Picsart Academy AI

17.01.2026

Sample Median

The Sample Median: Sample Median is, in some sense, the central value, the middle value, of our Dataset, when sorted in the increasing order.
The rigorous definition is: let $x : x_1, x_2, ..., x_n$ be our dataset.

- If n is **odd**, then we define

$$median(x) = x_{\left(\frac{n+1}{2}\right)};$$

- If n is **even**,

$$median(x) = \frac{1}{2} \cdot \left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right).$$

Sample Mode

Another measure of the Central Tendency is the Mode:
Sample Mode of the dataset is a value which occurs most frequently in our dataset.
In other words, Mode is the value with the maximum Frequency in the Frequency (or the RelFreq) Table.

Remark: Mode can be non-unique. One can have several Modes in the Dataset. If all elements in the Dataset are unique, then usually we say that we do not have a Mode (or all elements are Modes). If the Dataset has a unique Mode, we call it Unimodal. Bimodal Dataset has exactly 2 Modes. Similarly, one can talk about Multimodal Datasets.

Sample Mode: Remarks

Remark: If data comes from a Continuous Variable, then the Mode can be a non-meaningful measure - (almost) all Datapoints will have a Frequency equal to 1, so the Mode will consists of all elements of the Dataset. For this case, one is grouping Datapoints into bins, then calculating the most frequent bin.
Remark: Mode (but not the Mean or Median) can be calculated even for Nominal Scale Categorical Datasets. Say, you can find the Mode of all Armenians' First Names.
Remark: Sometimes, one considers also *local Modes* (local maximums of the Frequency Table) and call them just Modes.

The Sample Variance

The **Sample Variance** (with the denominator n) of our dataset x is defined by

$$var(x) = s^2 = \frac{\sum_{k=1}^{n} (x_k - \bar{x})^2}{n},$$

where $\bar{x}$ is the sample mean of our dataset:

$$\bar{x} = mean(x) = \frac{1}{n} \cdot \sum_{k=1}^{n} x_k.$$

In many textbooks, the **Sample Variance** of x is defined as

$$var(x) = s^2 = \frac{\sum_{k=1}^{n} (x_k - \bar{x})^2}{n-1}$$

with $n - 1$ in the denominator.

The Standard Deviation

The **Standard Deviation** of x is defined as

$$sd(x) = s = \sqrt{var(x)}.$$

Question: Which measure of the Spread/Variability is better: Variance or SD?

- $sd(x)$ is in the same units as x, but $var(x)$ is in the squared units of x
- $var(x)$ is easy to deal with, has some nice properties, but not $sd(x)$

Sample Quartiles

- Idea of the Median: a point on the axis dividing the Dataset into two equal-length portions
- Idea of Quartiles: 3 point on the axis dividing the Dataset into four equal-length portions

There are different methods to define Quartiles[1], and we will use the following.

Let $x: \quad x_1, x_2, ..., x_n$ be our Dataset. First we sort, by using Order Statistics, our Dataset into:

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n-1)} \leq x_{(n)}.$$

[1] See, for example, the Wiki page, https://en.wikipedia.org/wiki/Quartile

Sample Quartiles and IQR

Now,

- The **second (or middle) Quartile**, Q_2, is the Median of our dataset, $Q_2 = med(x)$;
- The **first (or lower) Quartile**, Q_1, is the Median of the ordered Dataset of all observations to the left of Q_2 (including Q_2, if it is a Datapoint);
- The **third (or upper) Quartile**, Q_3, is the Median of the ordered Dataset of all observations to the right of Q_2 (including Q_2, if it is a Datapoint)

Next, we define the **InterQuartile Range, IQR** to be

$$IQR = Q_3 - Q_1.$$

Quartiles and IQR

Remark: Note that the Quartiles Q_1, Q_2, Q_3 are not always Datapoints.

