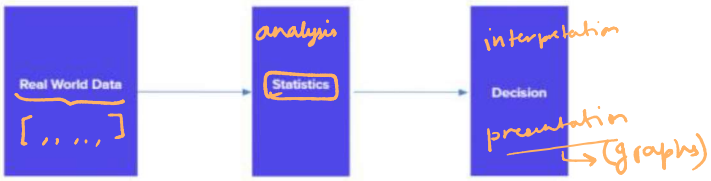
Revision on Intro to statistics and visualisation + Measures of central tendency and dispersion 17oct22

16 October 2022 15:37

Statistic

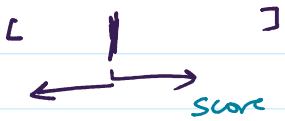
What is statistics?

Statistics is a branch of mathematics involving data collection, analysis, interpretation, and presentation. It takes data and turns it into insights we can use to make decisions.

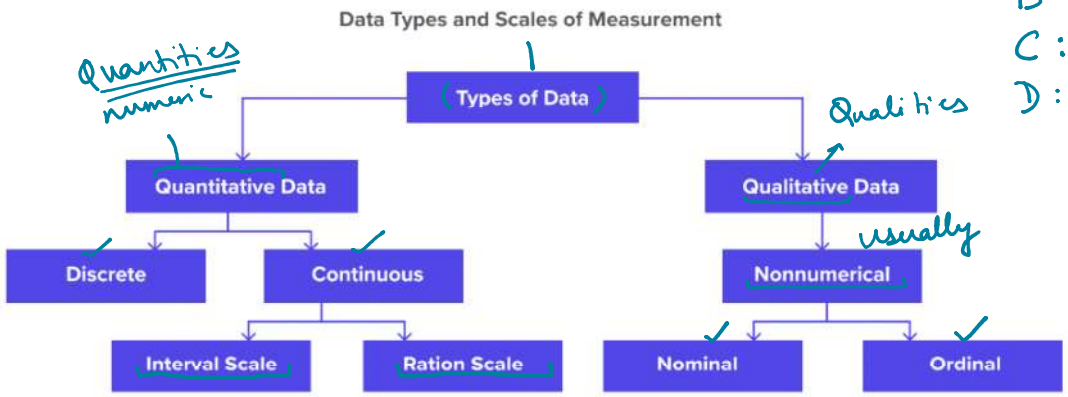


[a fixed set of data which we will be analysing.]

seeing & getting a feel of our dataset.



eg. A: (35, 40, 49, 48)
B: (4, 3, 1, 2)
C: (M, M, F, F)
D: (abc, def, ghi, jkl)



Quantities numeric

Qualities

usually

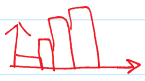
Qualitative data: usually non-numeric

Nominal

[all the levels of data are non-rankable or non-orderable]

Ordinal

[It can be compared within the levels, or the data is rankable]



eg. Excellent ✓
Great
Above avg
Avg. ✓
Below avg.
Poor.

Examples of Nominal Data

Are you married?

- Yes
- No

What languages do you speak?

- English
- French
- German
- Spanish

The first example has two categories. The other example has four categories.

Examples of Nominal Data

Are you married?

- Yes
- No

What languages do you speak?

- English
- French
- German
- Spanish

eg. Excellent ✓
Great
Above avg
Avg. ✓
Below avg.
Poor.

The first example has two categories. The other example has four categories.

Examples of Ordinal Data

at what level ~~What~~ languages do you speak?

- 1 - Elementary
- 2 - High School
- 3 - Undergraduate
- 4 - Graduate

it is rankable and measurable within the levels.

eg.	Marks (Quan.)	Rank (Qualitative)
	80%	1
	50%	2
	49%	3
	45%	4

exact difference on a scale is calculable while in Qualitative data is only comparable within itself.

A ✓ cont. B disc.

