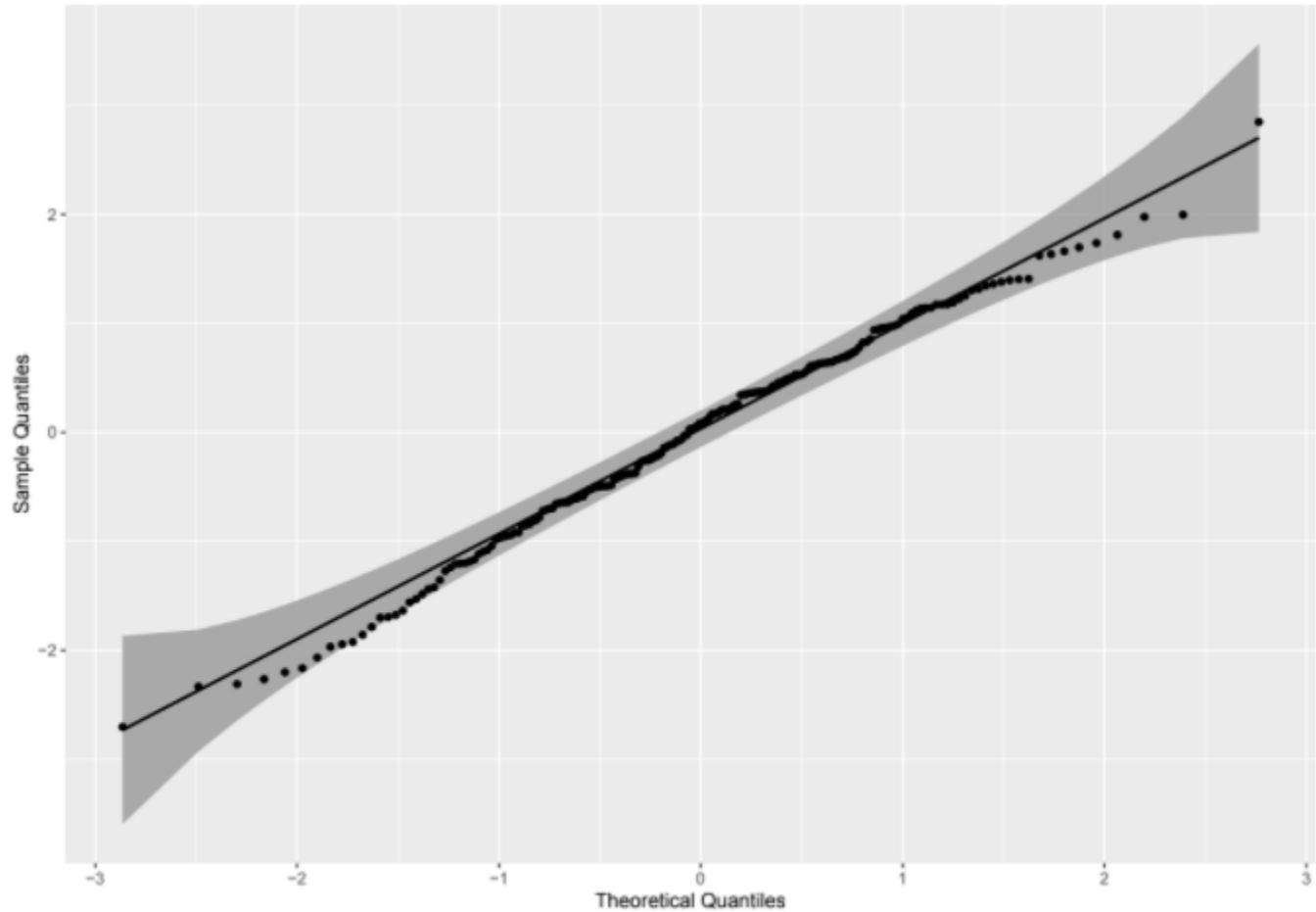
Contents

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")
```



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

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

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 $cov(x, y) = 0$.

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

Remark: For almost all numerical summaries for 1D data, first step was sorting the Dataset to obtain Order Statistics. But please note that for calculating Covariance or Correlation Coefficient (as well as for ScatterPlotting), sorting the Datasets will give incorrect results. This is because we want to find a relationship between x_1 and y_1, x_2 and y_2, ..., not the relationship between the minimal elements of Datasets etc.

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

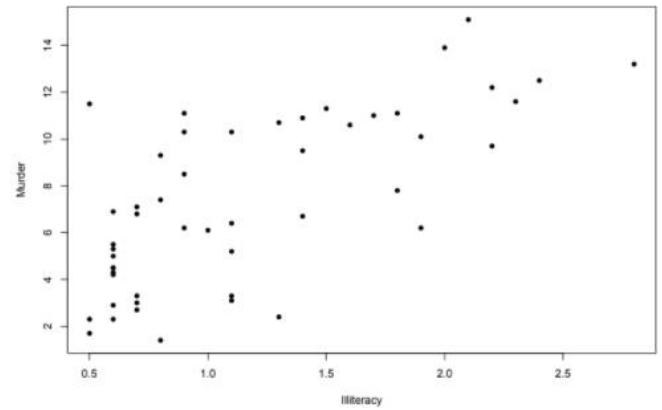
In both cases, when one calculates Standard Deviations and Covariance by using n simultaneously or $n-1$ simultaneously in the denominator, we will obtain

