical variable.
- Assume the categorical variable has two categories (dichotomous)

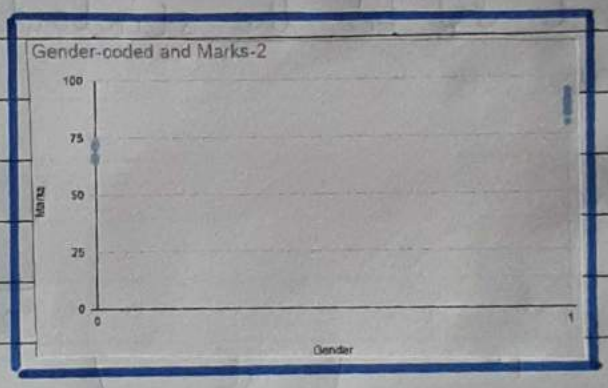
EXAMPLE 1 : GENDER VERSUS MARKS

A teacher was interested in knowing if female students performed better than male students in her class. She collected data from 20 students and the marks they obtained on 100 on the subject.

Sno.	Gender	Marks
1	F	71
2	F	67
3	F	65
4	M	69
5	M	75
6	M	83
7	F	91
8	F	85
9	F	69
10	F	75
11	M	92
12	F	79
13	M	71
14	M	94
15	F	86
16	F	75
17	F	90
18	M	84
19	F	91
20	M	90

SCATTER PLOT





Another Dataset

POINT BI-SERIAL CORRELATION COEFFICIENT

→ Let X be a numerical variable and Y be a categorical variable with 2 categories (a dichotomous var.)

→ The following steps are used for calculating the

Point Bi-Serial Correlation between these two variables:

STEP 1: Group the data into two sets based on the value of the dichotomous variable Y. That is, assume that the value of Y is either 0 or 1.

STEP 2: Calculate the mean values of two groups : Let $\bar{Y}_0$ and $\bar{Y}_1$ be the mean values of groups with $Y=0$ and $Y=1$ respectively.

STEP 3: Let p_0 and p_1 be the proportion of observations in a group with $Y=0$ and $Y=1$, respectively, and S_x be the standard deviation of the random variable X.

The correlation coefficient

$$r_{pb} = \left(\frac{\bar{Y}_0 - \bar{Y}_1}{S_x}\right)\sqrt{p_0 p_1}$$

$p_0 = \frac{\text{no. of obs. in 0}}{\text{total obs.}}$

S_x ← stand. dev

Week 5 & 6 Stats 1

(Arrange) (Select)

Permutations & Combinations

FUNDAMENTAL PRINCIPLE OF COUNTING

(1) **Addition principle**

We have x different events

No. of ways to do 1st event be n_1, 2nd event be n_2 and x event be n_x.

∴ Total no. of occurrence of x events = $n_1 + n_2 + \dots + n_x$

(2) **Multiplication Principle**

Suppose that x actions are to be performed in a definite order.

There are n_1 possibilities for first event, n_2 for second and so on and n_x for x^{th} event.

∴ No. of possibilities for x actions altogether = $n_1 \times n_2 \times \dots n_x$

FACTORIAL

→ $n! = n \times (n-1) \times (n-2) \times \dots \times 3 \times 2 \times 1$

→ $n! = n\,(n-1)!$

→ $0! = 1 \quad ; \quad 1! = 1$

→ Factorial of negative no. is not defined.

→ $(2n)! = 2^n \cdot n! \cdot (1 \cdot 3 \cdot 5 \cdot \dots (2n-1))$

PERMUTATIONS

- arrangement in definite order
- ORDER is important !

Theorem 1 : No. of permutations of 'n' distinct things taken all at a time is :

$$^{n}P_{n} = P(n, n) = A_{n}^{n} = n!$$

Theorem 2 : No. of permutations taken 'r' at a time where $0 \leq r \leq n$ is :

$$^{n}P_{r} = P(n, r) = A_{r}^{n} = \frac{n!}{(n-r)!}$$

Special Cases

- $^{n}P_{0} = \frac{n!}{(n-0)!} = \frac{n!}{n!} = 1$ { only 1 ordered arrangement of zero objects }
- $^{n}P_{1} = \frac{n!}{(n-1)!} = n$ { There are n ways of choosing one object from n objects }
- $^{n}P_{n} = \frac{n!}{(n-n)!} = n!$ { We can arrange n distinct objects in n! ways }

★ When Repetition Is Allowed

$n^r = n \times n \times n \times \ldots \times n$

The no. of possible permutations of r objects from a collection of n distinct objects when repetition is allowed is n^r.

★ When Objects Are 'Not Distinct'

No. of permutations of n objects where p_1 is one kind, p_2 is of second kind, and so on p_k is of k^{th} kind is given by

$$\frac{n!}{p_1!\, p_2! \ldots p_k!}$$

★ Circular Permutations

(1) If clockwise & anticlockwise are different, No. of permutations for n different objects to be arranged in circle $= (n-1)!$

(2) If clockwise & anticlockwise are same, No. of permutations for n different objects to be arranged in circle $= \frac{(n-1)!}{2}$

COMBINATIONS

→ selection of objects

→ ORDER does not matter

Theorem 1: No. of combinations / selections of n distinct things taken r at a time:

$$^nC_r = C(n,r) = \binom{n}{r} = \frac{n!}{r!\,(n-r)!}$$

Theorem 2: No. of combinations of n different things taken 'r' at a time when 'p' particular things are always included:

$$^{n-p}C_{r-p} = \frac{(n-p)!}{(r-p)!\,[(n-p)-(r-p)]!}$$

$$\Rightarrow {}^{n-p}C_{r-p} = \frac{(n-p)!}{(r-p)!\,(n-r)!}$$

Theorem 3: No. of combinations of 'n' different things taken 'r' at a time when 'p' particular things are always excluded.

$$^{n-p}C_r$$

* Important Results

→ ${}^nC_0 = 1$

→ ${}^nC_n = 1$

→ ${}^nC_{n-r} = {}^nC_r$

→ If ${}^nC_x = {}^nC_y$, either $x = y$ or $x + y = n$

→ ${}^nP_r = {}^nC_r \times r!$ { permutation is defined total no. of combination of object then arrangement of objects }

→ ${}^nC_r + {}^nC_{r+1} = {}^{n+1}C_{r+1}$

or ${}^nC_r + {}^nC_{r-1} = {}^{n+1}C_r \quad (0 \le r \le n)$

Date
July 28, 2021

classmate
Date
Page

WEEK 7

PROBABILITY

→ As a general rule, to be able to draw valid inferences about a population from a sample, one needs to know how likely it is that certain events will occur under various circumstances.

→ The determination of the likelihood, or chance, that an event will occur is the subject matter of probability.

RANDOM EXPERIMENT

Definition : EXPERIMENT
An experiment is any process that produces an observation or outcome.

Definition : RANDOM EXPERIMENT
A random experiment is an experiment whose outcome is not predictable with certainty.

REMARK : However, although the outcome of the experiment will not be known in advance, let us suppose that the set of all possible outcomes is known.

EXAMPLES OF RANDOM EXPERIMENT

(1) Experiment : Guessing answers to a four option multiple choice question.
Outcome : A, B, C, D

(2) Experiment : Order of finish in a race with six students A, B, C, D, E, F
Outcome : all possible permutations of A, B, C, D, E & F

(3) Experiment : Tossing two coins
Outcome : HH, HT, TH, TT

(4) Experiment : Measuring of lifetime (in hrs) of bulb
Outcome : 0, or 1 hr, or 2 hr or so on.

(5) Experiment : To throw a dart on a unit square & note the point where it lands.
Outcome : Any point in the square.

SAMPLE SPACE

Definition : A sample space (denoted by Ω or S) : collection of all basic outcomes

BASIC OUTCOMES → the possible outcomes that can occur must be :

(1) mutually exclusive : only 1 basic outcome can occur
(2) exhaustive : one basic outcome must occur

EXAMPLES OF SAMPLE SPACES

(1) Experiment : Guessing answers to a 4 option MCQ
Sample space, $S = \{A, B, C, D\}$

(2) Exp. : Order of finish in a race with 6 students A, B, C, D, E, F
$S = \{ABCDEF, ABCDFE, \ldots\ldots, EFDBAC\}$
$n(s) = 6!$

(3) Exp. : Tossing two coins
$S = \{HH, HT, TH, TT\}$

(4) Exp. : Measuring the lifetime (in hrs) of a bulb
$S = \{x : 0 \leq x < \infty\}$

(5) Exp. : To throw a dart on a unit square & note where it lands
$S = \{(x, y) : 0 \leq x \leq 1, \ 0 \leq y \leq 1\}$

EVENTS

An event E is a collection of basic outcomes.

→ That is, an event is a subset of the sample space.

→ We say an event has occurred if the outcome is contained in the subset.

EXAMPLE OF EVENTS

(1) Exp : Guessing answers to 4 option MCQ
Event : answer is A
E = {A}

(2) Exp. : Tossing two coins
Event : head on first toss
E = {HH, HT}

UNION OF EVENTS

→ For any two events E and F, we define the new event E ∪ F, called the union of events E and F, to consist of all outcomes that are in E or in F or in both E and F.

