

## Expected value and standard deviation of common discrete distributions (Lecture 6)

- If  $X$  has binomial distribution with order  $n$  and parameter  $p$ :

$$\mathbb{E}(X) = np; \quad D(X) = \sqrt{np(1-p)}.$$

- If  $X$  has hypergeometric distribution with parameters  $N, M, n$ :

$$\mathbb{E}(X) = \frac{M}{N}n; \quad D(X) = \sqrt{n \frac{M}{N} \left(1 - \frac{M}{N}\right) \frac{N-n}{N-1}}.$$

- If  $X$  has Poisson distribution with parameter  $\lambda$ :

$$\mathbb{E}(X) = \lambda; \quad D(X) = \sqrt{\lambda};$$

- If  $X$  has geometric distribution with parameter  $p$ :

$$\mathbb{E}(X) = \frac{1}{p}; \quad D(X) = \sqrt{\frac{1-p}{p^2}}.$$

## Independence: example

Which random variables can be considered independent? Anne is a randomly chosen participant of a survey.

number of Anne's cars

quantity of rain tomorrow in Buda

Anne's monthly income

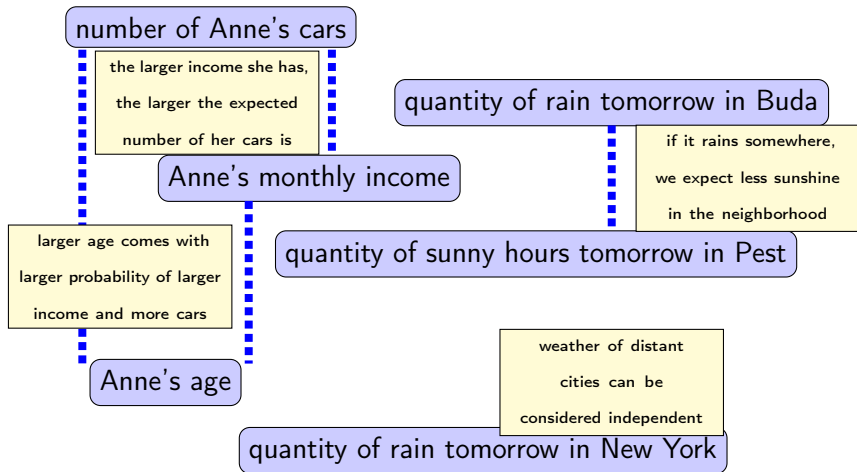
quantity of sunny hours tomorrow in Pest

Anne's age

quantity of rain tomorrow in New York

## Independence: example

Which random variables can be considered independent? Anne is a randomly chosen participant of a survey.



## Independence: example

Reminder: events  $A$  and  $B$  are **independent**, if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

If  $X$  is the quantity of rain in Buda tomorrow (in mm), and  $Y$  is the same in New York, then for events

$$A : X \leq 5; \quad B : Y \leq 5$$

this condition means that

$$\mathbb{P}(X \leq 5, Y \leq 5) = \mathbb{P}(X \leq 5) \cdot \mathbb{P}(Y \leq 5).$$

That is, by assuming that the weather of the two cities are independent, the probability that **there will be at most 5 mm rain in both cities**, is the **product of the probabilities**.

## Independence of random variables

- **for two random variables:** random variables  $X, Y : \Omega \rightarrow \mathbb{R}$  are **independent**, if

$$\mathbb{P}(X \leq t_1, Y \leq t_2) = \mathbb{P}(X \leq t_1) \cdot \mathbb{P}(Y \leq t_2)$$

holds for every real numbers  $t_1, t_2 \in \mathbb{R}$ .

- **for finitely many random variables:** random variables  $X_1, \dots, X_n : \Omega \rightarrow \mathbb{R}$  are **independent**, if

$$\begin{aligned}\mathbb{P}(X_1 \leq t_1, X_2 \leq t_2, \dots, X_n \leq t_n) &= \\ &= \mathbb{P}(X_1 \leq t_1) \cdot \mathbb{P}(X_2 \leq t_2) \dots \mathbb{P}(X_n \leq t_n)\end{aligned}$$

holds for every real numbers  $t_1, t_2, \dots, t_n$ .

- **for countably many random variables:** the random variables  $X_1, X_2, X_3, \dots$  are **independent**, if we get independent random variables with every choice of finitely many from  $X_1, X_2, \dots$

## Independence in the discrete case

If the random variables are **discrete**, that is, their range is finite or countable infinite, then independence is equivalent to the following condition.

