

## Discrete random variables (Lecture 5)

### Definition (Random variable)

A function  $X : \Omega \rightarrow \mathbb{R}$  is a **random variable** if for every real number  $t \in \mathbb{R}$  we have that

$$\{\omega \in \Omega : X(\omega) \leq t\} \in \mathcal{A},$$

that is, for all numbers  $t$ , the probability  $\mathbb{P}(X \leq t)$  is well-defined.

In this case  $\mathbb{P}(X = t)$  is also well-defined.

The random variable  $X : \Omega \rightarrow \mathbb{R}$  is **discrete** if **its range is finite or countably infinite**.

## Discrete random variables (Lecture 5)

### Definition (Random variable)

A function  $X : \Omega \rightarrow \mathbb{R}$  is a **random variable** if for every real number  $t \in \mathbb{R}$  we have that

$$\{\omega \in \Omega : X(\omega) \leq t\} \in \mathcal{A},$$

that is, for all numbers  $t$ , the probability  $\mathbb{P}(X \leq t)$  is well-defined.

In this case  $\mathbb{P}(X = t)$  is also well-defined.

The random variable  $X : \Omega \rightarrow \mathbb{R}$  is **discrete** if **its range is finite or countably infinite**.

Let  $X$  be a **discrete random variable** with range:

$$\{x_1, x_2, \dots\}, \quad \text{and } p_k = \mathbb{P}(X = x_k) \quad (k = 1, 2, \dots).$$

Then the sequence  $(x_1, p_1), (x_2, p_2), \dots$  is **distribution** of the random variable  $X$ . In this case, we have

$$p_k \geq 0 \text{ for every } k, \text{ and } \sum_{k=1}^{\infty} p_k = 1,$$

that is,  $(p_k)_{k \geq 0}$  is a probability distribution.

# Expected value of discrete random variables

## Definition (Expected value, discrete case)

Let  $X : \Omega \rightarrow \mathbb{R}$  be a discrete random variable with distribution  $(x_1, p_1), (x_2, p_2), \dots$ , that is,  $\mathbb{P}(X = x_i) = p_i$ , where  $i = 1, 2, \dots$ . Then  $X$  has expected value

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} x_i p_i, \quad \text{if } \sum_{i=1}^{\infty} |x_i| p_i < \infty.$$

# Expected value of discrete random variables

## Definition (Expected value, discrete case)

Let  $X : \Omega \rightarrow \mathbb{R}$  be a discrete random variable with distribution  $(x_1, p_1), (x_2, p_2), \dots$ , that is,  $\mathbb{P}(X = x_i) = p_i$ , where  $i = 1, 2, \dots$ . Then  $X$  has expected value

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} x_i p_i, \quad \text{if } \sum_{i=1}^{\infty} |x_i| p_i < \infty.$$

**Example: coin tosses.** Let  $X$  be the number of heads out of three fair coin tosses. Then

$$\mathbb{E}(X) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{12}{8} = \frac{3}{2} = 1.5.$$

**Example: dice rolls.** Let  $Y$  be the value of a die roll. Then

$$\mathbb{E}(Y) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = \frac{7}{2} = 3.5.$$

# Standard deviation

Possible motivation: error of an estimate or measurements, uncertainty of prediction

## Definition (Variance)

Let  $X : \Omega \rightarrow \mathbb{R}$  be a random variable for which  $\mathbb{E}(X^2)$  exists. Then the variance of  $X$  is defined by

$$\text{Var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right).$$

# Standard deviation

Possible motivation: error of an estimate or measurements, uncertainty of prediction

## Definition (Variance)

Let  $X : \Omega \rightarrow \mathbb{R}$  be a random variable for which  $\mathbb{E}(X^2)$  exists. Then the variance of  $X$  is defined by

$$\text{Var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right).$$

## Definition (Standard deviation)

Let  $X : \Omega \rightarrow \mathbb{R}$  be a random variable for which  $\mathbb{E}(X^2)$  exists. Then the standard deviation of  $X$  is defined by

$$D(X) = \sqrt{\mathbb{E}\left((X - \mathbb{E}(X))^2\right)}.$$

# Calculating standard deviation

## Proposition

Let  $X : \Omega \rightarrow \mathbb{R}$  be a random variable for which  $\mathbb{E}(X^2)$  exists. Then we have