→ i.e, the event E ∪ F will occur if either E or F occurs.

EXAMPLES OF UNION EVENTS

(1) Exp. : Guessing answers to a 4 option MCQ
Event : → answer is A ; E_1 = {A}
→ answer is B ; E_2 = {B}
→ answer is A or B ; $E_3 = E_1 \cup E_2$ = {A, B}

(2) Exp. : Order of finish in a race with 6 students - A, B, C, D, E, F.
Event : → A finishes first ; E_1 = {ABCDEF, ABCDFE ..... }
→ B finishes second; E_2 = {ABCDEF, ABDCEF ... }
→ A comes first or B comes second ;
$E_1 \cup E_2$ = {ABCDEF, ABCDFE, ... AFEDBC, CBDEFA}

INTERSECTION OF EVENTS

- For any 2 events E & F, we define the new event $E \cap F$ called the intersection of events E and F, to consist of all outcomes that are in E and in F.
- That is, the event $E \cap F$ will occur if both E & F occurs.

EXAMPLE

Experiment : Tossing 2 coins & noting the outcomes
Event : → head on first toss $E_1 = \{HH, HT\}$
→ head on second toss $E_2 = \{HH, TH\}$
→ head on first & second toss ; $E_3 = E_1 \cap E_2 = \{HH\}$

NULL EVENT & DISJOINT EVENT

NULL EVENT : We call the event without any outcomes the null event, and designate it as ϕ.

DISJOINT EVENT : If the intersection of E and F is the null event, then since E and F can't simultaneously occur, we say that E & F are disjoint or mutually exclusive.

EXAMPLE :

Experiment : Guessing answers to a four option multiple choice question.

Event : → answer is A ; $E_1 = \{A\}$
→ answer is B ; $E_2 = \{B\}$
→ answer is A & B ; $E_3 = E_1 \cap E_2 = \phi$

We say events E_1 & E_2 are mutually exclusive or disjoint. Occurence of E_1 disallows occurence of E_2. In another words, if my A(B) is my guess, then B(A) can't be my guess.

COMPLEMENT OF AN EVENT

The complement of E, denoted by E^c, consists of all outcomes in the sample spake S that are not in E.

→ That is, E^c will occur if & only if E does not occur.

→ The complement of the sample space is the null set, i-e! $S^c = \phi$.

EXAMPLE :

Experiment : Toss a coin once & note the outcomes
Sample Space : $S = \{H, T\}$
$E_1 = \{H\}$ (outcome is head)
$E_2 = \{T\}$ (outcome is tail)

Event 2 is complement of E_1. In other words $E_2 = E_1^c$

SUBSETS

For any 2 events E & F, if all the outcomes in E are also in F, then we say that E is contained in F, or E is a subset of F, and denote it as $E \subset F$.

Example :

Experiment : Tossing 2 coins & noting the outcomes
sample space : $S = \{HH, HT, TH, TT\}$

Event : head on first toss, $F = \{HH, HT\}$
Event : head on both tosses, $E = \{HH\}$

Here, $E \subset F$.

SECTION SUMMARY

What we learnt till now :

- Random Experiments
- Sample Space
- Events :
  - Union, Intersection, Complement
- Null & Disjoint Events
- Subsets

VENN DIAGRAMS

A graphical representation that is useful for illustrating logical relations among events is the VENN DIAGRAM.

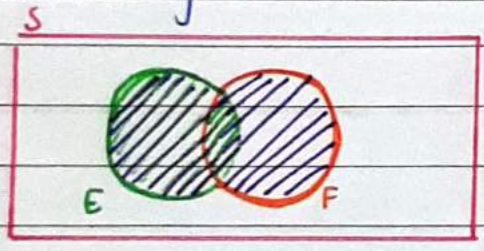
Representation Of A Sample Space

→ sample space consists of all possible outcomes and is represented by a large rectangle

Representation of An Event

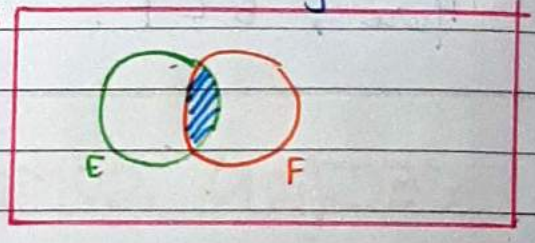
→ The events E, F, G ... are represented in circles within the rectangle

Union Of An Event



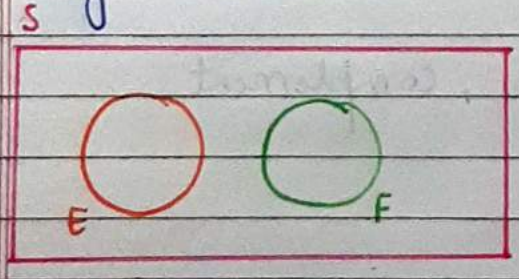
→ region shaded in purple is E ∪ F

Intersection of Events



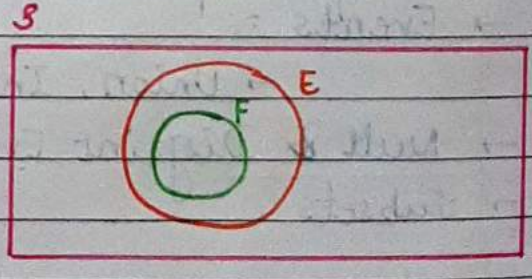
→ region shaded in blue is E ∩ F

Disjoint Events



→ E ∩ F = ϕ
Hence, disjoint events

Subsets



→ F is present in E
⇒ F ⊂ E

Properties of Probability

THE THREE MAIN INTERPRETATIONS OF PROBABILITY

1. Classical (Apriori or theoretical)
Let S be the sample space of a random experiment in which there are n equally likely outcomes, and the event E consists of exactly m of these outcomes, then we say the probability of the event E is m/n and represent it as $P(E) = \frac{m}{n}$.

2. Relative Frequency (Aposteriori or empirical)
The probability of an event in an experiment is the proportion (or fraction) of times the event occurs in a very long (theoretically infinite) series of (independent) repetitions of experiment. In other words, if $n(E)$ is the no. of times E occurs in n repetitions of the experiment, $P(E) = \lim_{n \to \infty} \frac{n(E)}{n}$

3. Subjective
The probability of an event is a 'best guess' by a person making the statement of the chances that the event will happen. The probability measures an individual's degree of belief in the event.

PROBABILITY AXIOMS

Consider an experiment whose sample space is S. We suppose that for each event E there is a number, denoted P(E) and called the probability of event E, that is in accord with the following three axioms :

(1) For any event E, the probability of E is a number between 0 and 1, i.e,

$$0 \leq P(E) \leq 1$$

(2) The probability of sample space S is 1. Symbolically, $P(S) = 1$. In other words, the outcome of the experiment will be an element of sample space S with probability 1.

(3) For a sequence of mutually exclusive (disjoint) events, $E_1, E_2, \ldots$

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

In simple words, the probability of the union of disjoint events is equal to the sum of the probabilities of these events.
For instance, if E_1 & E_2 are disjoint then

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

GENERAL PROPERTIES OF PROBABILITY

1. Probability of complement of an event

$$P(E^c) = 1 - P(E)$$ [E and E^c are disjoint events]

2. $P(\text{null event}) = P(\phi) = 0$ [$S^c = \phi$]

→ $P(S) = 1$

→ $S^c = \phi$

→ $P(S^c) = 1 - P(S) = 1 - 1 = 0$

→ $P(\phi) = 0$

because S and $S^c = \phi$ are disjoint events

ADDITION RULE OF PROBABILITY

We have two events E_1 and E_2 which are not mutually disjoint.

∴ $$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

Question: A customer that goes to the clothing store will purchase a shirt with probability 0.3. The customer will purchase a pant with probability 0.2 and will purchase both a shirt and a pant with probability 0.1. What proportion of customers purchases neither a shirt nor a pant?

Solution : E_1 = purchase a shirt
E_2 = purchase a pant
$E_1 \cap E_2$ = purchase both shirt & pant
$(E_1 \cup E_2)^c$ = neither purchase a shirt nor a pant

$\therefore P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$
($P(E_1)$: buy shirt, $P(E_2)$: buy pant, $P(E_1 \cap E_2)$: buy both shirt & pant)

$\Rightarrow P(E_1 \cup E_2) = 0.3 + 0.2 - 0.1$
(purchasing either shirt or a pant)

$\Rightarrow P(E_1 \cup E_2) = 0.4$

Thus, $P(E_1 \cup E_2)^c = 1 - P(E_1 \cup E_2)$
$= 1 - 0.4$
$= 0.6$ Ans.

Question : A student has a 40% chance of receiving an A Grade in statistics, 60% chance of receiving an A in mathematics, and an 86% chance of receiving an A in either statistics or maths. Find the probability that she

(a) does not receive an A grade in either stats or maths
(b) receives A's in both stats & maths

Solution : E_1 = receiving A grade in stats ; $P(E_1) = 0.4$
E_2 = receiving A grade in maths ; $P(E_2) = 0.6$
$E_1 \cup E_2$ = A in maths or in stats ; $P(E_1 \cup E_2) = 0.86$

(a) E_3 = does not receive A grade in maths or stats
$\Rightarrow E_3 = (E_1 \cup E_2)^c$

$\therefore P(E_3) = 1 - P(E_1 \cup E_2)$

$= 1 - 0.86$

$= 0.14$ Ans.

