Statistics week 1 summary

Descriptive statistics - The part of statistics concerned with the description and summarization of data is called descriptive statistics.

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

The concept of Population and Sample

- The percentage of all students in India who have passed their Class 12 exams and study engineering.
- The prices of all houses in Tamil Nadu.
- The total sales of all cars in India in the year 2019.
- The age distribution of people who visit a city Mall in a particular month.

Definition –

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.

Data - Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation.

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.

Unstructured data and 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.

- Case (observation): A unit from which data are collected.
- Variable:
- 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: Name, marks obtained, Board etc.
- Rows represent cases: for each case, same attribute is recorded • Columns represent variables: For each variables, same type of
- value for each case is recorded.

Classification of data –

- Categorical
- Numerical

- Categorical data
- Also called qualitative variables.
- Identify group membership
- Numerical data
- Also called quantitative variables.
- Describe numerical properties of cases
- Have measurement units
  Measurement units: Scale that de_nes the meaning of • numerical data, such as weights measured in kilograms, prices
- in rupees, heights in centimeters, etc.
- I The data that make up a numerical variable in a data table
- must share a common unit.

Time series - data recorded over time

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

- Cross-sectional - data observed at the same time.

Scales of measurement

Nominal scale - When the data for a variable consist 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.

- There is no ordering in the variable.
- Nominal – Name categories without variables.

Ordinal scale - Data exhibits properties of nominal data and the order or rank of data is meaningful, the scale of measurement is considered a ordinal scale.

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

Ordinal – Name categories that can be ordered.

Interval scale - 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.

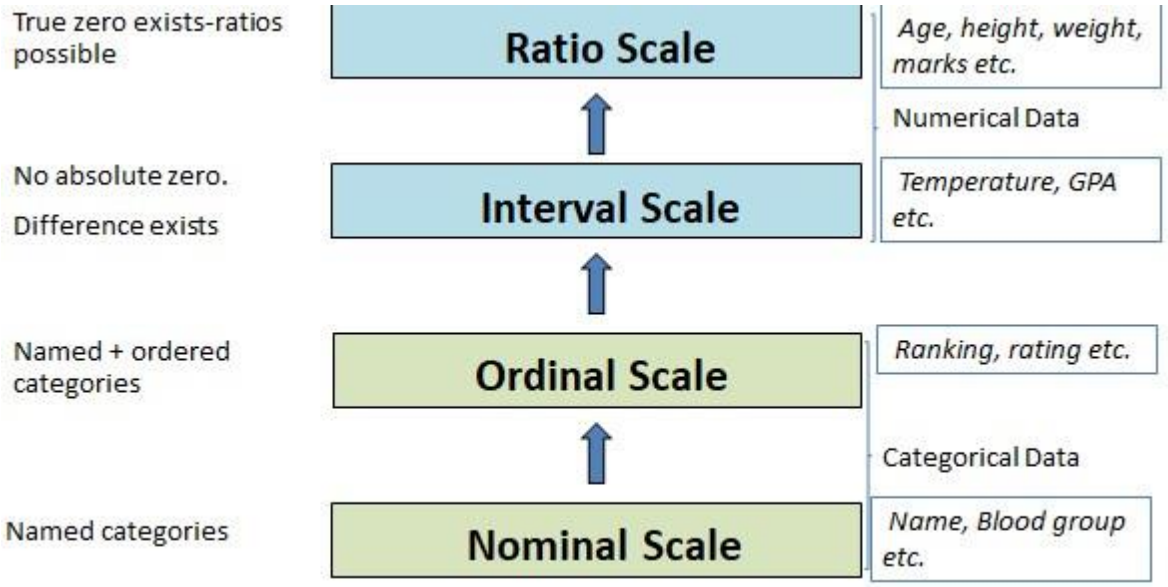
- Interval data are always numeric. Can _nd out di_erence between any two values.

- 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)

Ratio Scale - 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)



Statistics week 2

Frequency Distributions - A frequency distribution of qualitative data is a listing of the distinct values and their frequencies.

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

Ex –

Construct a frequency table for the given data

1. A,A,B,C,A,D,A,B,D,C
2. A,A,B,C,A,D,A,B,D,C,A,B,C,D,A
3. A,A,B,C,A,A,B,B,D,C,A,B,C,D,B
4. A, A, B, C,A ,D, A,B,D,C, A,B,C,D,A,C,D,D

The steps to construct a frequency distribution2

- 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.

Relative Frequency - The ratio of the frequency to the total number of observations 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 number of observations.