The **discrete** random variables  $X$  and  $Y$  are **independent** if and only if for **every possible value  $x_k$  of  $X$**  and for **every possible value  $y_l$  of  $Y$**  the following holds:

$$\mathbb{P}(X = x_k, Y = y_l) = \mathbb{P}(X = x_k) \cdot \mathbb{P}(Y = y_l) \quad (k, l = 1, 2, \dots).$$

That is, the probability that **the value of  $X$  is  $x_k$  and the value of  $Y$  is  $y_l$**  is equal to the **product of the corresponding probabilities**.

## Independence of random variables: example

We roll a fair die twice. Is it true that **the first number** and **the second number** are independent of each other?

## Independence of random variables: example

We roll a fair die twice. Is it true that **the first number** and **the second number** are independent of each other?

**Guess.** There is no connection between the rolls, we can forget about the first one at the second one  $\Rightarrow$  the two numbers are **independent**.

## Independence of random variables: example

We roll a fair die twice. Is it true that **the first number** and **the second number** are independent of each other?

**Guess.** There is no connection between the rolls, we can forget about the first one at the second one  $\Rightarrow$  the two numbers are **independent**.

**Proof.** Let  $X$  be the value of the first roll,  $Y$  the value of the second one. Let us choose  $x_k = 3, y_l = 5$ . Then the condition holds:

$$\frac{1}{36} = \mathbb{P}(X = 3, Y = 5) = \mathbb{P}(X = 3) \cdot \mathbb{P}(Y = 5) = \frac{1}{6} \cdot \frac{1}{6}.$$

## Independence of random variables: example

We roll a fair die twice. Is it true that **the first number** and **the second number** are independent of each other?

**Guess.** There is no connection between the rolls, we can forget about the first one at the second one  $\Rightarrow$  the two numbers are **independent**.

**Proof.** Let  $X$  be the value of the first roll,  $Y$  the value of the second one. Let us choose  $x_k = 3, y_l = 5$ . Then the condition holds:

$$\frac{1}{36} = \mathbb{P}(X = 3, Y = 5) = \mathbb{P}(X = 3) \cdot \mathbb{P}(Y = 5) = \frac{1}{6} \cdot \frac{1}{6}.$$

Similarly, for arbitrary possible values  $(x_k, y_l)$  (pairs of integers from 1 to 6) the following holds:

$$\frac{1}{36} = \mathbb{P}(X = x_k, Y = y_l) = \mathbb{P}(X = x_k) \cdot \mathbb{P}(Y = y_l) = \frac{1}{6} \cdot \frac{1}{6}.$$

Hence **the two rolls are independent**.

## Independence of random variables: example

We roll a fair die twice. Is it true that **the sum** and **the product** of the two numbers are independent of each other?

## Independence of random variables: example

We roll a fair die twice. Is it true that **the sum** and **the product** of the two numbers are independent of each other?

**Guess:** if the sum is larger, the product is larger with a higher probability  
 $\Rightarrow$  **they are not independent.**

**Proof:** let  $X$  be the sum, and  $Y$  the product. Then  $X = 2$  can happen only if both numbers are 1, and hence  $Y$  is 1 for sure. Hence if we choose  $x_1 = 2$  and  $y_2 = 2$ , then, since  $X = 2$  and  $Y = 2$  cannot occur at the same time:

$$\begin{aligned} 0 &= \mathbb{P}(X = 2, Y = 2) \neq \mathbb{P}(X = 2) \cdot \mathbb{P}(Y = 2) = \\ &= \mathbb{P}(11) \cdot \mathbb{P}(12 \text{ or } 21) = \frac{1}{36} \cdot \frac{1}{18} > 0. \end{aligned}$$

Hence the pair  $x_1 = 2, y_2 = 2$  does not satisfy the condition in the definition, the **sum and the product are not independent.**

## Properties of expected value

- (additivity) Let  $X, Y$  be random variables such that  $X, Y, X + Y$  all have finite expected value. Then we have

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

- (scaling) If the random variable  $X$  has finite expected value  $c \in \mathbb{R}$  is a real number, then we have

$$\mathbb{E}(c \cdot X) = c \cdot \mathbb{E}(X).$$

## Properties of expected value

- (additivity) Let  $X, Y$  be random variables such that  $X, Y, X + Y$  all have finite expected value. Then we have

