YSU ASDS, Statistics, Fall 2019

Lecture 05

Michael Poghosyan

11 Sep 2019

Descriptive Statistics

Contents

Last Lecture ReCap

- What is a **Frequency Histogram**?

Last Lecture ReCap

- What is a **Frequency Histogram**?
- What is a **Density Histogram**?

Last Lecture ReCap

- What is a **Frequency Histogram**?
- What is a **Density Histogram**?
- What is a **KDE**?

Last Lecture ReCap

- What is a **Frequency Histogram**?
- What is a **Density Histogram**?
- What is a **KDE**?
- What is it for?

Stem-n-Leaf Plot

Another method to visualize a (not-so-large) 1D Dataset is to give the Stem-and-Leaf plot:

Assume we have a 1D Dataset $x_1, x_2, ..., x_n$. We represent each number x_k in the form

$$Stem \mid Leaf$$

Stem-n-Leaf Plot

Another method to visualize a (not-so-large) 1D Dataset is to give the Stem-and-Leaf plot:

Assume we have a 1D Dataset $x_1, x_2, ..., x_n$. We represent each number x_k in the form

$$Stem \mid Leaf$$

The *Leaf* need to consist only of 1 digit. The rest is in Stem.

Stem-n-Leaf Plot

Another method to visualize a (not-so-large) 1D Dataset is to give the Stem-and-Leaf plot:

Assume we have a 1D Dataset $x_1, x_2, ..., x_n$. We represent each number x_k in the form

$$Stem \mid Leaf$$

The *Leaf* need to consist only of 1 digit. The rest is in Stem. Sometimes, we do a rounding before making the S-n-L Plot, but, for simplicity, let's assume we are not doing any roundings.

Example, S-n-L Plot

Example: Assume our Dataset is:

$$x : 14, 23, 5, 16, 32, 22$$

Example, S-n-L Plot

Example: Assume our Dataset is:

$$x : 14, 23, 5, 16, 32, 22$$

Now, for 14, the Leaf is the last digit, 4, and the rest is the Stem, i.e., the Stem is 1. So we represent 14 as

$$1 \mid 4$$

Example, S-n-L Plot

Example: Assume our Dataset is:

$$x : 14, 23, 5, 16, 32, 22$$

Now, for 14, the Leaf is the last digit, 4, and the rest is the Stem, i.e., the Stem is 1. So we represent 14 as

$$1 \mid 4$$

Similarly, 23 will give

$$2 \mid 3$$

and 5 will give

Example, S-n-L Plot

Example: Assume our Dataset is:

$$x : 14, 23, 5, 16, 32, 22$$

Now, for 14, the Leaf is the last digit, 4, and the rest is the Stem, i.e., the Stem is 1. So we represent 14 as

$$1 \mid 4$$

Similarly, 23 will give

$$2 \mid 3$$

and 5 will give

$$0 \mid 5$$

Example, S-n-L Plot

Next, 16 will be

$$1 \mid 6$$

Example, S-n-L Plot

Next, 16 will be

$$1 \mid 6$$

and we combine this with the S-n-L representation of 14 (because they both starts by 1) to write

$$1 \mid 46$$

Example, S-n-L Plot

Next, 16 will be

$$1 \mid 6$$

and we combine this with the S-n-L representation of 14 (because they both starts by 1) to write

$$1 \mid 46$$

Finally, our Dataset's S-n-L Plot will be

```
x <- c(14, 23, 5, 16, 32, 22)
stem(x)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 5
##   1 | 46
##   2 | 23
##   3 | 2
```

Notes

- Sometimes **R** will do some roundings before S-n-L Plotting

Notes

- Sometimes **R** will do some roundings before S-n-L Plotting
- Usually, Stems are ordered, and Leafs are sorted in the increasing order (ordered SnL Plot)

Notes

- Sometimes **R** will do some roundings before S-n-L Plotting
- Usually, Stems are ordered, and Leafs are sorted in the increasing order (ordered SnL Plot)
- The top row, the explanation about the position of |, is the **key**, is to recover the dataset.