(b) E_4 = receive A in both stats & maths

$E_4 = E_1 \cap E_2$

$\therefore P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$

$\Rightarrow 0.86 = 0.4 + 0.6 - P(E_4)$

$\Rightarrow P(E_4) = 1 - 0.86$

$= 0.14$ Ans.

EQUALLY LIKELY OUTCOMES

→ For certain experiments it is natural to assume that each outcome in the sample space S is equally likely to occur.

→ That is, if sample space S consists of N outcomes, say, $S = \{1, 2, 3, \ldots, N\}$, then it is often reasonable to suppose that:

$$P(\{1\}) = P(\{2\}) = \ldots = P(\{N\})$$

→ In this expression, $P(\{i\})$ is the probability of the event consisting of the single outcome i.

→ Using the properties of probability, we can show that the foregoing implies that the probability of any event A is equal to the proportion of the

outcomes in the sample space that is in A.

→ i.e, $P(A) = \frac{\text{number of outcomes in } S \text{ that are in } A}{N}$

EXAMPLE

1. **Rolling a fair dice**

Sample space : $S = \{1, 2, 3, 4, 5, 6\}$

Let E_i denote the event of outcome i. Since the dice is fair, $P(E_i) = \frac{1}{6}$

Define A to be the event : outcome is odd
i.e, $A = \{1, 3, 5\}$

$$P(A) = P(E_1) + P(E_3) + P(E_5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$$

Let B be the event : outcome > 4, i.e, $B = \{5, 6\}$

$$P(B) = P(E_5) + P(E_6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Let C be the event : outcome is either odd or greater than 4.
$C = \{1, 3, 5, 6\}$

$$P(C) = P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{3}{6} + \frac{2}{6} - \frac{1}{6} = \frac{4}{6}$$

Date
July 29, 2021

classmate
Date
Page

WEEK 8
CONDITIONAL PROBABILITY

LEARNING OBJECTIVES THIS WEEK

1. Understand the notion of conditional probability, i.e, find the probability of an event given another event has occurred.
2. Distinguish between independent & dependent events
3. Solve applications of probability

Contingency Tables

FROM TABLES TO PROBABILITY

- Recall the cell phone usage vs gender example when the association b/w categorical variables & the concept of relative frequencies was discussed.
- Percentages computed within rows or columns of a

contingency table correspond to CONDITIONAL probabilities

- Convert contingency tables into probabilities, we use the counts to define probabilities.

RELATIVE FREQUENCY

	OWN A SMARTPHONE		
GENDER	NO	YES	ROW TOTAL
Female	10	34	44
Male	14	42	56
COL. TOTAL	24	76	100

To get the relative frequency, divide each count by 100.

	OWN A SMARTPHONE		
GENDER	No	Yes	ROW TOTAL
Female	$10/100 = 0.10$	$34/100 = 0.34$	$44/100 = 0.44$
Male	$14/100 = 0.14$	$42/100 = 0.42$	$56/100 = 0.56$
COL. TOTAL	$24/100 = 0.24$	$76/100 = 0.76$	100

(1) Joint Probabilities

→ Displayed in cells of contingency table

→ Represent the probability of an intersection of 2 or more events.

- In the example : there are four joint probabilities, e.g.,
  - → P(Female and not owning smartphone) = 0.10
  - → P(Male and owning a smartphone) = 0.42

(2) Marginal Probabilities

- Displayed in the margins of a contingency table
- It is the probability of observing an outcome with a single attribute, regardless of its other attributes.
- In the example; There are 4 marginal probabilities, e.g.,
  - → P(Female) = 0.10 + 0.34 = 0.44
  - → P(owning a smartphone) = 0.34 + 0.42 = 0.76

Conditional Probability

Find conditional probabilities to answer questions like :	Recognize the answers
→ among female buyers, what is the chance someone owns phone?	→ Row relative frequency
→ among people who don't own a phone, how many are male	→ Column relative frequency

- We restrict the sample space to a row or a column.

For Example,

(1) 'Among female buyers, what is the chance that someone owns a phone?'

→ Restrict sample space to only 'females' - first row

$$P(\text{doesn't own a phone} \mid \text{Female}) = \frac{10}{44}$$

$$\Rightarrow \frac{P(\text{Female} \cap \text{doesn't own of phone})}{P(\text{Female})} = \frac{10/100}{44/100} = \frac{10}{44}$$

(2) 'Among people who don't own a phone, how many are male?'

→ Restrict sample space to only 'who don't own a phone' - first column

$$P(\text{Male} \mid \text{who don't own a phone}) = \frac{14}{24}$$

$$\Rightarrow \frac{P(\text{Don't own a phone} \cap \text{Male})}{P(\text{doesn't own phone})} = \frac{14/100}{24/100} = \frac{14}{24}$$

Date August 2, 2021

INTRO TO CONDITIONAL PROBAB.

We are often interested in determining probabilities when some partial information concerning the outcome of the experiment is available.
In such situations, the probabilities are called conditional probabilities.

EXAMPLE :

- Experiment → Roll a dice twice
- Sample Space → $S = \{(1,1), (1,2) \ldots (1,6) \ldots (6,6)\}$
- Each outcome is equally likely to occur with a probability of $\frac{1}{36}$.
- Suppose further that the first roll of the dice lands on 4.
  Given this information, what is the resulting probability that the sum of the dice is 10?

→ In other words, the restricted sample space if the first dice lands of a four $F = \{(4,1), (4,2) \ldots (4,6)\}$

If each outcome of a finite sample space S is equally likely, then, conditional on the event that the outcome lies in a subset F, all outcomes in F becomes equally likely. In such cases, it is often convinient to compute conditional probability of the form $P(E|F)$ by using F as sample space.

- Among the outcomes in the restricted sample space, the outcome that satisfies the sum of dice is 10 is outcome (4,6). And this happens with probability $\frac{1}{6}$.

Conditional Probability : formula

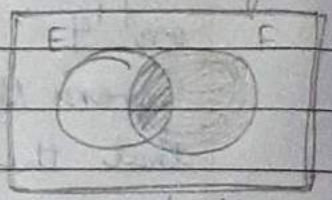
→ Let E denote the event that the sum of the dice is 10. Let F denote the event that the first die lands on 4, then the probability obtained is called the conditional probability of E given that F has occured.

It is denoted by : $P(E|F)$

→ The probability that event E occurs given that event F occurs (or conditional on event F occuring) is given by

$$P(E|F) = \frac{P(E \cap F)}{P(F)} \quad ; \quad P(F) > 0$$

Venn-diagram



Illustration

Apply the formula to example

$$\therefore P(E|F) = \frac{P(E \cap F)}{P(F)}$$

$$= \frac{P(\{(4,6)\})}{P(\{(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\})}$$

$$= \frac{1/36}{6/36} = \frac{1}{6} \quad \text{Ans.}$$

MULTIPLICATION RULE

Conditional probability formula: $P(E|F) = \frac{P(E \cap F)}{P(F)}$

Multiply both sides by $P(F)$,
we get,

$$P(E|F) \cdot P(F) = P(E \cap F)$$

→ This rule states that the probability that both E and F occur is equal to the probability that F occurs multiplied by the cond. prob. of E given that F occurs.

→ It is quite useful for computing the probability of an intersection.

GENERALISED MULTIPLICATION RULE

A generalisation of the multiplication rule, which provides an expression for the probability of the intersection of an arbitrary number of events, is refered to as the generalized multiplication rule and is given by,

$$P(E_1 \cap E_2 \cap E_3 \ldots \cap E_n) = P(E_1) \cdot P(E_2|E_1) \cdot P(E_3|E_1 \cap E_2) \ldots P(E_n|E_1 \cap E_2 \cap \ldots E_{n-1})$$

Example: Deck of Cards

An ordinary deck of 52 playing cards is randomly divided into 4 piles of 13 cards each. Compute the probability that each pile has exactly 1 ace.

→ Define events E_i ; $i = 1, 2, 3, 4$ as follows

(1) E_1 = { the ace of spades is in any one of the piles }

(2) E_2 = { the ace of spades and the ace of hearts are in different piles }

(3) E_3 = { the aces of spades, hearts and diamonds are in diffrent piles }

(4) E_4 = { all four aces are in different piles }

→ What we require is $P(E_1 \cap E_2 \cap E_3 \cap E_4)$.

Solution : 1. $P(E_1) = 1$

2. $P(E_2 | E_1) = 39/51$

3. $P(E_3 | E_1 \cap E_2) = 26/50$

4. $P(E_4 | E_1 \cap E_2 \cap E_3) = 13/49$

5. $P(E_1 \cap E_2 \cap E_3 \cap E_4) = \frac{39}{51} \times \frac{26}{50} \times \frac{13}{49} \approx 0.105$

INDEPENDENT EVENTS

In the cases where $P(E|F)$ is equal to $P(E)$, we say that E is independent of F

In other words, event E is independent of event F if knowing whether f occurs does not affect the probability of E.

