Session-15 Statistics

04 August 2024 01:37 PM

Agenda

- **Measure of Position**
- **Quartile vs Quantile vs Percentile**
- **Five Number Summary**
- **Interquartile Ranges**
- **Effect Of Outliers And Its Removal**
- **Outlier Detection using Boxplot**

Data

Mean, Median, Mode

Measure of Position

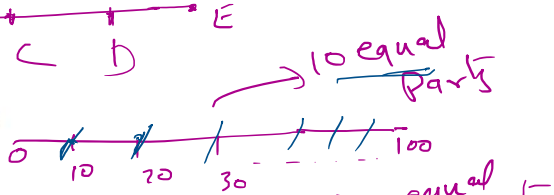
Measures of position give us a way to see where a certain data point or value falls in a sample or distribution. A measure can tell us whether a value is about the average, or whether it's unusually high or low. Measures of position are used for quantitative data that falls on some numerical scale. Sometimes, measures can be applied to ordinal variables— those variables that have an order, like first, second…fiftieth.

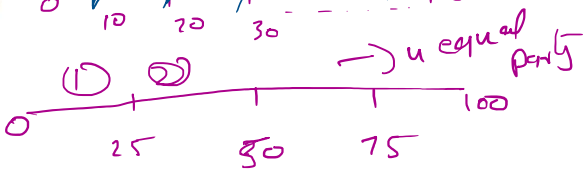
What are quantiles?

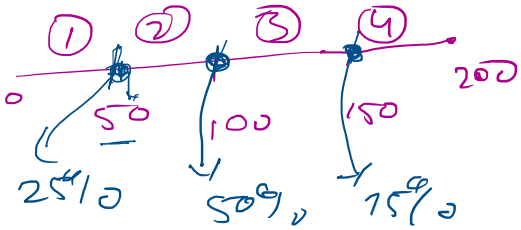
A quartile is a type of quantile.

Quantiles are values that split sorted data or a probability distribution into equal parts. In general terms, a *q*-quantile divides sorted data into *q* parts. The most commonly used quantiles have special names:

- **Quartiles (4-quantiles)**: Three quartiles split the data into four parts.
  - **Deciles (10-quantiles)**: Nine deciles split the data into 10 parts.
  - **Percentiles (100-quantiles):** 99 percentiles split the data into 100 parts.







What are quartiles?

Quartiles are a set of descriptive statistics. They summarize the central tendency and variability of a dataset or distribution.

Quartiles are a type of **percentile**. A percentile is a value with a certain percentage of the data falling below it. In general terms, *k*% of the data falls below the *k*th percentile.

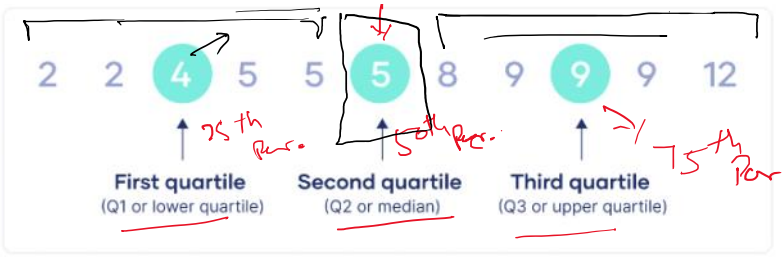
- The **first quartile** (Q1, or the lowest quartile) is the 25th percentile, meaning that 25% of the data falls below the first quartile.
- The **second quartile** (Q2, or the median) is the 50th percentile, meaning that 50% of the data falls below the second quartile.
- The **third quartile** (Q3, or the upper quartile) is the 75th percentile, meaning that 75% of the data falls below the third quartile.

By splitting the data at the 25th, 50th, and 75th percentiles, the quartiles divide the data into four equal parts.

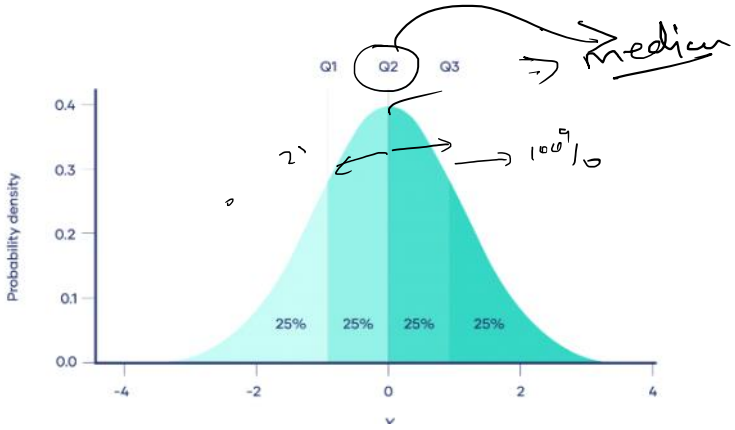
Position

median = 6

50th Percentile



- In a **sample** or dataset, the quartiles divide the data into four groups with equal numbers of observations.
- In a **probability distribution**, the quartiles divide the distribution's range into four intervals with equal probability.