A –	90% math	80% SST	85% english.

dataset represents ₹spent per order in a restaurant
(1036, 2053.8, 200, 6000.00, ...)
(0 to ∞)

Numerical Data – Discrete data

Discrete Data: Data is said to be discrete if its values are distinct and separate. In other words: Data is said to be discrete data if the data can only take on certain values. This data type **cannot be measured, but it can be counted.**

Countable or countably infinite.

n=10 possible denominator

dataset 1 : {1, 2, 5, 10, 20, 50, 100, 200, 500, 2000} countable

dataset 2 : {50, 100, 150, 200, 250, ... ∞} countably infinite

If I have only ₹50 notes

Numerical Data – Continuous Data

Since continuous data represent measurements, its values can be measured but not counted.
A person's height is something you could represent on a real number line by utilising intervals.

dataset 3 : (0, ∞) if I write a cheque

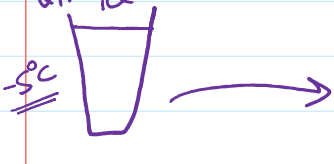
continuous data – infinite or uncountably infinite.

Interval vs. Ratio data.

true zero. means that the value is zero i.e. nothing is present
eg. 0kg rice means no rice

non-true zero. Zero does not mean the absence of characteristic.
eg. 0°c ≠ no heat. It doesn't mean absence of heat
Zero here is just a value in a measurement scale.

Q1. ice 0°c → 50°c water



A

from 0 to 50 is ice ∞ times colder that this water?

taking a ratio is not possible here

B

Q2. Is 2kg rice double heavier than 1kg rice? yes!
taking ratio is possible.

Interval	Ratio
★ True zero doesn't exist	★ True zero exists
★ Ratio measurement not possible	★ Ratio measurement is possible
★ Values below zero have a meaning	★ Values below zero don't have any meaning.

Graphs:
- Line Graphs ✓
- Bar Graphs ✓ (diff)
- Histograms ✓
- Pie charts ✓
- Scatterplot ✓ (imp.)

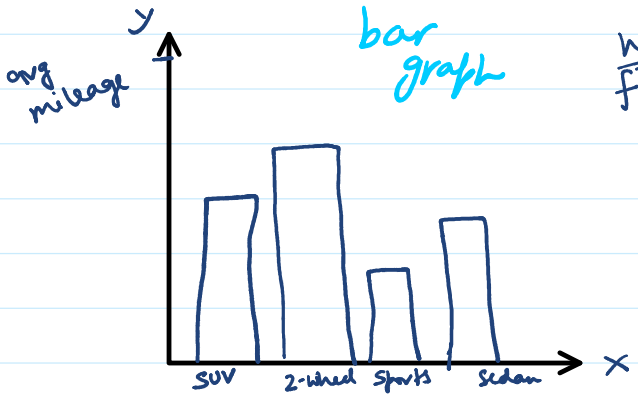
Graphs aid in a visual representation of our dataset.
→ humans are not good at reading large sets of nos. but can easily interpret figures.

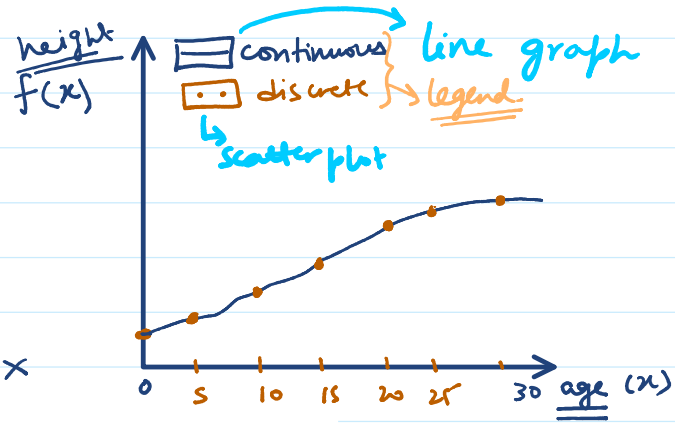
Parts of a Graph : Axes : x-axis : horizontal
y-axis : vertical one

value at x-axis is known as "abscissa"
value at y-axis is known as "ordinate"

