AUA CS108, Statistics, Fall 2020

Lecture 17

Michael Poghosyan

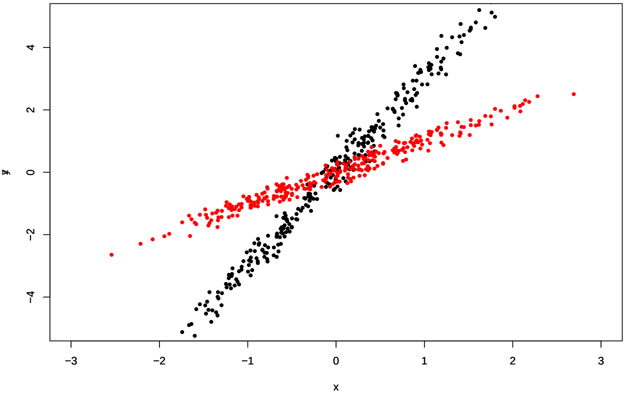
05 Oct 2020

Contents

- Sample Correlation Coefficient
- Quick reminder on R.V.s
- Important Discrete Distributions

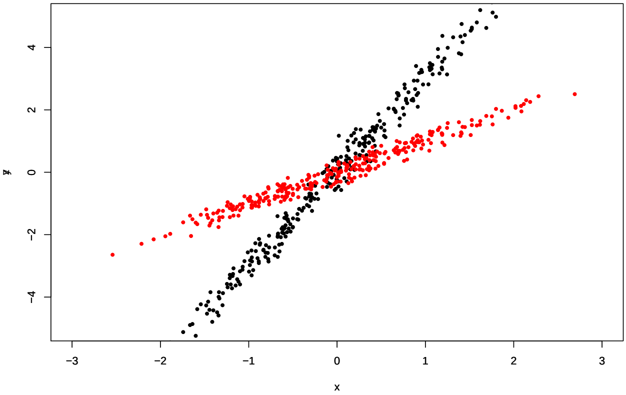
Example

For which of the following pairs the Correlation is higher ((x, y) pairs are in black, and (x, z) pairs are in red)?



Example

For which of the following pairs the Correlation is higher ((x, y) pairs are in black, and (x, z) pairs are in red)?

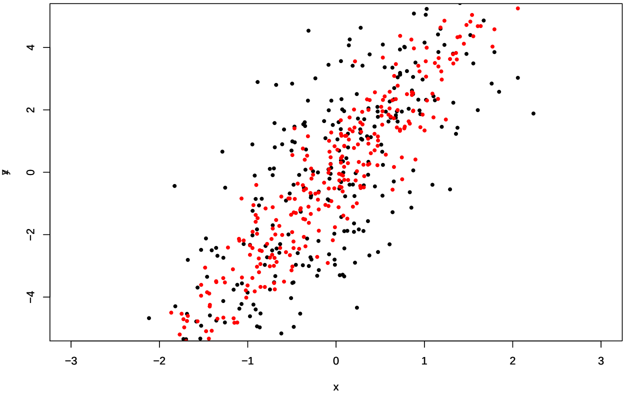


```
c(cor(x,y), cor(x,z))
```

```
## [1] 0.9941472 0.9831613
```

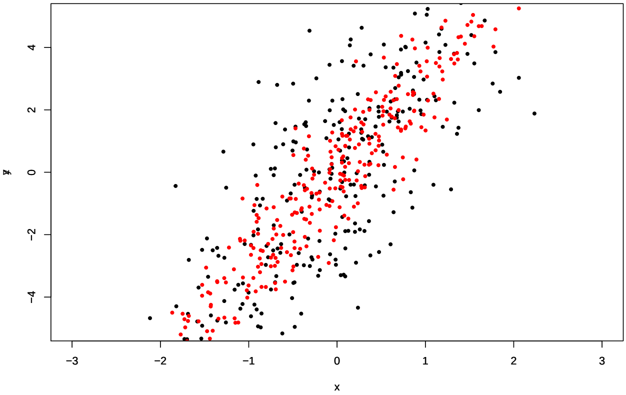
Example

For which of the following pairs the Correlation is higher ((x, y) pairs are in black, and (x, z) pairs are in red)?