How to find quartiles

To find the quartiles of a dataset or sample, follow the step-by-step guide below.

1. Count the number of observations in the dataset (n).
2. Sort the observations from smallest to largest.
3. Find the first quartile:
   - Calculate $n * (1 / 4)$.
   - If $n * (1 / 4)$ is an integer, then the first quartile is the mean of the numbers at positions $n * (1 / 4)$ and $n * (1 / 4) + 1$.
   - If $n * (1 / 4)$ is **not** an integer, then round it up. The number at this position is the first quartile.

Tip: An integer is a whole number—it can be written without any numbers after the decimal place.

4. Find the second quartile:
   - Calculate $n * (2 / 4)$.
   - If $n * (2 / 4)$ is an integer, the second quartile is the mean of the numbers at positions $n * (2 / 4)$ and $n * (2 / 4) + 1$.
   - If $n * (2 / 4)$ is **not** an integer, then round it up. The number at this position is the second quartile.
5. Find the third quartile:
   - Calculate $n * (3 / 4)$.
   - If $n * (3 / 4)$ is an integer, then the third quartile is the mean of the numbers at positions $n * (3 / 4)$ and $n * (3 / 4) + 1$.
   - If $n * (3 / 4)$ is **not** an integer, then round it up. The number at this position is the third quartile.

Step-by-step example

Imagine you conducted a small study on language development in children 1–6 years old. You're writing a paper about the study and you want to report the quartiles of the children's ages.

Age (years)	1	2	3	4	5	6
Frequency	2	3	4	1	2	2

Step 1: Count the number of observations in the dataset

$n = 2 + 3 + 4 + 1 + 2 + 2 = 14$

Step 2: Sort the observations in increasing order

1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 5, 5, 6, 6

Step 3: Find the first quartile

$n * (1 / 4) = 14 * (1 / 4) = 3.5$

3.5 is not an integer, so Q1 is the number at position 4.

1, 1, 2, **2**, 2, 3, 3, 3, 3, 4, 5, 5, 6, 6

Q1 = 2 years

$n \times \frac{1}{4} = 14 \times \frac{1}{4} = 3.5$

3-4

3 , 3.5 , 3.6 → 4

Step 4: Find the second quartile

$n * (2 / 4) = 14 * (2 / 4) = 7$

7 is an integer, so Q2 is the mean of the numbers at positions 7 and 8.

1, 1, 2, 2, 2, 3, **3**, **3**, 3, 4, 5, 5, 6, 6

Q2 = (3 + 3) / 2

Q2 = 3 years

Step 5: Find the third quartile

$n * (3 / 4) = 14 * (3 / 4) = 10.5$

10.5 is not an integer, so Q3 is the number at position 11.

1, 1, 2, 2, 2, 3, 3, 3, 3, 4, **5**, 5, 6, 6

Q3 = 5 years

Data
3
4
4
6
7
9
12
13
14
16
17
19
22
23
23
25
28

Percentile	Quartile	Value
0	0	3
25	1	8.5
50	2	16.5
75	3	23.5
100	4	37

min

25th → Q1

50th → Q2

75th → Q3

describe()