We can represent 2-variables together and how one changes with the other.

usually y-axis represents a function of the variable on x-axis.
x & y
x & f(x). x-axis may sometimes be categorical eg:





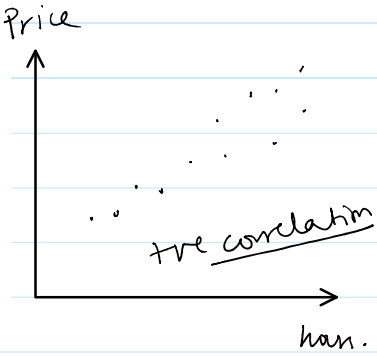
Scatterplot Correlation

Correlation is a statistical measure of the relationship between the two variables' relative movements. The points will fall along a line or curve if the variables are correlated. The better the correlation, the closer the points will touch the line. This cause examination tool is considered one of the seven essential quality tools.

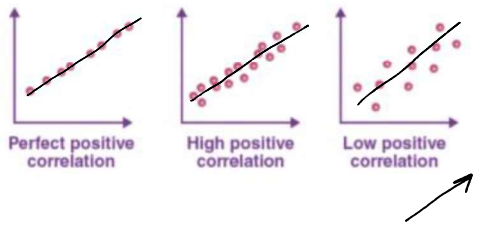
Types of correlation

The scatter plot explains the correlation between two attributes or variables. It represents how closely the two variables are connected. There can be three such situations to see the relation between the two variables –

1. Positive Correlation
2. Negative Correlation
3. No Correlation



+ve correlation happens when one variable moves in the same direction as the second variable



Similarly negative and 0 correlation

-ve correlation happens when one variable moves in the opposite direction of the other

---

Grouped data: It's basically a frequency type of data where the values could be discrete or in an interval format:

DISCRETE

In a quiz, the marks obtained by 20 students out of 30 are given as:
12,15,15,29,30,21,30,30,15,17,19,15,20,20,16,21,23,24,23,21

HW

X → discrete

f

Marks obtained in quiz	
12	
15	
16	
17	
19	
20	
21	
23	
24	
29	
30	
Total	

Frequency Distribution – Continuous (Example)

The heights of 50 students, measured to the nearest centimetres, have been found to be as follows:
161, 150, 154, 165, 168, 161, 154, 162, 150, 151, 162, 164, 171, 165, 158, 154, 156, 172, 160, 170, 153, 159, 161, 170, 162, 165, 166, 168, 165, 164, 154, 152, 153, 156, 158, 162, 160, 161, 173, 166, 161, 159, 162, 167, 168, 159, 158, 153, 154, 159

X- continuous

Frequency bin	Frequency
(150-154)	12
155-159	9
160-164	14
165-169	10
170-174	5

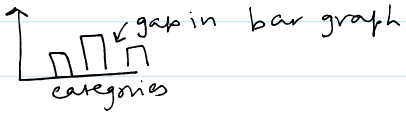
Generate data in Excel and now use these graphs:

- Line Graphs
- Bar Graphs
- Histograms
- Pie charts
- Scatterplot

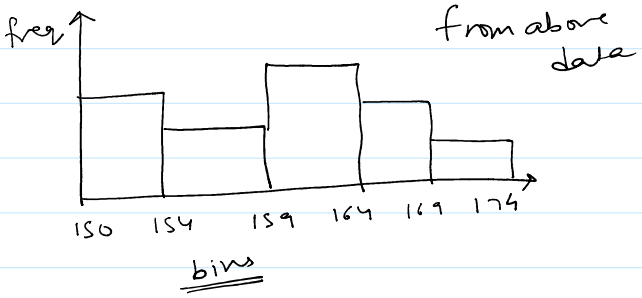
Histogram vs. Bar graph.

→ Bar graphs have a categorical x-axis.

→ In bar graph the bars have space within



Histogram is used to represent continuous frequency type distribution.



