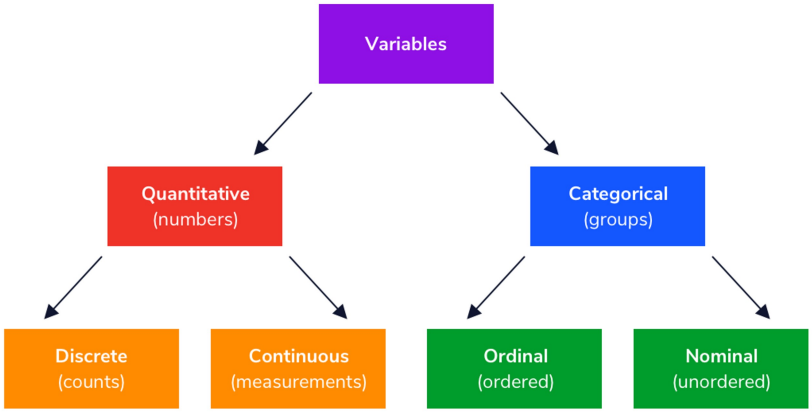
Descriptive Statistics

04 March 2024 10:27



There are 2 types of variables →

1→ Numerical / Quantitative → They represent quantities and are often the things that can be counted or measured.

2→ Categorical / Qualitative.

Quantitative Variables

* Discrete Variables → Discrete variables represent counts and are represented in whole numbers.
Ex→ Employees in a company, customers etc.

* Continuous Variables → The variables that make sense as parts of a whole. They are represented with decimal numbers.
Ex→ Height, weight, profits etc.

Categorical Variables →

We can group or separate our observations based on the values / categories.

* Ordinal variables→ When the groups have specific order or ranking.

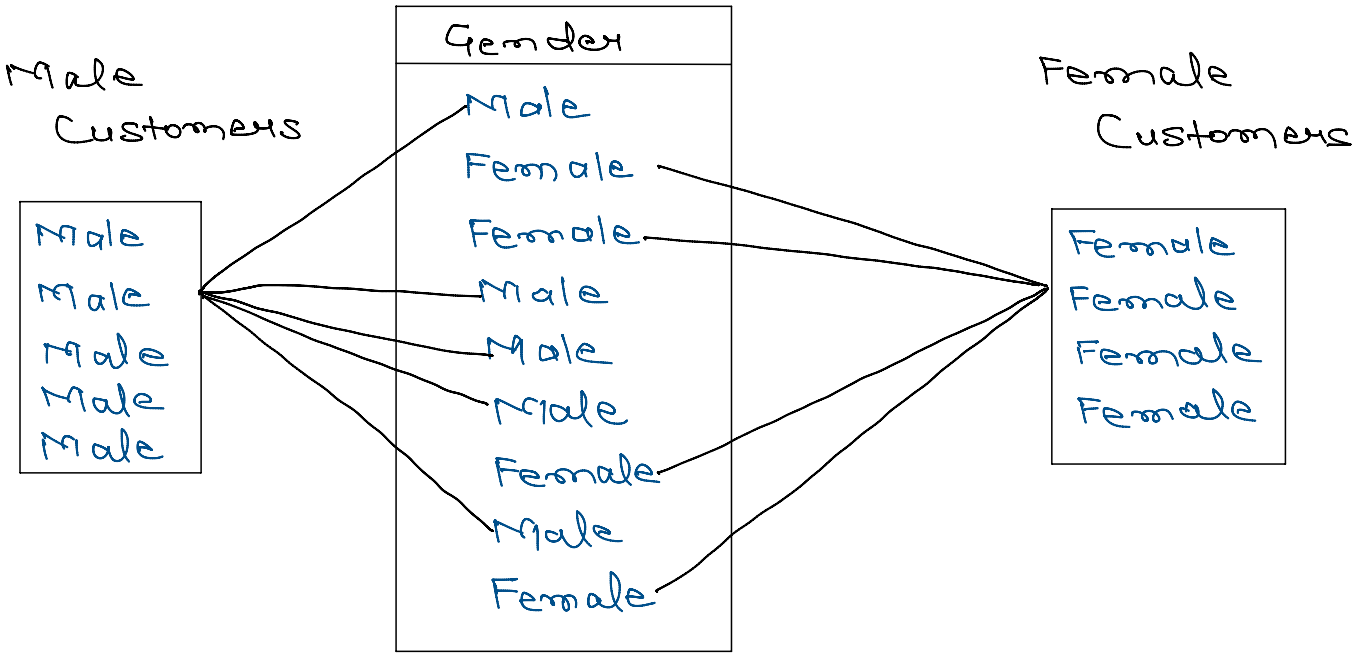
order or ranking.