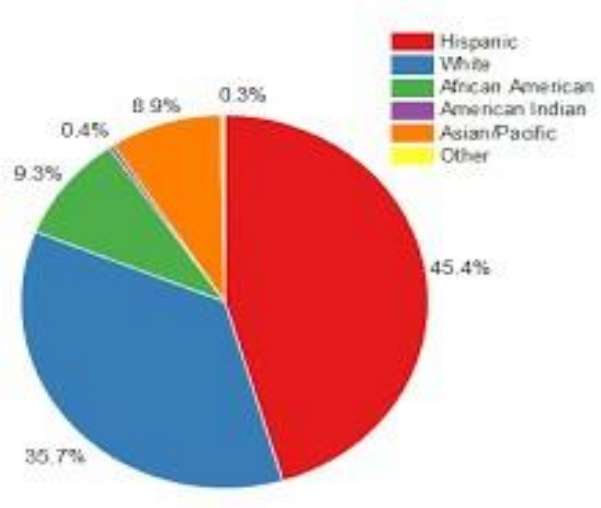
Why relative frequency?

I For comparing two data sets.
I Because relative frequencies always fall between 0 and 1, they provide a standard for comparison.

Pie chart - 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 the 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.



1. A pie chart is used to show the proportions of a categorical variable.

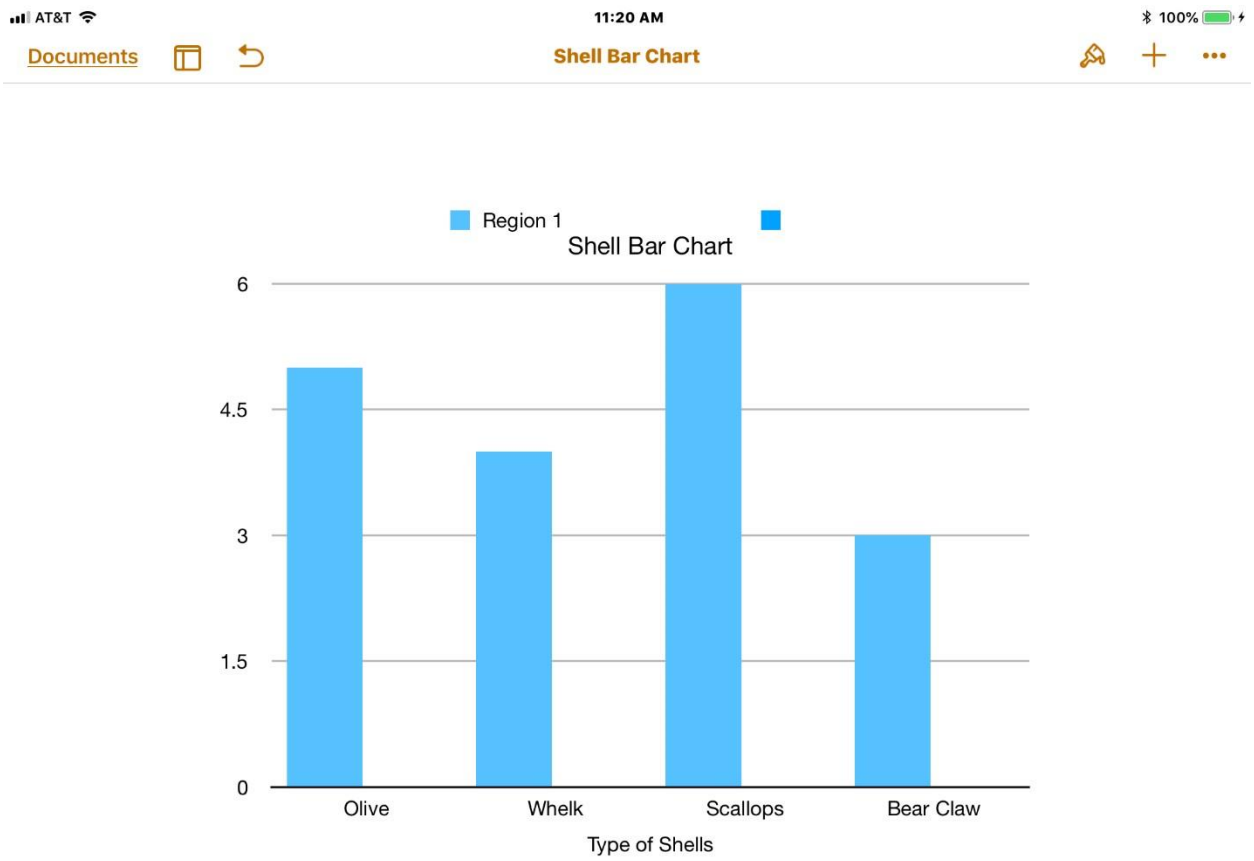
2. A pie chart is a good way to show that one category makes up more than half of the total.

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 bar 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.

Steps to construct a bar chart

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."



Pareto chart - 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.

