ASDS Statistics, YSU, Fall 2020

Lecture 07

Michael Poghosyan

30 Sep 2020

Contents

Last Lecture Recap

Numerical Summaries

Numerical Summaries

For 1D Datasets, we will consider the following Summaries:

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

Numerical Summaries

For 1D Datasets, we will consider the following Summaries:

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

Order Statistics

First we introduce the **Order Statistics**.

Order Statistics

First we introduce the **Order Statistics**.

Assume we have a 1D Numerical Dataset $x_1, x_2, ..., x_n$.

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.

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.

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

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

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,

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

Example

Example: Let x be the Dataset

$$-2, 1, 3, 2, 2, 1, 1$$

Find the 4-th and 5-th Order Statistics.

Statistical Measures for the Central Tendency/Location

Statistical Measures for the Central Tendency/Location

Here we want to answer to the 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

Some examples:

Example

Some examples:

- The average life expectancy for a person is

Example

Some examples:

- The average life expectancy for a person is 72 years, see WHO

Example

Some examples:

- The average life expectancy for a person is 72 years, see WHO
- On average, the number of times a person checks his/her phone is

Example

Some examples:

- The average life expectancy for a person is 72 years, see WHO
- On average, the number of times a person checks his/her phone is 58, see here

Example

Some examples:

- The average life expectancy for a person is 72 years, see WHO
- On average, the number of times a person checks his/her phone is 58, see here
- The average daily time spent on mobile phone is

Example

Some examples:

- The average life expectancy for a person is 72 years, see WHO
- On average, the number of times a person checks his/her phone is 58, see here
- The average daily time spent on mobile phone is 3h 15min, see here

Example

Some examples:

- The average life expectancy for a person is 72 years, see WHO
- On average, the number of times a person checks his/her phone is 58, see here
- The average daily time spent on mobile phone is 3h 15min, see here
- The average daily time spent on social media is

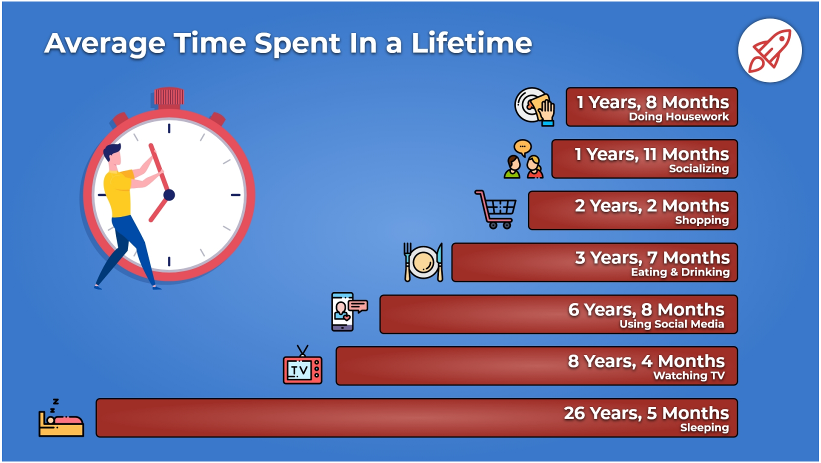
Example

Some examples:

- The average life expectancy for a person is 72 years, see WHO
- On average, the number of times a person checks his/her phone is 58, see here
- The average daily time spent on mobile phone is 3h 15min, see here
- The average daily time spent on social media is 144min, see here or here

Example

The average time spent during the lifetime, from this webpage:



Example

Sample Mean, Drawback: Sensitive to outliers (non-typical elements)

Example

Sample Mean, Drawback: Sensitive to outliers (non-typical elements)

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

Example

Sample Mean, Drawback: Sensitive to outliers (non-typical elements)

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

Example: Consider the following Dataset:

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

Example

Sample Mean, Drawback: Sensitive to outliers (non-typical elements)