Note: Recall the idea of Quartiles: the points Q_1, Q_2, Q_3 on the real axis divide our Dataset into (almost) four equal-length portions:

- almost 25% of our Datapoints are to the left to Q_1
- almost 25% of our Datapoints are between Q_1 and Q_2
- almost 25% of our Datapoints are between Q_2 and Q_3
- almost 25% of our Datapoints are to the right to Q_3

Note: The interval $[Q_1, Q_3]$ contains almost the half of the Datapoints. So the IQR shows the Spread of the middle half of our Dataset, it is a measure of the Spread/Variability.

Outlier

- the Lower and Upper Fences
- $W_1 = \min\{x_i : x_i \geq Q_1 - 1.5 \cdot IQR\}$ and $W_2 = \max\{x_i : x_i \leq Q_3 + 1.5 \cdot IQR\}$,i.e., the first and last observations lying in

$$\left[Q_1 - \frac{3}{2}IQR,\ \ Q_3 + \frac{3}{2}IQR\right];$$

  the lines joining that fences to corresponding quartiles are the *Whiskers*;
- the set of all Outliers

$$O = \left\{x_i : x_i \notin \left[Q_1 - \frac{3}{2}IQR, Q_3 + \frac{3}{2}IQR\right]\right\}$$

Numerical Summaries for Bivariate Data

Sample Covariance and the Correlation Coefficient

Assume now we have a bivariate Dataset

$$(x_1, y_1), ..., (x_n, y_n),$$

or just two 1D Datasets of the same size:

$$x: \ x_1, ..., x_n \qquad \textit{and} \qquad y: \ y_1, ..., y_n.$$

Our aim is to see if some linear relationship, association exists between x and y. Of course, the best way is to visualize our Dataset by a ScatterPlot.
Now we want to answer, numerically, how strong/week is the linear relationship between our variables x and y.

Sample Covariance

The **Sample Covariance** of Variables (1D Datasets) x and y is

$$cov(x,y) = s_{xy} = \frac{\sum_{k=1}^{n} (x_k - \bar{x}) \cdot (y_k - \bar{y})}{n}$$

or

$$cov(x,y) = s_{xy} = \frac{\sum_{k=1}^{n} (x_k - \bar{x}) \cdot (y_k - \bar{y})}{n-1}$$

Here $\bar{x}$ and $\bar{y}$ are the Sample Means for the Datasets x and y.

Sample Covariance

Definition: We say that the Variables (Datasets) x and y are **uncorrelated**, if $cov(x, y) = 0$.

Sample Correlation Coefficient

Another measure of the linear relationship between the Variables x and y of Bivariate Dataset is the *Pearson's Correlation Coefficient*:

Definition: The **Sample Correlation Coefficient** of x and y is

$$cor(x, y) = \rho_{xy} = \frac{cov(x, y)}{\sqrt{Var(x) \cdot Var(y)}} = \frac{cov(x, y)}{sd(x) \cdot sd(y)} = \frac{s_{xy}}{s_x \cdot s_y},$$

where s_x and s_y are the standard deviations for x and y, respectively.

If $s_x = 0$ or $s_y = 0$, then we take $cor(x, y) = 0$ by definition.

Sample Correlation Coefficient

In both cases, when one calculates Standard Deviations and Covariance by using n simultaneously or $n-1$ simultaneously in the denominator, we will obtain

$$cor(x,y) = \rho_{xy} = \frac{\sum_{k=1}^{n} (x_k - \overline{x}) \cdot (y_k - \overline{y})}{\sqrt{\sum_{k=1}^{n} (x_k - \overline{x})^2 \cdot \sum_{k=1}^{n} (y_k - \overline{y})^2}}$$

Another formula to calc the correlation coefficient is

$$cor(x,y) = \rho_{xy} = \frac{\sum_{k=1}^{n} x_k y_k - n \cdot \overline{x} \cdot \overline{y}}{\sqrt{\sum_{k=1}^{n} x_k^2 - n \cdot (\overline{x})^2} \cdot \sqrt{\sum_{k=1}^{n} y_k^2 - n \cdot (\overline{y})^2}}.$$