YSU ASDS, Statistics, Fall 2019

Lecture 09

Michael Poghosyan

21 Sep 2019

Descriptive Statistics

Happy Independence Day!

Contents

Last Lecture ReCap

- How to check (visually) if a Dataset is coming from a Normal Distribution?

Last Lecture ReCap

- How to check (visually) if a Dataset is coming from a Normal Distribution?
- How to check (visually) if a Dataset is coming from a Pareto Distribution?

Q-Q Plots, Theoretical vs Theoretical Distribution

Assume now we have two Theoretical Distributions (say, given by their CDFs F and G).

Q-Q Plots, Theoretical vs Theoretical Distribution

Assume now we have two Theoretical Distributions (say, given by their CDFs F and G). The Problem is to estimate visually which Distribution has fatter tails.

Q-Q Plots, Theoretical vs Theoretical Distribution

Assume now we have two Theoretical Distributions (say, given by their CDFs F and G). The Problem is to estimate visually which Distribution has fatter tails.

To answer this question, we again take some levels of quantiles, say, for some n,

$$\alpha = \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}$$

and then draw the points (q_α^F, q_α^G), where q_α^F is the α-quantile of the Theoretical Distribution with the CDF F, and q_α^G is the α-quantile of the Theoretical Distribution with the CDF G.

Q-Q Plots, Theoretical vs Theoretical Distribution

Assume now we have two Theoretical Distributions (say, given by their CDFs F and G). The Problem is to estimate visually which Distribution has fatter tails.

To answer this question, we again take some levels of quantiles, say, for some n,

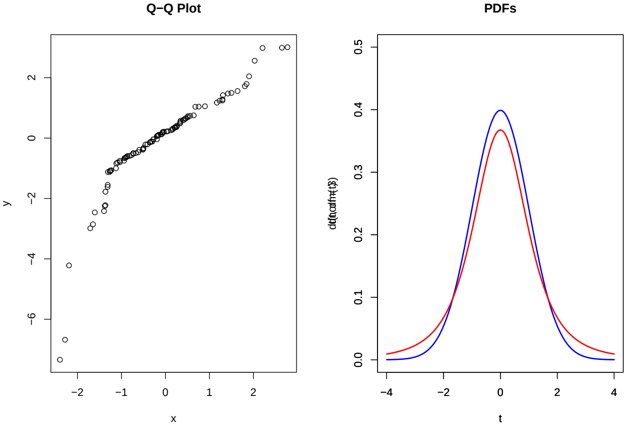
$$\alpha = \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}$$

and then draw the points (q_α^F, q_α^G), where q_α^F is the α-quantile of the Theoretical Distribution with the CDF F, and q_α^G is the α-quantile of the Theoretical Distribution with the CDF G.

Idea: If G has fatter tails on both sides than F, then we will have graphically some cubic-function graph shape Quantiles.

Some Experiments

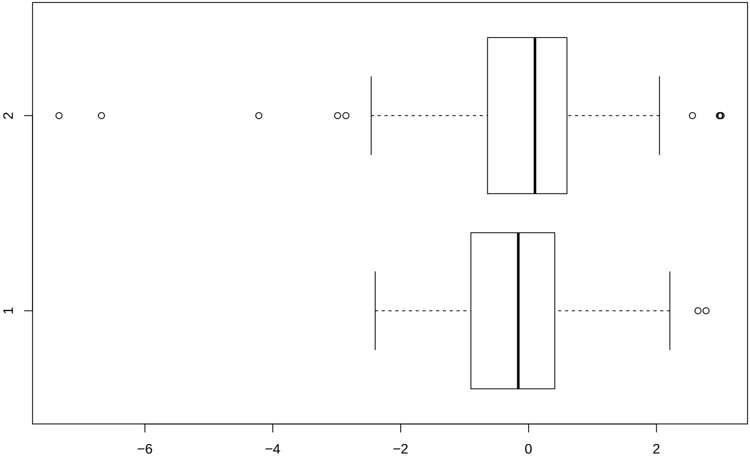
```
par(mfrow = c(1,2))
x <- rnorm(100, mean=0, sd=1); y <- rt(100, df = 3)
qqplot(x,y, main = "Q-Q Plot")
t <- seq(-4,4,0.01)
plot(t, dnorm(t), type = "l", xlim = c(-4,4), ylim = c(0, 0.5), col ="blue", lwd = 2, main = "PDFs")
par(new = TRUE)
plot(t, dt(t, df = 3), type = "l", xlim = c(-4,4), ylim = c(0, 0.5), col ="red", lwd = 2)
```



Some Experiments

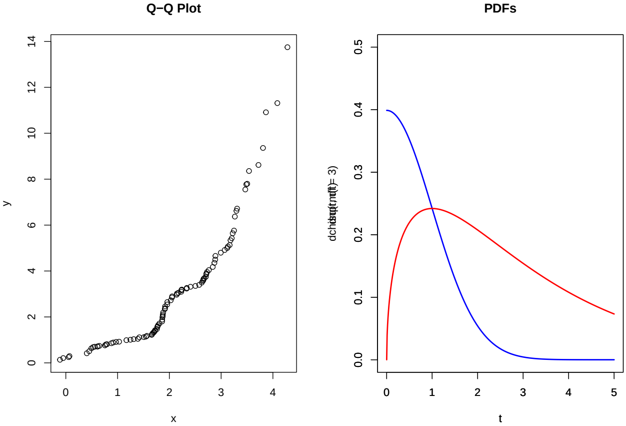
The above Datasets, using BoxPlots:

```
boxplot(x,y, horizontal = T)
```