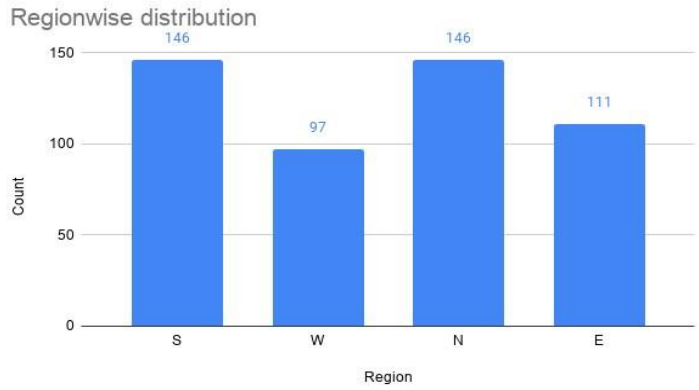
1. A bar chart is used to show the frequencies/relative frequencies of a categorical variable.
2. If ordinal, the order of categories is preserved.
3. The bars can be oriented either horizontally or vertically.
4. A Pareto chart is a bar chart where the categories are sorted by frequency.

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: violating area principle

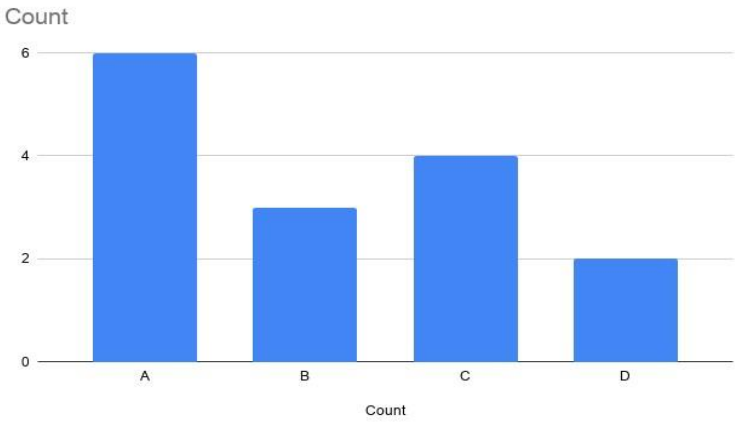
Decorated graphics: Charts decorated to attract attention
often violate the area principle.
No baseline and the chart shows bottles on top of labeled boxes of various sizes and shapes. I Obeys area principle and accurate.



- Another common violation is when the baseline of a bar chart is not at zero.
- Left graph exaggerates the number coming from the South and North. Graph on right shows same data with the baseline at zero.

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

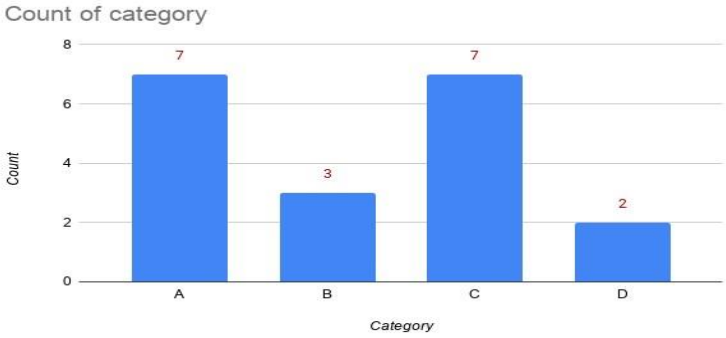
- The longest bar in a bar chart
- The widest slice in a pie chart.
- In a Pareto chart, the mode is the first category shown.



- Let consider the example A,A,B,C,A,D,A,B,C,C, A,B,C,D,A
- The longest bar in a bar chart
- The most common category is "A"

Bimodal and multimodal data

- If two or more categories tie for the highest frequency, the • data are said to be bimodal (in the case of two) or multimodal (more than two).
- Let 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 - The median of an ordinal variable is the category of the middle
- observation of the sorted values.

- 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
- The ordered data is A,A,A,A,B,B,B,B,B,C,C,C,D,D

- The median grade is the category associated with the 8 observation which is "B".
- Consider the grades 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,B,C,C,C,D,D
- The median grade is the category associated with the 7 or 8 observation which is "B".

Statistics Week 3

Types of variables-

1) Categorical
2) Numerical
   I) Discrete
   II) Continuous

Organizing numerical data

- Recall, 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 relativefrequency distributions 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 di_erent, values, it is convenient 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 number 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, number of people in each household, takes is 1,2,3,4,5.