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

Another formula to calc the correlation coefficient is

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

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

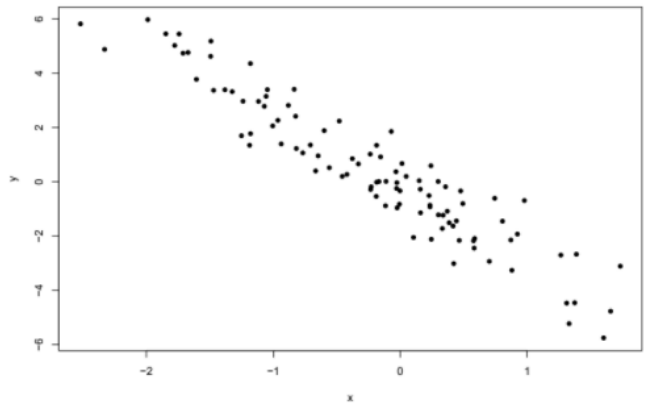



```
cor(state$Illiteracy, state$Murder)
```

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

Some simulations:

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




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

```
## [1] -0.9447191 -2.2926135
```

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

Answer: 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

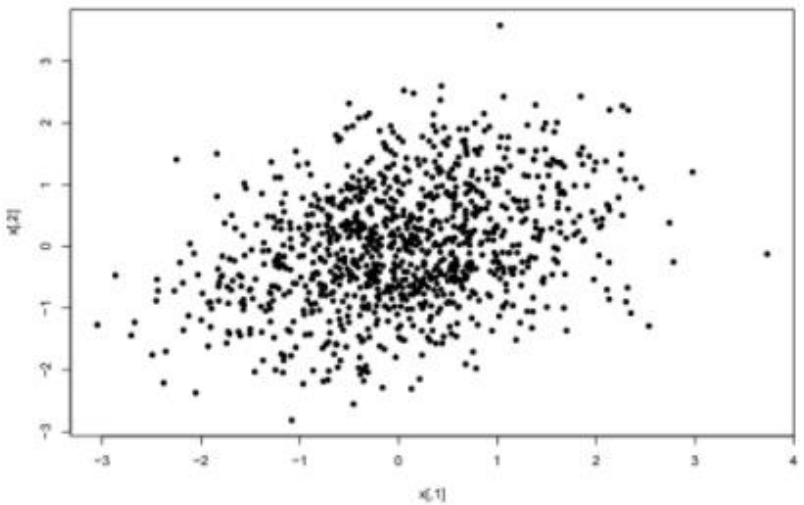
$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

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

```
rho <- 0.35
covmatrix <- matrix(c(1,rho, rho, 1), nrow = 2)
mu <- c(0,0)
x <- mvtnorm::rmvnorm(1000, mean = mu, sigma = covmatrix)
plot(x, pch = 16)
```