Example, SnL Plot

Here is another example: we use again the *airquality* Dataset, but now, the *Wind* Variable:

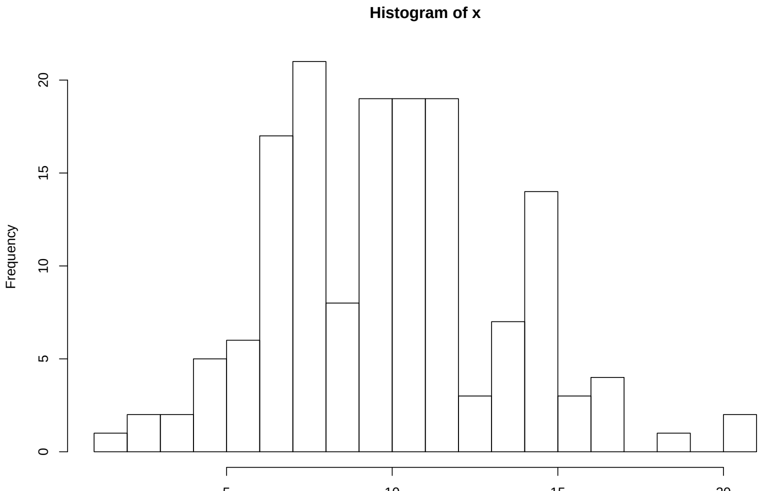
```
x <- airquality$Wind
stem(x)
```

```
##
##   The decimal point is at the |
##
##    1 | 7
##    2 | 38
##    3 | 4
##    4 | 016666
##    5 | 111777
##    6 | 3333333399999999
##    7 | 4444444444
##    8 | 0000000000066666666
##    9 | 2222222277777777777
##   10 | 3333333333399999999
##   11 | 555555555555555
##   12 | 0000666
##   13 | 2288888
##   14 | 33333399999999
##   15 | 555
##   16 | 1666
##   17 |
##   18 | 4
##   19 |
##   20 | 17
```

Example, SnL Plot

Let's draw the Histogram of the same Dataset:

```r
x <- airquality$Wind
hist(x, breaks = 15)
```



Notes

- From the previous 2 graphs, you can see that SnL Plot is, in some sense, an inverted Histogram

Notes

- From the previous 2 graphs, you can see that SnL Plot is, in some sense, an inverted Histogram
- Pros of SnL is that

Notes

- From the previous 2 graphs, you can see that SnL Plot is, in some sense, an inverted Histogram
- Pros of SnL is that we can recover the Dataset from it (if no rounding was made), but not from the Histogram

Notes

- From the previous 2 graphs, you can see that SnL Plot is, in some sense, an inverted Histogram
- Pros of SnL is that we can recover the Dataset from it (if no rounding was made), but not from the Histogram
- Cons of SnL is that it is for a small-size Dataset

Notes

- From the previous 2 graphs, you can see that SnL Plot is, in some sense, an inverted Histogram
- Pros of SnL is that we can recover the Dataset from it (if no rounding was made), but not from the Histogram
- Cons of SnL is that it is for a small-size Dataset

Say, you can try

```
x <- rnorm(10000)
stem(x)
```

Some Parameters of the SnL Plot

Let's run the following code:

```
set.seed(77777)
x <- sample(1:30, size = 20, replace = T)
stem(x)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 1113
##   0 | 6689
##   1 | 0023333
##   1 | 9
##   2 | 0124
```

Some Parameters of the SnL Plot

Let's runt the following code:

```
set.seed(77777)
x <- sample(1:30, size = 20, replace = T)
stem(x, scale = 2)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 1113
##   0 | 6689
##   1 | 0023333
##   1 | 9
##   2 | 0124
```

Some Parameters of the SnL Plot

Let's runt the following code:

```
set.seed(77777)
x <- sample(1:30, size = 20, replace = T)
stem(x, scale = 0.5)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 11136689
##   1 | 00233339
##   2 | 0124
```

Some Additions: Comparing 2 Groups, Back-to-Back Histograms and SnL Plots

Sometimes we want to compare the values of the same variable for two different groups, say, the Height Variable for the Man and Woman groups.