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

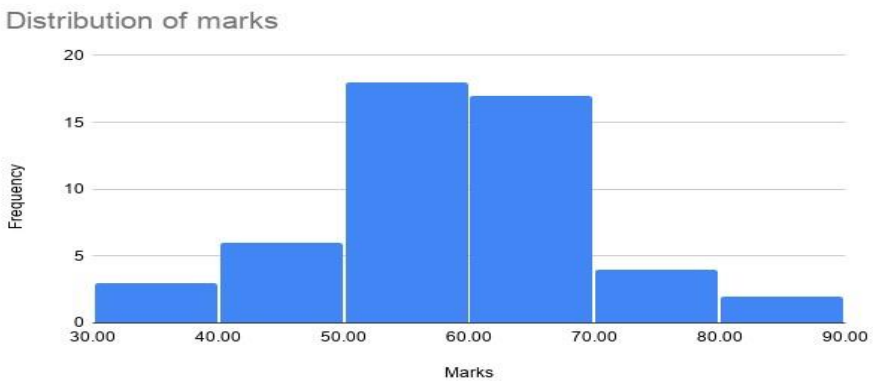
1. Number of classes: The appropriate number is a subjective choice, the rule of thumb is to have between 5 and 20 classes.
2. Each observation should belong to some class and no observation should belong to more than one class.
3. It is common, although not essential, to choose class intervals of equal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

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 (relative frequencies).
- Step 3 For each class, construct a vertical bar whose height equals the frequency (relative 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" (\Relative frequency").



Stem-and-leaf diagram

Definition

- In a stem-and-leaf diagram (or stemplot)1, 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 stem plot

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.

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.

Measures of central tendency: These are measures that indicate the most typical value or center of a data set.

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

The mean - The mean of a data set is the sum of the observations divided by the number of observations.

The mean is usually referred to as average.

- Arithmetic average; divide the sum of the values by the number of values (another typical value)
- For discrete observations:
- Sample mean: $x = x_1 + x_2 + : : : + x_{n/n}$

- Population mean: $= x_1 + x_2 + : : : + x_{N/N}$

Mean for grouped data: discrete single value data

$x = f_1x_1 + f_2x_2 + : : : + f_nx_{n/n}$

Mean for grouped data: continuous data

$\_x = f_1m_1 + f_2m_2 + : : : + f_nm_{n/n}$

Here m = midpoint

Adding a constant

Let $y_i = x_i + c$ where c is a constant then $\_y = \_x + c$ (_y is y bar _x is x bar)

Multiplying a constant

Let $y_i = x_i c$ where c is a constant then $\_y = \_xc$

Median - 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 number of observations in the dataset.

1. If the number 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 number of obsevations is even, then the median is the mean of the two middle observations in the ordered list, i.e. mean of $n2$ and $n2 + 1$ observation

Adding a constant

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

Multiplying a constant

I Let $y_i = x_i\ c$ where c is a constant then
new median = old median *c

Mode - The mode of a data set is its most frequently occurring value.

Steps to obtain mode

1. If no value occurs more than once, then the data set has no mode.
2. 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

Multiplying a constant

I Let $y_i = x_i\ c$ where c is a constant then
new mode = old mode *c

Measures of dispersion

I 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 referred to as

- measures of dispersion, or • measures of variation, or
- measures of spread.

Range - The range of a data set is the difference between its largest and smallest values.

The range of a data set is given by the formula Range = Max - Min where Max and Min denote the maximum and minimum observations, respectively.

Range sensitive to outliers

- Range is sensitive to outliers.
- Though the two datasets di_er only in one datapoint, we can see that this contributes to the value of Range significantly. This happens because the range takes into consideration only the Min and Max of the dataset.

Variance –

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

Population variance and 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 dataset has n observations. I The variance is computed using the following formulae I The denominator for computing population variance is N, the total number of observations.
I The denominator for computing sample variance is (n 1).
The reason for this will be clear in forthcoming courses on statistics.

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$ Population Variance

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

$$s = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n - 1}}$$ Sample Standard Deviation

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 *old variance.

Standard definition – The quantity 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.

Adding a constant

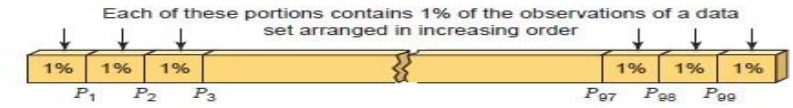
I Let $y_i = x_i + c$ where c is a constant then new variance = old variance.

Multiplying a constant

I Let $y_i = x_i c$ where c is a constant then
new variance = C2 _ old variance, (C2 is c square)

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 average of these values.
- Median is the 50th 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 average of the values in positions np and np + 1 is the sample 100p percentile.