Since, $P(E \cap F) = P(F) \times P(E|F)$
we see that E is independent of F if

$$P(E \cap F) = P(F) \times P(E)$$

Definition : Two events E and F are independent if
$$P(E \cap F) = P(E) \times P(F)$$

Multiplication Rule for Independent Events

For any 2 events if, E and F, are independent events then
$$P(E \cap F) = P(E) \cdot P(F)$$
and conversely if,
$$P(E \cap F) = P(E) \cdot P(F)$$
then E & F are independent.

Example of Independent Events

Consider the experiment of randomly selecting one card from a deck of 52 playing cards.

Define the following events :

→ E_1 : A face card is selected
→ E_2 : A king is selected
→ E_3 : A heart is selected

Are E_1 & E_2 independent ?
Are E_2 & E_3 independent ?

So, for first question, i-e, Are E_1 & E_2 independent, let's see:

$E_1 \cap E_2$ is the event that a face card and a king is selected which is the event a king is selected.

→ $P(E_1 \cap E_2) = P(\{KH, KC, KS, KD\}) = \frac{4}{52}$

→ $P(E_1) = P(\{JC, JH, JS, JD, QC, QH, QS, QD, KH, KC, KD, KS\}) = \frac{12}{52}$

→ $P(E_2) = P(\{KH, KC, KS, KD\}) = \frac{4}{52}$

Since, $P(E_1 \cap E_2) \neq P(E_1) \times P(E_2)$, ie,

$$\frac{4}{52} \neq \frac{12}{52} \times \frac{4}{52}$$

So, the events, E_1 & E_2 are not independent.

For the second question, i-e, Are E_2 & E_3 independent,

$E_2 \cap E_3$ is the event that a king & a heart is selected which is the event a king of hearts is selected.

- $P(E_2 \cap E_3) = P(\{KH\}) = 1/52$
- $P(E_2) = P(\{KH, KC, KS, KD\}) = 4/52$
- $P(E_3) = P(\{AH, 2H, 3H, 4H, 5H, 6H, 7H, 8H, 9H, 10H, JH, QH, KH\})$
  $= \frac{13}{52}$

Now, $P(E_2 \cap E_3) = P(E_2) \times P(E_3)$

$\Rightarrow \quad \frac{1}{52} = \frac{4}{52} \times \frac{13}{52}$

$\Rightarrow \quad \frac{1}{52} = \frac{1}{52}$

LHS = RHS

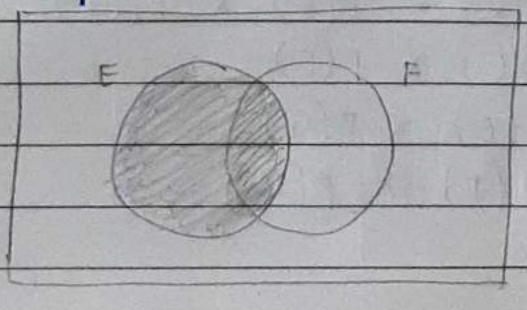
Thus, E_2 & E_3 are independent events.

INDEPENDENCE OF E & F^c

Proposition: If E and F are independent, then so are E and F^c.

Proof: Assume E & F are independent

- E can be expressed as : $E = (E \cap F) \cup (E \cap F^c)$



$\Rightarrow E = (E \cap F) \cup (E \cap F^c)$

$E \cap F$ and $E \cap F^c$ are mutually exclusive, hence

$\Rightarrow P(E) = P(E \cap F) + P(E \cap F^c)$

$\Rightarrow P(E) = P(E) \cdot P(F) + P(E \cap F^c)$ {$\because$ E & F are independent}

$\Rightarrow P(E \cap F^c) = -P(E) \cdot P(F) + P(E)$

$\Rightarrow P(E \cap F^c) = P(E)\,[1 - P(F)]$

$\Rightarrow P(E \cap F^c) = P(E) \cdot P(F^c)$

which means that E & F^c are independent events.

Thus, if E is independent of F, then the probability of E's occurrence is unchanged by information as to whether or not F has occurred.

INDEPENDENCE OF 3 EVENTS

Three events E, F and G are said to be independent if

1. $P(E \cap F \cap G) = P(E) \times P(F) \times P(G)$
2. $P(E \cap F) = P(E) \times P(F)$
3. $P(E \cap G) = P(E) \times P(G)$
4. $P(F \cap G) = P(F) \times P(G)$

For independent events, the probability that they all occur equals the product of their individual probabilities.

Application: A couple is planning on having three children. Assuming that each child is equally likely to be of either sex and that the sexes of the children are independent, find the probability that all the three children are girls.

Solution: Define E_i to be the event that the i^{th} child is a girl. The event all three children are girls is $(E_1 \cap E_2 \cap E_3)$.

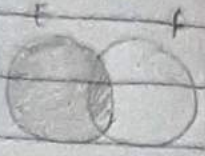
- Given ~~eg:~~ each child is equally likely to be of either sex $\Rightarrow$ $P(E_i) = \frac{1}{2}$

- the sexes of the children are independent $\Rightarrow$
$P(E_1 \cap E_2 \cap E_3) = P(E_1) \times P(E_2) \times P(E_3)$

$\Rightarrow$ Hence, the probability all three children are girls $\rightarrow$

$$P(E_1 \cap E_2 \cap E_3) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}.$$

LAW OF TOTAL PROBABILITY



→ Let E and F be events.

→ E can be refered as $(E \cap F) \cup (E \cap F^c)$

→ In order words, for, in order for an outcome to be in E, it must ~~be~~ either be in both E and F or be in E but not F.

Formula and Interpretation

$$P(E) = P(E \cap F) + P(E \cap F^c)$$
$$= P(F) \cdot P(E|F) + P(F^c) \cdot P(E|F^c) \quad - (1)$$

Eq(1) states that the probability of event E is a weighted average of the conditional probability of E given that F occurs and the conditional probability of E given that F does not occur.

Each conditional probability is weighted by the probability of the event on which it is conditioned.

Rule of Total Probability

Suppose that events $F_1, F_2, \ldots, F_k$, are mutually exclusive and exhaustive; that is, exactly one of the events must occur. Then for any event E,

$$P(E) = P(E|F_1)P(F_1) + P(E|F_2)P(F_2) + \ldots + P(E|F_k)P(F_k)$$

i.e, $P(E) = \sum_{i=1}^{k} P(E|F_i) \cdot P(F_i)$

Application : Insurance Policy

Question : An insurance company believes that people can be divided into two classes - those who are prone to have accidents and those who are not. The data indicate that an accident prone person will have an accident in a 1-year period with probability 0.1 ; the probability ofor all others is 0.05. Suppose that the probability is 0.2 that a new policyholder is accident-prone. What is the probability that a new policy holder will have an accident in first year?

Solution : Define events

→ E : a new policyholder will have accident in 1st year

→ F : a new policyholder is accident prone

Given

→ an accident prone person will have an accident in a 1-year period with probability 0.1, i.e, $P(E|F) = 0.1$

→ probability for all others is 0.05, i.e, $P(E|F^c) = 0.05$

→ probability is 0.2 that a new policyholder is accident prone ; $P(F) = 0.2$

$\therefore P(E) = P(F) \cdot P(E|F) + P(F^c) \cdot P(E|F^c)$

$= 0.2 \times 0.1 + 0.8 \times 0.05$

$= 0.06$

→ There is a 6 percent chance that a new policyholder will have an accident in the first year.

BAYES' RULE

Suppose we are now interested in the conditional probability of event F conditioned on E.
We know,

$$P(F|E) = \frac{P(F \cap E)}{P(E)}$$

From definition,

$$P(F|E) = \frac{P(F \cap E)}{P(E)} = \frac{P(E|F) \cdot P(F)}{P(F)\,P(E|F) + P(F^c)\,P(E|F^c)}$$

Bayes' Rule : Suppose that events $F_1, F_2, \ldots, F_k$ are mutually exclusive and exhaustive; Then for any event E,

$$P(F_i|E) = \frac{P(E|F_i) \cdot P(F_i)}{\sum_{i=1}^{k} P(E|F_i) \cdot P(F_i)}$$

Application : Insurance policy (..contd)

Already computed probability that a new policy holder will have an accident in the first year, $P(E) = 0.06$

Now, 'if a new policy holder has an accident in the first year' implies occurence of event E.

# What is the probability that he or she is accident-prone? In other words, what is
P(accident prone | policy holder had accident in 1st year)
This is equivalent to $P(F|E)$.

Applying Bayes' Rule, we get

$$P(F|E) = \frac{P(F) \cdot P(E|F)}{P(F) \cdot P(E|F) + P(F^c) \cdot P(E|F^c)} = \frac{0.02}{0.06} = \frac{1}{3}$$

∴ given that a new policyholder has an accident in the first year, the conditional probability that the policyholder is prone to accidents is $\frac{1}{3}$.

Dat
August 3, 2021

classmate
Date
Page

WEEK 9

Random Variables

INTRODUCTION

When a probability experiment is performed, often we are not interested in all the details of the experimental results, but rather are interested in the value of some numerical quantity determined by the result.

For example, in rolling a dice twice, often we care about only their sum of outcomes & are not concerned about the values on the individual dice.
i.e., we may be interested in knowing that the sum is 7 and may not be concerned over whether the actual outcome was (1,6), (2,5), (3,4), (4,3), (5,2) or (6,1).

