28/10/21

Basic Statistics

What is statistics ?

It is measurable

It means summary of data or inference of data rather than entire data.

What is data ?

Structured data - excel, table, csv i.e tabular form

We know the maximum info we can get from the data.

Unstructured data - Video, audio, email

Semi structured data - Dictionary or JSON files → this can be both unstructured or structured

i.e key : value pairs

In machine Learning we mainly work with structured data

With NLP we learn to convert unstructured text data to structured data.

Different data types

Number / numerical data
- continuous (ex any measurements temp, height)
- discrete (ex no of students in the class)

Categorical data → Ordinal data
Categorical data → Nominal / categorical / Non ordinal data.

→ When there is no sense of numbers. That data is called categorical data.

→ these are not number, some sort of category like yes or NO, M or F

→ There are few categorical data that can be measured.

Ordinal data

→ These are categorical data that can be ordered.

ex: Movie reviews in star, food experiences, Wealth of a family.

exercise
Volume of corflakes - continuous
Speed of car - continuous
population of a town - discrete
length of crocodile - continuous
No of mathes in a box = discrete
temp of oven - continuous
# of t-shirts - discrete
No of goals - discreet

Measures in Statistics

Descriptive Statistics : Those statistics that contain some description

of the data.

ex: Rate of employment is 75%

Ratio of female is 1:2

→ It gives general description of data

① Min of data
② Max of data
③ Count of data - no of data points
④ Median of data
⑤ Mean of data
⑥ Var of data
⑦ Standard deviation of the data
⑧ qu

a) Mean ⇒ average

$$\frac{\text{sum of data}}{\text{count or number of data points}}$$

b) Median : middle value of ordered data (ascending / descending order)

**

If mean and median are same then its uniformly distributed otherwise data is not uniformly distributed.

For non uniformly distributed data, if we only look at the mean we get wrong data.

It is less trustworthy.

However Median is not affected by Outlier

a data point that's very far away from the rest of the data.

> Median is more trustworthy because it is not affected by outlier.
>
> Median is more robust then mean
>
> Mean is affected by outlier ∴ not trustworthy

** If mean and median are nearly equal then there is no outlier in the data.

c) Variance / dispersion / volatility

→ This is the variation in the data

→ When most of data is close to centre, there is less variance or when data is away from centre there is more variance.

Calculate Variance

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

→ Robust version of Variance

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

↓

We use median instead of mean.

ex: 1, 3, 5, 7, 9

$\bar{x} = 5$

$$\text{Var} = \frac{(1-5)^2 + (3-5)^2 + (5-5)^2 + (7-5)^2 + (9-5)^2}{5-1}$$

$$= \frac{4^2 + 2^2 + 0^2 + 2^2 + 4^2}{4}$$

$$= \frac{16+4+0+4+16}{4} = \frac{40}{4} = 10$$

$$\rightarrow \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

$$\sum_{i=1}^{n} x_i = \bar{x}n$$

data : $x_1, x_2, x_3, \ldots x_n \rightarrow$ mean is m

then mean of $x_1+c, x_2+c, x_3+c, \ldots, x_n+c \rightarrow$ $m+c$

Ex : 1, 2, 3, 4, 5

Mean : $\frac{15}{5} = 3$

Var : $\frac{2^2 + 1^2 + 0 + 1^2 + 2^2}{4} = \frac{10}{4} = 5/2 = 2.5$

2, 4, 6, 8, 10

Mean : $30/5 = 6$ (3×2=6)

Var : $\frac{(6-2)^2 + (4-6)^2 + (6-6)^2 + (8-6)^2 + (10-6)^2}{(5-1)=4}$

$= \frac{16+4+4+16}{4} = 40/4 = 10$

When data is doubled, the new variance is 4 times of old variance.

Standard deviation

$$SD = \sqrt{\text{Variance}}$$

Mode : The data point that occurs maximum number of times in the data.

→ discrete or categorical data

Range : the value between min and max.

quantile :

Tool to detect outliers in the data using visualisation

90 percentile ⇒ means more than 90% and less than 10% of the people.

25% quantile ⇒ The number that is more than 25% of the data and less than 75% of the data.

