AUA CS 108, Statistics, Fall 2019

Lecture 08

Michael Poghosyan
YSU, AUA
michael@ysu.am, mpoghosyan@aua.am

11 Sep 2019

Contents

Question

Who wants to have a Slack Channel for our Stat course?

Last Lecture ReCap

- What are Numerical Summaries (of a Dataset) for?

Last Lecture ReCap

- What are Numerical Summaries (of a Dataset) for?
- What is the drawback of the Sample Mean?

Last Lecture ReCap

- What are Numerical Summaries (of a Dataset) for?
- What is the drawback of the Sample Mean?
- What is the Sample Trimmed Mean?

Last Lecture ReCap

- What are Numerical Summaries (of a Dataset) for?
- What is the drawback of the Sample Mean?
- What is the Sample Trimmed Mean?
- What is the Winsorized Mean?

Last Lecture ReCap

- What are Numerical Summaries (of a Dataset) for?
- What is the drawback of the Sample Mean?
- What is the Sample Trimmed Mean?
- What is the Winsorized Mean?
- Is Sample Mean a Weighed Mean?

Last Lecture ReCap

- What are Numerical Summaries (of a Dataset) for?
- What is the drawback of the Sample Mean?
- What is the Sample Trimmed Mean?
- What is the Winsorized Mean?
- Is Sample Mean a Weighed Mean?
- Is Trimmed Mean a Weighted Mean?

Last Lecture ReCap

- What are Numerical Summaries (of a Dataset) for?
- What is the drawback of the Sample Mean?
- What is the Sample Trimmed Mean?
- What is the Winsorized Mean?
- Is Sample Mean a Weighed Mean?
- Is Trimmed Mean a Weighted Mean?
- Is Winsorized Mean a Weighted Mean?

Last Lecture ReCap

- What are Numerical Summaries (of a Dataset) for?
- What is the drawback of the Sample Mean?
- What is the Sample Trimmed Mean?
- What is the Winsorized Mean?
- Is Sample Mean a Weighed Mean?
- Is Trimmed Mean a Weighted Mean?
- Is Winsorized Mean a Weighted Mean?
- What is the Median?

Last Lecture ReCap

- What are Numerical Summaries (of a Dataset) for?
- What is the drawback of the Sample Mean?
- What is the Sample Trimmed Mean?
- What is the Winsorized Mean?
- Is Sample Mean a Weighed Mean?
- Is Trimmed Mean a Weighted Mean?
- Is Winsorized Mean a Weighted Mean?
- What is the Median?
- What is the Mode?

Range

Recall that we were talking about the Range last time:

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

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:

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

and run

```r
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 $sd(x)$

Example

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

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

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

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

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 $x_1, x_2, ..., x_n$, 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;
- $var(\alpha \cdot x) = \alpha^2 \cdot var(x)$;
- $var(x + \beta) =$

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

Another measure for the Spread of a Dataset is the **Mean Absolute Deviation** from the Mean/Median.

MAD

Another measure for the Spread of a Dataset is the **Mean Absolute Deviation** from the Mean/Median.

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

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

MAD

Another measure for the Spread of a Dataset is the **Mean 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 Mode, 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

Another measure for the Spread of a Dataset is the **Mean Absolute Deviation** from the Mean/Median.

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

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

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

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

Note: MAD is in the same units as x, like sd!

Quartiles, Quantiles and BoxPlots

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 of

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

Example:

Example: Find the Quartiles of

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

Example: Find the Quartiles of

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

Quartiles and IQR

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

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

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

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

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

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.

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.

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
```

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
```

Or, you can use the following commands:

```
x <- 1:10
fivenum(x)
```

```
## [1]  1.0  3.0  5.5  8.0 10.0
```

```
summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    3.25    5.50    5.50    7.75   10.00
```

Note

Note: Please note that **R** is not using our definition of the Quartiles, so sometimes we will get different results when calculating by a hand or by **R**.