Opinions
Neutral
Agree
Strongly Agree
Strongly Disagree
Agree
Disagree
Neutral
Strongly Agree
Strongly Disagree

This column values have ranking or order

strongly disagree < disagree < neutral < agree < strongly agree

* Normal variables → When the groups or categories have no relational order.



The number of possible values for a nominal variable can be quite large such as **Name** or **Emails**.

Measures of Central Tendency →

It is defined as a single value that represents

Measures of Central Tendency

It is defined as a single value that represents the entire data. It is very helpful while dealing with large data.

* **Mean:** It is the average of the data.

$$\text{Mean}(\bar{x}) = \frac{x_1 + x_2 + x_3 + x_4 + \cdots + x_n}{n}$$

Ex: 29, 49, 42, 43

$$\frac{29+49+42+43}{4} = \frac{163}{4} = 40.75$$

Mean is very frequently used measure of Central tendency but it is not reliable in case of outliers.

Let's see that with an example →

The average of 29, 49, 42, 43 is 40.75

Now let's add another number to this →

29, 49, 42, 43, 50

↓

Average = 213/5 = 42.6

Now instead of taking 50 as 5th number if we take **250** (an outlier).

29, 49, 42, 43, 250

↓

Average: 413/5 = **82.6**

Can we still say the mean (82.6) is representing our data? **No.**

So when we have outliers in our data then it fails to represent the **centre** of the data.

This is when we move on to our next measure of central tendency → **Median**.

* **Median:** Median is the mid-value of the sorted data.

Ex 1: 29, 49, 42, 43, 50

(i) 29, 42, 43, 49, 50

↑ Median

Now if we replace 50 with 250 then,

29, 42, 43, 49, 250

↑ Median

Note: So if notice, the mean of the data changed from **42.6** to **82.6** when we replaced 50 with an outlier (250) whereas on the other hand, the median value for the given data was not affected because of the outlier.

Hence, we would prefer using **Median** over **Mean** for data with extreme values.

* **Mode:** Mode is the most frequently occuring value. We generally use this measure in the case of categorical data.

Gender
Male
Female
Female
Male
Male
Female
Male
Female
Female
Female

Most Customers:

Female

Purpose of Loan
Home
Car
Personal
Home
Home
Car
Personal
Home
Car
Home

Most frequent loan taken by customers:

Home Loan

These measures of central tendency are a very good starting point to understand. But they don't tell the full story. So for that we will move on to the next topic.

Measures of Spread / Dispersion:

1. Variance
2. Standard Deviation
3. InterQuartile Range.

Let's understand how these are important:

Data A = -4, -2, 0, 2, 4

Data B = -400, -200, 0, 200, 400

So we just present the mean or median of both the data then they are same (0). But we are missing out on meaningful difference b/w both the data.

This is where the spread of the data comes into play.

1. Variance:

Suppose you have invested in two different stocks, Stock A and Stock B, over a period of time, and you want to assess their riskiness using variance.

Here's a hypothetical scenario:

- Stock A: Returns over five days: 2%, 3%, -1%, 4%, 0%
- Stock B: Returns over five days: 1%, 2%, -2%, 3%, 1%

To calculate the variance of each stock's returns, we'll follow these steps:

1. Calculate the mean return for each stock.
2. Find the squared difference between each return and the mean.
3. Calculate the average of these squared differences.