Quartiles

Defnition

The sample 25th percentile is called the _rst quartile. The sample50th percentile is called the median or the second quartile. Thesample 75th percentile is called the third quartile.In other words, the quartiles break up a data set into four partswith about 25 percent of the data values being less than the first(lower) quartile, about 25 percent being between the _rst andsecond quartiles, about 25 percent being between the second and 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 of Median
- Q_3: Third Quartile or upper quartile
- Maximum

The Interquartile Range (IQR) - 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

Contingency table –

- 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: Gender versus use of smartphone

- A market research _rm is interested in _nding out whether ownership of a smartphone is associated with gender of a student. In other words, they want to _nd out whether more females own 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 smart phone or not.
- The categorical variables in this example are

Gender: Male, Female (2 categories)- Nominal variable

- ❖ Own a smartphone: Yes,
- ❖ No (2 categories)- Nominal variable

Example 2: Income versus use of smartphone

- A market research _rm is interested in _nding out whether ownership of a smartphone is associated with income of an individual. In other words, they want to find out whether income is associated with ownership of a smartphone.
- To answer this question, a group of 100 randomly picked individuals were surveyed about whether they owned a smart phone or not.
- The categorical variables in this example are

❖ Income: Low, Medium, High (3 categories) -Ordinal variable

- ❖ Own a smartphone: Yes, No (2 categories) - Nominal variable

Row relative frequencies

- What proportion of total participants own a smart phone?
- What proportion of female participants own a smart phone?

Row relative frequency: Divide each cell frequency in a row by its row total.

Column relative frequencies

- What proportion of total participants are female?
- What proportion of smart phone owners are females?

Column relative frequency: Divide each cell frequency in a column by its column total.

Association between two variables

Knowing information about one variable provides information

about the other variable.

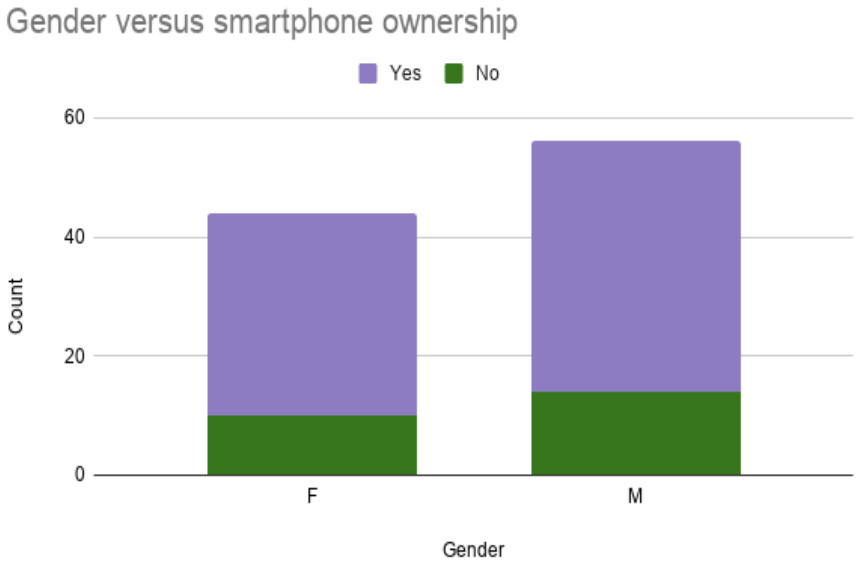
- To determine if two categorical variables are associated, we use the notion of relative row frequencies and relative column frequencies described earlier.

Association between two variables

- If the row relative frequencies (the column relative frequencies) are the same for all rows (columns) then we saythat the two variables are not associated with each other.
- If the row relative frequencies (the column relative frequencies) are different for some rows (some columns) then we say that the two variables are associated with each other.