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

- (scaling) If the random variable  $X$  has finite expected value  $c \in \mathbb{R}$  is a real number, then we have

$$\mathbb{E}(c \cdot X) = c \cdot \mathbb{E}(X).$$

- (multiplication and independence) If  $X$  and  $Y$  are **independent** random variables, and  $X, Y, X \cdot Y$  all have finite expected value, then we have

$$\mathbb{E}(XY) = \mathbb{E}(X) \cdot \mathbb{E}(Y).$$

## Properties of expected value

- (additivity) Let  $X, Y$  be random variables such that  $X, Y, X + Y$  all have finite expected value. Then we have

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

- (scaling) If the random variable  $X$  has finite expected value  $c \in \mathbb{R}$  is a real number, then we have

$$\mathbb{E}(c \cdot X) = c \cdot \mathbb{E}(X).$$

- (multiplication and independence) If  $X$  and  $Y$  are **independent** random variables, and  $X, Y, X \cdot Y$  all have finite expected value, then we have

$$\mathbb{E}(XY) = \mathbb{E}(X) \cdot \mathbb{E}(Y).$$

- (expected value of a function) If  $g : \mathbb{R} \rightarrow \mathbb{R}$  is a function,  $\mathbb{E}(X)$  exists, and the range of  $X$  is  $\{x_1, x_2, \dots\}$ , then we have

$$\mathbb{E}(g(X)) = \sum_{k=1}^{\infty} g(x_k) \mathbb{P}(X = x_k).$$

## Additivity

### Proposition

Let  $X, Y$  be random variables such that  $X, Y, X + Y$  all have finite expected value. Then we have

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

*Proof.* Let  $\{x_1, x_2, \dots\}$  be the range of  $X$ , and  $\{y_1, y_2, \dots\}$  be the range of  $Y$ . Then all possible values of  $X + Y$  are of the form  $x_k + y_m$ , and

$$\begin{aligned}\mathbb{E}(X + Y) &= \sum_{k,m} (x_k + y_m) \mathbb{P}(X = x_k, Y = y_m) = \\ &= \sum_{k,m} x_k \mathbb{P}(X = x_k, Y = y_m) + \sum_{k,m} y_m \mathbb{P}(X = x_k, Y = y_m) = \\ &= \sum_k x_k \mathbb{P}(X = x_k) + \sum_m y_m \mathbb{P}(Y = y_m) = \mathbb{E}(X) + \mathbb{E}(Y),\end{aligned}$$

where we used that the events  $\{X = x_k, Y = y_m\}$  are pairwise disjoint, their union is  $\{X = x_k\}$ , and similar argument works for the term corresponding to  $Y$ .

## Additivity

### Proposition

Let  $X, Y$  be random variables such that  $X, Y, X + Y$  all have finite expected value. Then we have

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

*Proof.* Let  $\{x_1, x_2, \dots\}$  be the range of  $X$ , and  $\{y_1, y_2, \dots\}$  be the range of  $Y$ . Then all possible values of  $X + Y$  are of the form  $x_k + y_m$ , and

$$\begin{aligned}\mathbb{E}(X + Y) &= \sum_{k,m} (x_k + y_m) \mathbb{P}(X = x_k, Y = y_m) = \\ &= \sum_{k,m} x_k \mathbb{P}(X = x_k, Y = y_m) + \sum_{k,m} y_m \mathbb{P}(X = x_k, Y = y_m) = \\ &= \sum_k x_k \mathbb{P}(X = x_k) + \sum_m y_m \mathbb{P}(Y = y_m) = \mathbb{E}(X) + \mathbb{E}(Y),\end{aligned}$$

where we used that the events  $\{X = x_k, Y = y_m\}$  are pairwise disjoint, their union is  $\{X = x_k\}$ , and similar argument works for the term corresponding to  $Y$ .