Example

For which of the following pairs the Correlation is higher ((x, y) pairs are in black, and (x, z) pairs are in red)?



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

```
## [1] 0.7955134 0.9440403
```

Moral

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

Moral

Moral: Correlation coefficient is not about the slope of the Linear Relationship! It is about how close to the linear is the relationship between two Datasets.

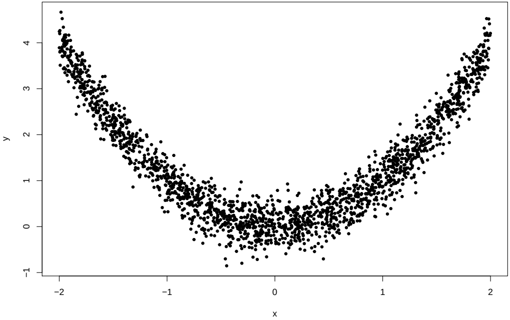
Moral

Moral: Correlation coefficient is not about the slope of the Linear Relationship! It is about how close to the linear is the relationship between two Datasets.

Note: We will talk about this and about the relationship of slope with the Correlation Coefficient 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)
```

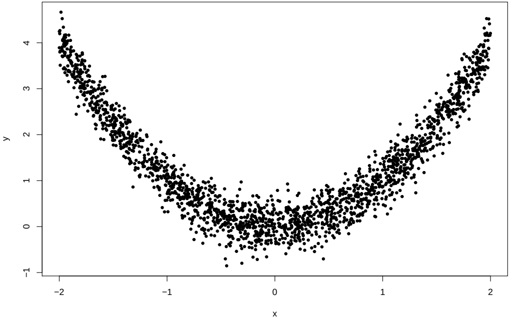


```
cor(x,y)
```

```
## [1] -0.01353751
```

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] -0.01353751
```

See more at Wiki

Another Relationship between the Correlation and Covariance

Assume we have two datasets x and y of the same size. We standardize them, i.e., we consider

$$\frac{x-\bar{x}}{s_x}, \qquad \frac{y-\bar{y}}{s_y},$$

then the Correlation Coefficient is just the Covariance between these standardized daatasets:

$$cor(x,y) = cov\left(\frac{x-\bar{x}}{s_x}, \frac{y-\bar{y}}{s_y}\right).$$

Correlation is not Causation

- Some Examples: Spurious Correlations

Supplements, Other Measures of Correlation

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

Supplements, Other Measures of Correlation

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

For example, if you will run **R**-s `cor` function over a DataFrame, it will calculate the Correlation Matrix of the DataFrame Variables.

Supplements, Other Measures of Correlation

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

For example, if you will run **R**-s `cor` function over a DataFrame, it will calculate the Correlation Matrix of the DataFrame Variables.

- If working with multiple variables, one can calculate the Multiple correlation

Supplements, Other Measures of Correlation

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

For example, if you will run **R**-s `cor` function over a DataFrame, it will calculate the Correlation Matrix of the DataFrame Variables.

- 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 pairwise Correlations (Correlation Matrix) and plot the HeatMap

For example, if you will run **R**-s `cor` function over a DataFrame, it will calculate the Correlation Matrix of the DataFrame Variables.

- 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

In **R**, the `cor` function has a parameter *method*, where you can change the Correlation Coefficient type.

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

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

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

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

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):

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

Main Partial Characteristics of a r.v.

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

- the Expected Value (Mean):

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

Note:

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

Main Partial Characteristics of a r.v.