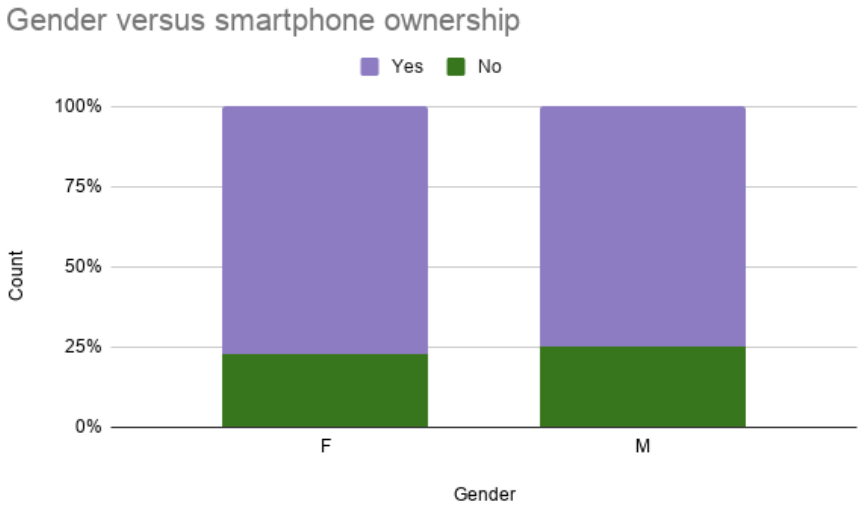
Stacked bar chart

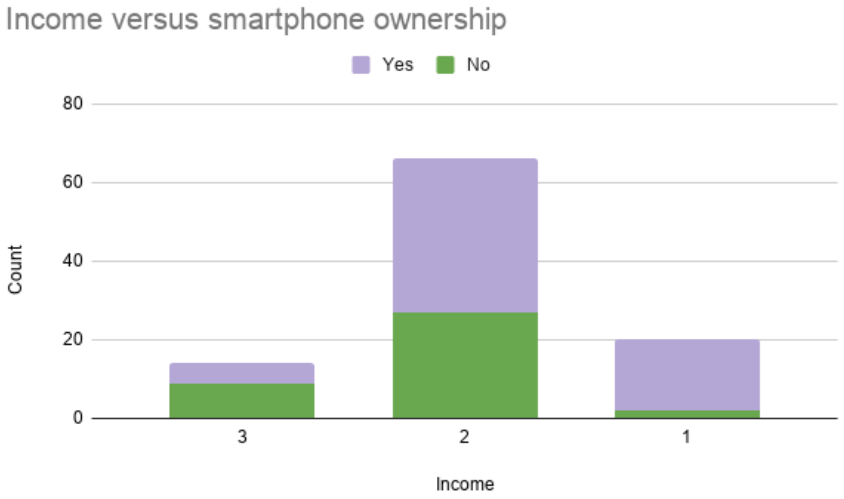
- Recall, a bar chart summarized the data for a categorical variable. It presented a graphical summary of the categorical variable under consideration, with the length of the bars representing the frequency of occurrence of a particular category.
- A stacked bar chart represents the counts for a particular 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: 100% Stacked bar chart

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





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

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

- To decide which variable to put on the x-axis and which to put on the 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.)

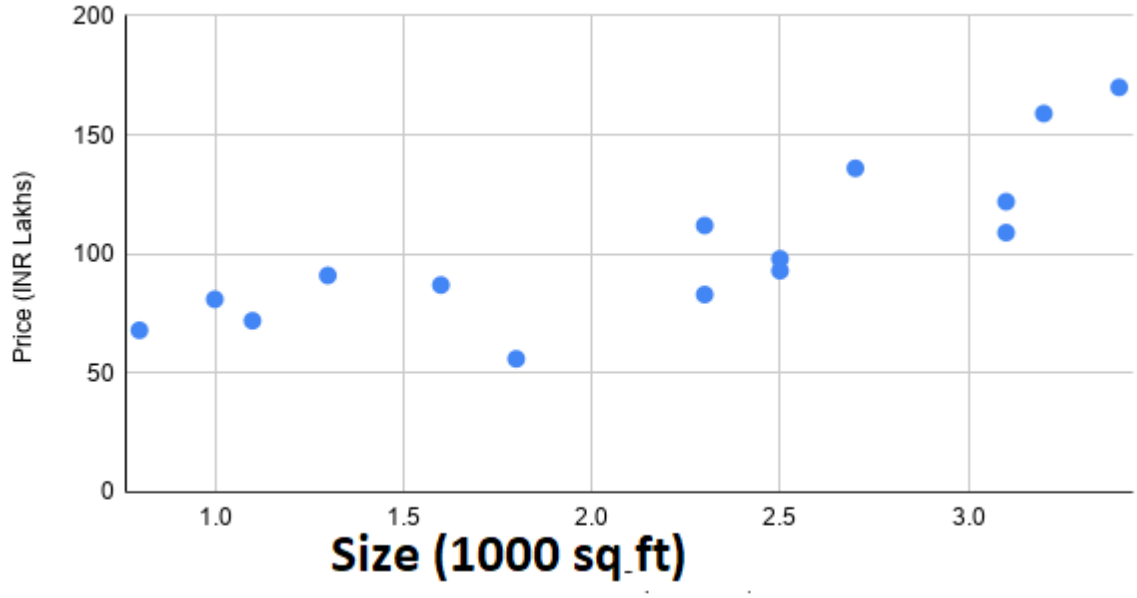
Example 1: Prices of homes

A real estate agent collected the prices of di_erent sizes of homes. He wanted to see what was the relationship between the price of a home and size of a home. In particular, he wanted to know if the prices of homes increased linearly with the size or in any other way? To answer the question, he collected data on 15 homes. The data he recorded was

	Size (100sq feet)	Price INR Lakh
1	0.1	
2	1.1	68
3	1.3	87
4	1.6	45
5	1.8	72
6	2.3	69
7	2.3	36
8	2.5	52
9	2.5	47
10	2.7	85
11	3.1	69
12	3.1	69
13	3.2	52
14	3.4	88
15	3.6	56
16	3.6	66