These quantities of interest, or, more formally, these real-valued functions defined on the sample space, are known as <u>random variables</u>.

Because the value of a random variable is determined by the outcome of the experiment, we may assign probabilities to the possible values of the random variable.

EXAMPLES

(1) ROLLING A DICE : Sample Space

- Experiment : Roll a dice twice
- Sample Space, $S = \{(1,1), (1,2) \ldots (1,6) \ldots (6,1), \ldots (6,6)\}$
- Consider the problems associated with the two questions

(1) Of the outcomes, how many outcomes will result in a sum of outcomes as 7?

(2) Of the outcomes, how many outcomes will have the smaller of the outcomes as 3?

Solution : → Let X denote the sum of outcomes of two rolls
→ Let Y denote the lesser of the two outcomes. (If the outcomes are same, the value of the outcome is taken as value of Y)

OUTCOME	X	Y	OUTCOME	X	Y	OUTCOME	X	Y
(1,1)	2	1	(2,3)	5	2	(3,5)	8	3
(1,2)	3	1	(2,4)	6	2	(3,6)	9	3
(1,3)	4	1	(2,5)	7	2	(4,1)	5	1
(1,4)	5	1	(2,6)	8	2	(4,2)	6	2
(1,5)	6	1	(3,1)	4	1	(4,3)	7	3
(1,6)	7	1	(3,2)	5	2	(4,4)	8	4
(2,1)	3	1	(3,3)	6	3	(4,5)	9	4
(2,2)	4	2	(3,4)	7	3	(4,6)	10	4

OUTCOME	X	Y	OUTCOME	X	Y	OUTCOME	X	Y
(5,1)	6	1	(5,5)	10	5	(6,3)	9	3
(5,2)	7	2	(5,6)	11	5	(6,4)	10	4
(5,3)	8	3	(6,1)	7	1	(6,5)	11	5
(5,4)	9	4	(6,2)	8	2	(6,6)	12	6

Now, for the first question :

→ Let X denote the sum of outcomes of 2 rolls
→ X takes the values 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12.

Value of X	Relevant Event	Probability
2	{(1,1)}	$P(X=2) = 1/36$
3	{(2,1), (1,2)}	$P(X=3) = 2/36$
4	{(1,3), (3,1), (2,2)}	$P(X=4) = 3/36$
5	{(1,4), (4,1), (2,3), (3,2)}	$P(X=5) = 4/36$
6	{(1,5), (5,1), (2,4), (4,2), (3,3)}	$P(X=6) = 5/36$
7	{(1,6), (6,1), (2,5), (5,2), (3,4), (4,3)}	$P(X=7) = 6/36$
8	{(2,6), (6,2), (3,5), (5,3), (4,4)}	$P(X=8) = 5/36$
9	{(3,6), (6,3), (4,5), (5,4)}	$P(X=9) = 4/36$
10	{(4,6), (6,4), (5,5)}	$P(X=10) = 3/36$
11	{(5,6), (6,5)}	$P(X=11) = 2/36$
12	{(6,6)}	$P(X=12) = 1/36$

→ We say X is a random variable taking on one of the values 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 & 12 with the respective probabilities given above.

Similarly, goes with second question too.

(2) TOSSING A COIN THREE TIMES

- Experiment → Tossing a coin 3 times
- $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
- Consider the probabilities associated with the two questions :

(1) Of the 3 tosses, how many tosses will be heads?
(2) Of the 3 tosses, which toss results in a heads? i.e first, second or third toss is a head?

Solution : Let X denote the no. of heads that appear.
Let Y denote the toss in which a head appear first.

OUTCOME	X	Y
H H H	3	1
H H T	2	1
H T H	2	1
H T T	1	1

OUTCOME	X	Y
T H H	2	2
T H T	1	2
T T H	1	3
T T T	0	NIL

Now, for the first question (No. of toss that will be head) :

Value of X	Relevant Event	Probability
0	$\{(TTT)\}$	$1/8$ $(P(X=0))$
1	$\{(HTT), (THT), (TTH)\}$	$3/8$ $(P(X=1))$
2	$\{(HHT), (HTH), (THH)\}$	$3/8$ $(P(X=2))$
3	$\{(HHH)\}$	$1/8$ $(P(X=3))$

Same goes with the second question, too.

APPLICATION

(1) LIFE INSURANCE

→ A life insurance agent has 2 elderly clients, each of whom has a life insurance policy that pays ₹1 lakh per death. Let A be the event that the younger one dies in the following year, and let B be the event that the older one dies in the following year. Assume that A and B are independent with probabilities as $P(A) = 0.05$ and $P(B) = 0.10$. Let X denote the total amount of money (in units of ₹ lakhs) that will be paid out this year to any of these clients' beneficiaries. Find $P(X=0)$, $P(X=1)$, $P(X=2)$.

Solution : X is a random variable that takes on one of the possible values 0, 1, 2 with respective probabilities.

We have three possibilities :
→ Both dies : A and B happens ⇒ $A \cap B$
→ Younger one dies : A happens & B doesn't ⇒ $A \cap B^c$
→ Older one dies : A doesn't but B happens ⇒ $A^c \cap B$
→ Both survives : A & B doesn't happen ⇒ $A^c \cap B^c$

Value of X	Relevant Event	Probability
0	$A^c \cap B^c$	$P(A^c \cap B^c) = P(X=0)$
1	$(A^c \cap B) \cup (A \cap B^c)$	$P(A^c \cap B) \cup (P(A \cap B^c) = P(X=1)$
2	$(A \cap B)$	$P(A \cap B) = P(X=2)$

We say X takes 0,1,2 values with respective probabilities:

- $P(X=0) = P(A^c \cap B^c) = (1-0.05) \times (1-0.10) = 0.95 \times 0.9 = 0.855$
- $P(X=1) = P(A^c \cap B) \cup P(A \cap B^c) = (1-0.05 \times 0.1) + (0.0.5 \times 0.9) = 0.140$
- $P(X=2) = P(A \cap B) = 0.05 \times 0.1 = 0.005$

DISCRETE RANDOM VARIABLE

Definition

A random variable that can take on at most a countable no. of possible values is said to be a discrete random variable.

→ Thus, any random variable that can take on only a finite number or countably infinite number of different values is discrete.

→ There also exist random variables whose set of possible values is uncountable.

CONTINUOUS RANDOM VARIABLE

Definition

When outcomes for random event are numerical, but cannot be counted and are infinitely divisible, we have continuous random variables.

DISCRETE Vs CONTINUOUS RANDOM VARIABLES

A discrete random variable is one that has possible values that are discrete points along the real no. line.
→ Discrete random variables typically involve counting.

A continuous random variable is one that has possible values that form an interval along the real no. line.
→ Continuous random variables typically involve measuring.

EXAMPLE : Apartment Complex

Apartment No.	Floor No.	No. of Bedrooms	Size of Apartment (sq ft)	Distance of Apartment from lift (in mtr)
1	1	1	900.23	503.5
2	1	2	1175.34	325.6
3	1	3	1785.85	450.8
4	2	1	900.48	500.1
5	2	2	1175.23	324.5
6	2	3	1785.85	456.7
7	3	1	900.48	502.5
8	3	2	1175.23	325.6
9	3	3	1785.85	450.8
10	4	1	900.48	500.1
11	4	2	1176.03	325.4
12	4	3	1784.85	455.7

Random Experiment : Randomly selecting an apartment in an apartment complex of 12 apartments

$S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

Questions :

(a) Let the random variable be no. of bedrooms, what are the possible values that might be observed ?

Answer : 1, 2, 3

(b) Let the random variable be floor no. of the apartment. What are the possible values that might be observed ?

Answer : 1, 2, 3, 4

(c) Let the random variable be size of apartment. What are the possible values that might be observed ?

Answer : [900, 1800] sq. ft

(d) Let the random variable be distance of the apartment from the lift. What are the possible values that might be observed ?

Answer : [324, 505] meters

Which variables are discrete random variables ?

→ No. of bedrooms, floor number.

Which variables are continuous random variables ?

→ Size, Distance from the lift

More Examples Of Discrete & Continuous Variables

(a) DISCRETE

- → No. of people in a household
- → No. of languages a person can speak
- → No. of times a person takes a particular test before qualifying.
- → No. of accidents in an intersection
- → No. of spelling mistakes in a report.

(b) CONTINUOUS

- → Temperature of a person
- → Height of a person
- → Speed of a vehicle
- → Time Taken by a person to write an exam.

DISCRETE RANDOM VAR

Probability mass function (p.m.f.)