Some Experiments

```
par(mfrow = c(1,2))
x <- rnorm(100, mean=2, sd=1); y <- rchisq(200, df = 3)
qqplot(x,y, main = "Q-Q Plot")
t <- seq(0,5,0.01)
plot(t, dnorm(t), type = "l", xlim = c(0,5), ylim = c(0, 0.5), col ="blue", lwd = 2,  main = "PDFs")
par(new = TRUE)
plot(t, dchisq(t, df = 3), type = "l", xlim = c(0,5), ylim = c(0, 0.5), col ="red", lwd = 2)
```

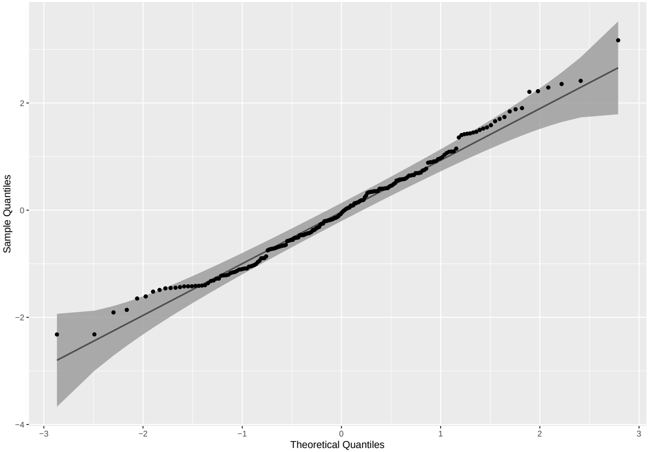


Addition, Q-Q Plot

here you can find some interpretaitons of different shapes of Q-Q Plots: StackExchange Page.

Addition, Q-Q Plot with a Confidence Band

```
require(qqplotr)
x <- data.frame(variable = rnorm(200))
ggplot(data = x, mapping = aes(sample = variable)) + stat_qq_band() +
  stat_qq_line() + stat_qq_point() + labs(x = "Theoretical Quantiles", y = "Sample Quantiles")
```



Numerical Summaries for Bivariate Data

Sample Covariance and the Correlation Coefficient

Assume now we have a bivariate Dataset

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

or just two 1D Datasets of the same size:

$$x: \; x_1, ..., x_n \qquad and \qquad y: \; y_1, ..., y_n.$$

Sample Covariance and the Correlation Coefficient

Assume now we have a bivariate Dataset

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

or just two 1D Datasets of the same size:

$$x: \; x_1, ..., x_n \qquad and \qquad y: \; y_1, ..., y_n.$$

Our aim is to see if some linear relationship, association exists between x and y.

Sample Covariance and the Correlation Coefficient

Assume now we have a bivariate Dataset

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

or just two 1D Datasets of the same size:

$$x: \; x_1, ..., x_n \qquad and \qquad y: \; y_1, ..., y_n.$$

Our aim is to see if some linear relationship, association exists between x and y. Of course, the best way is to visualize our Dataset by a ScatterPlot.

Sample Covariance and the Correlation Coefficient

Assume now we have a bivariate Dataset

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

or just two 1D Datasets of the same size:

$$x: \ x_1, ..., x_n \qquad \textit{and} \qquad y: \ y_1, ..., y_n.$$

Our aim is to see if some linear relationship, association exists between x and y. Of course, the best way is to visualize our Dataset by a ScatterPlot.

Now we want to answer, numerically, how strong/week is the linear relationship between our variables x and y.

Sample Covariance

The **Sample Covariance** of Variables (1D Datasets) x and y is

$$cov(x, y) = s_{xy} = \frac{\sum_{k=1}^{n} (x_k - \overline{x}) \cdot (y_k - \overline{y})}{n}$$

Sample Covariance

The **Sample Covariance** of Variables (1D Datasets) x and y is

$$cov(x, y) = s_{xy} = \frac{\sum_{k=1}^{n} (x_k - \overline{x}) \cdot (y_k - \overline{y})}{n}$$

or

$$cov(x, y) = s_{xy} = \frac{\sum_{k=1}^{n} (x_k - \overline{x}) \cdot (y_k - \overline{y})}{n - 1}$$

Sample Covariance

The **Sample Covariance** of Variables (1D Datasets) x and y is

$$cov(x, y) = s_{xy} = \frac{\sum_{k=1}^{n} (x_k - \overline{x}) \cdot (y_k - \overline{y})}{n}$$

or

$$cov(x, y) = s_{xy} = \frac{\sum_{k=1}^{n} (x_k - \overline{x}) \cdot (y_k - \overline{y})}{n - 1}$$

Here $\overline{x}$ and $\overline{y}$ are the Sample Means for the Datasets x and y.

Sample Covariance

The **Sample Covariance** of Variables (1D Datasets) x and y is

$$cov(x,y) = s_{xy} = \frac{\sum_{k=1}^{n}(x_k - \overline{x}) \cdot (y_k - \overline{y})}{n}$$

or

$$cov(x,y) = s_{xy} = \frac{\sum_{k=1}^{n}(x_k - \overline{x}) \cdot (y_k - \overline{y})}{n-1}$$

Here $\overline{x}$ and $\overline{y}$ are the Sample Means for the Datasets x and y.

Note: Recall that for a r.v. X, $Cov(X,X) = Var(X)$.

Sample Covariance

The **Sample Covariance** of Variables (1D Datasets) x and y is

$$cov(x, y) = s_{xy} = \frac{\sum_{k=1}^{n} (x_k - \overline{x}) \cdot (y_k - \overline{y})}{n}$$

or

$$cov(x, y) = s_{xy} = \frac{\sum_{k=1}^{n} (x_k - \overline{x}) \cdot (y_k - \overline{y})}{n - 1}$$

Here $\overline{x}$ and $\overline{y}$ are the Sample Means for the Datasets x and y.