Categories of Statistics

Statistics is majorly categorised into two types:

1. **Descriptive statistics:** Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures.
For example;- reporting a general trend about the collection of people in a city using the internet or television. This reporting would come under descriptive analytics.
2. **Inferential statistics:** With the help of inferential statistics, we can use data from a sample to extrapolate conclusions about the population. It enables us to make claims beyond the scope of the facts or data at hand.
For example, deriving inferences from our report on the trends in people using the internet or TV, we can infer that people below 40 years of age use the internet more than TV. In contrast, people above 40 years of age use TV more.

Measures of central tendency:
Mean, median and mode

Central tendency is a number which represents your set of data and has certain characteristics to measure centrality or other desirable properties.

AVERAGE : {Mean, Median, Mode, Harmonic mean, geometric mean}
Positional

Mean

Mean is the sum of all the components in a group or collection, divided by the number of components. It is also known as average.

The formula to calculate the mean for ungrouped data to represent it as the measure is given as,
For a set of observations: Mean = Sum of the terms/Number of terms
For a set of grouped data: Mean, $\bar{x} = \Sigma fx/\Sigma f$
where,

- $\bar{x}$ = the mean Value of the set of given data.
- f = frequency of each observation
- x = Value of each observation (in case we go for a frequency bin distribution like we did for the continuous data

Mean

Mean is the sum of all the components in a group or collection, divided by the number of components. It is also known as average.

The formula to calculate the mean for ungrouped data to represent it as the measure is given as,
For a set of observations: Mean = Sum of the terms/Number of terms
For a set of grouped data: Mean, $\bar{x} = \Sigma fx/\Sigma f$
where,

- $\bar{x}$ = the mean Value of the set of given data.
- f = frequency of each observation
- x = Value of each observation (in case we go for a frequency bin distribution like we did for the continuous data in the last class, x will take the mid-interval value)

Physical Significance: If all the values in the dataset were to be the same and still the overall sum was to remain intact then that same value is the arithmetic mean.

eg. {7, 6, 5, 2} $\Sigma = 20$
{5, 5, 5, 5} $\Sigma = 20$

A.M. is equally redistributed data.

Median

The value of the provided data-set that is the middle-most observation after the data are arranged in ascending order is known as the median, another measure of central tendency. The median is less impacted by outliers and skewed data, which is a primary benefit of using it as a central tendency. Using the median formula, we can determine the median for various data types, grouped data, and ungrouped data.

Median is a positional measure.

Def^: Any number which can divide your dataset into two equal halves is your median.

eg. {14, 16, 16, 20, 34, 11, 7}

what is the median value in this data

{34, 20, 16, 16, 14, 11, 7}

A: 20
B: 16
C: 7
D: 14

median: $\left\{\frac{n+1}{2}^{th} \text{ value of your dataset.}\right\}$

eg. {5, 9, 11, 17, 2, 4, 8, 1}

A: 8
B: 6.5 ✓
C: 5
D: 7 ✓

order {1, 2, 4, 5, 8, 9, 11, 17}

$\left\{\frac{n+1}{2}^{th} \text{ value}\right\} \rightarrow \frac{8+1}{2} = \frac{9}{2} = 4.5^{th}$ value

ie. in b/w 4 & 5

in general we take a.m. of 4^{th} & 5^{th}

Calculating Median

Calculating Median

For ungrouped data:

- For odd number of observations,

Median = [(n + 1)/2]th term

- For even number of observations,

Median = [(n/2)th term + ((n/2) + 1)th term]/2

For grouped data:

Median = l + [((n/2) - c)/f] × h

where,

l = Lower limit of the median class

c = Cumulative frequency of prev. class

h = Class size

n = Number of observations

Median class = Class where n/2 lies

ie. in b/w 4th & 5th

in general we take a.m. of 4th & 5th value.

eg.

LL	UL	freq.	c.f.	
1	5	5	5	
6	10	18	23	
11	15	15	38	
16	20	23	61	←
21	25	16	77	
		77		

39th value will be median

$$16 + \left[\frac{38.5 - 38}{23}\right] 5 = \left(\frac{0.5}{23} \times 5\right) + 16 = 16.10$$

Mode

The value in your dataset which repeats the most frequently.