- A random variable that can take on at most a countable no. of possible values is said to be a discrete random variable.
- Let X be a discrete random variable, and suppose that it has possible values, which will be label $x_1, x_2, x_3 \ldots\ldots x_n$.
- For a discrete random variable X, we define the probability mass function $p(x)$ of X by

$$p(x_i) = P(X = x_i)$$

- Tabular Form :

X	x_1	x_2	x_3	...	...	x_n
$P(X = x_i)$	$p(x_1)$	$p(x_2)$	$p(x_3)$	...	...	$p(x_n)$

PROPERTIES OF p.m.f

The probability mass function $p(x)$ is positive for at most a countable number of values of x. That is, if X must assume one of the values $x_1, x_2, \ldots$, then

1. $p(x_i) \geq 0$, $i = 1, 2, \ldots$
2. $p(x) = 0$ for all other values of x

TABULAR FORM :

X	x_1	x_2	x_3
$P(X = x_i)$	$p(x_1)$	$p(x_2)$	$p(x_3)$

Since X must take one of the values of x_i, we have

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

For Example, suppose X is a random variable that takes three values 0, 1, 2 with probabilities,

$p(0) = P(X=0) = 1/4$
$p(1) = P(X=1) = 1/2$
$p(2) = P(X=2) = 1/2$

X	0	1	2
$P(X = x_i)$	1/4	1/2	1/4

Now, since all probabilities are greater than equal to 0, first condition checked, i.e., $p(x_i) \geq 0$.

Now, $\sum_{i=1}^{3} p(x_i) = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$

Hence, second condition verified, i.e., $\sum_{i=1}^{\infty} p(x_i) = 1$

Thus, it is a probability mass function.

QUESTION: Let X be a random variable that takes values 1, 2, 3, 4, 5. Which of the following are probability mass functions?

1.

X	1	2	3	4	5
$P(X = x_i)$	0.4	0.1	0.2	0.1	0.3

Step I : $p(x_i) \geq 0$; verified

Step II : $\Sigma p(x_i) = 1$

$\Rightarrow 0.4 + 0.1 + 0.2 + 0.1 + 0.3 = 1$

$\Rightarrow 1.1 \neq 1$

$\therefore$ Given is NOT a p.m.f.

2.

X	1	2	3	4	5
$P(x_i)$	0.2	0.3	0.4	-0.1	0.2

Since $p(x_4) < 0$, thus given is NOT a p.m.f.

3.

X	1	2	3	4	5
$P(X = x_i)$	0.3	0.1	0.2	0.4	0.0

Step I : $p(x_i) \geq 0$; verified

Step II : $\Sigma p(x_i) = 1$

$\Rightarrow 0.3 + 0.1 + 0.2 + 0.4 + 0.0 = 1$

$\Rightarrow 1 = 1$

Hence, the given is a p.m.f.

X having infinite values

Suppose X is a random variable that takes values 0, 1, 2, ... with probabilities,

$$p(i) = c\frac{\lambda^i}{i!}, \text{ for some +ve } \lambda$$

What is the value of c?

We know, $\sum_{i=0}^{\infty} p(x_i) = 1$

$$\Rightarrow \sum_{i=0}^{\infty} c\frac{\lambda^i}{i!} = 1$$

$$\Rightarrow c\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = 1$$

Recall, $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!} \Rightarrow c\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = ce^{\lambda}$

$$\therefore c\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = 1$$

$$\Rightarrow ce^{\lambda} = 1$$

$$\Rightarrow c = e^{-\lambda}$$

EXAMPLE : Tossing A Coin Three Times

$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$

X is a random variable which counts the no. of heads in the tosses.

P.M.F :

X	0	1	2	3
$P(X=x_i)$	1/8	3/8	3/8	1/8

Verify : $\sum_{i=0}^{3} p(x_i) = \frac{8}{8} = 1$

Y is a random variable which counts the toss in which heads appears first

P.M.F.

Y	1	2	3	NIL
$P(Y=y_i)$	4/8	2/8	1/8	1/8

Verify : $\sum_{i=1}^{4} p(x_i) = \frac{8}{8} = 1$

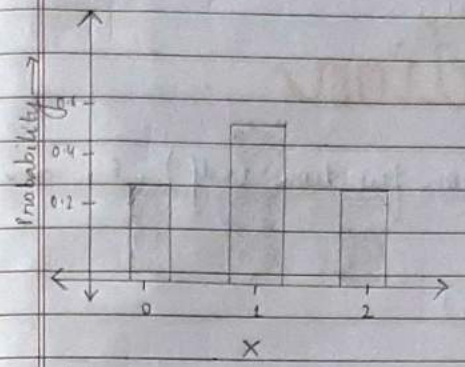
GRAPH OF P.M.F

→ It is helpful to illustrate the p.m.f in a graphical format by plotting $P(X=x_i)$ on the y axis against x_i on the x-axis.

→ Let's look some examples :

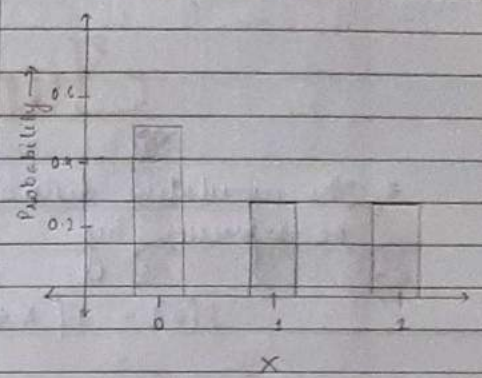
(1)

X	0	1	2
$P(X=x_i)$	0.25	0.5	0.25



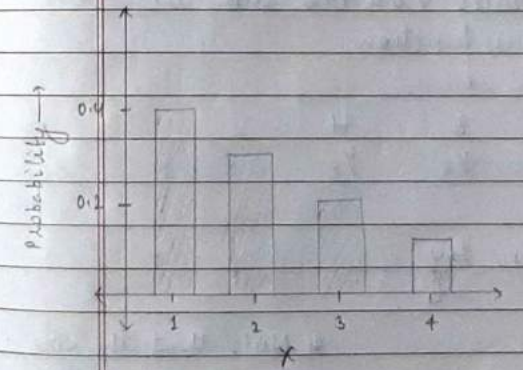
(2)

X	0	1	2
$P(X=x_i)$	0.5	0.25	0.25



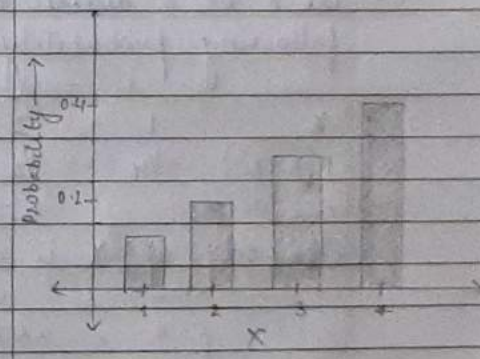
Positive Skewed Distribution

X =	1	2	3	4
$P(X=x_i)$	0.4	0.3	0.2	0.1



Negative Skewed Distribution

X	1	2	3	4
$P(X=x_i)$	0.1	0.2	0.3	0.4



Cumulative Distribution Function

The cumulative distribution function (cdf), F, can be expressed by

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

If X is a discrete random variable whose possible values are $x_1, x_2, x_3, \ldots$ where $x_1 < x_2 < x_3 \ldots$, then the distribution function F of X is a step function.

STEP FUNCTION

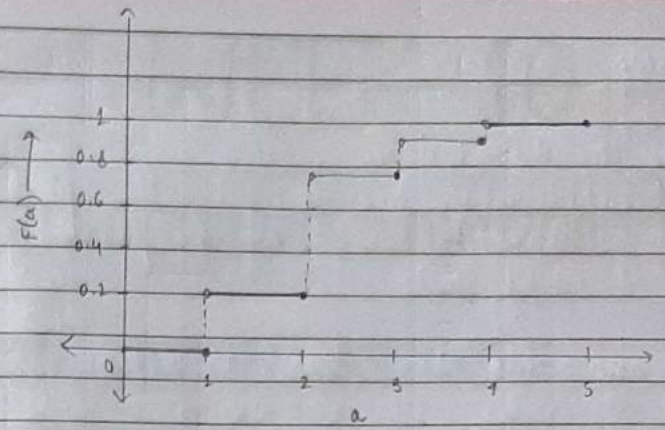
- Let X be a discrete random variable with the following probability mass function.

X	1	2	3	4
$P(X=x_i)$	1/4	1/2	1/8	1/8

- The cdf of X is given by

$$F(a) = \begin{cases} 0, & a < 1 \\ 1/4, & 1 \leq a < 2 \\ 3/4, & 2 \leq a < 3 \\ 7/8, & 3 \leq a < 4 \\ 1, & 4 \leq a \end{cases}$$

Note that the size of the step at any of the values 1, 2, 3, and 4 is equal to the probability that X assumes that particular value.



Summary of Week 9

(1) Define what is a random variable

(2) Types of Random Variables:
- Discrete : countable
- Continuous : measurable

(3) Probability Mass Function → GRAPH
- $p(x_i) \geq 0$
- $\Sigma p(x_i) = 1$
- skewed, symmetric, constant, uniform.