Note: Recall that for a r.v. X, $Cov(X, X) = Var(X)$. Here, for Datasets, we have two definitions for the Sample Variance $var(x)$. And we give two definitions of the Sample Covariance, so the property $cov(x, x) = var(x)$ will hold in both cases.

Sample Covariance

Definition: We say that the Variables (Datasets) x and y are **uncorrelated**, if $cov(x, y) = 0$.

Sample Covariance

Definition: We say that the Variables (Datasets) x and y are **uncorrelated**, if $cov(x, y) = 0$.

Remark: In Probability, we have 2 notions: *Independence* and *Corelation*. Here, in the case of Datasets, we do not have the notion of *Independence*

Example

Here is the **R** code to calculate the Covariance between the `Speed` and `Dist` variables in the `cars` Dataset:

```r
cov(cars$speed, cars$dist)
```

```
## [1] 109.9469
```

Sample Correlation Coefficient

Another measure of the linear relationship between the Variables x and y of Bivariate Dataset is the *Pearson's Correlation Coefficient*:

Sample Correlation Coefficient

Another measure of the linear relationship between the Variables x and y of Bivariate Dataset is the *Pearson's Correlation Coefficient*:

Definition: The **Sample Correlation Coefficient** of x and y is

$$cor(x,y) = \rho_{xy} = \frac{cov(x,y)}{\sqrt{Var(x) \cdot Var(y)}} = \frac{cov(x,y)}{sd(x) \cdot sd(y)} = \frac{s_{xy}}{s_x \cdot s_y},$$

where s_x and s_y are the standard deviations for x and y, respectively.

Sample Correlation Coefficient

Another measure of the linear relationship between the Variables x and y of Bivariate Dataset is the *Pearson's Correlation Coefficient*:

Definition: The **Sample Correlation Coefficient** of x and y is

$$cor(x,y) = \rho_{xy} = \frac{cov(x,y)}{\sqrt{Var(x) \cdot Var(y)}} = \frac{cov(x,y)}{sd(x) \cdot sd(y)} = \frac{s_{xy}}{s_x \cdot s_y},$$

where s_x and s_y are the standard deviations for x and y, respectively.

If $s_x = 0$ or $s_y = 0$, then we take $cor(x,y) = 0$ by definition.

Sample Correlation Coefficient

Another measure of the linear relationship between the Variables x and y of Bivariate Dataset is the *Pearson's Correlation Coefficient*:

Definition: The **Sample Correlation Coefficient** of x and y is

$$cor(x,y) = \rho_{xy} = \frac{cov(x,y)}{\sqrt{Var(x) \cdot Var(y)}} = \frac{cov(x,y)}{sd(x) \cdot sd(y)} = \frac{s_{xy}}{s_x \cdot s_y},$$

where s_x and s_y are the standard deviations for x and y, respectively.

If $s_x = 0$ or $s_y = 0$, then we take $cor(x,y) = 0$ by definition.

Note: Please note that we need to calculate the Standard Deviations and Covariance by using the same denominator: either everywhere take n, or take everywhere $n-1$.

Sample Correlation Coefficient

In both cases, when one calculates Standard Deviations and Covariance by using n simultaneously or $n-1$ simultaneously in the denominator, we will obtain

$$cor(x,y) = \rho_{xy} = \frac{\sum_{k=1}^{n}(x_k - \overline{x}) \cdot (y_k - \overline{y})}{\sqrt{\sum_{k=1}^{n}(x_k - \overline{x})^2 \cdot \sum_{k=1}^{n}(y_k - \overline{y})^2}}$$

Sample Correlation Coefficient

In both cases, when one calculates Standard Deviations and Covariance by using n simultaneously or $n-1$ simultaneously in the denominator, we will obtain

$$cor(x,y) = \rho_{xy} = \frac{\sum_{k=1}^{n}(x_k - \overline{x}) \cdot (y_k - \overline{y})}{\sqrt{\sum_{k=1}^{n}(x_k - \overline{x})^2 \cdot \sum_{k=1}^{n}(y_k - \overline{y})^2}}$$

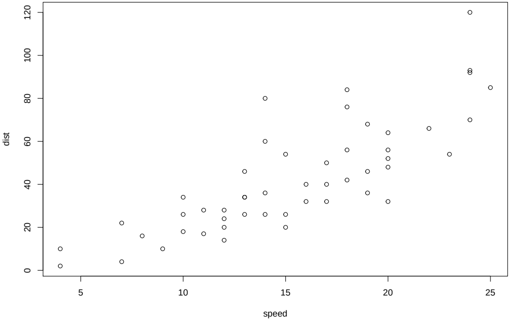
Another formula to calc the correlation coefficient is

$$cor(x,y) = \rho_{xy} = \frac{\sum_{k=1}^{n} x_k y_k - n \cdot \overline{x} \cdot \overline{y}}{\sqrt{\sum_{k=1}^{n} x_k^2 - n \cdot (\overline{x})^2} \cdot \sqrt{\sum_{k=1}^{n} y_k^2 - n \cdot (\overline{y})^2}}.$$

Examples:

Now, the **R** code to calculate the Covariance between the `Speed` and `Dist` variables in the `cars` Dataset:

```
plot(dist~speed, data = cars)
```



