AUA CS 108, Statistics, Fall 2019

Lecture 13

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

23 Sep 2019

Contents

About my OH: Is it OK to split the OH into 1h on Monday, 11:30 - 12:30 and 1h on Wednesday, 11:30 - 12:30 ?

Last Lecture ReCap

- Give the definition of the Sample Covariance

Last Lecture ReCap

- Give the definition of the Sample Covariance
- Give the definition of the Sample Correlation Coefficient

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

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 show 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 show 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 show the direction of the relationship: if

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

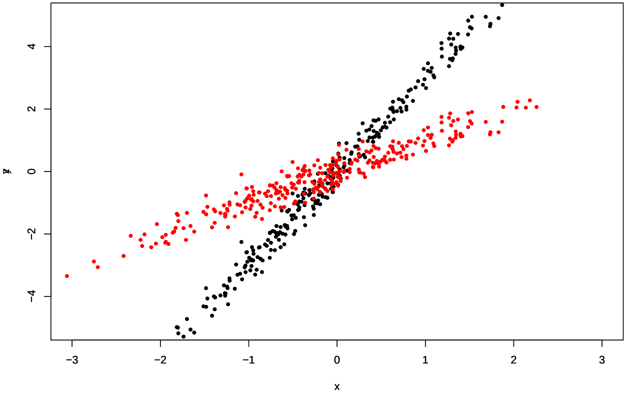
  then if x is increasing, then y also tends to be larger. And if

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

  then if x is increasing, then y tends to be smaller.
- The magnitude of the Correlation Coefficient shows the strength of the Linear Relationship.

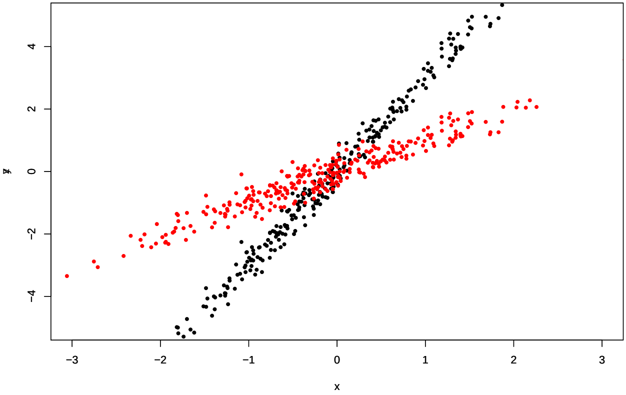
Example

For which of the following pairs the Correlation is higher?



Example

For which of the following pairs the Correlation is higher?

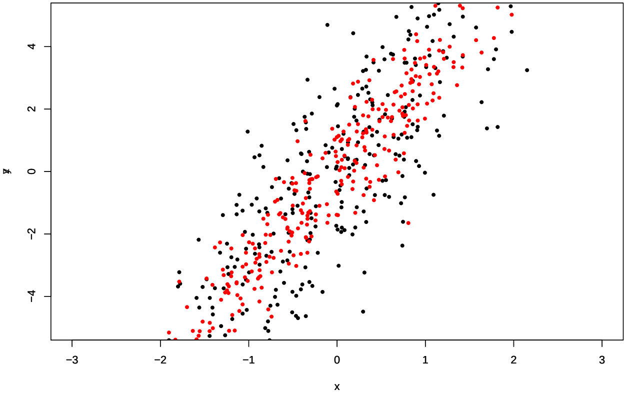



```r
c(cor(x,y), cor(x,z))
```

```
## [1] 0.9954556 0.9610155
```

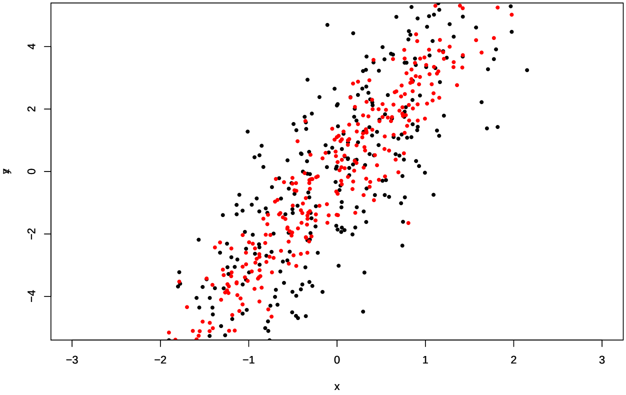
Example

For which of the following pairs the Correlation is higher?