(4) Cumulative Distribution Function
- $F(a) = P(X \leq a)$
- Step function

Date August 6, 2021

WEEK 10

Discrete Random Variable

APPLICATION OF C.D.F

(1) CREDIT CARDS

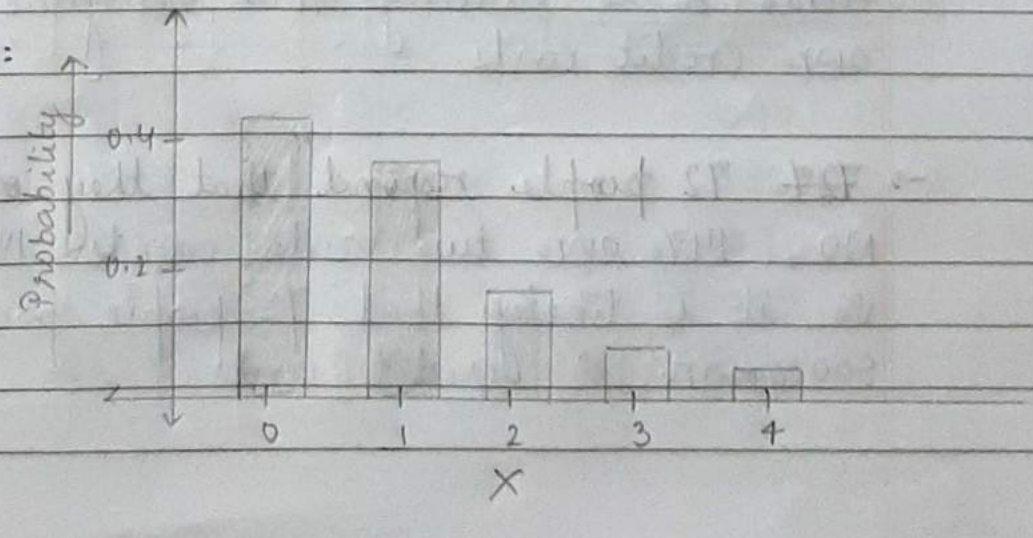
Random Experiment : Consider the random experiment of selecting an adult at random from the sample.

Random Variable : No. of credit cards owned by person.

The following table summarizes the probability distribution of the number of credit cards per person based on the relative frequencies :

x	0	1	2	3	4
$P(X = x_i)$	0.42	0.36	0.14	0.06	0.02

PMF Graph :



QUESTIONS :

(1) Describe the distribution.

→ · The distribution is skewed right with a peak at 0.
· No. of credit cards owned by people vary b/w 0 to 4 credit cards.

(2) Choose an adult at random. Is he/she more likely to have 0 credit cards or 2 or more cards?

→ The probability that an adult has no credit cards is 0.42, while the probability of having 2 or more credit cards is about 0.22.
So $p(0) > p(2) + p(3) + p(4)$

i.e. $p(0) > p(x \geq 2)$
$0.42 > 0.22$

(3) You take a random sample of 500 people and ask them how many credit cards they own. Would you be surprised at the following:

→ Everyone owns a credit card.
YES, 42% of adults do not own credit cards.
Hence it is unlikely that everyone of the 500 would own credit cards.

→ ~~72%~~ 72 people respond that they own 2 credit cards.
NO, 14% own two credit cards 14% of 500 = 70.
So, it is likely that 72 people from the sample of 500 own 2 credit cards.

Expectation of Random Var.

INTRODUCTION

Consider the following game of rolling a dice once.
→ If the outcome is even - you lose an amount equal to the outcome.
→ If the outcome is odd - you win an amount equal to the outcome.

In other words, the gains/losses as per table

OUTCOME	1	2	3	4	5	6
WINNING	+1	-2	+3	-4	+5	-6

A winning of $-x$ indicates a loss of x amount.

Question: Would you play this game?

→ Rolling 100 times (Average Winnings = -0.09)

OUTCOME	WINNING	FREQUENCY	RELATIVE FREQ.
1	+1	16	0.16
2	-2	10	0.10
3	+3	16	0.16
4	-4	21	0.21
5	+5	23	0.23
6	-6	14	0.14
		100	1

→ Rolling a 1000 times

OUTCOME	WINNING	FREQUENCY	RELATIVE FREQUENCY
1	+1	177	0.177
2	-2	177	0.177
3	+3	167	0.167
4	-4	153	0.153
5	+5	163	0.163
6	-6	163	0.163
		1000	1

Average Winnings : -0.451

OBSERVATIONS

The relative frequency of each of the six possible outcomes is close to the probability of 1/6 for the respective outcomes.

Hence, it suggests, that if I repeat rolling the dice for a very larger no. of times, our average gain should be

$$1\left(\frac{1}{6}\right) - 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) - 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) - 6\left(\frac{1}{6}\right) = -0.5$$

This is close to what we got as the average winning for 1000 rolls of the dice.

Definition :

Let X be a discrete random variable taking values $x_1, x_2, \ldots$ The expected value of X denoted by $E(X)$ are referred to as Expectation of X is given by

$$E(X) = \sum_{i=1}^{\infty} x_i \, P(X = x_i)$$

The Expectation of a random variable can be considered the 'long-run average' value of the random variable in repeated independent obs.

EXAMPLE :

(1) → Random Experiment : Roll a dice once

→ Sample space, $S = \{1, 2, 3, 4, 5, 6\}$

→ Random variable X is the outcome of the roll.

→ The probability distribution is given by,

X	1	2	3	4	5	6
$P(X = x_i)$	1/6	1/6	1/6	1/6	1/6	1/6

$$E(X) = 1(1/6) + 2(1/6) + 3(1/6) + 4(1/6) + 5(1/6) + 6(1/6) = 3.5$$

Does this mean that if we roll a dice once, should we expect the outcome to be 3.5 ?

→ NO ! the expected value tells us that we would expect the average of a large no. of rolls to be in the long run.

	100 rolls		1000 rolls		
OUTCOME	Freq.	Rel. Freq.	Freq.	Rel. Freq.	Probability
1	16	0.16	177	0.177	0.166667
2	10	0.1	177	0.177	0.166667
3	16	0.16	167	0.167	0.166667
4	21	0.21	153	0.153	0.166667
5	23	0.23	163	0.163	0.166667
6	14	0.14	163	0.163	0.166667
AVERAGE:		3.67		3.437	3.5

Notice that average of the rolls need not be exactly 3.5. However we can expect it to be close to 3.5.

The expected value of X is a theoretical average.

BERNOULLI RANDOM VARIABLE

- A random variable that takes on either the value 1 or 0 is called a Bernoulli random variable.
- Let X be a Bernoulli random variable that takes on the value 1 with probability p.

Probability Distribution:

X	0	1
$P(X = x_i)$	$1-p$	p

Expected value of Bernoulli random variable:

$$E(X) = 0(1-p) + 1(p) = p$$

DISCRETE UNIFORM RANDOM VAR.

Let X be a random variable that is equally likely to take any of the values $1, 2, \ldots n$

P.m.f :

X	1	2	...	n
$P(X = x_i)$	$1/n$	$1/n$	...	$1/n$

$$\therefore E(X) = \sum_{i=1}^{n} x_i\, p(x_i) = \frac{(1\times1) + (2\times1) + \ldots + (n\times1)}{n}$$

$$= \frac{n(n+1)}{2n} = \frac{n+1}{2}$$

Expectation of a function of a Random Variable

PROPOSITION

Let X be a discrete random variable which takes values x_i along with its p.m.f, $P(X = x_i)$.
Let g be any real valued function.

The expected value of $g(X)$ is

$$E(g(X)) = \sum_i g(x_i)\, P(X = x_i)$$

COROLLARY

If a and b are constants,

$$E(aX + b) = aE(X) + b$$

EXAMPLE 1: Let X be a discrete random variable with the following distribution

X	-1	0	1
$P(X = x_i)$	0.2	0.5	0.3

Let $Y = g(X) = X^2$. What is $E(Y)$?

Solution: $E(Y) = [(-1)^2 \times 0.2] + [0 \times 0.5] + [1^2 \times 0.3]$
$= 0.5$

Distribution of Y :

Y	0	1
$P(Y = y_i)$	0.5	0.5

Note → $0.5 = E(X^2) \neq [E(X)]^2 = 0.01$

EXAMPLE 2: Sanjay & Anitha work for the same company. Anitha's diwali bonus is random variable whose expected value is 15,000.

(a) Sanjay's bonus is set equal to 90% of Anita's, find the expected value of Sanjay's bonus.

Sol. Let X denote Anitha's bonus.
Given, $E(X) = 15,000$

Let Y denote Sanjay's bonus
Given, $Y = 0.9X$

$\therefore E(Y) = E(0.9X) = 0.9\, E(X) = 0.9 \times 15000 = ₹13,500$

(b) If Sanjay's is set to equal ₹1000 more than Anitha's, find his expected bonus.

Given, $Y = X + 1000$; Hence $E(Y) = E(X + 1000)$
$= E(X) + 1000$
$= 15000 + 1000 = 16000$ Rs.

The expected value of the sum of random variables is equal to the sum of the individual expected values, i.e. Let X and Y be two random variables. Then,

$$E(X+Y) = E(X) + E(Y)$$

EXAMPLE : Rolling A Dice

Let X be the outcome of a fair dice.
Let Y be the outcome of another fair dice.

We know, $E(X) = E(Y) = 3.5$

X + Y is the sum of outcomes of both the dice rolled together. Then,

$$E(X+Y) = E(X) + E(Y) = 3.5 + 3.5 = 7$$

This is the same expectation of the sum of outcomes of rolling a dice twice.

EXPECTATION OF SUM OF MANY RANDOM VARIABLES

The result that the expected value of the sum of random variables is equal to the sum of the expected values holds for not only two but any number of random variables.

