ASDS Statistics, YSU, Fall 2020

Lecture 13

Michael Poghosyan

14 Oct 2020

Contents

Properties of the Sample Correlation Coefficient

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

---

[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

- $cor(x, y) = cor(y, x)$;
- If $\alpha > 0$ and $\beta \in \mathbb{R}$, then $cor(\alpha \cdot x + \beta, y) = cor(x, y)$

---

[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

- $cor(x, y) = cor(y, x)$;
- If $\alpha > 0$ and $\beta \in \mathbb{R}$, then $cor(\alpha \cdot x + \beta, y) = cor(x, y)$
- If $\alpha < 0$ and $\beta \in \mathbb{R}$, then $cor(\alpha \cdot x + \beta, y) = -cor(x, y)$

[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

- $cor(x, y) = cor(y, x)$;
- If $\alpha > 0$ and $\beta \in \mathbb{R}$, then $cor(\alpha \cdot x + \beta, y) = cor(x, y)$
- If $\alpha < 0$ and $\beta \in \mathbb{R}$, then $cor(\alpha \cdot x + \beta, y) = -cor(x, y)$
- 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

- $cor(x, y) = cor(y, x)$;
- If $\alpha > 0$ and $\beta \in \mathbb{R}$, then $cor(\alpha \cdot x + \beta, y) = cor(x, y)$
- If $\alpha < 0$ and $\beta \in \mathbb{R}$, then $cor(\alpha \cdot x + \beta, y) = -cor(x, y)$
- For any Datasets x, y,

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

- $\rho_{xy} = 1$ 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

- $cor(x, y) = cor(y, x)$;
- If $\alpha > 0$ and $\beta \in \mathbb{R}$, then $cor(\alpha \cdot x + \beta, y) = cor(x, y)$
- If $\alpha < 0$ and $\beta \in \mathbb{R}$, then $cor(\alpha \cdot x + \beta, y) = -cor(x, y)$
- 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

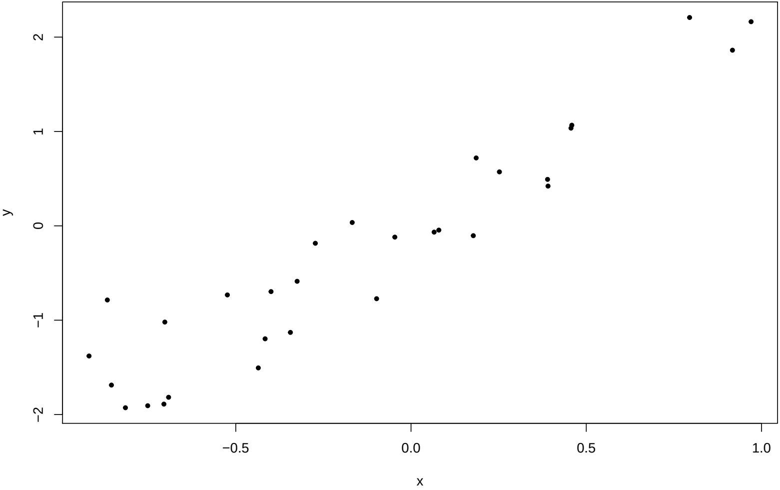
$$cov(x, y) < 0, \quad \text{equivalently, if} \quad cor(x, y) < 0,$$