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

## Proposition (Variance in the integer valued case)

Let  $X$  be a discrete random variable for which  $\mathbb{E}(X^2)$  exists, and whose range consists only of nonnegative integers. Then

$$\text{Var}(X) = \sum_{k=0}^{\infty} k^2 \mathbb{P}(X = k) - \mathbb{E}(X)^2 = \sum_{k=0}^{\infty} k^2 \mathbb{P}(X = k) - \left[ \sum_{k=0}^{\infty} k \mathbb{P}(X = k) \right]^2.$$

## Standard deviation in the discrete case

Let  $X$  be the number of heads out of three fair coin tosses:

$$\mathbb{P}(X = 0) = 1/8; \quad \mathbb{P}(X = 1) = 3/8; \quad \mathbb{P}(X = 2) = 3/8; \quad \mathbb{P}(X = 3) = 1/8.$$

We can calculate as follows:

$$\mathbb{E}(X^2) = \sum_{k=0}^3 k^2 \mathbb{P}(X = k) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 4 \cdot \frac{3}{8} + 9 \cdot \frac{1}{8} = \frac{24}{8} = 3.$$

This and the earlier results imply that

$$\text{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = 3 - 1.5^2 = 3 - 2.25 = 0.75 = \frac{3}{4}.$$

Finally, the standard deviation of the number of heads:

$$D(X) = \sqrt{\frac{3}{4}} = 0.866.$$

## Fair die

Let  $X$  be the value of a roll with a fair die. Then we have

$$\mathbb{E}(X^2) = \frac{1}{6} \cdot 1^2 + \frac{1}{6} \cdot 2^2 + \frac{1}{6} \cdot 3^2 + \frac{1}{6} \cdot 4^2 + \frac{1}{6} \cdot 5^2 + \frac{1}{6} \cdot 6^2 = \frac{91}{6}.$$

On the other hand,

$$\mathbb{E}(X) = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = \frac{7}{2}.$$

Hence

$$\text{Var}^2(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = 2.92.$$

Standard deviation:  $D(X) = \sqrt{2.92} = 1.71$ .

In general, if there are  $n$  "sides":  $D(X) = \sqrt{\frac{n^2-1}{12}}$ .

## Binomial distribution: example

**Six people** work together on a project.

Suppose that on each day, **independently**  
all of them are **out of office** with probability  $p = 0.03$

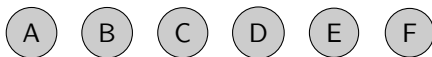
What is the probability that on a working day  
**exactly two members of the team are absent?**

# Binomial distribution: example

**Six people** work together on a project.

Suppose that on each day, **independently**  
all of them are **out of office** with probability  $p = 0.03$

What is the probability that on a working day  
**exactly two members of the team are absent?**



a few good outcomes and their probabilities:

0.03	0.03	0.97	0.97	0.97	0.97	$\rightarrow 0.03^2 \cdot 0.97^4$
0.03	0.97	0.03	0.97	0.97	0.97	$\rightarrow 0.03^2 \cdot 0.97^4$
0.03	0.97	0.97	0.03	0.97	0.97	$\rightarrow 0.03^2 \cdot 0.97^4$
...						
0.97	0.03	0.03	0.97	0.97	0.97	$\rightarrow 0.03^2 \cdot 0.97^4$
...						
0.97	0.97	0.97	0.97	0.03	0.03	$\rightarrow 0.03^2 \cdot 0.97^4$

multiplication

# Binomial distribution: example

**Six people** work together on a project.

Suppose that on each day, **independently**  
all of them are **out of office** with probability  $p = 0.03$

What is the probability that on a working day  
**exactly two members of the team are absent?**

$$\boxed{0.97} \boxed{0.03} \boxed{0.03} \boxed{0.97} \boxed{0.97} \boxed{0.97} \rightarrow 0.03^2 \cdot 0.97^4$$

number of good outcomes:

number of ways to choose the two absent colleagues:

probability of a good outcome:

hence the probability:

# Binomial distribution: example

**Six people** work together on a project.

Suppose that on each day, **independently**  
all of them are **out of office** with probability  $p = 0.03$

What is the probability that on a working day  
**exactly two members of the team are absent?**

$$\boxed{0.97} \quad \boxed{0.03} \quad \boxed{0.03} \quad \boxed{0.97} \quad \boxed{0.97} \quad \boxed{0.97} \quad \rightarrow 0.03^2 \cdot 0.97^4$$

number of good outcomes:

number of ways to choose the two absent colleagues:  $\binom{6}{2}$

probability of a good outcome:  $0.03^2 \cdot 0.97^4$

hence the probability:

$$\mathbb{P}(\text{exactly 2 people missing}) = \binom{6}{2} \cdot 0.03^2 \cdot 0.97^4 = 1.2\%.$$

## Binomial distribution: definition

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

The random variable  $X$  has **binomial distribution** with order  $n$  and parameter  $p$ , if its range is

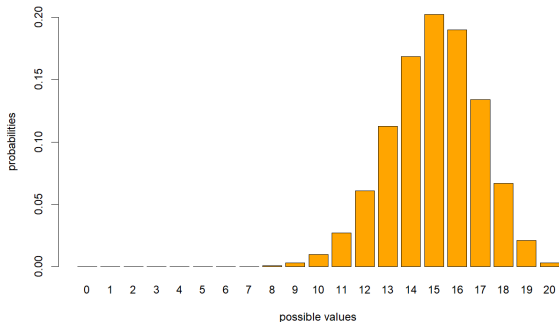
$$0, 1, 2, \dots, n,$$

and for every integer  $0 \leq k \leq n$  we have

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

( $n \geq 1$  is an integer,  $0 < p < 1$ .) Notation:  $\text{Bin}(n, p)$ .

## Example: binomial distribution



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

# Binomial distribution

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

For example:

- Sampling with replacement,  $n$  draws,  $p$  is the proportion of red balls in the urn.
- In a survey, a question is asked from  $n = 1500$  people, everyone answers independently with probability  $p = 0.8$ . The number of answers has binomial distribution.
- Each of the  $n = 60000$  customers of an insurance company causes accident with probability  $p = 0.0001$ , independently. The number of customers causing accident has binomial distribution.

## Binomial distribution

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

What is the probability that there are **exactly  $k$  successful experiments**, that is,  $X = k$ ? As in the earlier example:

- Number of ways to choose the  $k$  successful experiments:  $\binom{n}{k}$ .
- The probability of such a good outcome:  $p^k(1-p)^{n-k}$ , because the experiments are independent, the probability of joint occurrence (intersection) is the product of the probabilities, and there are  $k$  successful experiments, the remaining  $n - k$  are not successful.
- We can take the product to get the probability of  $X = k$ :

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

## Example: binomial distribution

In a survey,  $n = 1500$  people were involved. For one particular question, each participant answers with probability  $p = 0.8$ , **independently** of each other. Let  $X$  be the number of participants who answered this question. Then

- $X$  has **binomial distribution** with order  $n = 1500$  and parameter  $p = 0.8$ .
- For every integer  $0 \leq k \leq 1500$  we have

$$\begin{aligned}\mathbb{P}(k \text{ answers}) &= \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \\ &= \binom{1500}{k} 0.8^k \cdot 0.2^{1500-k}.\end{aligned}$$

- For example, the probability that we get exactly  $k = 1200$  answers:

$$\mathbb{P}(1200 \text{ answers}) = \mathbb{P}(X = 1200) = \binom{1500}{1200} 0.8^{1200} \cdot 0.2^{300} = 2.57\%.$$

## 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

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

## Binomial distribution: expected value and variance

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

In a survey,  $n = 1500$  people were involved. For one particular question, each participant answers with probability  $p = 0.8$ , **independently** of each other. Let  $X$  be the number of participants who answered this question. Then