- One of the measures of central tendency is the mode, which is defined as the value that occurs the most frequently in the supplied data or, more specifically, as the observation with the highest frequency. Using the mode formulas provided below, one can determine the mode for grouped or ungrouped data.
- The most frequent observation in the data set serves as the mode for ungrouped data.
- Data grouping mode:

$$L + h \frac{(f_m - f_1)}{(f_m - f_1) + (f_m - f_2)}$$

grouped formula

f1 fm f2

- where L is the modal class lower limit.
- FM is the frequency of the modal class, and h is the size of the class interval.
- The frequency of the class that comes before the modal class is f1, and the frequency of the class that follows the modal class is f2. Do not worry about this terminology; we will understand them all.
- modal class: class w/ max frequency.

The following table of grouped data represents the weight (in pounds) of 100 computer towers. Calculate the mean weight for a computer.

Calculate all 3

Weight (pounds)	Number of Computers
[3 - 5)	8
[5 - 7)	25
[7 - 9)	45 ← modal class
[9 - 11)	18
[11 - 13)	4

to calculate mode.

$$7 + 2 \times \left\{\frac{45 - 25}{(45 - 25) + (45 - 18)}\right\}$$

$$= 7 + 2 \times \left\{\frac{20}{20 + 27}\right\}$$

$$= 7 + \frac{40}{47} = 7.85$$

41 mode ✓

HW calculate mean & median.

Measure of position: (75 mins)

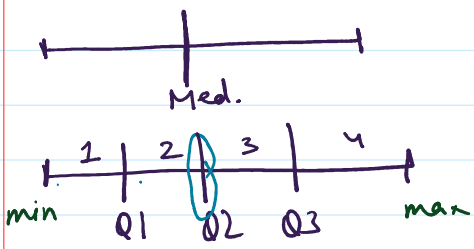
Provide an overview on:
Deciles
Quartiles
Percentiles
Boxplot
Detecting Outlier through Box Plot
Choosing right measures of Central Tendency

median is a classic measure of position:

- median : ÷ into 2 parts
- quartile : ÷ into 4 parts
- decile : ÷ into 10 parts
- percentile : ÷ into 100 parts.

remember the median formula: $\frac{1\cdot(n+1)}{2}^{th}$ value

for quartile: $\frac{1(n+1)}{4}^{th}$ 1st quartile $\frac{2(n+1)}{4}^{th}$ 2nd quartile

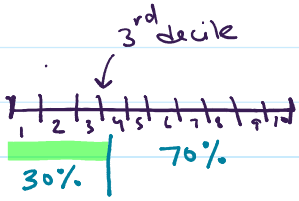


$\frac{3(n+1)}{4}$ value 3rd quartile.

99th %ile is last %ile.

Decile

3rd decile : $\frac{3(n+1)}{10}^{th}$ value



Percentile: d^{th} %ile : $\frac{d(n+1)}{100}^{th}$ value.

Interquartile range : Q3 − Q1

Percentile – Example: 1

Example 1: The scores obtained by 10 students are 38, 47, 49, 58, 60, 65, 70, 79, 80, 92. Using the percentile formula, calculate the percentile for score 70.

Solution:

Given:

Scores obtained by students are 38, 47, 49, 58, 60, 65, 70, 79, 80, 92

Number of scores below 70 = 6

Using percentile formula,

Percentile = (Number of Values Below "x" / Total Number of Values) × 100

Percentile of 70

= (6/10) × 100

= 0.6 × 100 = 60

Therefore, the percentile for score 70 = 60

Decile

A decile is a quantile that uses 9 data points to divide the dataset into ten equal subsections. Each part of the sorted data represents one-tenth of the initial sample or population. Decile aids in arranging massive volumes of data in ascending or descending order. The scale used for this ordering ranges from 1 to 10, with each value after that denoting an increase of 10 percentage points.

Decile – Example: 2

Find the 7th decile for the following frequency distribution table.