Let's compute the variance for both stocks:

Stock A:

- **Mean return** = (2% + 3% - 1% + 4% + 0%) / 5 = 1.6%
- **Squared differences from the mean:**
  - $(2\% - 1.6\%)^2 = 0.0256$
  - $(3\% - 1.6\%)^2 = 0.1936$
  - $(-1\% - 1.6\%)^2 = 6.76$
  - $(4\% - 1.6\%)^2 = 4.84$
  - $(0\% - 1.6\%)^2 = 2.56$
- **Variance** = (0.0256 + 0.1936 + 6.76 + 4.84 + 2.56) / 5 ≈ 2.78%

Stock B:

- **Mean return** = (1% + 2% - 2% + 3% + 1%) / 5 = 1%
- **Squared differences from the mean:**
  - $(1\% - 1\%)^2 = 0$
  - $(2\% - 1\%)^2 = 0.01$
  - $(-2\% - 1\%)^2 = 9$
  - $(3\% - 1\%)^2 = 4$
  - $(1\% - 1\%)^2 = 0$
- **Variance** = (0 + 0.01 + 9 + 4 + 0) / 5 = 2.002%

Now, let's interpret the results:

- Stock A has a variance of approximately 2.78%, indicating higher variability or risk compared to Stock B.
- Stock B has a variance of approximately 2.002%, suggesting relatively lower variability or risk compared to Stock A.

In investment risk analysis, variance provides a measure of how much the returns of an investment vary around their mean. Higher variance implies higher risk, as there's more uncertainty about the returns. Lower variance indicates lower risk, as the returns are more stable and predictable.

2. Standard Deviation:

When we calculate variance we get a numeric representation of the spread of the data. We can use it to compare the spread of two data.

representation of the spread of the data. We can use it to compare the spread of two data.

→ But what does that number really mean alone?

→ How do we interpret the spread?

What does variance of 13.97, 108.55, 3812.87 mean?

$$\text{Standard Deviation} = \sqrt{\text{variance}}$$

The variance value is squared distance from the mean, so its unit is squared which is not directly related to the data and not easily interpretable.

Whereas the SD represents the values in the same unit as the data so it is directly interpretable.

Ex: Above the variance of stock A was 2.78%.

$$SD(\text{stock A}) = \sqrt{2.78}$$

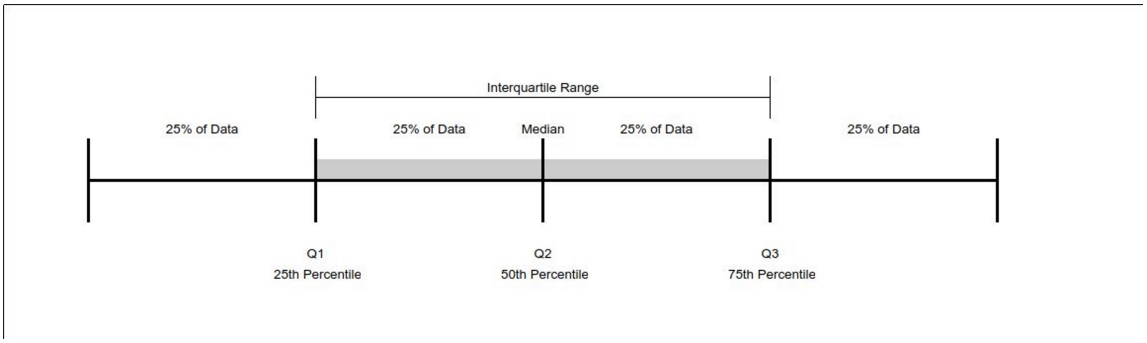
$$= 1.67\%.$$

This indicates that on an average the returns of stock A is 1.67% less or more than the average return.

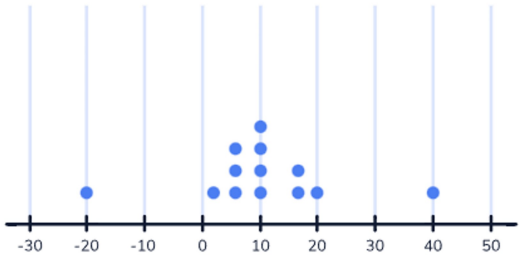
Note: Standard Deviation and Variance both are unreliable in the case of outliers present in the data since they use mean in the calculation.