Example

For which of the following pairs the Correlation is higher?

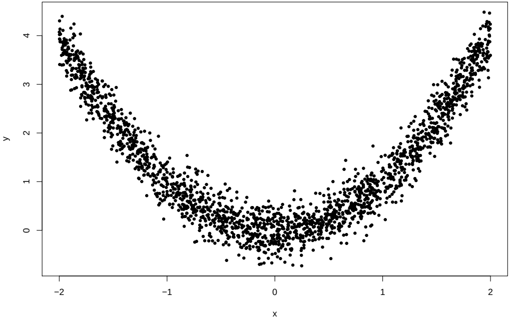



```r
c(cor(x,y), cor(x,z))
```

```
## [1] 0.8243984 0.9493276
```

Correlation is a Measure of Linear Relationship

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

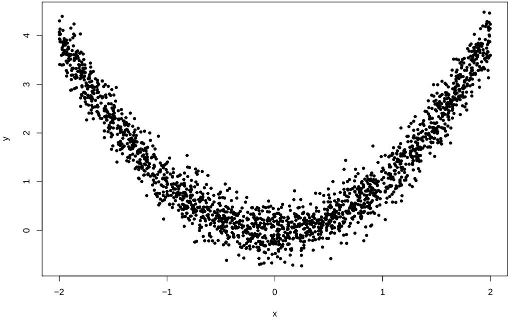



```
cor(x,y)
```

```
## [1] 0.006338811
```

Correlation is a Measure of Linear Relationship

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




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

```
## [1] 0.006338811
```

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

Correlation is not Causation

- Some Examples: Spurious Correlations

Anscombe Quartet

Wiki

Reminder on Random Variables and Distributions

Random Variables

Everything starts at the Probability Space (Experiment, Model): we are given

$$(\Omega, \mathcal{F}, \mathbb{P}) \qquad \text{or, we usually use} \qquad (\Omega, \mathbb{P}),$$

where

Random Variables

Everything starts at the Probability Space (Experiment, Model): we are given

$$(\Omega, \mathcal{F}, \mathbb{P}) \qquad \text{or, we usually use} \qquad (\Omega, \mathbb{P}),$$

where

- Ω is the Sample Space

Random Variables

Everything starts at the Probability Space (Experiment, Model): we are given

$$(\Omega, \mathcal{F}, \mathbb{P}) \qquad \text{or, we usually use} \qquad (\Omega, \mathbb{P}),$$

where

- Ω is the Sample Space
- $\mathcal{F}$ is the set of all Events

Random Variables

Everything starts at the Probability Space (Experiment, Model): we are given

$$(\Omega, \mathcal{F}, \mathbb{P}) \qquad \text{or, we usually use} \qquad (\Omega, \mathbb{P}),$$

where

- Ω is the Sample Space
- $\mathcal{F}$ is the set of all Events
- $\mathbb{P}$ is a Probability Measure

Random Variables

Everything starts at the Probability Space (Experiment, Model): we are given

$$(\Omega, \mathcal{F}, \mathbb{P}) \qquad \text{or, we usually use} \qquad (\Omega, \mathbb{P}),$$

where

- Ω is the Sample Space
- $\mathcal{F}$ is the set of all Events
- $\mathbb{P}$ is a Probability Measure

Definition: Any (measurable) function $X : \Omega \to \mathbb{R}$ is called a r.v. on the Probability Space $(\Omega, \mathbb{P})$.

Random Variables

Everything starts at the Probability Space (Experiment, Model): we are given

$$(\Omega, \mathcal{F}, \mathbb{P}) \qquad \text{or, we usually use} \qquad (\Omega, \mathbb{P}),$$

where

- Ω is the Sample Space
- $\mathcal{F}$ is the set of all Events
- $\mathbb{P}$ is a Probability Measure

Definition: Any (measurable) function $X : \Omega \to \mathbb{R}$ is called a r.v. on the Probability Space $(\Omega, \mathbb{P})$.

So $X = X(\omega)$, but usually we forget about ω, and use X.

Main Complete Characteristics of a r.v.