Let $X_1, X_2, \ldots, X_k$ be k discrete random variables. Then,

$$E\left(\sum_{i=1}^{k} X_i\right) = \sum_{i=1}^{k} E(X_i)$$

Hypergeometric Random Variable

- Suppose that a sample of size n is to be chosen randomly (without replacement) from a bowl containing N balls, of which m are red and N - m are blue.
- Let X denote the no. of red balls selected, then

$$P(X=i) = \frac{({}^{m}C_i)\cdot({}^{N-m}C_{n-i})}{{}^{N}C_n}, \quad i = 0, 1, 2, 3, \ldots n$$

- X is said to be a hypergeometric variable for some values of n, m and N.

$$\Rightarrow E(X) = \frac{nm}{N}$$

Example : Two students are randomly chosen from a group of 20 boys and 10 girls. Let X denote the no. of boys chosen, and let Y denote the no. of girls chosen.

(1) Find $E(X)$.

Sol. $n = 2 \quad N = 30 \quad m = 20$

$$\therefore E(X) = \frac{nm}{N} = \frac{2 \times 20}{30} = \frac{4}{3}$$

(2) Find $E(Y)$

Sol. $n = 2 \quad N = 30 \quad m = 10$

$$\therefore E(Y) = \frac{nm}{N} = \frac{2 \times 10}{30} = \frac{2}{3}$$

(3) Find $E(X+Y)$

Sol. $$E(X+Y) = E(X) + E(Y) = \frac{4}{3} + \frac{2}{3} = \frac{6}{3} = 2$$

Variance of Random Var.

INTRODUCTION

The expected value of a random variable gives the weighted average of the possible values of the random variable, it does not tell us anything about the variation, or spread, of these values.

For instance, consider random variables X, Y & Z, whose values and probabilities are as follows :

$X = 0$ with probability 1

$$Y = \begin{cases} -2 & \text{with probability } 1/2 \\ 2 & \text{with probability } 1/2 \end{cases}$$

$$Z = \begin{cases} -20 & \text{with probability } 1/2 \\ 20 & \text{with probability } 1/2 \end{cases}$$

$E(X) = E(Y) = E(Z) = 0$. However we noticed the spread of Z is greater than spread of Y which is greater than spread of X.

Let us denote expected value of a random variable X by the greek alphabet μ.

Definition:

Let X be a random variable with expected value μ, then the variance of X, denoted by $Var(X)$ or $V(X)$ is defined by.

$$Var(X) = E(X-\mu)^2$$

In other words, the Variance of a random variable X measures the square of the difference of the random variable from its mean, μ, on the average.

Computational Formula for $Var(X)$:

$$Var(X) = E(X-\mu)^2 \quad - (1)$$

$$(X-\mu)^2 = X^2 + \mu^2 - 2X\mu \quad - (2)$$

Using properties of expectation we know,

$$E(X^2+\mu^2-2X\mu) = E(X^2) + \mu^2 - 2\mu E(X)$$

which is same as

$$= E(X^2) - \mu^2$$

Example 1: Rolling a dice once

Random Experiment: Roll a dice once

Sample Space, $S = \{1, 2, 3, 4, 5, 6\}$

Random variable X is the out come of the roll.

The probability distribution is given by

X	1	2	3	4	5	6
X^2	1	4	9	16	25	36
$P(X=x_i)$	1/6	1/6	1/6	1/6	1/6	1/6

$$\therefore E(X) = 1(1/6) + 2(1/6) + 3(1/6) + 4(1/6) + 5(1/6) + 6(1/6)$$
$$= 3.5$$

$$E(X^2) = 1(1/6) + 4(1/6) + 9(1/6) + 16(1/6) + 25(1/6) + 36(1/6)$$
$$= 15.167$$

$$Var(X) = E(X^2) - \mu^2 = 15.167 - (3.5)^2 = 2.917$$

Example 2: Tossing A Coin Thrice

$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$

X is the random variable which counts the no. of heads in the tosses.

P.m.f.

X	0	1	2	3
X^2	0	1	4	9
$P(X=x_i)$	1/8	3/8	3/8	1/8

$$E(X) = 0(1/8) + 1(3/8) + 2(3/8) + 3(1/8) = 3/2$$
$$E(X^2) = 0(1/8) + 1(3/8) + 4(3/8) + 9(1/8) = 3$$

$$Var(X) = E(X^2) - [E(X)]^2 = 3 - (3/2)^2 = 3 - \frac{9}{4} = 0.75$$

Variance in Bernoulli Random Variable

Pmf :

X	0	1
X^2	0	1
$P(X=x_i)$	$1-p$	p

Expected value of a Bernoulli Random Variable :

$E(X) = p$

Variance of Bernoulli Random Variable

$Var(X) = p - p^2 = p(1-p)$

Variance in Discrete Uniform Random Variable

Let X be a random variable that is equally likely to take any of the values $1, 2, \ldots n$.

P.mf :

X	1	2	...	n
X^2	1	4	...	n^2
$P(X=x_i)$	$1/n$	$1/n$	...	$1/n$

$E(X) = (n+1)/2$

$E(X^2) = \frac{(n+1)(2n+1)}{6}$

$\therefore Var(X) = \frac{n^2-1}{12}$

VARIANCE OF A FUNCTION OF A RANDOM VARIABLE

[PROPERTIES OF VARIANCE]

Proposition

Let X be a Random Variable, let c be a constant, then

→ $Var(cX) = c^2 Var(X)$

→ $Var(X+c) = Var(X)$

Corollary

If a and b are constants, $V(aX+b) = a^2 Var(X)$

PROOF : We know $E(aX+b) = a E(X) + b$

Hence, $Var(aX+b) = E(aX+b-a\mu-b)^2$

$= E(a^2(X-\mu)^2)$

$= a^2 E(X-\mu)^2$

$= a^2 Var(X)$

Variance of Sum of Two Random Variables

We know, $E(X+Y) = E(X) + E(Y)$

$\therefore Var(X+X) = Var(2X) = 4\,Var(X) \neq Var(X) + Var(X)$

INDEPENDENT RANDOM VARIABLES

Definition :
Random variables X and Y are independent if knowing the value of one of them does not change the probabilities of the other.

Example :
→ Roll a dice twice
→ ~~S = {HH, HT, TH, TT}~~
→ $S = \{(1,1), (1,2) \ldots\ldots (6,6)\}$

→ X is the outcome of the first dice.
→ Y be the outcome of second dice.

→ Knowing X = i does not change the probability of Y taking any value 1, 2 ... 6.

→ X and Y are independent random variables.

Variance of Sum of Independent Random Variables

Let X and Y be independent random variables.
Then,

$$Var(X+Y) = Var(X) + Var(Y)$$

Example : Rolling a dice twice

→ Let X be the outcome of a fair dice.
Let Y be the outcome of another fair dice.

→ We know, $E(X) = E(Y) = 3.5$

→ X+Y is the sum of the outcomes of both dice rolled together.

Thus, $E(X+Y) = E(X) + E(Y) = 3.5 + 3.5 = 7$

We also know, $Var(X) = Var(Y) = 2.917$

→ X & Y are independent, hence,

$$Var(X+Y) = Var(X) + Var(Y)$$
$$= 2.917 + 2.917 \approx 5.83$$

which is the same as what we obtained earlier applying the computational formula.

Variance of Sum of many Independent Random Var.

Let $X_1, X_2, X_3 \ldots X_k$ be k discrete random (independent) variables. Then

$$Var\left(\sum_{i=1}^{k} X_i\right) = \sum_{i=1}^{k} Var(X_i)$$

HYPERGEOMETRIC RANDOM VAR.

Suppose that a sample of size n is to be chosen randomly (without replacement) from a box containing N balls of which m are red and $N-m$ are blue.

Let X denote the no. of red balls selected, then

$$P(X=i) = \frac{({}^{m}C_{i}) \cdot ({}^{N-m}C_{n-i})}{{}^{N}C_{n}}, \quad i = 0,1,2,\ldots n$$

X is said to be a hypergeometric variable for some values of n, m and N.

$$E(X) = \frac{mn}{N}$$

$$\therefore \text{Var}(X) = \frac{nm}{N}\left[\frac{(n-1)(m-1)}{N-1} + 1 - \frac{mn}{N}\right]$$

Standard Deviation of a Random Variable

Definition:

The quantity $SD(X) = \sqrt{\text{Var}(X)}$ is called the standard deviation of X.

Hence, the SD is the +ve square root of variance.

#REMARK, The standard deviation, like the expected value, is measured in the same units as is random variable.

Properties of Standard Deviation:

Let X be a random variable, let c be a constant, then

- $SD(cX) = c \cdot SD(X)$
- $SD(X+c) = SD(X)$

Example: If $\text{Var}(X) = 4$, what is $SD(3X)$?

$SD(3X) = 3 \cdot SD(X) = 3\sqrt{4} = 3 \times 2 = 6$

Example: If $\text{Var}(2X+3) = 16$, what is $SD(X)$?

$\text{Var}(2X+3) = 16 \Rightarrow 4 \cdot \text{Var}(X) = 16 \Rightarrow \text{Var}(X) = 4$

$\therefore SD(X) = \sqrt{\text{Var}(X)} = \sqrt{4} = 2$