AUA CS108, Statistics, Fall 2020

Lecture 09

Michael Poghosyan

14 Sep 2020

Contents

Statistical Measures for the Spread/Variability

Statistical Measures for the Spread/Variability

Here we want to answer to the questions: how spread/concentrated are our Datapoints, how much is the variability of our Data?

Deviations from the Mean (or from the Median)

We consider a 1D Numerical Dataset

$$x: \ x_1, x_2, ..., x_n.$$

Deviations from the Mean (or from the Median)

We consider a 1D Numerical Dataset

$$x: \ x_1, x_2, ..., x_n.$$

The differences

$$x_k - \bar{x} = x_k - mean(x), \qquad k = 1, ..., n$$

are called **Deviations of** x **from the Mean**.

Deviations from the Mean (or from the Median)

We consider a 1D Numerical Dataset

$$x: \ x_1, x_2, ..., x_n.$$

The differences

$$x_k - \bar{x} = x_k - mean(x), \qquad k = 1, ..., n$$

are called **Deviations of x from the Mean**.

Absolute Deviations of x from its Mean are defined as

$$|x_k - \bar{x}|, \qquad k = 1, ..., n.$$

Deviations from the Mean (or from the Median)

We consider a 1D Numerical Dataset

$$x: \; x_1, x_2, ..., x_n.$$

The differences

$$x_k - \bar{x} = x_k - mean(x), \qquad k = 1, ..., n$$

are called **Deviations of x from the Mean**.

Absolute Deviations of x from its Mean are defined as

$$|x_k - \bar{x}|, \qquad k = 1, ..., n.$$

Similarly, **Deviations of x from the Median** are defined as the differences

$$x_k - median(x), \qquad k = 1, ..., n$$

Example

Consider the Dataset `islands` from **R**:

```r
head(islands, 3)
```

```
##      Africa Antarctica       Asia
##       11506       5500      16988
```

Example

Consider the Dataset `islands` from **R**:

```
head(islands, 3)
```

```
##     Africa Antarctica       Asia
##      11506       5500      16988
```

To calculate Deviations from the Mean for this Dataset, we just use

```
x.bar <- mean(islands)
deviations <- islands - x.bar
head(deviations)
```

```
##       Africa   Antarctica          Asia    Australia Axel
##    10253.271     4247.271    15735.271     1715.271     -
```

Range

The simplest measure of the Spread is the Range:

The **Range** of the Dataset x is

$$Range(x) = x_{(n)} - x_{(1)} = \max_k x_k - \min_k x_k.$$

Range

The simplest measure of the Spread is the Range:

The **Range** of the Dataset x is

$$Range(x) = x_{(n)} - x_{(1)} = \max_k x_k - \min_k x_k.$$

In **R**, the command `range` gives the pair $(x_{(1)}, x_{(n)})$, not their difference.

Range

The simplest measure of the Spread is the Range:

The **Range** of the Dataset x is

$$Range(x) = x_{(n)} - x_{(1)} = \max_k x_k - \min_k x_k.$$

In **R**, the command `range` gives the pair $(x_{(1)}, x_{(n)})$, not their difference.

Say,

```r
range(islands)
```

```
## [1]    12 16988
```

Example, **R** code to Calculate the Range

We can define our custom function to calculate the Range as the difference:

```
my.range <- function(x){
  return(max(x)-min(x))
}
```

Example,**R** code to Calculate the Range

We can define our custom function to calculate the Range as the difference:

```
my.range <- function(x){
  return(max(x)-min(x))
}
```

and run

```
my.range(1:10)
```

```
## [1] 9
```

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:

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

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 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.

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?

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

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 $sd(x)$

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.

Example

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

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

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

```r
sd(x)
```

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

Some Properties of the Variance

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

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

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 $x_1^2, x_2^2, ..., x_n^2$.

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.

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 $\alpha \cdot x_1, \alpha \cdot x_2, ..., \alpha \cdot x_n$,

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 $x_1 + \beta, x_2 + \beta, ..., x_n + \beta$.

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

- $var(x) \geq 0$;

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

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;

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;
- $var(\alpha \cdot x) =$

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;
- $var(\alpha \cdot x) = \alpha^2 \cdot var(x)$;

Some Properties of the Variance

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

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

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.

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

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

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.

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**.

Quartiles, IQR and the BoxPlot

Sample Quartiles

- Idea of the Median:

Sample Quartiles

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

Sample Quartiles

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

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

[1] See, for example, the Wiki page

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 $x: \ x_1, x_2, ..., x_n$ 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

Sample Quartiles and IQR

Now,

- The **second (or middle) Quartile**, Q_2, is the Median of our dataset, $Q_2 = med(x)$;

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);

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)

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:

Example: Find the Quartiles and IQR of

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

Example:

Example: Find the Quartiles and IQR of

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

Example: Find the Quartiles and IQR of

$$x: \ 1, 1, 2, 3, 1, 1, 3, 4, 5, 2$$