```
cor(x)
```

```
##           [,1]      [,2]
## [1,] 1.0000000 0.3262176
## [2,] 0.3262176 1.0000000
```

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,

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

$\mathrm{Cov}(2X+3Y, X-2Y)$,

$$\mathrm{Cov}(2X+3Y, X-2Y) = 2\,\mathrm{Var}(X) - \mathrm{Cov}(X,Y) - 6\,\mathrm{Var}(Y).$$

$\mathrm{Var}(2X+3Y)$,

$$\mathrm{Var}(2X+3Y) = \mathrm{Cov}(2X+3Y, 2X+3Y)$$
$$= 4\,\mathrm{Var}(X) + 12\,\mathrm{Cov}(X,Y) + 9\,\mathrm{Var}(Y).$$

Example: Assume $sd(x) = 1$ and $sd(y) = 2$ and $cov(x, y) = 1.5$. Calculate

$$cov(3x - 2y + 7, y - 5x).$$

Properties of the Sample Correlation Coefficient

- $cor(x, y) = cor(y, x)$;
- $cor(x, x) = 1$;
- 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).

Calculate

- $cor(2x - 5, x)$;
- $cor(2x - 5y, -4x + 9)$, if x and y are uncorrelated and $sd(x) = 2$, $sd(y) = 10$.

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:

- 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

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

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

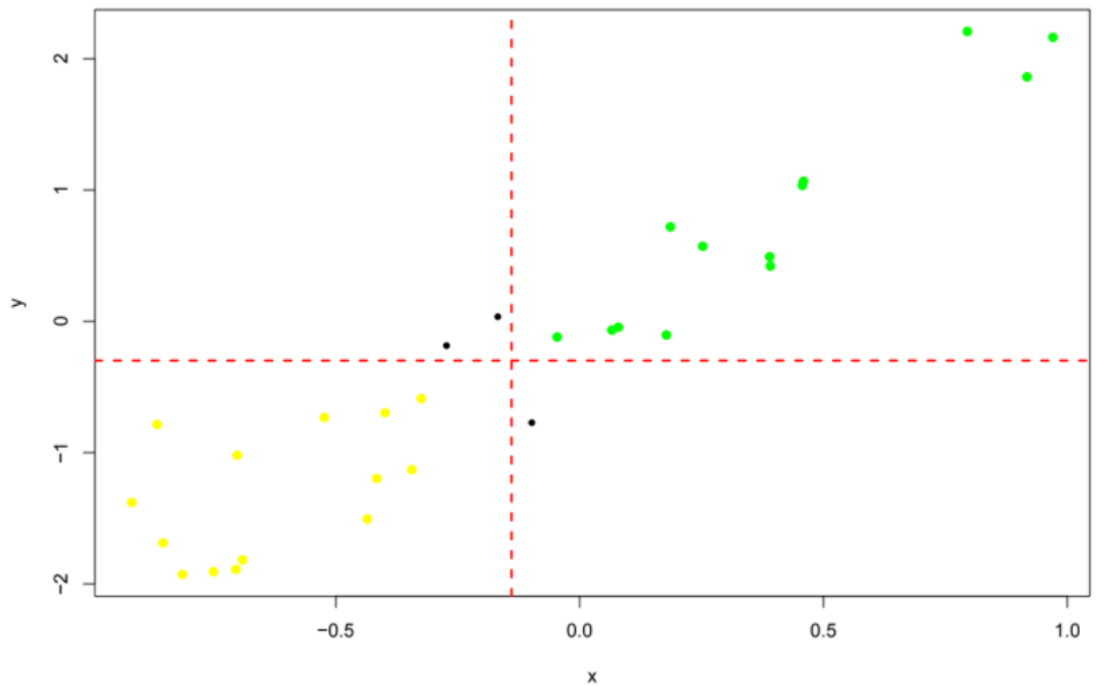
Similarly, Points in the 3rd quadrant, yellow points, again contribute positive terms to $cov(x, y)$, since in this case

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

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.

Explanation

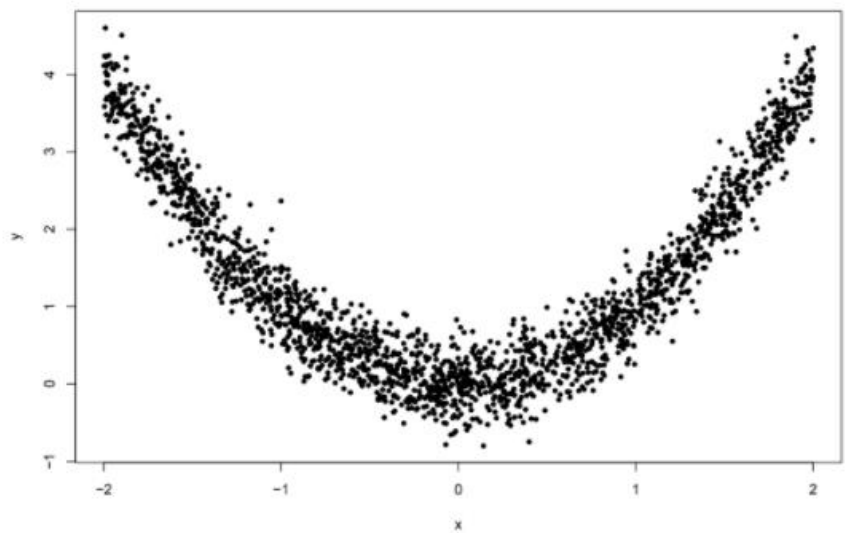
We color the points in the first and third quadrants:



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.

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




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

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

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

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.

- You can think about how to define the measure of Association when working with (N, C) or (C, C) type of Variable pairs.

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.

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. 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:

- 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

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