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

- the Expected Value (Mean):

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

Note:

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

- The Variance

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

Important Discrete Distributions

Bernoulli Distribution

- Parameter: $p \in [0, 1]$ (usually, $p \in (0, 1)$)

Bernoulli Distribution

- Parameter: $p \in [0, 1]$ (usually, $p \in (0, 1)$)
- Notation: $X \sim \mathit{Bernoulli}(p)$;

Bernoulli Distribution

- Parameter: $p \in [0, 1]$ (usually, $p \in (0, 1)$)
- Notation: $X \sim \mathit{Bernoulli}(p)$;
- Support: $\{0, 1\}$

Bernoulli Distribution

- Parameter: $p \in [0, 1]$ (usually, $p \in (0, 1)$)
- Notation: $X \sim \mathit{Bernoulli}(p)$;
- Support: $\{0, 1\}$
- PMF:

Values of X	0	1
$\mathbb{P}(X = x)$	$1 - p$	p

Bernoulli Distribution

- Parameter: $p \in [0, 1]$ (usually, $p \in (0, 1)$)
- Notation: $X \sim \mathit{Bernoulli}(p)$;
- Support: $\{0, 1\}$
- PMF:

Values of X	0	1
$\mathbb{P}(X = x)$	$1 - p$	p

Note: This can be written in the form:

$$f(x) = f(x; p) = f(x|p) = p^x \cdot (1 - p)^{1-x}, \qquad x \in \{0, 1\}.$$

Bernoulli Distribution

- Parameter: $p \in [0, 1]$ (usually, $p \in (0, 1)$)
- Notation: $X \sim Bernoulli(p)$;
- Support: $\{0, 1\}$
- PMF:

Values of X	0	1
$\mathbb{P}(X = x)$	$1 - p$	p

Note: This can be written in the form:

$$f(x) = f(x; p) = f(x|p) = p^x \cdot (1 - p)^{1-x}, \qquad x \in \{0, 1\}.$$

- Mean and Variance: $\mathbb{E}(X) = p$, $Var(X) = p(1 - p)$.

Bernoulli Distribution

- Parameter: $p \in [0,1]$ (usually, $p \in (0,1)$)
- Notation: $X \sim \mathit{Bernoulli}(p)$;
- Support: $\{0,1\}$
- PMF:

Values of X	0	1
$\mathbb{P}(X = x)$	$1-p$	p

Note: This can be written in the form:

$$f(x) = f(x;p) = f(x|p) = p^x \cdot (1-p)^{1-x}, \qquad x \in \{0,1\}.$$

- Mean and Variance: $\mathbb{E}(X) = p$, $\mathit{Var}(X) = p(1-p)$.
- Models: Models binary output, “success-failure” type Experiments, a lot of examples.

Bernoulli Distribution

- Parameter: $p \in [0,1]$ (usually, $p \in (0,1)$)
- Notation: $X \sim Bernoulli(p)$;
- Support: $\{0,1\}$
- PMF:

Values of X	0	1
$\mathbb{P}(X = x)$	$1 - p$	p

Note: This can be written in the form:

$$f(x) = f(x;p) = f(x|p) = p^x \cdot (1-p)^{1-x}, \qquad x \in \{0,1\}.$$

- Mean and Variance: $\mathbb{E}(X) = p$, $Var(X) = p(1-p)$.
- Models: Models binary output, “success-failure” type Experiments, a lot of examples.
- **R** name: `binom` with the parameters `size=1` and `prob`

Bernoulli Distribution

- Parameter: $p \in [0,1]$ (usually, $p \in (0,1)$)
- Notation: $X \sim Bernoulli(p)$;
- Support: $\{0,1\}$
- PMF:

Values of X	0	1
$\mathbb{P}(X = x)$	$1-p$	p

Note: This can be written in the form:

$$f(x) = f(x;p) = f(x|p) = p^x \cdot (1-p)^{1-x}, \qquad x \in \{0,1\}.$$

- Mean and Variance: $\mathbb{E}(X) = p$, $Var(X) = p(1-p)$.
- Models: Models binary output, "success-failure" type Experiments, a lot of examples.
- **R** name: `binom` with the parameters `size=1` and `prob`