Class	Frequency
10 - 20	15
20 - 30	10
30 - 40	12
40 - 50	8
50 - 60	7
60 - 70	18
70 - 80	5
80 - 90	25

Decile – Example: 2 (Solution)

Class	Frequency	Cumulative Frequency (cf)
10 - 20	15	15
20 - 30	10	25
30 - 40	12	37
40 - 50	8	45
50 - 60	7	52
60 - 70	18	70
70 - 80	5	75
80 - 90	25	100

Quartile: Example-1

Calculate the median, lower quartile, upper quartile, and interquartile range of the following data set of values: 20, 19, 21, 22, 23, 24, 25, 27, 26

Quartile: Example-1 (Solution)

Arranging the values in ascending order: 19, 20, 21, 22, 23, 24, 25, 26, 27
Putting the values in the formulas above, we get,
Median(Q2) = 5th Term = 23
Lower Quartile (Q1) = Mean of 2nd and 3rd term = (20 + 21)/2 = 20.5
Upper Quartile(Q3) = Mean of 7th and 8th term = (25 + 26)/2 = 25.5
IQR = Upper Quartile–Lower Quartile
IQR = 25.5 – 20.5
IQR = 5

Q1 Q2
19, 20 | 21, 22, (23), 24,
25 | 26, 27
Q3

$Q2 = \frac{9+1}{2}^{th} = 5^{th}$ value = 23

$Q1 \quad 1 \cdot \frac{(n+1)}{4}^{th} = \frac{10}{4} = 2.5^{th}$ value
i.e. b/w 20 & 21 = 20.5

$Q3 \quad 3 \cdot \frac{(n+1)}{4}^{th} = \frac{30}{4} = 7.5^{th}$
25.5

Intro to Boxplot

Boxplot is a

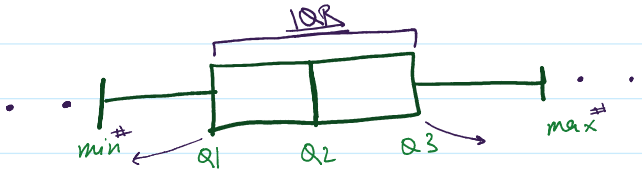
Intro to Boxplot

A box plot or box plot graphically represents groups of numerical data through their quartiles in descriptive statistics. Box-and-whisker plots and box-and-whisker diagrams refer to box plots that can additionally feature vertical lines (whiskers) demonstrating variability outside the upper and lower quartiles.

Using a five-number summary (the minimum, first quartile (Q1), median, third quartile (Q3), and "maximum"), boxplots are a common approach to depict data distribution.

Boxplot is a 5 point plot

→ min , Q1 , Q2 , Q3 , max. calculate all 5 & plot



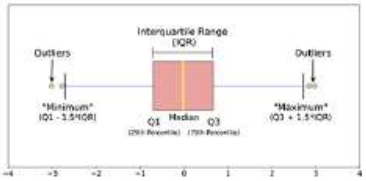
→ min & max after removal of outliers.

Outliers: Basically any point lying outside the range

✓ Q1 − 1.5 (IQR) , ✓ Q3 + 1.5 (IQR) is outlier

Detecting Outlier Through Boxplot

Box plot is used to detect **outliers** easily. It can also tell us if your data is symmetrical, how tightly your information is grouped, and if and how your data is skewed.



- 25th to 75th percentiles are the **interquartile range** (IQR). IQR reveals how widely distributed the middle values are.
- **Outliers**: "maximum": Q3 + 1.5*IQR "minimum": Q1 -1.5*IQR (shown as green circles) An observation point that is far from other observations is referred to as an outlier in statistics.

Central Tendency Measure	Pros	Cons
Mean	Sensitive as it takes all data values into account (reliable)	Biased output if outliers/ extreme values exist in the data set
Median	Not affected by extreme values	Less sensitive than Mean as it only focuses on giving out the middle data point irrespective of how far the other values are from the middle Needs the data to be arranged in the ascending order before computing
Mode	Not affected by extreme values and can be used with non-numerical data	There may be more than one mode or no mode at all and it may not reflect data summary accurately