1. Size of a home measured in 1000 of square feet.
2. Price of a home measured in lakh of rupees.

Scatter plot -



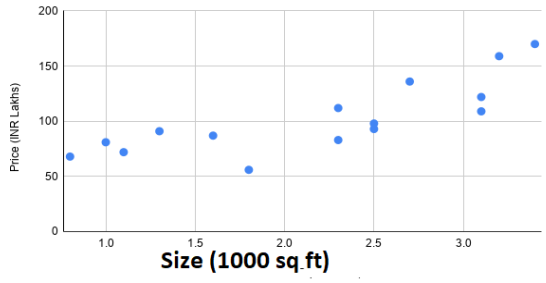
Describing association

When describing association between varaibles in a scatter plot, there are four key questions[1] that need to be answered

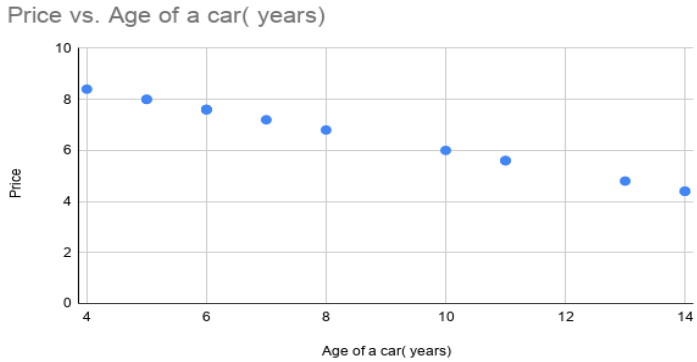
1. Direction: Does the pattern trend up, down, or both?
2. Curvature: Does the pattern appear to be linear or does it curve?
3. Variation: Are the points tightly clustered along the pattern?
4. Outliers: Did you find something unexpected?

Describing association: Direction

Does the pattern trend up, down, or both?



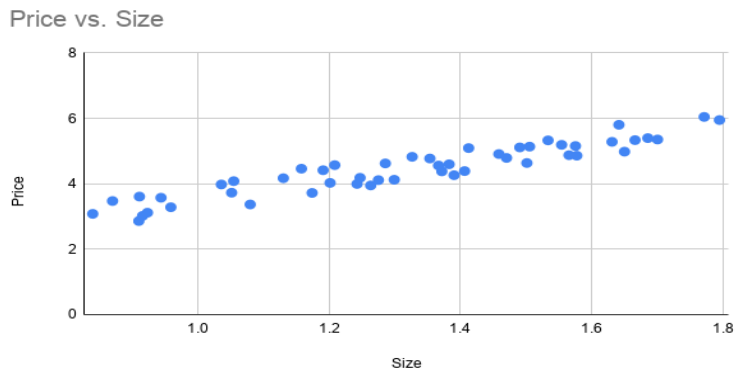
i) UP



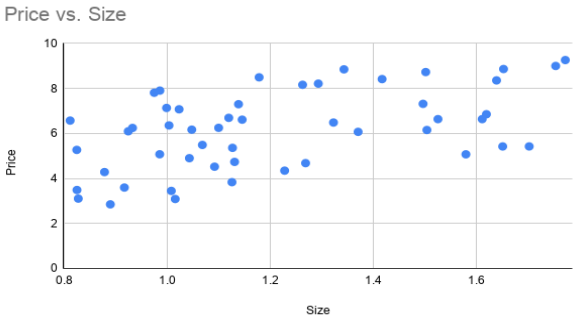
ii) Down

Describing association: Variation

Are the points tightly clustered along the pattern?

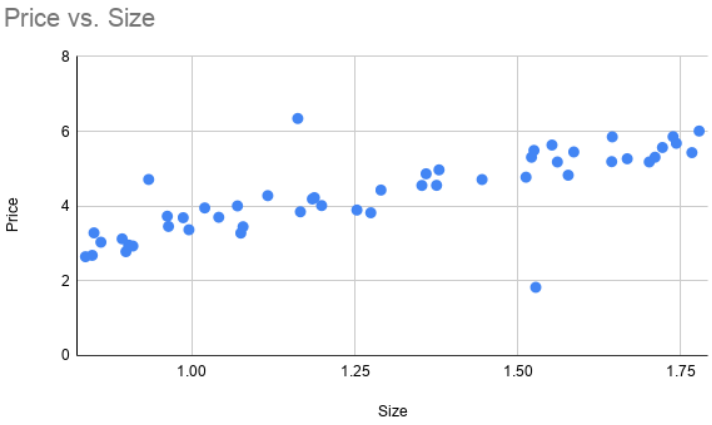


i) Tightly clustered



ii) Variable

Describing association: Outliers



Did you find something unexpected ?

Covariance

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

Key observation

I When large (small) values of x tend to be associated with large (small) values of y- the signs of the deviations, (x_i - _x)and (y_i - _y) will also tend to be same.
I When large (small) values of x tend to be associated with small (large) values of y- the signs of the deviations, (x_i - _x)and (y_i - _y) will also tend to be different.