## Multiplication and independence

### Proposition

Let  $X, Y$  be **independent** random variables. If the expected value of  $X$  and  $Y$  exist, then we have

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

*Proof.* Let  $\{x_1, x_2, \dots\}$  be the range of  $X$ , and  $\{y_1, y_2, \dots\}$  be the range of  $Y$ . Then all possible values of  $X + Y$  are of the form  $x_k \cdot y_m$ , and

$$\begin{aligned}\mathbb{E}(XY) &= \sum_{k,m} x_k \cdot y_m \mathbb{P}(X = x_k, Y = y_m) = \\ &\stackrel{(*)}{=} \sum_{k,m} x_k y_m \mathbb{P}(X = x_k) \cdot \mathbb{P}(Y = y_m) \\ &= \left( \sum_k x_k \mathbb{P}(X = x_k) \right) \cdot \left( \sum_m y_m \mathbb{P}(Y = y_m) \right) = \mathbb{E}(X) \cdot \mathbb{E}(Y),\end{aligned}$$

where in equality (\*) we used the consequence of independence for the discrete case.

## Properties of variance

- (nonnegativity)  $D^2(X) \geq 0$  and  $D(X) \geq 0$  always hold
- (scaling and translation) for real numbers  $a, b$  and random variable  $X$  with finite variance we have

$$D^2(aX + b) = a^2 D^2(X) \quad \Rightarrow \quad D(aX + b) = |a|D(X).$$

## Properties of variance

- (nonnegativity)  $D^2(X) \geq 0$  and  $D(X) \geq 0$  always hold
- (scaling and translation) for real numbers  $a, b$  and random variable  $X$  with finite variance we have

$$D^2(aX + b) = a^2 D^2(X) \quad \Rightarrow \quad D(aX + b) = |a| D(X).$$

- (standard deviation of the sum) if  $X, Y$  are **independent** random variables with finite variance, then we have

$$D^2(X+Y) = D^2(X) + D^2(Y) \quad \Rightarrow \quad D(X+Y) = \sqrt{D^2(X) + D^2(Y)}.$$

- there exists a random variable whose expected value exist, but whose variance does not exist (for example:  $\mathbb{P}(X = k) = c/k^3$  with an arbitrary  $c$ )

## Calculating the expected value and the standard deviation

**Example.** Let  $X$  be a random variable with  $\mathbb{E}(X) = 4$ ,  $D(X) = 1$ , and  $Y$  be a random variable with  $\mathbb{E}(Y) = 6$  and  $D(Y) = 2$ . Suppose furthermore that  $X$  and  $Y$  are **independent**. Then we have

- $\mathbb{E}(X + Y) =$

## Calculating the expected value and the standard deviation

**Example.** Let  $X$  be a random variable with  $\mathbb{E}(X) = 4$ ,  $D(X) = 1$ , and  $Y$  be a random variable with  $\mathbb{E}(Y) = 6$  and  $D(Y) = 2$ . Suppose furthermore that  $X$  and  $Y$  are **independent**. Then we have

- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 10$ ;
- $\mathbb{E}(X - Y) =$

## Calculating the expected value and the standard deviation

**Example.** Let  $X$  be a random variable with  $\mathbb{E}(X) = 4$ ,  $D(X) = 1$ , and  $Y$  be a random variable with  $\mathbb{E}(Y) = 6$  and  $D(Y) = 2$ . Suppose furthermore that  $X$  and  $Y$  are **independent**. Then we have

- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 10$ ;
- $\mathbb{E}(X - Y) = \mathbb{E}(X) - \mathbb{E}(Y) = -2$ ;
- $\mathbb{E}(2X + 3Y) =$

## Calculating the expected value and the standard deviation

**Example.** Let  $X$  be a random variable with  $\mathbb{E}(X) = 4$ ,  $D(X) = 1$ , and  $Y$  be a random variable with  $\mathbb{E}(Y) = 6$  and  $D(Y) = 2$ . Suppose furthermore that  $X$  and  $Y$  are **independent**. Then we have

- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 10$ ;
- $\mathbb{E}(X - Y) = \mathbb{E}(X) - \mathbb{E}(Y) = -2$ ;
- $\mathbb{E}(2X + 3Y) = 2\mathbb{E}(X) + 3\mathbb{E}(Y) = 2 \cdot 4 + 3 \cdot 6 = 26$ ;
- $\mathbb{E}(2X - 3Y) =$

## Calculating the expected value and the standard deviation

**Example.** Let  $X$  be a random variable with  $\mathbb{E}(X) = 4$ ,  $D(X) = 1$ , and  $Y$  be a random variable with  $\mathbb{E}(Y) = 6$  and  $D(Y) = 2$ . Suppose furthermore that  $X$  and  $Y$  are **independent**. Then we have

- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 10$ ;
- $\mathbb{E}(X - Y) = \mathbb{E}(X) - \mathbb{E}(Y) = -2$ ;
- $\mathbb{E}(2X + 3Y) = 2\mathbb{E}(X) + 3\mathbb{E}(Y) = 2 \cdot 4 + 3 \cdot 6 = 26$ ;
- $\mathbb{E}(2X - 3Y) = 2\mathbb{E}(X) - 3\mathbb{E}(Y) = 2 \cdot 4 - 3 \cdot 6 = -10$ ;
- $D(X + Y) =$

## Calculating the expected value and the standard deviation

**Example.** Let  $X$  be a random variable with  $\mathbb{E}(X) = 4$ ,  $D(X) = 1$ , and  $Y$  be a random variable with  $\mathbb{E}(Y) = 6$  and  $D(Y) = 2$ . Suppose furthermore that  $X$  and  $Y$  are **independent**. Then we have

- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 10$ ;
- $\mathbb{E}(X - Y) = \mathbb{E}(X) - \mathbb{E}(Y) = -2$ ;
- $\mathbb{E}(2X + 3Y) = 2\mathbb{E}(X) + 3\mathbb{E}(Y) = 2 \cdot 4 + 3 \cdot 6 = 26$ ;
- $\mathbb{E}(2X - 3Y) = 2\mathbb{E}(X) - 3\mathbb{E}(Y) = 2 \cdot 4 - 3 \cdot 6 = -10$ ;
- $D(X + Y) = \sqrt{D^2(X) + D^2(Y)} = \sqrt{1 + 2^2} = \sqrt{5}$ ;
- $D(X - Y) =$

## Calculating the expected value and the standard deviation

**Example.** Let  $X$  be a random variable with  $\mathbb{E}(X) = 4$ ,  $D(X) = 1$ , and  $Y$  be a random variable with  $\mathbb{E}(Y) = 6$  and  $D(Y) = 2$ . Suppose furthermore that  $X$  and  $Y$  are **independent**. Then we have

- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 10$ ;
- $\mathbb{E}(X - Y) = \mathbb{E}(X) - \mathbb{E}(Y) = -2$ ;
- $\mathbb{E}(2X + 3Y) = 2\mathbb{E}(X) + 3\mathbb{E}(Y) = 2 \cdot 4 + 3 \cdot 6 = 26$ ;
- $\mathbb{E}(2X - 3Y) = 2\mathbb{E}(X) - 3\mathbb{E}(Y) = 2 \cdot 4 - 3 \cdot 6 = -10$ ;
- $D(X + Y) = \sqrt{D^2(X) + D^2(Y)} = \sqrt{1 + 2^2} = \sqrt{5}$ ;
- $D(X - Y) = \sqrt{D^2(X) + (-1)^2 D^2(Y)} = \sqrt{1 + 2^2} = \sqrt{5}$ ;
- $D(2X + 3Y) =$

## Calculating the expected value and the standard deviation

**Example.** Let  $X$  be a random variable with  $\mathbb{E}(X) = 4$ ,  $D(X) = 1$ , and  $Y$  be a random variable with  $\mathbb{E}(Y) = 6$  and  $D(Y) = 2$ . Suppose furthermore that  $X$  and  $Y$  are **independent**. Then we have

- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 10$ ;
- $\mathbb{E}(X - Y) = \mathbb{E}(X) - \mathbb{E}(Y) = -2$ ;
- $\mathbb{E}(2X + 3Y) = 2\mathbb{E}(X) + 3\mathbb{E}(Y) = 2 \cdot 4 + 3 \cdot 6 = 26$ ;
- $\mathbb{E}(2X - 3Y) = 2\mathbb{E}(X) - 3\mathbb{E}(Y) = 2 \cdot 4 - 3 \cdot 6 = -10$ ;
- $D(X + Y) = \sqrt{D^2(X) + D^2(Y)} = \sqrt{1 + 2^2} = \sqrt{5}$ ;
- $D(X - Y) = \sqrt{D^2(X) + (-1)^2 D^2(Y)} = \sqrt{1 + 2^2} = \sqrt{5}$ ;
- $D(2X + 3Y) = \sqrt{2^2 D^2(X) + 3^2 D^2(Y)} = \sqrt{4 \cdot 1 + 9 \cdot 4} = \sqrt{40}$ ;
- $D(2X - 3Y) =$