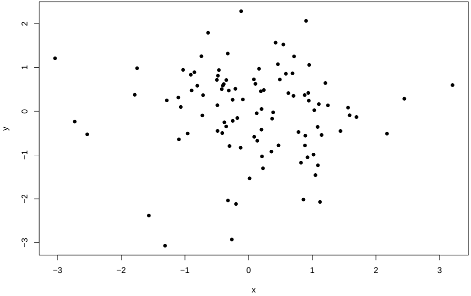

```
cor(cars$speed, cars$dist)
```

```
## [1] 0.8068949
```

Examples:

Some simulations:

```
x <- rnorm(100); y <- rnorm(100);
plot(x,y, pch=16)
```



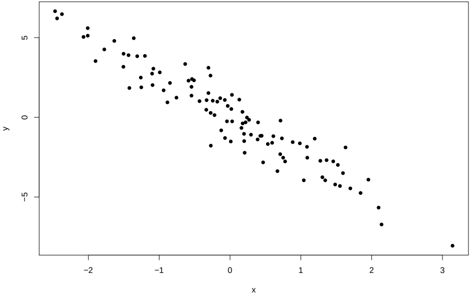

```
c(cor(x,y), cov(x,y))
```

```
## [1] -0.04749377 -0.04916903
```

Examples:

Some simulations:

```
x <- rnorm(100); y <- -2.4*x + rnorm(100);
plot(x,y, pch=16)
```




```
c(cor(x,y), cov(x,y))
```

```
## [1] -0.9512799 -3.2496763
```

Examples:

Let us now use the `state.x77` Dataset from **R**:

```r
head(state.x77)
```

```
##            Population Income Illiteracy Life Exp Murder HS Gr
## Alabama          3615   3624        2.1    69.05   15.1    41
## Alaska            365   6315        1.5    69.31   11.3    66
## Arizona          2212   4530        1.8    70.55    7.8    58
## Arkansas         2110   3378        1.9    70.66   10.1    39
## California      21198   5114        1.1    71.71   10.3    62
## Colorado         2541   4884        0.7    72.06    6.8    63
##              Area
## Alabama     50708
## Alaska     566432
## Arizona    113417
## Arkansas    51945
## California 156361
## Colorado   103766
```

Examples:

Let us now use the `state.x77` Dataset from **R**:

```
head(state.x77)
```

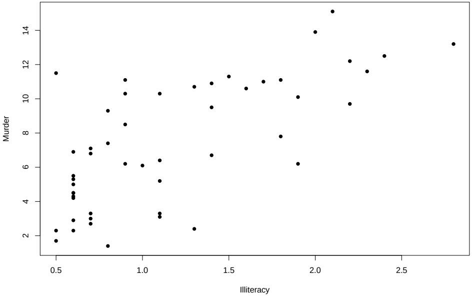
```
##            Population Income Illiteracy Life Exp Murder HS Gr
## Alabama          3615   3624        2.1    69.05   15.1    41
## Alaska            365   6315        1.5    69.31   11.3    66
## Arizona          2212   4530        1.8    70.55    7.8    58
## Arkansas         2110   3378        1.9    70.66   10.1    39
## California      21198   5114        1.1    71.71   10.3    62
## Colorado         2541   4884        0.7    72.06    6.8    63
##              Area
## Alabama     50708
## Alaska     566432
## Arizona    113417
## Arkansas    51945
## California 156361
## Colorado   103766
```

It is not of the DataFrame format, so we change it to DataFrame:

```
state <- as.data.frame(state.x77)
```

Examples:

```
plot(Murder~Illiteracy, data = state, pch=16)
```




```
cor(state$Illiteracy, state$Murder)
```

```
## [1] 0.7029752
```

Examples:

Question: How to generate samples x, y with some given Correlation Coefficient?

Examples:

Question: How to generate samples x, y with some given Correlation Coefficient?

Answer:

Examples:

Question: How to generate samples x, y with some given Correlation Coefficient?

Answer: Exactly. I do not know it too $\ddot\smile$

Examples:

Question: How to generate samples x, y with some given Correlation Coefficient?

Answer: Exactly. I do not know it too $\ddot\smile$ Kidding, of course, I know.

Examples:

Question: How to generate samples x, y with some given Correlation Coefficient?

Answer: Exactly. I do not know it too $\ddot\smile$ Kidding, of course, I know.

Say, we want to have Datasets x, y of size n with $cor(x, y) = \rho \in (-1, 1)$.

Examples:

Question: How to generate samples x, y with some given Correlation Coefficient?

Answer: Exactly. I do not know it too $\ddot{\smile}$ Kidding, of course, I know.

Say, we want to have Datasets x, y of size n with $cor(x, y) = \rho \in (-1, 1)$.

One of the possible methods: take a Matrix

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

which is **Positive Definite**, take any 2D vector, say $\mu = [0, 0]^T$, and generate a Sample of size n from the Bivariate Normal Distribution $\mathcal{N}(\mu, \Sigma)$.

Examples:

Question: How to generate samples x, y with some given Correlation Coefficient?

Answer: Exactly. I do not know it too $\ddot\smile$ Kidding, of course, I know.

Say, we want to have Datasets x, y of size n with $cor(x, y) = \rho \in (-1, 1)$.

One of the possible methods: take a Matrix

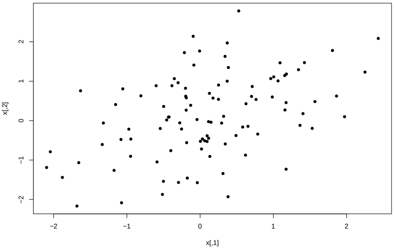
$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

which is **Positive Definite**, take any 2D vector, say $\mu = [0, 0]^T$, and generate a Sample of size n from the Bivariate Normal Distribution $\mathcal{N}(\mu, \Sigma)$.