Covariance

Definition

Let x_i denote the i th observation of variable x, and y_i denote thei th observation of variable y. Let (x_i ; y_i) be the i th pairedobservation 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: $Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

Covariance: Example 1

Age x	Height y	Deviation of x $(x_i - \bar{x})$	Deviation 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
				82

- Population covariance: $\frac{82}{5} = 16.4$
- Sample covariance: $\frac{82}{4} = 20.5$

Units of Covariance

- The size of the covariance, however, is di_cult to interpret because the covariance has units.
- The units of the covariance are those of the x-variable times those of the y-variable.

Correlation

- A more easily intepreted 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 standard deviations of x and y. The Pearson correlation
    - coefficient, r , between x and y is given by

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

Remark

The units of the standard deviations cancel out the units of covariance.

Remark

It can be shown that the correlation measure always lies between -1 and +1.

Correlation: Example 1

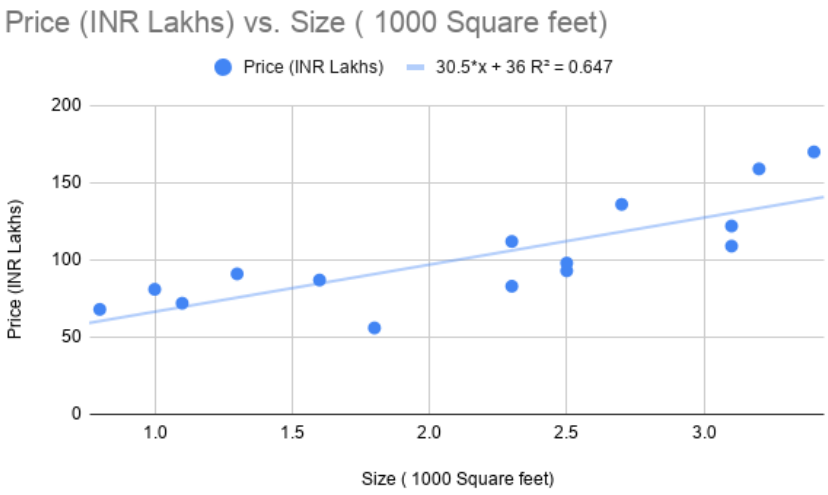
Age x	Height y	sq.Devn of x $(x_i - \bar{x})^2$	sq.Devn of y $(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	75	4	309.76	35.2
2	85	1	57.76	7.6
3	94	0	1.96	0
4	101	1	70.56	8.4
5	108	4	237.16	30.8
		10	677.2	82

- $s_x = 1.58,\ s_y = 13.01$
- $r = \frac{82}{\sqrt{10 \times 677.2}}$ OR $\frac{20.5}{1.58 \times 13.01} = 0.9964$

Summarizing the association with a line

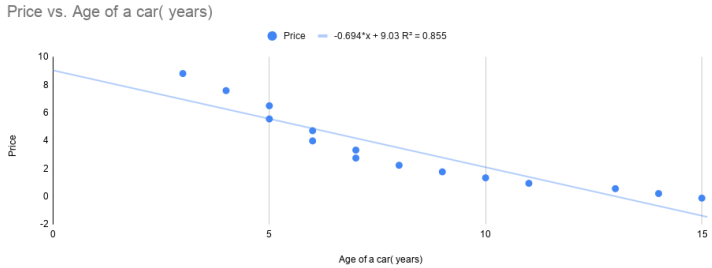
- The strength of linear association between the variables was measured using the measures of Covariance and Correlation.
- The linear association can be described using the equation of a line.

Example 1: Size versus Price of homes: Equation



Equation of the line: Price = 30:5 *Size + 36;
R_2 = 0:647; r = 0:804

Example 2: Age versus Price of cars: Equation



Equation of the line: Price = -0:694 * Age + 9:03;
R_2 = 0:855; r = -0:9247

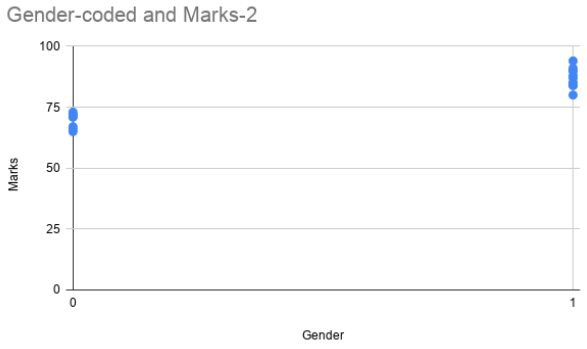
- Understand the association between a categorical variable and numerical variable.
- Assume the categorical variable has two categories (dichotomous).

Example 1: Gender versus marks DATA

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

	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

Example 1: Scatter plot



Point Bi-serial Correlation Coe_cient

Let X be a numerical variable and Y be a categorical variable with two categories (a dichotomous variable).

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 Y_0 and 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.

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