AUA CS 108, Statistics, Fall 2019

Lecture 07

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

09 Sep 2019

Contents

Last Lecture ReCap

- What is a ScatterPlot? What is it for?

Last Lecture ReCap

- What is a ScatterPlot? What is it for?
- What was Hans Rosling's short talk about?

Last Lecture ReCap

- What is a ScatterPlot? What is it for?
- What was Hans Rosling's short talk about?
- What is the j-th Order Statistics

Numerical Summaries

Numerical Summaries

- Summaries (Statistics) about the Center, Mean, Location

Numerical Summaries

- Summaries (Statistics) about the Center, Mean, Location
- Summaries (Statistics) about the Spread, Variability

Statistical Measures for the Central Tendency/Location

Statistical Measures for the Central Tendency/Location

Here we want to answer to trhe questions: what are the typical values of our Dataset, where is our Data located at?

Sample Mean

Assume we are given a 1D numerical Dataset $x : x_1, x_2, ..., x_n$.

Sample Mean

Assume we are given a 1D numerical Dataset $x : x_1, x_2, ..., x_n$. We want to describe its typical value, its center.

Sample Mean

Assume we are given a 1D numerical Dataset $x : x_1, x_2, ..., x_n$. We want to describe its typical value, its center.

- **The Sample Mean:**

$$\bar{x} = mean(x) = \frac{x_1 + x_2 + ... + x_n}{n}$$

Sample Mean

Assume we are given a 1D numerical Dataset $x : x_1, x_2, ..., x_n$. We want to describe its typical value, its center.

- **The Sample Mean:**

$$\bar{x} = mean(x) = \frac{x_1 + x_2 + ... + x_n}{n}$$

Drawback: Sensitive to outliers (non-typical elements)

Sample Mean

Assume we are given a 1D numerical Dataset $x : x_1, x_2, ..., x_n$. We want to describe its typical value, its center.

- **The Sample Mean:**

$$\bar{x} = mean(x) = \frac{x_1 + x_2 + ... + x_n}{n}$$

Drawback: Sensitive to outliers (non-typical elements)

Note: Sometimes this property is a plus, not a drawback! Say, if we want to have an estimator which is sensitive to outliers.

Example

When we talk about, say, that the mean/average Midterm grade is 68, we think about this like the grades are 68 plyus-minus something.

Example

When we talk about, say, that the mean/average Midterm grade is 68, we think about this like the grades are 68 plyus-minus something. But . . .

Example: Consider the following Dataset:

$$1, 2, 3, 4, 5, 6, 789$$

Example

When we talk about, say, that the mean/average Midterm grade is 68, we think about this like the grades are 68 plyus-minus something. But ...

Example: Consider the following Dataset:

$$1, 2, 3, 4, 5, 6, 789$$

The mean of this is

```
mean(c(1,2,3,4,5,6, 789))
```

```
## [1] 115.7143
```

Example

When we talk about, say, that the mean/average Midterm grade is 68, we think about this like the grades are 68 plyus-minus something. But ...

Example: Consider the following Dataset:

$$1, 2, 3, 4, 5, 6, 789$$

The mean of this is

```r
mean(c(1,2,3,4,5,6, 789))
```

```
## [1] 115.7143
```

Can we say here that the elements of our Dataset are 115.7143 plyus-minus something?

Example

When we talk about, say, that the mean/average Midterm grade is 68, we think about this like the grades are 68 plyus-minus something. But . . .

Example: Consider the following Dataset:

$$1, 2, 3, 4, 5, 6, 789$$

The mean of this is

```r
mean(c(1,2,3,4,5,6, 789))
```

```
## [1] 115.7143
```

Can we say here that the elements of our Dataset are 115.7143 plyus-minus something? Not exactly.

Example

When we talk about, say, that the mean/average Midterm grade is 68, we think about this like the grades are 68 plyus-minus something. But ...

Example: Consider the following Dataset:

$$1, 2, 3, 4, 5, 6, 789$$

The mean of this is

```
mean(c(1,2,3,4,5,6, 789))
```

```
## [1] 115.7143
```

Can we say here that the elements of our Dataset are 115.7143 plyus-minus something? Not exactly.