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

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

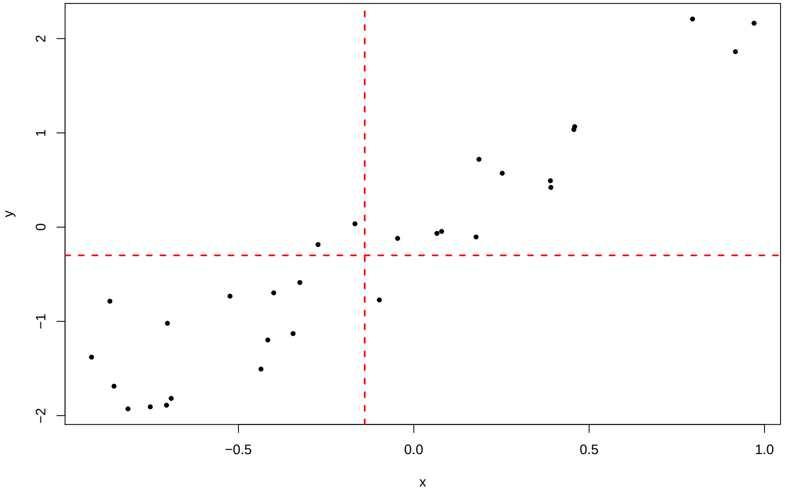
Explanation

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



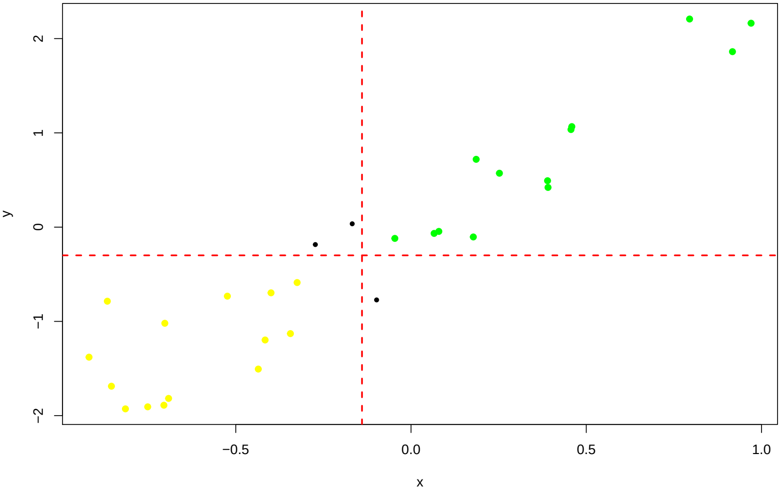
Explanation

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

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

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

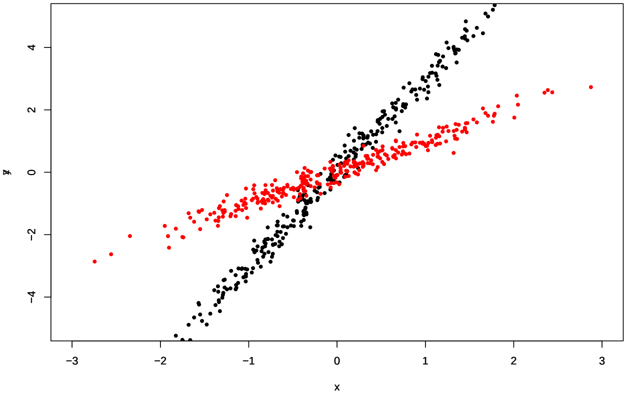
$$x_k < \bar{x} \quad and \quad y_k < \bar{y}, \quad hence, \quad (x_k - \bar{x}) \cdot (y_k - \bar{y}) > 0.$$

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.

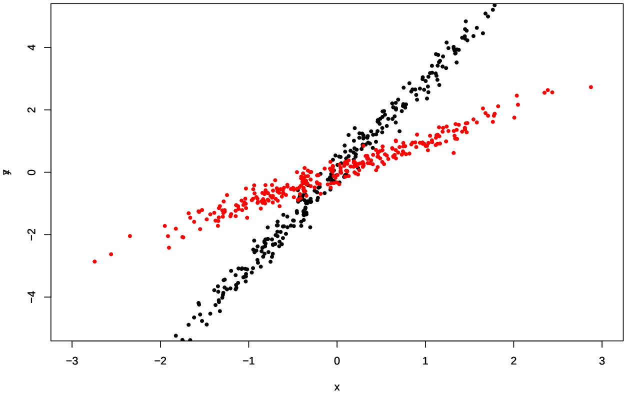
Example

For which of the following pairs the Correlation is higher ((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)?

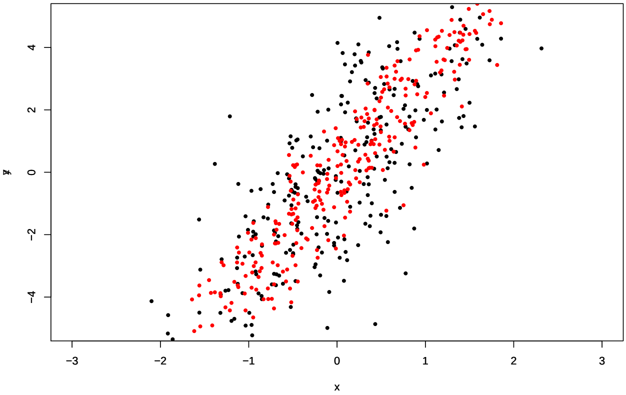



```r
c(cor(x,y), cor(x,z))
```

```
## [1] 0.9949983 0.9775781
```