If X is a r.v., then we get the **complete information** (everything we can get) about X from either its CDF or PDF/PMF.

Main Complete Characteristics of a r.v.

If X is a r.v., then we get the **complete information** (everything we can get) about X from either its CDF or PDF/PMF.

Definition: The CDF of X is defined as

$$F(x) = F_X(x) = \mathbb{P}(X \leq x), \qquad x \in \mathbb{R}.$$

Main Complete Characteristics of a r.v.

If X is a r.v., then we get the **complete information** (everything we can get) about X from either its CDF or PDF/PMF.

Definition: The CDF of X is defined as

$$F(x) = F_X(x) = \mathbb{P}(X \leq x), \qquad x \in \mathbb{R}.$$

Definition: We say that X is a *Continuous r.v.*, if it has a PDF: a function $f(x)$ such that

$$F(x) = \int_{-\infty}^{x} f(t)dt, \qquad \forall x \in \mathbb{R}.$$

Main Complete Characteristics of a r.v.

If X is a r.v., then we get the **complete information** (everything we can get) about X from either its CDF or PDF/PMF.

Definition: The CDF of X is defined as

$$F(x) = F_X(x) = \mathbb{P}(X \leq x), \qquad x \in \mathbb{R}.$$

Definition: We say that X is a *Continuous r.v.*, if it has a PDF: a function $f(x)$ such that

$$F(x) = \int_{-\infty}^{x} f(t)dt, \qquad \forall x \in \mathbb{R}.$$

So for a Continuous r.v., another complete characteristic, besides the CDF, is its PDF.

Discrete r.v.s

Definition: We say that X is a *Discrete r.v.*, if the set of values of X is finite or countably infinite.

Discrete r.v.s

Definition: We say that X is a *Discrete r.v.*, if the set of values of X is finite or countably infinite. And if the possible values are x_k, $k = 1, 2, ...$, then we define the PMF of X as

$$f(x_k) = \mathbb{P}(X = x_k), \qquad k = 1, 2, ...,$$

Discrete r.v.s

Definition: We say that X is a *Discrete r.v.*, if the set of values of X is finite or countably infinite. And if the possible values are x_k, $k = 1, 2, ...$, then we define the PMF of X as

$$f(x_k) = \mathbb{P}(X = x_k), \qquad k = 1, 2, ...,$$

or, in a table form,

Values of X	x_1	x_2	...
$\mathbb{P}(X = x)$	p_1	p_2	...

Main Partial Characteristics of a r.v.

Main partial characteristics of a r.v. X are:

Main Partial Characteristics of a r.v.

Main partial characteristics of a r.v. X are:

- the Expected Value (Mean):

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} x \cdot f(x) dx \, (cont.) \quad | \quad \mathbb{E}(X) = \sum_k x_k \cdot \mathbb{P}(X = x_k) \, (disc.).$$

Main Partial Characteristics of a r.v.

Main partial characteristics of a r.v. X are:

- the Expected Value (Mean):

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} x \cdot f(x) dx \, (cont.) \quad | \quad \mathbb{E}(X) = \sum_k x_k \cdot \mathbb{P}(X = x_k) \, (disc.).$$

Note:

$$\mathbb{E}(g(X)) = \int_{-\infty}^{+\infty} g(x) \cdot f(x) dx \, (cont.) \quad | \quad \mathbb{E}(g(X)) = \sum_k g(x_k) \cdot \mathbb{P}(X = x_k) \, ($$

Main Partial Characteristics of a r.v.

Main partial characteristics of a r.v. X are:

- the Expected Value (Mean):

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} x \cdot f(x) dx \, (cont.) \quad | \quad \mathbb{E}(X) = \sum_k x_k \cdot \mathbb{P}(X = x_k) \, (disc.).$$

Note:

$$\mathbb{E}(g(X)) = \int_{-\infty}^{+\infty} g(x) \cdot f(x) dx \, (cont.) \quad | \quad \mathbb{E}(g(X)) = \sum_k g(x_k) \cdot \mathbb{P}(X = x_k) \, ($$

- The Variance

$$Var(X) = \mathbb{E}\Big((X - \mathbb{E}(X))^2\Big) = \mathbb{E}(X^2) - \Big[\mathbb{E}(X)\Big]^2.$$