Well, 115.7143 is not the typical value/center of our Dataset. This number gives us a wrong information about the elements of the Dataset.

Trimmed Sample Mean

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

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.

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 to be deleted), we calculate $p = [r \cdot n]$.

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 to be deleted), 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,

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

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

Trimmed Sample Mean

Mathematically,

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

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

Example: See, for example, Scoring the Dive Competition.

Trimmed Sample Mean

Mathematically,

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

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

Example: See, for example, Scoring the Dive Competition.

Idea of Trimming: Reduce the influence of outliers.

Trimmed Sample Mean

Mathematically,

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

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

Example: See, for example, Scoring the Dive Competition.

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.

Example

```r
x <- c(1, 10, 20, 30, 4, 50)
mean(x)
```

```
## [1] 19.16667
```

```r
mean(x, trim = 0.4)
```

```
## [1] 15
```

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

$$\text{winsorized sample mean}(x) =$$

$$\frac{x_{(p+1)} + \ldots + x_{(p+1)} + x_{(p+2)} + x_{(p+3)} + \ldots + x_{(n-p-2)} + x_{(n-p-1)} + \ldots + x_{(n-p-1)}}{n}$$

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

Weighted Sample Mean

Assume we want to calculate the mean of the dataset $x : x_1, x_2, ..., x_n$.

Weighted Sample Mean

Assume we want to calculate the mean of the dataset $x : x_1, x_2, ..., x_n$.We take nonnegative *weights* w_k's, such that $\sum_{k=1}^n w_k \neq 0$, and we calculate

$$\text{weighted sample mean}(x; w) = \bar{x}_w = \frac{\sum_{k=1}^n w_k x_k}{\sum_{k=1}^n w_k}.$$

Weighted Sample Mean

Assume we want to calculate the mean of the dataset $x : x_1, x_2, ..., x_n$.We take nonnegative *weights* w_k's, such that $\sum_{k=1}^{n} w_k \neq 0$, and we calculate

$$\text{weighted sample mean}(x; w) = \bar{x}_w = \frac{\sum_{k=1}^{n} w_k x_k}{\sum_{k=1}^{n} w_k}.$$

The weight of data x_k is then $\dfrac{w_k}{\sum_{i=1}^{n} w_i}$.

Example

```r
x <- c(-1,2,3,2,3,1,4,5, 10)
w <- c(0,1.2,1,1,5,3,2,3, 1)
weighted.mean(x, w)
```

```
## [1] 3.395349
```

Example

```r
x <- c(-1,2,3,2,3,1,4,5, 10)
w <- c(0,1.2,1,1,5,3,2,3, 1)
weighted.mean(x, w)
```

```
## [1] 3.395349
```

We can check:

```r
sum(x*w)/sum(w)
```

```
## [1] 3.395349
```

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.

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 Median

So to calculate the Median of x, first we sort x in the increasing order.

Sample Median

So to calculate the Median of x, first we sort x in the increasing order. Then

- If n is odd: we take the number at the center of the sorted list.

Sample Median

So to calculate the Median of x, first we sort x in the increasing order. Then

- If n is odd: we take the number at the center of the sorted list.

Example: For

$$x : -1, 2, 3, 1, 2, 4, 9,$$

the Median is: OTB

Sample Median

So to calculate the Median of x, first we sort x in the increasing order. Then

- If n is odd: we take the number at the center of the sorted list.

Example: For

$$x : -1, 2, 3, 1, 2, 4, 9,$$

the Median is: OTB

- If n is even: then, in the sorted list, we have 2 elements at the center. We take the average of these two elements.

Sample Median

So to calculate the Median of x, first we sort x in the increasing order. Then

- If n is odd: we take the number at the center of the sorted list.

Example: For

$$x : -1, 2, 3, 1, 2, 4, 9,$$

the Median is: OTB

- If n is even: then, in the sorted list, we have 2 elements at the center. We take the average of these two elements.

Example: For

$$x : -1, 2, 3, 1,$$

the Median is: OTB

Example

Calculation of the Median is simple in **R**: just use the `median` function.