Then, the $cor(x, y)$ will be approximately ρ (and it will approach ρ as $n \to +\infty$).

Example

```r
rho <- 0.35
covmatrix <- matrix(c(1,rho, rho, 1), nrow = 2)
mu <- c(0,0)
x <- mvtnorm::rmvnorm(100, mean = mu, sigma = covmatrix)
plot(x, pch = 16)
```




```r
cor(x)
```

```
##             [,1]      [,2]
## [1,] 1.0000000 0.4445266
## [2,] 0.4445266 1.0000000
```

Properties of the Sample Covariance

- $cov(x, y) = cov(y, x)$;

Properties of the Sample Covariance

- $cov(x, y) = cov(y, x)$;
- For any Datasets x, y, z and real numbers α, β,

$$cov(\alpha \cdot x + \beta \cdot y, z) = \alpha \cdot cov(x, z) + \beta \cdot cov(y, z);$$

Properties of the Sample Covariance

- $cov(x, y) = cov(y, x)$;
- For any Datasets x, y, z and real numbers α, β,

$$cov(\alpha \cdot x + \beta \cdot y, z) = \alpha \cdot cov(x, z) + \beta \cdot cov(y, z);$$

- For any Dataset x,

$$cov(x, x) = var(x)$$

Properties of the Sample Correlation Coefficient

- For any Datasets x, y,

$$-1 \leq \rho_{xy} \leq 1;$$

[1] Or $x_i = a \cdot y_i + b$ for any $i = 1, ..., n$ (maybe for another a and b).
[2] Or $x_i = a \cdot y_i + b$ for any $i = 1, ..., n$ (maybe for another a and b).

Properties of the Sample Correlation Coefficient

- For any Datasets x, y,

$$-1 \leq \rho_{xy} \leq 1;$$

- $\rho_{xy} = 1$ iff there exists a constant $a > 0$ and $b \in \mathbb{R}$ such that[1] $y_i = a \cdot x_i + b$ for any $i = 1, ..., n$.

[1] Or $x_i = a \cdot y_i + b$ for any $i = 1, ..., n$ (maybe for another a and b).

[2] Or $x_i = a \cdot y_i + b$ for any $i = 1, ..., n$ (maybe for another a and b).

Properties of the Sample Correlation Coefficient

- For any Datasets x, y,
$$-1 \leq \rho_{xy} \leq 1;$$

- $\rho_{xy} = 1$ iff there exists a constant $a > 0$ and $b \in \mathbb{R}$ such that[1] $y_i = a \cdot x_i + b$ for any $i = 1, ..., n$.
- $\rho_{xy} = -1$ iff there exists a constant $a < 0$ and $b \in \mathbb{R}$ such that[2] $y_i = a \cdot x_i + b$ for any $i = 1, ..., n$.

[1] Or $x_i = a \cdot y_i + b$ for any $i = 1, ..., n$ (maybe for another a and b).
[2] Or $x_i = a \cdot y_i + b$ for any $i = 1, ..., n$ (maybe for another a and b).

Pros/Cons of Sample Covariance and Correlation Coefficient

- Covariance is *linear*, correlation is not

Pros/Cons of Sample Covariance and Correlation Coefficient

- Covariance is *linear*, correlation is not
- Correlation is scale-invariant: if we will change the scale of one or both Datasets, then the Correlation Coefficient will not be changed (but the Covariance will be).

Pros/Cons of Sample Covariance and Correlation Coefficient

- Covariance is *linear*, correlation is not
- Correlation is scale-invariant: if we will change the scale of one or both Datasets, then the Correlation Coefficient will not be changed (but the Covariance will be).

Say, if x is a Dataset of heights of some persons, in centimeters, y their weights in grams, and if x' will be the same heights Dataset using meters as units, and y' will be the weights in Kg-s, then $cov(x, y)$ and $cov(x', y')$ will not be the same, but $cor(x, y) = cor(x', y')$.

Pros/Cons of Sample Covariance and Correlation Coefficient

- Covariance is *linear*, correlation is not
- Correlation is scale-invariant: if we will change the scale of one or both Datasets, then the Correlation Coefficient will not be changed (but the Covariance will be).

Say, if x is a Dataset of heights of some persons, in centimeters, y their weights in grams, and if x' will be the same heights Dataset using meters as units, and y' will be the weights in Kg-s, then $cov(x, y)$ and $cov(x', y')$ will not be the same, but $cor(x, y) = cor(x', y')$.

- If $cov(x, y) > cov(z, t)$, we cannot state that the relationship between x and y is stronger than the relationship between z and t.

Pros/Cons of Sample Covariance and Correlation Coefficient

- Covariance is *linear*, correlation is not
- Correlation is scale-invariant: if we will change the scale of one or both Datasets, then the Correlation Coefficient will not be changed (but the Covariance will be).

Say, if x is a Dataset of heights of some persons, in centimeters, y their weights in grams, and if x' will be the same heights Dataset using meters as units, and y' will be the weights in Kg-s, then $cov(x, y)$ and $cov(x', y')$ will not be the same, but $cor(x, y) = cor(x', y')$.

- If $cov(x, y) > cov(z, t)$, we cannot state that the relationship between x and y is stronger than the relationship between z and t. But if $cor(x, y) > cor(z, t)$, we can.

Pros/Cons of Sample Covariance and Correlation Coefficient

- Covariance is *linear*, correlation is not
- Correlation is scale-invariant: if we will change the scale of one or both Datasets, then the Correlation Coefficient will not be changed (but the Covariance will be).