50% quantile ⇒ Median i.e 50% of data is less than this value and 50% of data is more than this value.

doubt

ex: 101, 102, 103, 104, ---- , 200

25% = 1/4

25th data = 125

26th data = 126

∴ 25 quantile = 125 + 126 = 125.5

50% = 1/2 = 150 + 151 = 150.5

75% = 175.5

ex: 101, 102, 103, ... , 199

25% quantile: 99 × 1/4 = 24.75 ≃ 25

= 125

50% quantile: 99 × 1/2 = 49.5 ≃ 50 = 150

75% quantile: 99 × 3/4 = 74.75 ≃ 75 = 175

09/10/21

Boxplot : Its a tool to detect an outlier

→ data

→ Calculate 25% and 75% quantile

1st quantile 3rd quantile

→ 50% is the 2nd quantile

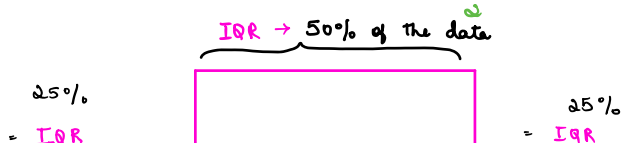
ex: 101, 102, ... , 200

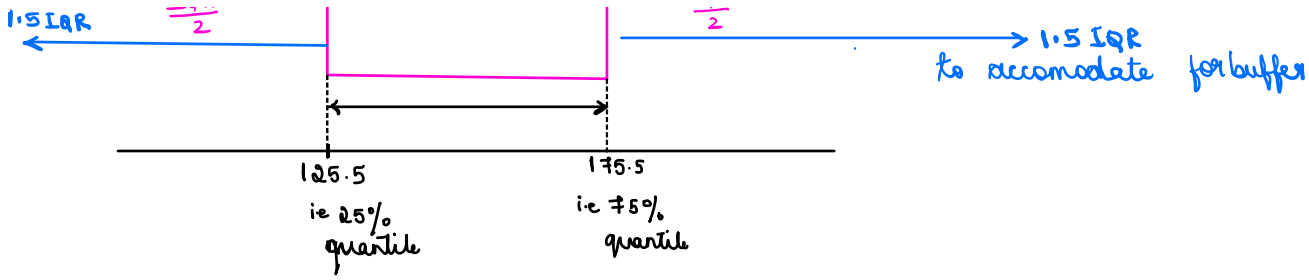
25% quantile | 1st quantile = 100 × 1/4 = 25

= (125 + 126)/2 = 125.5

75% quantile | 3rd quantile = 100 × 3/4 = 75

= (175 + 176)/2 = 175.5





IQR → Inter quartile range. Its the distance between 1st and 3rd quartile.
→ It has 50% of the data

on left & right we should have allocated only $\frac{IQR}{2}$ but to give buffer space to accomodate non-uniform data we give 3/2 IQR

outlier

Any data outside of this buffer space ie $\frac{3}{2}$ IQR is considered to be an outlier.

ex: 0, 101, 102, 103, ... , 198, 199, 300

Count = 101

25% quantile / 1st quartile = 101 x 1/4 = 25.25 ≃ 26

= 125

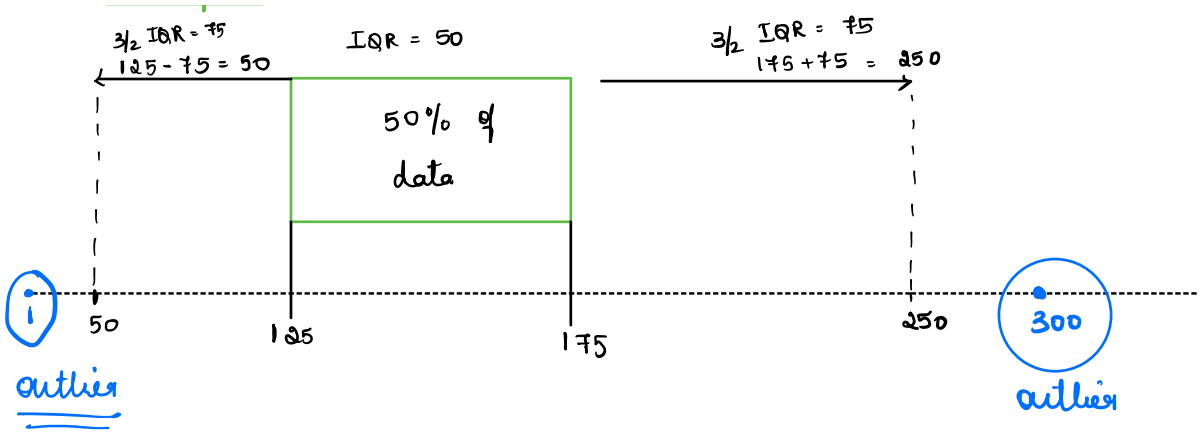
75% quantile / 3rd quartile = 101 x 3/4 = 75.75 ≃ 76

= 175

IQR = 175 - 125 = 50

$\frac{3}{2}$ IQR = 50 x $\frac{3}{2}$ = 75

Box plot


3/2 IQR = 75
125 - 75 = 50
IQR = 50
50% of data
3/2 IQR = 75
175 + 75 = 250
50
125
175
250
300
outlier
outlier

Python