Example

Calculation of the Median is simple in **R**: just use the `median` function.

```
x <- c(1,3,2, 4,2,3,2,2,1)
mean(x)
```

```
## [1] 2.222222
```

```
median(x)
```

```
## [1] 2
```

Example

Calculation of the Median is simple in **R**: just use the `median` function.

```
x <- c(1,3,2, 4,2,3,2,2,1)
mean(x)
```

```
## [1] 2.222222
```

```
median(x)
```

```
## [1] 2
```

Now, let's add an outlier:

```
x <- c(x, 1000)
mean(x)
```

```
## [1] 102
```

```
median(x)
```

```
## [1] 2
```

Important Property of the Median

- Half of the Datapoints are to the left of the Median, and half of the Datapoints are to the right

Example: Give OTB

Sample Mode

Another measure of the Central Tendency is the Mode:

Definition: Sample Mode of the dataset is a value which occurs most frequently in our dataset.

Sample Mode

Another measure of the Central Tendency is the Mode:

Definition: Sample Mode of the dataset is a value which occurs most frequently in our dataset.

Example: The Sample Mode of the following Dataset:

$$x : 0, -1, 2, 0, 0, 2, 3, 2, 1, 2$$

is

Sample Mode

Another measure of the Central Tendency is the Mode:

Definition: Sample Mode of the dataset is a value which occurs most frequently in our dataset.

Example: The Sample Mode of the following Dataset:

$$x : 0, -1, 2, 0, 0, 2, 3, 2, 1, 2$$

is 2.

Remark: Mode can be non-unique. One can have several Modes in the Dataset.

Sample Mode

Another measure of the Central Tendency is the Mode:

Definition: Sample Mode of the dataset is a value which occurs most frequently in our dataset.

Example: The Sample Mode of the following Dataset:

$$x : 0, -1, 2, 0, 0, 2, 3, 2, 1, 2$$

is 2.

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).

Sample Mode

Another measure of the Central Tendency is the Mode:

Definition: Sample Mode of the dataset is a value which occurs most frequently in our dataset.

Example: The Sample Mode of the following Dataset:

$$x : 0, -1, 2, 0, 0, 2, 3, 2, 1, 2$$

is 2.

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.

Sample Mode

Another measure of the Central Tendency is the Mode:

Definition: Sample Mode of the dataset is a value which occurs most frequently in our dataset.

Example: The Sample Mode of the following Dataset:

$$x : 0, -1, 2, 0, 0, 2, 3, 2, 1, 2$$

is 2.

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.

Sample Mode

Another measure of the Central Tendency is the Mode:

Definition: Sample Mode of the dataset is a value which occurs most frequently in our dataset.

Example: The Sample Mode of the following Dataset:

$$x : 0, -1, 2, 0, 0, 2, 3, 2, 1, 2$$

is 2.

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

Another measure of the Central Tendency is the Mode:

Definition: Sample Mode of the dataset is a value which occurs most frequently in our dataset.

Example: The Sample Mode of the following Dataset:

$$x : 0, -1, 2, 0, 0, 2, 3, 2, 1, 2$$

is 2.

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: Mode can be calculated even for the Nominal Scale Categorical Datasets

Mode Calculation in **R**

We do not have a simple command in basic **R** to calculate all Modes in **R**. Suggesion: write it by yourself!

Other Measures of the Central Tendency

In Stat, one also considers the following Measures of the Central Tendency:

- Midrange,
$$midrange(x) = \frac{x_{(1)} + x_{(n)}}{2}$$

Other Measures of the Central Tendency

In Stat, one also considers the following Measures of the Central Tendency:

- Midrange,

$$midrange(x) = \frac{x_{(1)} + x_{(n)}}{2}$$

- Hodges–Lehmann statistic,

$$HLS(x) = median\Big(mean(x_i, x_j) : j = 1, ..., n, i = 1, ..., j\Big).$$

Other Measures of the Central Tendency

In Stat, one also considers the following Measures of the Central Tendency:

- Midrange,

$$midrange(x) = \frac{x_{(1)} + x_{(n)}}{2}$$

- Hodges–Lehmann statistic,

$$HLS(x) = median\Big(mean(x_i, x_j) : j = 1, ..., n, i = 1, ..., j\Big).$$

- See others at Wiki

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

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

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:

```r
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
```