## Calculating the expected value and the standard deviation

**Example.** Let  $X$  be a random variable with  $\mathbb{E}(X) = 4$ ,  $D(X) = 1$ , and  $Y$  be a random variable with  $\mathbb{E}(Y) = 6$  and  $D(Y) = 2$ . Suppose furthermore that  $X$  and  $Y$  are **independent**. Then we have

- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 10$ ;
- $\mathbb{E}(X - Y) = \mathbb{E}(X) - \mathbb{E}(Y) = -2$ ;
- $\mathbb{E}(2X + 3Y) = 2\mathbb{E}(X) + 3\mathbb{E}(Y) = 2 \cdot 4 + 3 \cdot 6 = 26$ ;
- $\mathbb{E}(2X - 3Y) = 2\mathbb{E}(X) - 3\mathbb{E}(Y) = 2 \cdot 4 - 3 \cdot 6 = -10$ ;
- $D(X + Y) = \sqrt{D^2(X) + D^2(Y)} = \sqrt{1 + 2^2} = \sqrt{5}$ ;
- $D(X - Y) = \sqrt{D^2(X) + (-1)^2 D^2(Y)} = \sqrt{1 + 2^2} = \sqrt{5}$ ;
- $D(2X + 3Y) = \sqrt{2^2 D^2(X) + 3^2 D^2(Y)} = \sqrt{4 \cdot 1 + 9 \cdot 4} = \sqrt{40}$ ;
- $D(2X - 3Y) = \sqrt{2^2 D^2(X) + (-3)^2 D^2(Y)} = \sqrt{4 \cdot 1 + 9 \cdot 4} = \sqrt{40}$ .

# Properties of binomial distribution

- $n$  independent experiments;
- each of them is successful with probability  $p$ ;
- $X$  is the number of successful experiments.

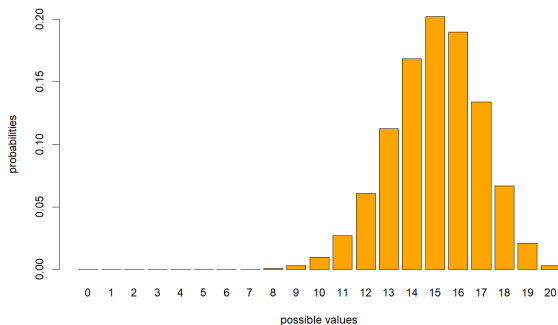
If  $X$  has **binomial distribution** with order  $n$  and parameter  $p$ , that is,

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad (k = 0, 1, \dots, n),$$

then the **expected value**, and **standard deviation** of  $X$  are as follows:

$$\mathbb{E}(X) = np; \quad D(X) = \sqrt{np(1 - p)}.$$

## Example: binomial distribution



Binomial distribution,  $n = 20$ ,  $p = 0.75$ . Horizontal axis: possible values,  $k = 0, 1, \dots, 20$ , height of the columns: probabilities  $\mathbb{P}(X = k)$ .

# Expected value of binomial distribution: proof

## Expected value of binomial distribution: proof

### Proposition

*Suppose that  $X$  has binomial distribution with order  $n$  and parameter  $p$ . Then the expected value of  $X$  is equal to  $np$ .*

## Expected value of binomial distribution: proof

### Proposition

*Suppose that  $X$  has binomial distribution with order  $n$  and parameter  $p$ . Then the expected value of  $X$  is equal to  $np$ .*

$X$  can be constructed as the number of successful experiments out of  $n$  trials, each being successful with probability  $p$ , independently. Let us take the following indicator random variables for  $j = 1, 2, \dots, n$ :

$$\mathbb{I}_j = \begin{cases} 1, & \text{if the } j\text{th trial is successful;} \\ 0, & \text{otherwise.} \end{cases}$$

## Expected value of binomial distribution: proof

### Proposition

*Suppose that  $X$  has binomial distribution with order  $n$  and parameter  $p$ . Then the expected value of  $X$  is equal to  $np$ .*

$X$  can be constructed as the number of successful experiments out of  $n$  trials, each being successful with probability  $p$ , independently. Let us take the following indicator random variables for  $j = 1, 2, \dots, n$ :

$$\mathbb{I}_j = \begin{cases} 1, & \text{if the } j\text{th trial is successful;} \\ 0, & \text{otherwise.} \end{cases}$$