mean, std

min

25%

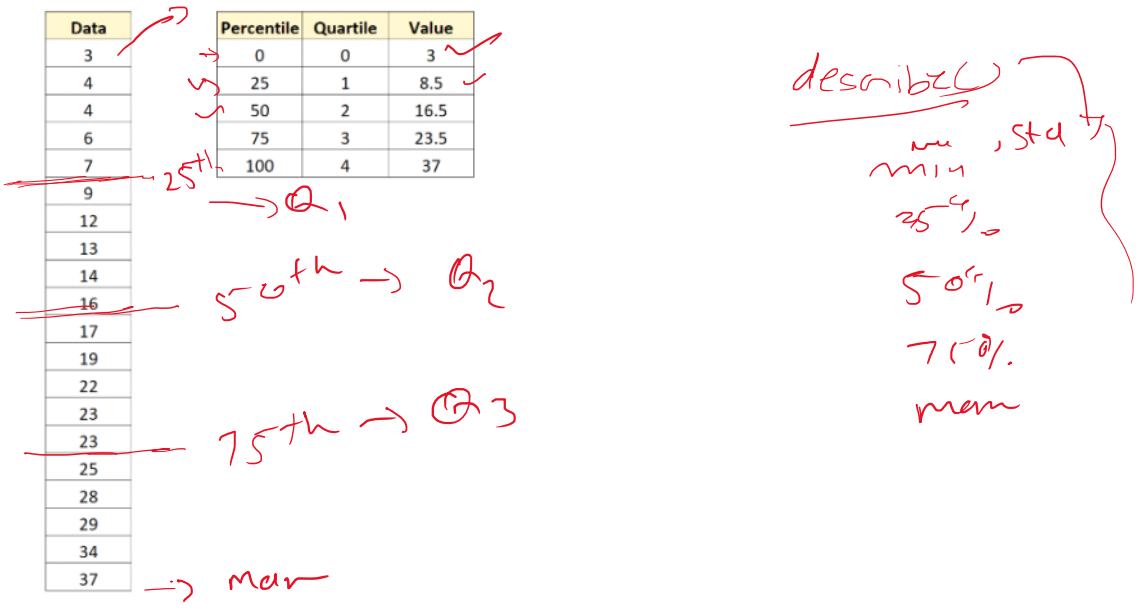
50%

75%

max

Data
3
4
4
6
7
9
12
13
14
16
17
19
22
23
23
25
28
29
34
37

Percentile	Quartile	Value
0	0	3
25	1	8.5
50	2	16.5
75	3	23.5
100	4	37



- The 0 percentile and 0 quartile is **3**.
- The 25th percentile and 1st quartile is **8.5**.
- The 50th percentile and 2nd quartile is **16.5**.
- The 75th percentile and 3rd quartile is **23.5**.
- The 100th percentile and 4th quartile is **37**.

What is Five Number Summary

Descriptive Statistics involves understanding the distribution and nature of the data. Five number summary is a part of descriptive statistics and consists of five values and all these values will help us to describe the data.

- The minimum value (the lowest value)
- 25th Percentile or Q1
- 50th Percentile or Q2 or Median
- 75th Percentile or Q3
- Maximum Value (the highest value)

How to calculate Five Number Summary

Let's understand this with the help of an example . Suppose we have some data such as :
11,23,32,26,16,19,30,14,16,10

Here, in the above set of data points our Five Number Summary are as follows :

Here, in the above set of data points our Five Number Summary are as follows :

First of all , we will arrange the data points in ascending order and then calculate the summary :

10,11,14,16,16,19,23,26,30,32

- Minimum value: 10
- 25th Percentile: 14

Calculation of 25th Percentile : (25/100)*(n+1) = (25/100)*(11) = 2.75 i.e 3rd value of the data

- 50th Percentile : 17.5

Calculation of 50th Percentile : (16+19)/2 = 17.5

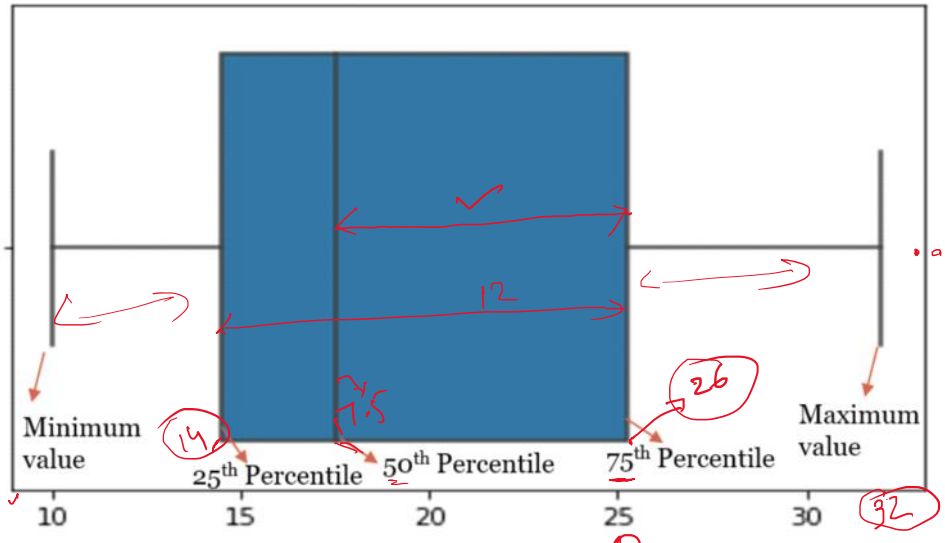
- 75th Percentile : 26

Calculation of 75th Percentile : (75/100)*(n+1) = (75/100)*(11) = 8.25 i.e 8th value of the data

- Maximum value: 32

Box plots and how they are constructed?