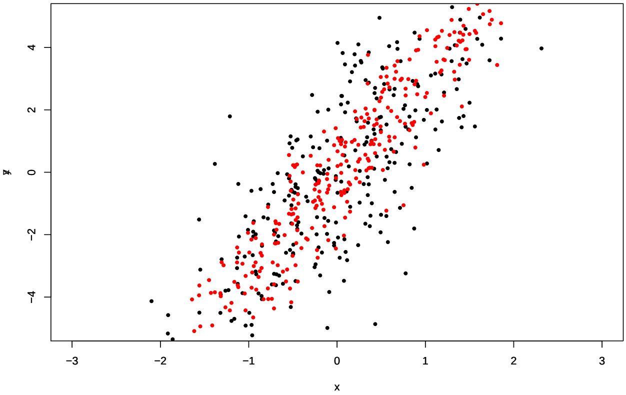
Example

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] 0.8206651 0.9428407
```

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.

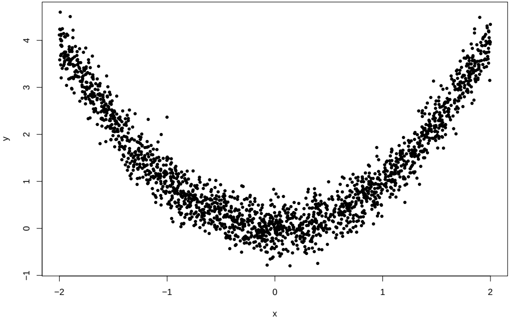
Moral

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)
```

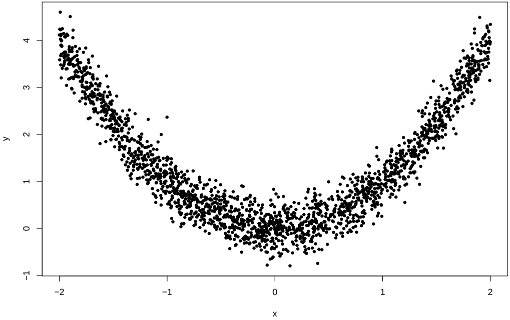



```
cor(x,y)
```

```
## [1] -0.01703987
```

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] -0.01703987
```

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:

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

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.

Correlation is not Causation

- Some Examples: Spurious Correlations

Anscombe Quartet

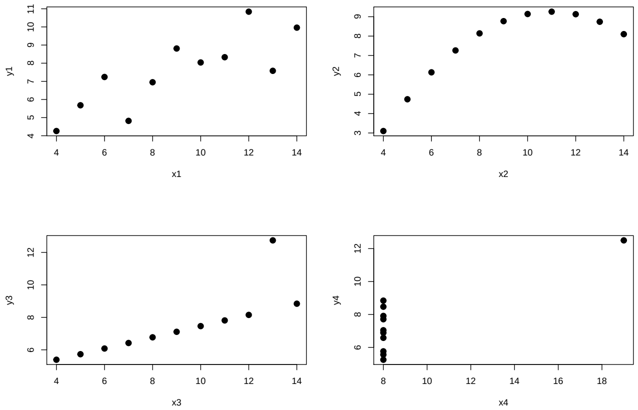
See Wiki

```
anscombe
```

```
##    x1 x2 x3 x4    y1   y2    y3    y4
## 1  10 10 10  8  8.04 9.14  7.46  6.58
## 2   8  8  8  8  6.95 8.14  6.77  5.76
## 3  13 13 13  8  7.58 8.74 12.74  7.71
## 4   9  9  9  8  8.81 8.77  7.11  8.84
## 5  11 11 11  8  8.33 9.26  7.81  8.47
## 6  14 14 14  8  9.96 8.10  8.84  7.04
## 7   6  6  6  8  7.24 6.13  6.08  5.25
## 8   4  4  4 19  4.26 3.10  5.39 12.50
## 9  12 12 12  8 10.84 9.13  8.15  5.56
## 10  7  7  7  8  4.82 7.26  6.42  7.91
## 11  5  5  5  8  5.68 4.74  5.73  6.89
```

Anscombe Quartet

```
rm(x1, x2, x3, x4, y1, y2, y3, y4); attach(anscombe)
par(mfrow=c(2,2))
plot(y1~x1, pch=19, cex=1.4); plot(y2~x2, pch=19, cex=1.4);
plot(y3~x3, pch=19, cex=1.4); plot(y4~x4, pch=19, cex=1.4);
```



Anscombe Quartet

```
c(mean(x1), mean(x2), mean(x3), mean(x4))
```

```
## [1] 9 9 9 9
```

```
c(mean(y1), mean(y2), mean(y3), mean(y4))
```

```
## [1] 7.500909 7.500909 7.500000 7.500909
```

```
c(var(x1), var(x2), var(x3), var(x4))
```

```
## [1] 11 11 11 11
```

```
c(var(y1), var(y2), var(y3), var(y4))
```

```
## [1] 4.127269 4.127629 4.122620 4.123249
```

```
c(cor(x1,y1), cor(x2,y2), cor(x3,y3), cor(x4,y4))
```

```
## [1] 0.8164205 0.8162365 0.8162867 0.8165214
```

Anscombe Quartet

Moral: Just calculating numbers (summary statistics) is not enough, visualize your Data if possible.

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.