Then the sum of indicators  $\mathbb{I}_j$  is equal to  $X$ . Hence, based on the additivity property of the expected value, we obtain

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{j=1}^n \mathbb{I}_j\right) = \sum_{j=1}^n \mathbb{E}(\mathbb{I}_j) = \sum_{j=1}^n 1 \cdot \mathbb{P}(\mathbb{I}_j = 1) = np. \quad \square$$

## Standard deviation of binomial distribution

*Proof.* We consider the same indicator random variables:

$$\mathbb{I}_j = \begin{cases} 1 & \text{if the } j\text{th experiment is successful;} \\ 0 & \text{if the } j\text{th experiment is not successful.} \end{cases}$$

## Standard deviation of binomial distribution

*Proof.* We consider the same indicator random variables:

$$\mathbb{I}_j = \begin{cases} 1 & \text{if the } j\text{th experiment is successful;} \\ 0 & \text{if the } j\text{th experiment is not successful.} \end{cases}$$

Now  $\mathbb{I}_j = \mathbb{I}_j^2$ , since  $0^2 = 0$  and  $1^2 = 1$ , and we have already seen that  $\mathbb{E}(\mathbb{I}_j) = p$ . Therefore

$$D^2(\mathbb{I}_j) = \mathbb{E}(\mathbb{I}_j^2) - \mathbb{E}(\mathbb{I}_j)^2 = \mathbb{E}(\mathbb{I}_j) - \mathbb{E}(\mathbb{I}_j)^2 = p - p^2 = p(1 - p).$$

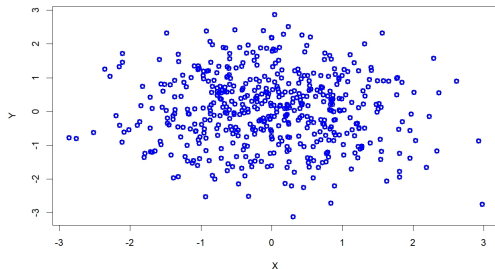
Since the indicators  $\mathbb{I}_j$  are **independent**, and they sum up to  $X$ :

$$\begin{aligned} D(X) &= \sqrt{D^2(\mathbb{I}_1 + \mathbb{I}_2 + \dots + \mathbb{I}_n)} = \sqrt{D^2(\mathbb{I}_1) + D^2(\mathbb{I}_2) + \dots + D^2(\mathbb{I}_n)} = \\ &= \sqrt{p(1 - p) + p(1 - p) + \dots + p(1 - p)} = \sqrt{np(1 - p)}. \end{aligned}$$

# Covariance and correlation coefficient

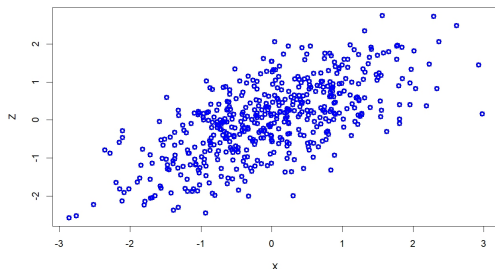
- two random variables can be
  - ▶ **independent**: the income of two randomly chosen people, or
  - ▶ **not independent**: the monthly income of a given person now, and in next April
- the **strength of the connection** can be different:
  - ▶ the age and monthly income of a randomly chosen person has a "strong connection", young and elderly people often have significantly less income;
  - ▶ the age and the height of an adult can have a "weak connection".
- **covariance** and **correlation coefficient** measure the strength of the connection (among other possibilities)

# Independent random variables



500 random points on the plane, whose coordinates follow **independent** standard normal distribution. Both the covariance and correlation coefficient will be equal to **0**.

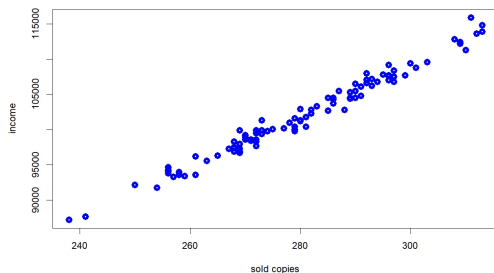
## Positive correlation



A sample of size 500 from the following two-dimensional distribution:  $(X, \frac{X+Z}{\sqrt{2}})$ , where  $X, Z \sim N(0, 1)$  are independent.