Boxplots are the graphical representation of the distribution of the data using Five Number summary values. **It is one of the most efficient ways to detect outliers in our dataset.**



Interquartile Range (IQR)

Interquartile range is the amount of spread in the middle 50% of a dataset.

In other words, it is the distance between the first quartile (Q_1) and the third

Interquartile Range (IQR)

Interquartile range is the amount of spread in the middle 50% of a dataset.

In other words, it is the distance between the first quartile (Q_1) and the third quartile (Q_3).

$$IQR = Q_3 - Q_1$$

Here's how to find the IQR:

Step 1: Put the data in order from least to greatest.

Step 2: Find the median. If the number of data points is odd, the median is the middle data point. If the number of data points is even, the median is the average of the middle two data points.

Step 3: Find the first quartile (Q_1). The first quartile is the median of the data points to the left of the median in the ordered list.

Step 4: Find the third quartile (Q_3). The third quartile is the median of the data points to the right of the median in the ordered list.

Step 5: Calculate IQR by subtracting $Q_3 - Q_1$.

Example: Consider the following dataset of exam scores for a class tenth:

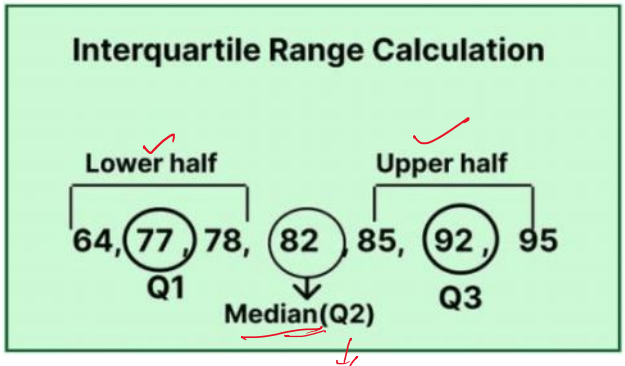
77, 85, 92, 64, 78, 95, 82

Find the Interquartile Range of the above data

Solution:

Now to calculate the Interquartile Range steps involved are:

1. First, we need to arrange in ascending order
2. Count the given values i.e is 7 ,so count is odd,then median is middle value =82
3. Next Divide into two halfs ,Lower half and Upper half
4. Next identify median value in lower half as Q_1 and upper half as Q_3



Now,Q_1 = 77 and Q_3 = 92

⇒ IQR = $Q_3 - Q_1$ = 92-77 = 15

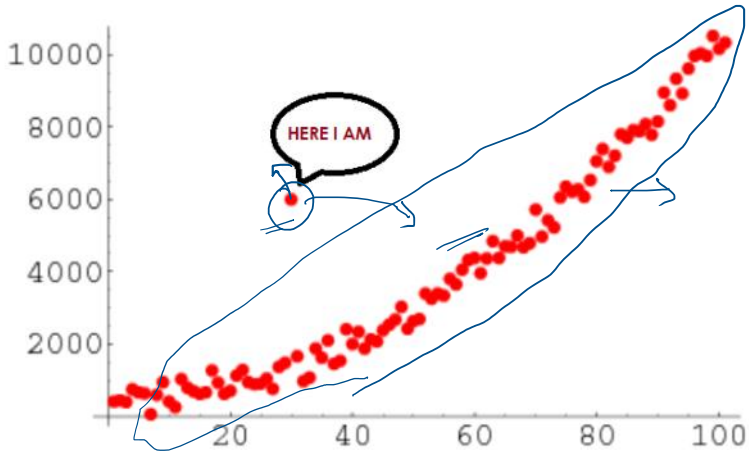
What are Outliers?

In terms of statistics, Outliers can be defined as,

"An Outlier is that observation which is significantly different from all other observations."

From this definition, we can conclude that an outlier is something that is an **odd-one-out** or the one that is different from the crowd. Some statisticians formally define outliers as **'Observations having a different underlying behavior than the rest of the observations'**.

Alternatively, outliers are those observations that are significantly different from other observations.



Source link:
https://www.analyticsvidhya.com/blog/2021/05/five-number-summary-for-analysis/
https://www.scribbr.com/statistics/quartiles-quantiles/
https://www.khanacademy.org/math/cc-sixth-grade-math/cc-6th-data-statistics/cc-6th/a/interquartile-range-review
https://www.geeksforgeeks.org/interquartile-range/
https://www.analyticsvidhya.com/blog/2021/05/why-you-shouldnt-just-delete-outliers/
https://www.geeksforgeeks.org/detect-and-remove-the-outliers-using-python/