Say, if x is a Dataset of heights of some persons, in centimeters, y their weights in grams, and if x' will be the same heights Dataset using meters as units, and y' will be the weights in Kg-s, then $cov(x, y)$ and $cov(x', y')$ will not be the same, but $cor(x, y) = cor(x', y')$.

- If $cov(x, y) > cov(z, t)$, we cannot state that the relationship between x and y is stronger than the relationship between z and t. But if $cor(x, y) > cor(z, t)$, we can.

So it is not easy to interpret the magnitude of the covariance, but the magnitude of the correlation coefficient is the strength of the linear relationship.

Pros/Cons of Sample Covariance and Correlation Coefficient

- An important drawback of the Sample Correlation Coefficient is that it is sensitive to outliers.

Covariance and Correlation Coefficient, again

So what are showing Covariance and Correlation Coefficient:

Covariance and Correlation Coefficient, again

So what are showing Covariance and Correlation Coefficient:

- The sign of Covariance and Corelation Coefficient shows the direction of the relationship: if

$$cov(x, y) > 0, \quad \text{equivalently, if} \quad cor(x, y) > 0,$$

  then if x is increasing, then y also tends to be larger.

Covariance and Correlation Coefficient, again

So what are showing Covariance and Correlation Coefficient:

- The sign of Covariance and Corelation Coefficient shows the direction of the relationship: if

  $$cov(x, y) > 0, \quad \text{equivalently, if} \quad cor(x, y) > 0,$$

  then if x is increasing, then y also tends to be larger. And if

  $$cov(x, y) < 0, \quad \text{equivalently, if} \quad cor(x, y) < 0,$$

  then if x is increasing, then y tends to be smaller.

Covariance and Correlation Coefficient, again

So what are showing Covariance and Correlation Coefficient:

- The sign of Covariance and Corelation Coefficient shows the direction of the relationship: if

$$cov(x,y) > 0, \quad \text{equivalently, if} \quad cor(x,y) > 0,$$

  then if x is increasing, then y also tends to be larger. And if

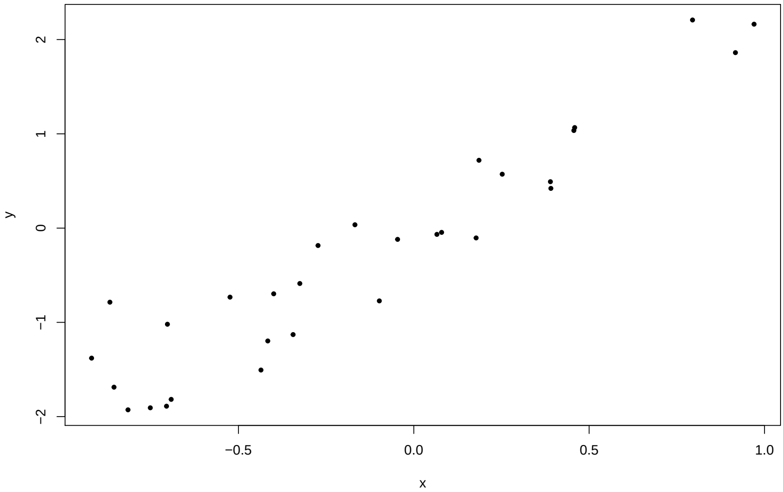
$$cov(x,y) < 0, \quad \text{equivalently, if} \quad cor(x,y) < 0,$$

  then if x is increasing, then y tends to be smaller.

- The magnitude of the Correlation Coefficient shows the strength of the Linear Relationship.

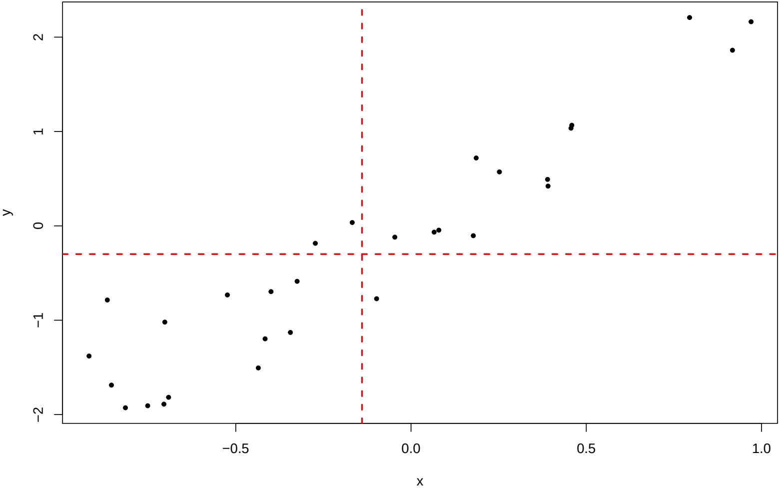
Explanation

Here is a Bivariate Dataset (x, y) with $cov(x, y) > 0$:



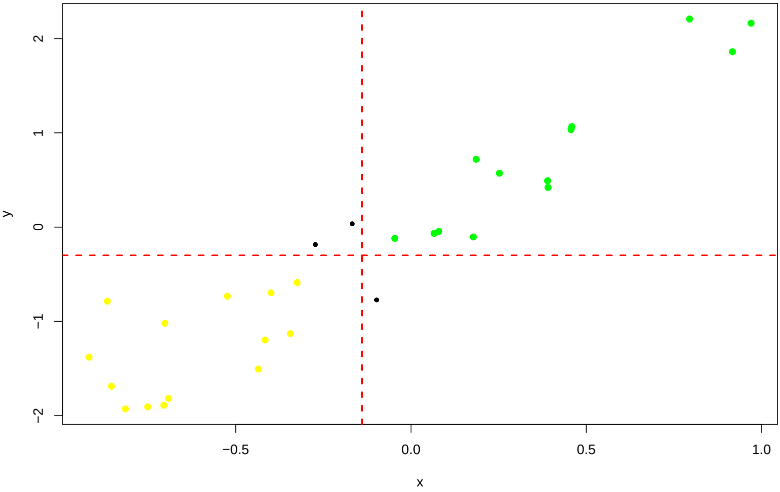
Explanation

Now we add a vertical line through $\bar{x}$ and a horizontal line through $\bar{y}$



Explanation

We color the points in the first and third quadrants:



Explanation

The points in the 1st quadrant (of the dotted coordinate system, with the center at $(\bar{x}, \bar{y})$), green points, satisfy

$$x_k > \bar{x} \quad and \quad y_k > \bar{y},$$

so

$$(x_k - \bar{x}) \cdot (y_k - \bar{y}) > 0,$$

so green points contribute positive terms to

$$cov(x, y) = \frac{1}{n} \cdot \sum_{k=1}^{n} (x_k - \bar{x}) \cdot (y_k - \bar{y}).$$

Explanation

The points in the 1st quadrant (of the dotted coordinate system, with the center at $(\bar{x}, \bar{y})$), green points, satisfy

$$x_k > \bar{x} \quad and \quad y_k > \bar{y},$$

so

$$(x_k - \bar{x}) \cdot (y_k - \bar{y}) > 0,$$

so green points contribute positive terms to

$$cov(x, y) = \frac{1}{n} \cdot \sum_{k=1}^{n} (x_k - \bar{x}) \cdot (y_k - \bar{y}).$$

Similarly, Points in the 3rd quadrant, yellow points, again contribute positive terms to $cov(x, y)$, since in this case

$$x_k < \bar{x} \quad and \quad y_k < \bar{y}, \quad hence, \quad (x_k - \bar{x}) \cdot (y_k - \bar{y}) > 0.$$

Explanation

The points in the 1st quadrant (of the dotted coordinate system, with the center at $(\bar{x}, \bar{y})$), green points, satisfy

$$x_k > \bar{x} \quad and \quad y_k > \bar{y},$$

so

$$(x_k - \bar{x}) \cdot (y_k - \bar{y}) > 0,$$

so green points contribute positive terms to

$$cov(x, y) = \frac{1}{n} \cdot \sum_{k=1}^{n} (x_k - \bar{x}) \cdot (y_k - \bar{y}).$$

Similarly, Points in the 3rd quadrant, yellow points, again contribute positive terms to $cov(x, y)$, since in this case

$$x_k < \bar{x} \quad and \quad y_k < \bar{y}, \quad hence, \quad (x_k - \bar{x}) \cdot (y_k - \bar{y}) > 0.$$

In the same way, the points in the 2nd and 4th quadrants give negative terms to $cov(x, y)$, as in this case $(x_k - \bar{x}) \cdot (y_k - \bar{y}) < 0$.

Explanation

The points in the 1st quadrant (of the dotted coordinate system, with the center at $(\bar{x}, \bar{y})$), green points, satisfy

$$x_k > \bar{x} \quad and \quad y_k > \bar{y},$$

so

$$(x_k - \bar{x}) \cdot (y_k - \bar{y}) > 0,$$

so green points contribute positive terms to

$$cov(x, y) = \frac{1}{n} \cdot \sum_{k=1}^{n} (x_k - \bar{x}) \cdot (y_k - \bar{y}).$$

Similarly, Points in the 3rd quadrant, yellow points, again contribute positive terms to $cov(x, y)$, since in this case

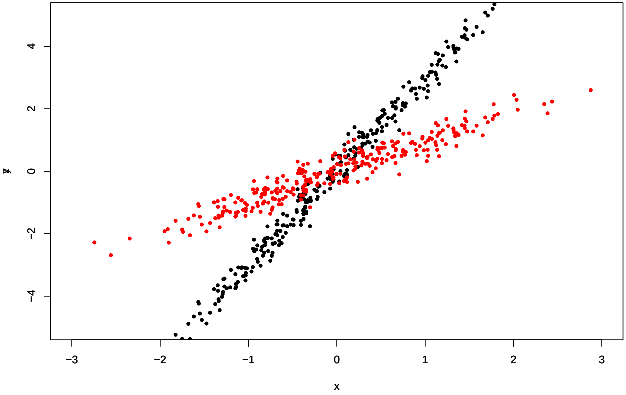
$$x_k < \bar{x} \quad and \quad y_k < \bar{y}, \quad hence, \quad (x_k - \bar{x}) \cdot (y_k - \bar{y}) > 0.$$

In the same way, the points in the 2nd and 4th quadrants give negative terms to $cov(x, y)$, as in this case $(x_k - \bar{x}) \cdot (y_k - \bar{y}) < 0$. And positive covariance means that the terms for points in the 1st and 3rd quadrants dominate to the ones from 2nd and fourth ones.

Note: Of course, we can have a negative trend and just one strong outlier in the 1st quadrant resulting in a positive covariance.