So we prefer using a robust measure called Inter-quartile Range (IQR) in such cases.

3 → Interquartile Range:



We divide our data into 4 equal parts (more or less of equal size) called as quartiles.



Look at the above graph, if we represent the spread of the data using Range (simplest measure of spread).

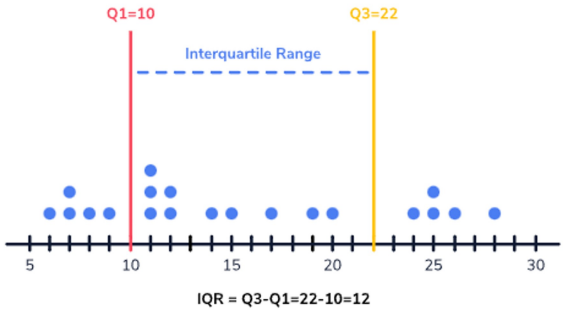
$$\text{Range} = \text{Max} - \text{Min}$$
$$= 40 - (-20)$$
$$= 60$$

In the above plot we can clearly see that most of our data is scattered between 0 and 15 but according to formula we consider the outliers too because of that the Range of 60 is not correct representative of the data.

And also since variance & SD are also not reliable in this case because of outliers, we have to bring in IQR for the rescue.

IQR simply tries to ignore the tails of the dataset or the extreme values, so that we

IQR simply [illegible] dataset or the extreme values, so that we capture the range where most of our data lies or is centered.



First quartile is the value that separates the first 25% of the data from the remaining 75%.

Third quartile (Q_3) is the value that separates the starting 75% data from remaining 25%.

IQR can also be for outlier detection and removal.

Example:

18, 27, 68, 75, 80, 82, 85, 88, 90, 92, 95, 98, 102, 127, 148

(i) → Arrange the data in ascending order.

(ii) → Median = 88

First half of the data

18 27 68 75 80 82 85

↑

Q_1

Second half of the data

90 92 95 98 102 127 148

↑

Q_2

Q_3

$$IQR = Q_3 - Q_1$$
$$= 98 - 75$$
$$= 23$$

Outlier Detection:

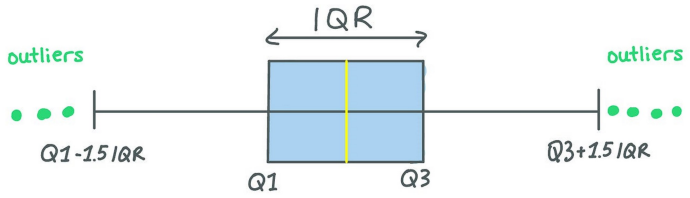
Lower fence $= Q_1 - 1.5 \times IQR$

Upper fence $= Q_3 + 1.5 \times IQR$

$LF = 75 - 1.5 \times 23 = 40.5$

$UF = 98 + 1.5 \times 23 = 132.5$

So any value less than lower fence and values higher than upper fence are considered as outliers and can be removed from the data.



Association between variables →

Two quantitative Variables:

Covariance: Measures the degree to which two variables change together. It indicates the direction of linear relationship b/w the variables.

Covariance can be any value, positive, negative or zero. A positive value indicates that both variable either increase together or decrease together.

It does not give a clear idea about the

either increase together or decrease ...

But it does not give a clear idea about the strength of relationship b/w two variables. That is where correlation comes into picture.

Correlation: Correlation indicates both the strength and direction of relationship b/w two numerical variables. Also called Pearson Correlation.

Correlation value ranges from -1 to 1.

→ Strong positive relations will have value close to 1.

→ strong negative relations will have value close to -1.

→ No relationship b/w variables is indicated by correlation close to 0.

Generally, a correlation larger than about 0.4 indicates a linear association. A correlation going from 0.6 to 1 or -0.6 to -1.0 indicates moderate to strong relationship.