Contents

Order Statistics

First we introduce the **Order Statistics**.

Assume we have a 1D Numerical Dataset $x_1, x_2, ..., x_n$. We sort this Dataset in the increasing order, and denote by $x_{(j)}$ the j-th element in the sorted array. $x_{(j)}$ is called the ***j*-th Order Statistics** of our Dataset.

In other word, $x_{(1)}, x_{(2)}, ..., x_{(n)}$ is just a reordering of our Dataset with

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

In particular,

$$x_{(1)} = \min\{x_1, x_2, ..., x_n\} \qquad \text{and} \qquad x_{(n)} = \max\{x_1, x_2, ..., x_n\}.$$

Trimmed Sample Mean

Usually, one considers other measures for the Central Tendency, which are less sensitive to outliers.

- **The Trimmed (Truncated) Sample Mean:** First we take a real number $r \in (0, 0.5)$ (or, in percents, from 0 to 50%). We will drop the *lowest r percent and largest r percent* of our data, and then we will calculate the Sample Mean of the rest.

So we take r (ratio, fraction of points to be deleted from the both ends), we calculate $p = [r \cdot n]$. Then we sort our x in the acsending order, delete first p and last p values from this sorted array, and calculate the mean of the remaining Dataset.

Trimmed Sample Mean

Mathematically,

$$\text{trimmed sample mean}(x) = \bar{x}_{trimmed} =$$

$$= \frac{x_{(p+1)} + x_{(p+2)} + \ldots + x_{(n-p-1)} + x_{(n-p)}}{n-2p} = \frac{\sum\limits_{k=p+1}^{n-p} x_{(k)}}{n-2p}.$$

Idea of Trimming: Reduce the influence of outliers. This *Statistics* for the Central Tendency, Center, is more *robust* to outliers, extremes, than the ordinary mean.

Winsorized Sample Mean

- **Winsorized Sample Mean:** Again, to reduce the influence of outliers, one can calculate the *Winsorized Sample Mean*. Here we again take $r \in (0, 0.5)$, take $p = [n \cdot r]$, and calculate

$$\text{winsorized sample mean}(x) = \frac{x_{(p+1)} + \ldots + x_{(p+1)} + x_{(p+2)} + x_{(p+3)} + \ldots + x_{(n-p-1)} + x_{(n-p)} + \ldots + x_{(n-p)}}{n}$$

$$= \frac{(p+1) \cdot x_{(p+1)} + \sum_{k=p+2}^{n-p-1} x_{(k)} + (p+1) \cdot x_{(n-p)}}{n}.$$

Remarks

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.

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, people are 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. Just like in Calculus: when saying *extremum*, we think about a *Local*

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.

We will use both, and later we will talk about the difference between these two - there are reasons to prefer one over the other.

The Standard Deviation

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

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

So we will have 2 formulas to calculate the Standard Deviation: with n or $n-1$ in the denominator.

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 $sd(x)$

So, like in the Probability Theory, `var` is easy to deal with, `sd` is the measure to report.

R is calculating Var and SD by using $n-1$ in the denominator:

```
x <- 1:5
var(x)
```

```
## [1] 2.5
```

```
sd(x)
```

```
## [1] 1.581139
```

Some Properties of the Variance

The Sample Variance (with the denominator n) can be calculated by the following formula

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

We can write this, using an analogy with the r.v. Variance,

$$var(x) = mean(x^2) - \Big(mean(x)\Big)^2 = \overline{x^2} - (\bar{x})^2,$$

where x^2 is the dataset $x_1^2, x_2^2, ..., x_n^2$. Just remember to use this in the case when the Sample Variance is with the denominator n !

Some Properties of the Variance

Assume x is the dataset $x_1, x_2, ..., x_n$, and $\alpha, \beta \in \mathbb{R}$ are constants. We will denote by $\alpha \cdot x$ the dataset $\alpha \cdot x_1, \alpha \cdot x_2, ..., \alpha \cdot x_n$, and by $x + \beta$ the dataset $x_1 + \beta, x_2 + \beta, ..., x_n + \beta$. Then

- $var(x) \geq 0$;
- $var(x) = 0$ if and only if $x_k = x_j$ for any k, j;
- $var(\alpha \cdot x) = \alpha^2 \cdot var(x)$;
- $var(x + \beta) = var(x)$.

MAD

Other measures for the Spread of a Dataset are the **Mean/Median Absolute Deviation** from the Mean/Median.

The Mean Absolute Deviation (MAD) from the Mean for the Dataset $x_1, ..., x_n$ is

$$mad(x) = mad(x, mean) = \frac{\sum_{k=1}^{n} |x_k - \bar{x}|}{n}.$$

By replacing the Mean by the Median, we will obtain the **Mean Absolute Deviation from the Median**:

$$mad(x) = mad(x, median) = \frac{\sum_{k=1}^{n} |x_k - median(x)|}{n}$$

MAD

The idea of the **Median Absolute Deviation from the Mean/Median** is to calculate first the Absolute Deviations from the Mean/Median, then find the Median of that Absolute Deviations. See, for example, the description of the `mad` function in **R**.

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[2], and we will use the following.

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

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

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

$$IQR = Q_3 - Q_1.$$

Example: Find the Quartiles and IQR of

$$x: \ -2, 1, 3, 0, 5, 7, 5, 2, 0$$

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.

Quartiles in **R**

In **R**, one can use the commands `quantile(x, 0.25)` and `quantile(x, 0.75)` to find Q_1 and Q_3. For example,

```
x <- 1:10
quantile(x,0.25)
```

```
##  25%
## 3.25
```