When we talk about, say, that the mean/average Midterm grade is 68, we think about this like the grades are 68 plus/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
```

Example

Sample Mean, Drawback: Sensitive to outliers (non-typical elements)

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

Example

Sample Mean, Drawback: Sensitive to outliers (non-typical elements)

When we talk about, say, that the mean/average Midterm grade is 68, we think about this like the grades are 68 plus/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 plus-minus something? Not exactly.

Example

Sample Mean, Drawback: Sensitive to outliers (non-typical elements)

When we talk about, say, that the mean/average Midterm grade is 68, we think about this like the grades are 68 plus/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 plus-minus something? Not exactly.

Well, 115.7143 is not describing well 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 of points to be deleted from the both ends), 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 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,

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

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

Trimmed Sample Mean

Mathematically,

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

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

Idea of Trimming: Reduce the influence of outliers.

Trimmed Sample Mean

Mathematically,

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

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

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.

Examples

Example: Scores for the Figure Skating Competition is calculated using the Trimmed Mean, see, e.g., Wiki.

Examples

Example: Scores for the Figure Skating Competition is calculated using the Trimmed Mean, see, e.g., Wiki.

Example: *LIBOR* is calculated using a 22% Trimmed Mean, see, e.g., Wiki

Examples

Example: Scores for the Figure Skating Competition is calculated using the Trimmed Mean, see, e.g., Wiki.

Example: *LIBOR* is calculated using a 22% Trimmed Mean, see, e.g., Wiki

Example: People are calculating Trimmed CPI (Consumer Price Index), see here

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

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

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

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

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 $\frac{w_k}{\sum_{i=1}^n w_i}$.

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 $\frac{w_k}{\sum_{i=1}^{n} w_i}$.

Example: CPI (Consumer Price Index) is percentage change of a weighted average market basket of consumer goods and services purchased by households , see Wiki

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

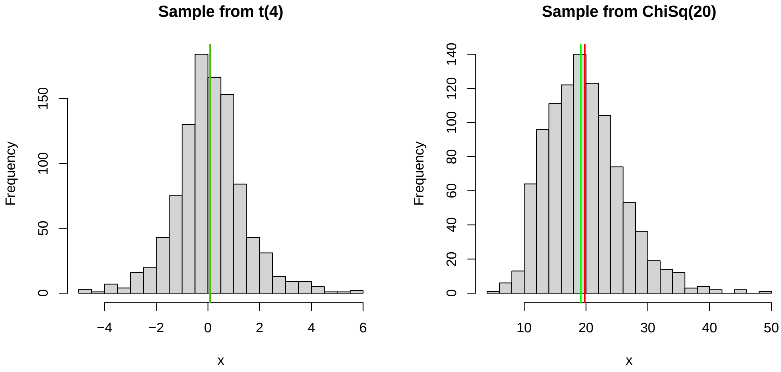
Mean and the Median

- If the Dataset is Symmetric, then the Mean and the Median of that Dataset coincide[1].

[1] Try to define the Symmetry of a Dataset and prove the above statement.

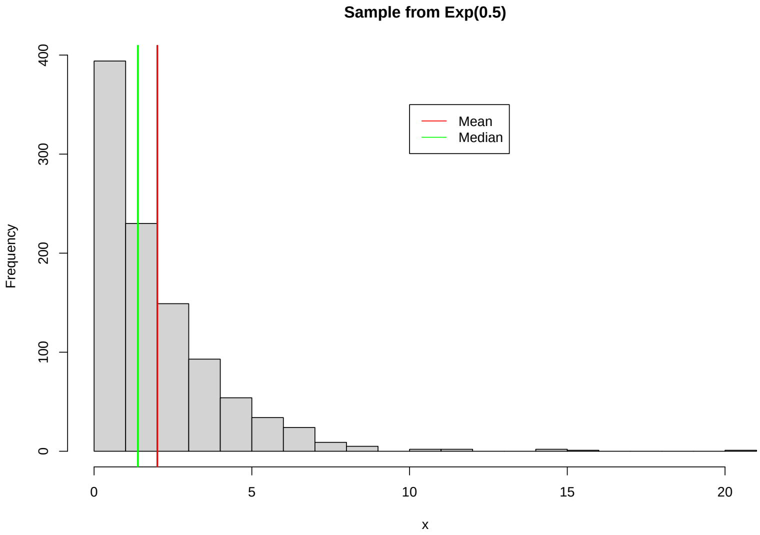
Mean and the Median

- If the Dataset is Symmetric, then the Mean and the Median of that Dataset coincide[1].
- If the Dataset is Skewed, then the Mean and the Median can be very different (Mean is in Red, and the Median is in Green):

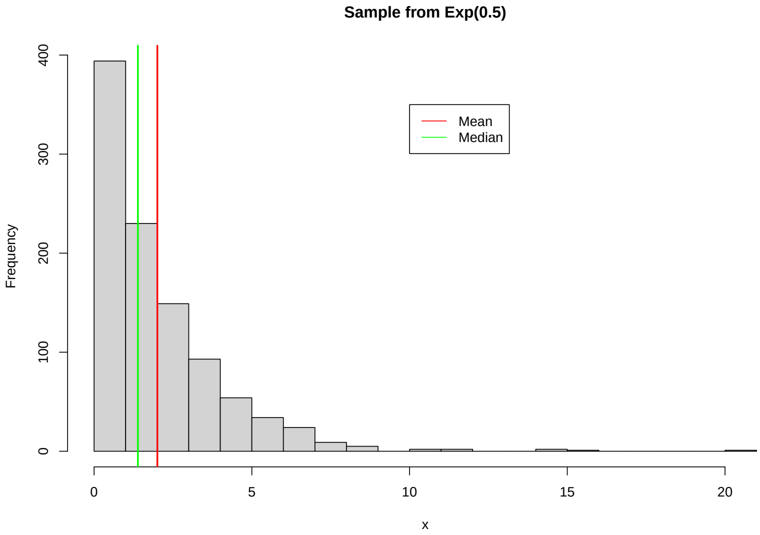


[1] Try to define the Symmetry of a Dataset and prove the above statement.

Mean and the Median



Mean and the Median



Example: (another one) See, e.g., the Distribution of Wealth article in Wikipedia.

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.

In other words, Mode is the value with the maximum Frequency in the Frequency (or the RelFreq) Table.

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.

In other words, Mode is the value with the maximum Frequency in the Frequency (or the RelFreq) Table.

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.

In other words, Mode is the value with the maximum Frequency in the Frequency (or the RelFreq) Table.

Example: The Sample Mode of the following Dataset:

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

is 2.

Remarks

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

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).

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.

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.

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.

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.

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.

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.

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.

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*