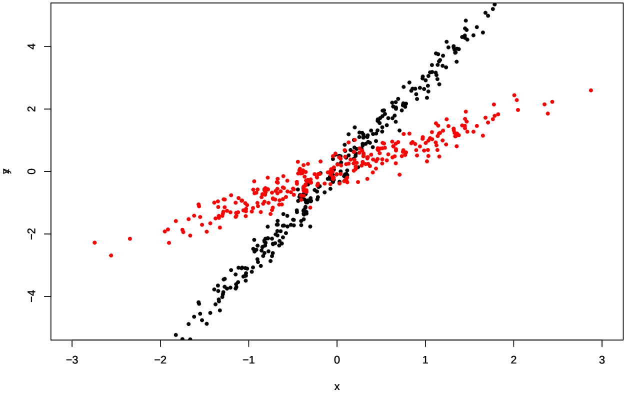
Example

For which of the following pairs the Correlation is higher?



Example

For which of the following pairs the Correlation is higher?

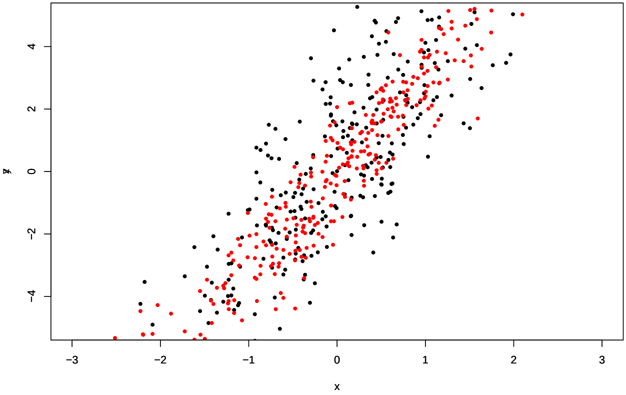



```r
c(cor(x,y), cor(x,z))
```

```
## [1] 0.9949983 0.9558994
```

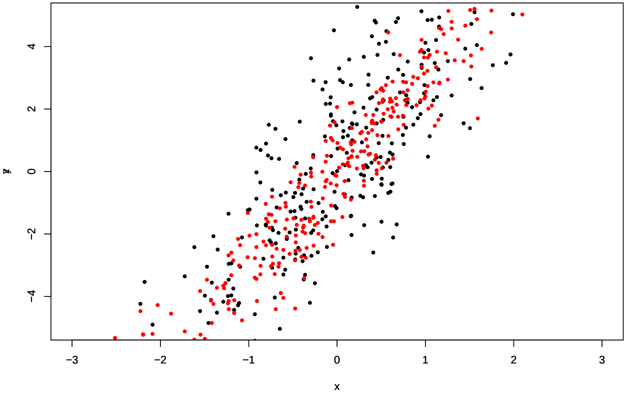
Example

For which of the following pairs the Correlation is higher?



Example

For which of the following pairs the Correlation is higher?




```r
c(cor(x,y), cor(x,z))
```

```
## [1] 0.8594477 0.9577039
```

Moral

Moral: Correlation is not about the slope of the Linear Relationship!

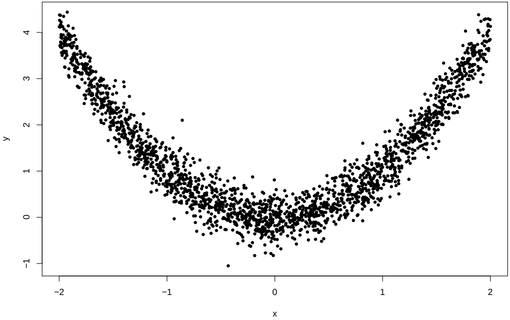
Moral

Moral: Correlation is not about the slope of the Linear Relationship!

Note: We will talk about this during the Linear Regression lectures.

Correlation is a Measure of Linear Relationship

```
x <- runif(2000, -2,2)
y <- x^2 + 0.3*rnorm(2000)
plot(x,y, pch = 20)
```

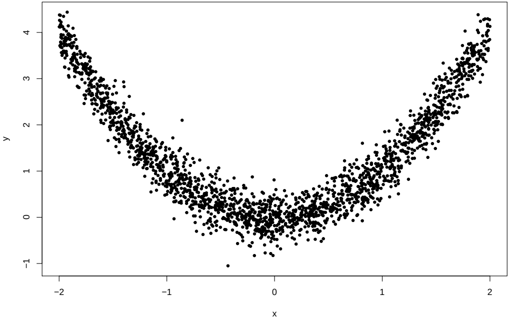



```
cor(x,y)
```

```
## [1] -0.03034234
```

Correlation is a Measure of Linear Relationship

```
x <- runif(2000, -2,2)
y <- x^2 + 0.3*rnorm(2000)
plot(x,y, pch = 20)
```




```
cor(x,y)
```

```
## [1] -0.03034234
```

See more at Wiki

Supplements, Other Measures of Correlation

- if working with several variables, we can calculate parwise Correlations (Correlation Matrix) and plot the HeatMap

Supplements, Other Measures of Correlation

- if working with several variables, we can calculate parwise Correlations (Correlation Matrix) and plot the HeatMap
- If working with multiple variables, one can calculate the Multiple correlation

Supplements, Other Measures of Correlation

- if working with several variables, we can calculate parwise Correlations (Correlation Matrix) and plot the HeatMap
- If working with multiple variables, one can calculate the Multiple correlation
- One can interpret the Correlation Coefficient as a Cosine of the angle between the r.v.s (or observations), see Wiki

Supplements, Other Measures of Correlation

- if working with several variables, we can calculate parwise Correlations (Correlation Matrix) and plot the HeatMap
- If working with multiple variables, one can calculate the Multiple correlation
- One can interpret the Correlation Coefficient as a Cosine of the angle between the r.v.s (or observations), see Wiki
- There are other measures of Association between variables, such as Rank Correlations, say, Kendal's τ