The larger  $X$  is, „probably” the larger  $(X+Z)/\sqrt{2}$  is  $\rightarrow$  both the **covariance** and the **correlation coefficient** is **positive**.

## Strong positive correlation



Sample of size 100 from distribution  $(X + Y, 300X + 400Y)$ , where  $X \sim \text{Poisson}(100)$  and  $Y \sim \text{Poisson}(180)$  are independent. The points fit very well to a line with positive slope  $\rightarrow$  the **correlation coefficient** is **positive** and **close to 1**, which is the largest possible value.

# Covariance

Let  $X$  and  $Y$  be random variables whose standard deviation exist. Then the **covariance** of  $X$  and  $Y$  is defined by

$$\text{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))).$$

# Covariance

Let  $X$  and  $Y$  be random variables whose standard deviation exist. Then the **covariance** of  $X$  and  $Y$  is defined by

$$\text{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))).$$

- **Calculating covariance:**

$$\text{cov}(X, Y) = \mathbb{E}(X \cdot Y) - \mathbb{E}(X)\mathbb{E}(Y).$$

- **Symmetry.**  $\text{cov}(X, Y) = \text{cov}(Y, X)$ .
- **Relationship with variance.**  $\text{cov}(X, X) = D^2(X)$ .

# Covariance

Let  $X$  and  $Y$  be random variables whose standard deviation exist. Then the **covariance** of  $X$  and  $Y$  is defined by

$$\text{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))).$$

- **Calculating covariance:**

$$\text{cov}(X, Y) = \mathbb{E}(X \cdot Y) - \mathbb{E}(X)\mathbb{E}(Y).$$

- **Symmetry.**  $\text{cov}(X, Y) = \text{cov}(Y, X)$ .
- **Relationship with variance.**  $\text{cov}(X, X) = D^2(X)$ .
- **Relationship with independence.** If random variables  $X$  and  $Y$  are **independent**, then  $\text{cov}(X, Y) = 0$ .

**The other direction is not true:**  $\text{cov}(X, Y) = 0$  does not imply that  $X$  and  $Y$  are independent.

## Properties of covariance

- Covariance with a constant.  $\text{cov}(X, c) = 0$ , if  $c$  is a real number.
- **Linearity.** We have

$$\text{cov}(X + Y, Z) = \text{cov}(X, Z) + \text{cov}(Y, Z),$$

and furthermore, for every  $c \in \mathbb{R}$  we have

$$\text{cov}(c \cdot X, Y) = c \cdot \text{cov}(X, Y).$$

- **Variance of a sum.**  $D^2(X + Y) = D^2(X) + D^2(Y) + 2\text{cov}(X, Y)$ . In addition, we have

$$D^2\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n D^2(X_i) + 2 \sum_{i < j} \text{cov}(X_i, X_j).$$

- Variance of a difference.  $D^2(X - Y) = D^2(X) + D^2(Y) - 2\text{cov}(X, Y)$ .

## Correlation coefficient

Let  $X$  and  $Y$  be random variables whose standard deviation exist. Then the (Pearson) **correlation coefficient** of  $X$  and  $Y$  is defined by

$$R(X, Y) = \begin{cases} \frac{\text{cov}(X, Y)}{D(X)D(Y)}, & \text{if } D(X) > 0, D(Y) > 0; \\ 0, & \text{if } D(X) = 0 \text{ or } D(Y) = 0. \end{cases}$$

## Correlation coefficient

Let  $X$  and  $Y$  be random variables whose standard deviation exist. Then the (Pearson) **correlation coefficient** of  $X$  and  $Y$  is defined by

$$R(X, Y) = \begin{cases} \frac{\text{cov}(X, Y)}{D(X)D(Y)}, & \text{if } D(X) > 0, D(Y) > 0; \\ 0, & \text{if } D(X) = 0 \text{ or } D(Y) = 0. \end{cases}$$

- **Possible values.** The value of the correlation coefficient is always between  $-1$  and  $1$ :

$$|R(X, Y)| \leq 1.$$

- **Linear dependence.** Let  $a > 0$  and  $b$  be real numbers. Then we have

$$R(X, aX + b) = 1 \quad \text{and} \quad R(X, -aX + b) = -1.$$

- Suppose that  $|R(X, Y)| = 1$ . Then there exist real numbers  $a$  and  $b$  such that  $Y = aX + b$  holds with probability 1. That is, the extreme values of  $R$  can be achieved in the case of linear dependence.