YSU ASDS, Statistics, Fall 2019

Lecture 06

Michael Poghosyan

14 Sep 2019

Descriptive Statistics

Contents

Last Lecture ReCap

- What is the drawback of the Sample Mean?

Last Lecture ReCap

- What is the drawback of the Sample Mean?
- What is the Sample Trimmed Mean?

Last Lecture ReCap

- What is the drawback of the Sample Mean?
- What is the Sample Trimmed Mean?
- What is the Winsorized Mean?

Last Lecture ReCap

- 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 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 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?

Statistical Measures for the Central Tendency/Location

Statistical Measures for the Central Tendency/Location

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**:

```
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    -
##        Baffin
##     -1068.729
```

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 $sd(x)$

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

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

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**:

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

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,

```r
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,

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

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

Or, you can use the following commands:

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

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

```r
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**.

BoxPlot

BoxPlot (or Box and Whiskers Plot) is another very common method of visualisation.

BoxPlot

BoxPlot (or Box and Whiskers Plot) is another very common method of visualisation. To draw the BoxPlot, we calculate the following:

BoxPlot

BoxPlot (or Box and Whiskers Plot) is another very common method of visualisation. To draw the BoxPlot, we calculate the following:

- The Quartiles $Q_1, Q_2 = Median, Q_3$

BoxPlot

BoxPlot (or Box and Whiskers Plot) is another very common method of visualisation. To draw the BoxPlot, we calculate the following:

- The Quartiles $Q_1, Q_2 = Median, Q_3$
- the Lower and Upper Fences
  $W_1 = \min\{x_i : x_i \geq Q_1 - 1.5 \cdot IQR\}$ and
  $W_2 = \max\{x_i : x_i \leq Q_3 + 1.5 \cdot IQR\}$,

BoxPlot

BoxPlot (or Box and Whiskers Plot) is another very common method of visualisation. To draw the BoxPlot, we calculate the following:

- The Quartiles $Q_1, Q_2 = Median, Q_3$
- the Lower and Upper Fences $W_1 = \min\{x_i : x_i \geq Q_1 - 1.5 \cdot IQR\}$ and $W_2 = \max\{x_i : x_i \leq Q_3 + 1.5 \cdot IQR\}$,i.e., the first and last observations lying in
$$\left[Q_1 - \frac{3}{2}IQR,\ Q_3 + \frac{3}{2}IQR\right];$$

BoxPlot

BoxPlot (or Box and Whiskers Plot) is another very common method of visualisation. To draw the BoxPlot, we calculate the following:

- The Quartiles $Q_1, Q_2 = Median, Q_3$
- the Lower and Upper Fences $W_1 = \min\{x_i : x_i \geq Q_1 - 1.5 \cdot IQR\}$ and $W_2 = \max\{x_i : x_i \leq Q_3 + 1.5 \cdot IQR\}$,i.e., the first and last observations lying in
$$\left[Q_1 - \frac{3}{2}IQR,\ Q_3 + \frac{3}{2}IQR\right];$$

the line joining that fences to corresponding quartiles are the *Whiskers*;

BoxPlot

BoxPlot (or Box and Whiskers Plot) is another very common method of visualisation. To draw the BoxPlot, we calculate the following:

- The Quartiles $Q_1, Q_2 = Median, Q_3$
- the Lower and Upper Fences $W_1 = \min\{x_i : x_i \geq Q_1 - 1.5 \cdot IQR\}$ and $W_2 = \max\{x_i : x_i \leq Q_3 + 1.5 \cdot IQR\}$,i.e., the first and last observations lying in

$$\left[Q_1 - \frac{3}{2}IQR,\ Q_3 + \frac{3}{2}IQR\right];$$

the line joining that fences to corresponding quartiles are the *Whiskers*;

- the set of all Outliers

$$O = \left\{x_i : x_i \notin \left[Q_1 - \frac{3}{2}IQR, Q_3 + \frac{3}{2}IQR\right]\right\}$$

BoxPlot, Example

Then we draw the points W_1, Q_1, Q_2, Q_3, W_2 on the real line and add all outliers, and make a box over $[Q_1, Q_3]$.

BoxPlot, Example

Then we draw the points W_1, Q_1, Q_2, Q_3, W_2 on the real line and add all outliers, and make a box over $[Q_1, Q_3]$.

Example: Draw the Boxplot of

$$x: \; 0, -2, 2, 1, 5, 6, 4, 1, 2, 1, 12$$

BoxPlot, Example

Then we draw the points W_1, Q_1, Q_2, Q_3, W_2 on the real line and add all outliers, and make a box over $[Q_1, Q_3]$.

Example: Draw the Boxplot of

$$x: \ 0, -2, 2, 1, 5, 6, 4, 1, 2, 1, 12$$

Solution: OTB;

BoxPlot, Example

Then we draw the points W_1, Q_1, Q_2, Q_3, W_2 on the real line and add all outliers, and make a box over $[Q_1, Q_3]$.

Example: Draw the Boxplot of

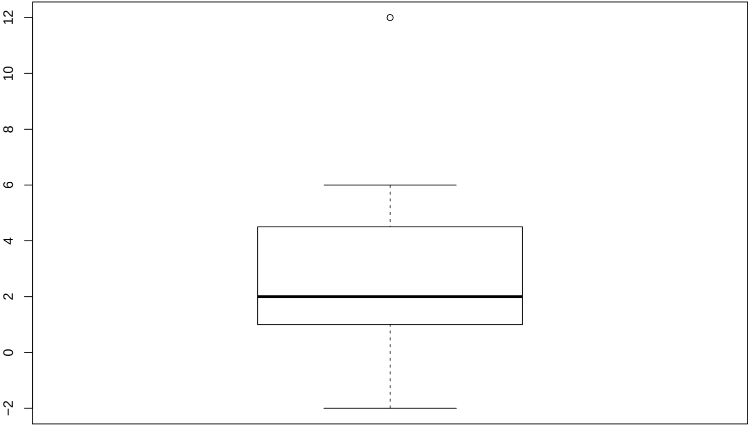
$$x: \; 0, -2, 2, 1, 5, 6, 4, 1, 2, 1, 12$$

Solution: OTB;

BoxPlot, Example

Now, using **R**:

```
x <- c(0, -2, 2, 1, 5, 6, 4, 1, 2, 1, 12)
boxplot(x)
```



BoxPlot, Example

Another view:

```r
x <- c(0, -2, 2, 1, 5, 6, 4, 1, 2, 1, 12)
boxplot(x, horizontal = T, col = "magenta")
```