Some Additions: Comparing 2 Groups, Back-to-Back Histograms and SnL Plots

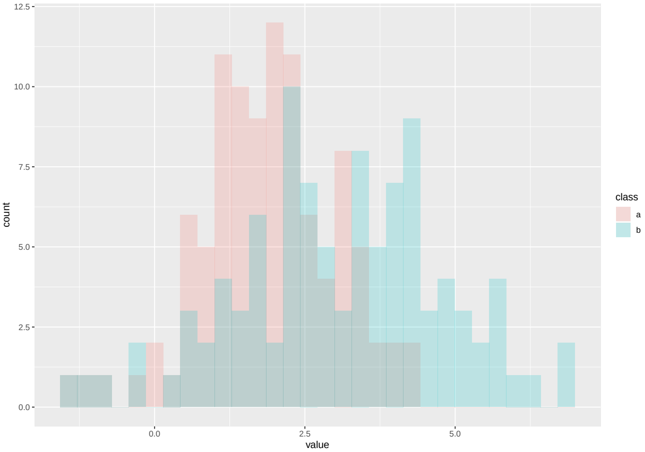
Sometimes we want to compare the values of the same variable for two different groups, say, the Height Variable for the Man and Woman groups. Then, we can use different colors to visualize the difference.

Example

Here is a synthetic (artifical) example:

```
library(ggplot2)
v1 <- rnorm(100,2,1); v2 <- rnorm(100,3,2)
df <- data.frame(value = c(v1, v2), class = rep(c("a","b"), each=100))
ggplot(df, aes(x=value, fill=class)) + geom_histogram(alpha=0.2, position="identity")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Example

And sometimes Back-to-back Histograms, SnL Plots or Barplots can help.

Example

We do not have a command to draw a Back-to-Back SnL Plot, so we load the *aplpack* package:

```
x <- sample(1:30, size = 50, replace = T);
y <- sample(1:30, size = 50, replace = T);
aplpack::stem.leaf.backback(x,y,  rule.line = "Sturges")
```

```
## ________________________________________________
##   1 | 2: represents 12, leaf unit: 1
##                      x      y
## ________________________________________________
##    4            3211| 0* |11124            5
##   18   99877776666555| 0. |56677778888999  19
##   22            4321| 1* |011234          (6)
##   (9)      988766555| 1. |5667789         (7)
##   19      4443322100| 2* |1222233444      18
##    8         9998777| 2. |56778            7
##    1               0| 3* |00               2
## ________________________________________________
## n:                50      50
## ________________________________________________
```

Example

Here is a real Back-to-Back Histogram Plot: Selfiecity.

Visualizing 2D Data

In case we have a 2D numerical Dataset

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n),$$

we usually do the ScaterPlot - the plot of all points (x_i, y_i), $i = 1, ...n$.

Example

Say, consider again the *cars* Dataset:

```r
head(cars, 3)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
```

```r
str(cars)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
##  $ dist : num  2 10 4 22 16 10 18 26 34 17 ...
```

Example

Say, consider again the *cars* Dataset:

```r
head(cars, 3)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
```

```r
str(cars)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
##  $ dist : num  2 10 4 22 16 10 18 26 34 17 ...
```

It has 2 Variables: *Speed* and *Distance*, and 50 Observations.

Example

Say, consider again the *cars* Dataset:

```
head(cars, 3)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
```

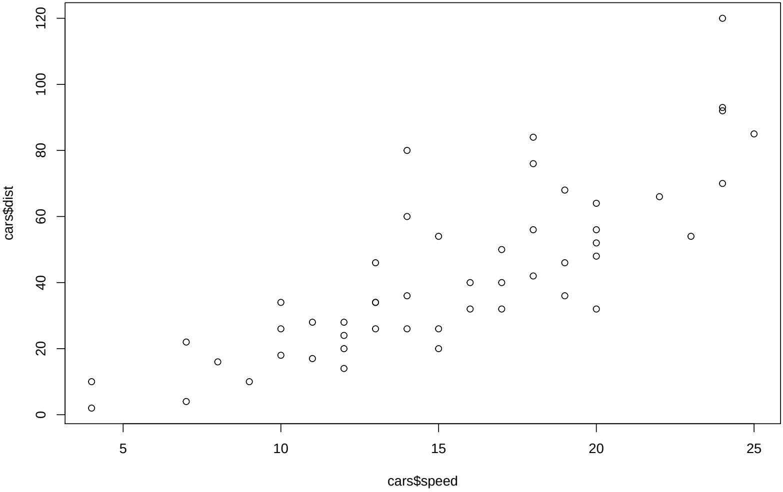
```
str(cars)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
##  $ dist : num  2 10 4 22 16 10 18 26 34 17 ...
```

It has 2 Variables: *Speed* and *Distance*, and 50 Observations. Let us do the ScatterPlot of Observations:

ScatterPlot

```
plot(cars$speed, cars$dist)
```



Notes

- In this graph you can see that there is some relationship between the *Speed* and *Distance*, there is a *trend*: if the speed gets larger, the (stopping) distance is tending to increase.

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one.

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one.
Some ideas for multidimensional Visualizations:

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one. Some ideas for multidimensional Visualizations:

- One can draw 3D in 3D $\ddot\smile$,

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one. Some ideas for multidimensional Visualizations:

- One can draw 3D in 3D $\ddot\smile$, give some 3D Histograms and KDEs

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one. Some ideas for multidimensional Visualizations:

- One can draw 3D in 3D $\ddot\smile$, give some 3D Histograms and KDEs
- One can draw 3D in 2D, using the 3rd variable as the Color (not in all cases, of course)

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one. Some ideas for multidimensional Visualizations:

- One can draw 3D in 3D $\ddot\smile$, give some 3D Histograms and KDEs
- One can draw 3D in 2D, using the 3rd variable as the Color (not in all cases, of course)
- One can add the 4th Dimension by using the Size of Points
- And add the 5-th one by using the Shape of Points,...

Examples

See, for example, beautiful visualizations by **Hans Rosling**.

Examples

See, for example, beautiful visualizations by **Hans Rosling**. Say, this short one: Hans Rosling's 200 Countries, 200 Years, 4 Minutes - The Joy of Stats - BBC Four

Examples

See, for example, beautiful visualizations by **Hans Rosling**. Say, this short one: Hans Rosling's 200 Countries, 200 Years, 4 Minutes - The Joy of Stats - BBC Four

Or, the following one: Gender Gap in Earnings per University

Additions: Multidimensional Graphs

- One can do the Pairs Plot

Additions: Multidimensional Graphs

- One can do the Pairs Plot
- One can draw the HeatMap

Additions: Multidimensional Graphs

- One can do the Pairs Plot
- One can draw the HeatMap
- One can use a Dimensionality Reduction Methods to Visualize some high dimensional Data

Additions: Multidimensional Graphs

- One can do the Pairs Plot
- One can draw the HeatMap
- One can use a Dimensionality Reduction Methods to Visualize some high dimensional Data
- etc ...

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

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 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

```r
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

```
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,

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

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

Trimmed Sample Mean

Mathematically,

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

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

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

Trimmed Sample Mean

Mathematically,

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

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

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

Idea of Trimming: Reduce the influence of outliers.

Trimmed Sample Mean

Mathematically,

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

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

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

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

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

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

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
```