- the **expected value** of the number of answers:

$$\mathbb{E}(X) = np = 1500 \cdot 0.8 = 1200.$$

- the **standard deviation** of the number of answers:

$$D(X) = \sqrt{np(1-p)} = \sqrt{1500 \cdot 0.8 \cdot 0.2} = 15.5.$$

## Approximation of binomial distribution

An insurance company has  $n = 100000$  customers, and each of them causes accident with probability  $p = 0.0001$ , independently of each other. The **expected value** of the number of customers causing accident (denoted by  $X$ ):

$$\mathbb{E}(X) = np = 100000 \cdot 0.0001 = 10.$$

The probability that **exactly  $k$  customers cause accident**:

$$\mathbb{P}(X = k) = \binom{100000}{k} \cdot 0.0001^k \cdot 0.9999^{100000-k} =$$

## Approximation of binomial distribution

An insurance company has  $n = 100000$  customers, and each of them causes accident with probability  $p = 0.0001$ , independently of each other. The **expected value** of the number of customers causing accident (denoted by  $X$ ):

$$\mathbb{E}(X) = np = 100000 \cdot 0.0001 = 10.$$

The probability that **exactly  $k$  customers cause accident**:

$$\begin{aligned} \mathbb{P}(X = k) &= \binom{100000}{k} \cdot 0.0001^k \cdot 0.9999^{100000-k} = \\ &= \frac{100000 \cdot 99999 \cdot \dots \cdot (100001 - k)}{k!} \cdot 0.0001^k \cdot 0.9999^{100000-k} \approx \end{aligned}$$

## Approximation of binomial distribution

An insurance company has  $n = 100000$  customers, and each of them causes accident with probability  $p = 0.0001$ , independently of each other. The **expected value** of the number of customers causing accident (denoted by  $X$ ):

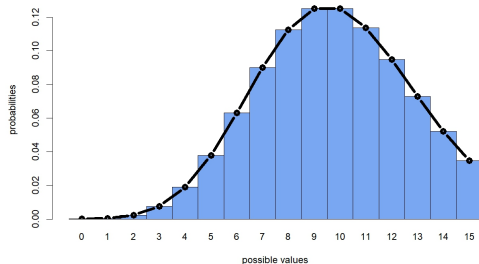
$$\mathbb{E}(X) = np = 100000 \cdot 0.0001 = 10.$$

The probability that **exactly  $k$  customers cause accident**:

$$\begin{aligned}\mathbb{P}(X = k) &= \binom{100000}{k} \cdot 0.0001^k \cdot 0.9999^{100000-k} = \\ &= \frac{100000 \cdot 99999 \cdot \dots \cdot (100001 - k)}{k!} \cdot 0.0001^k \cdot 0.9999^{100000-k} \approx \\ &\approx \frac{100000^k \cdot 0.0001^k}{k!} \left(1 - \frac{10}{100000}\right)^{100000} \approx \frac{10^k}{k!} e^{-10},\end{aligned}$$

by using that  $\lim_{n \rightarrow \infty} \left(1 - \frac{x}{n}\right)^n = e^{-x}$  holds for every  $x > 0$ .

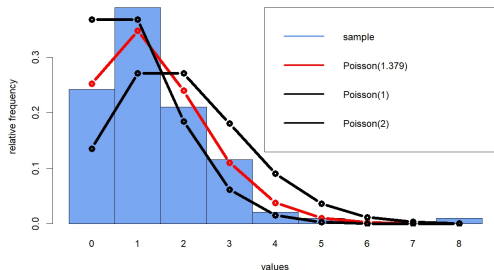
# Approximation of binomial distribution



Binomial distribution with order  $n = 100000$  and parameter  $p = 0.0001$  ( $x$  axis:  $k$ , height of columns:  $\mathbb{P}(X = k)$ ).

The function  $\frac{10^k}{k!} e^{-10}$  is in black (this is the Poisson(10) distribution).

# Poisson distribution



Histogram of the number of goals at  $n = 95$  soccer matches, and Poisson distributions with different parameters

# Applications of Poisson distribution

- **occurences of rare event within a given time period:**
  - ▶ victims of horse kick in the Prussian army during a year (this was the first statistical example)
  - ▶ number of accidents in a city during a month or a year;
- incoming requests in a system (see also queueing theory):
  - ▶ number of customers in a shop in an hour
  - ▶ number of downloads of a website in an hour

# Applications of Poisson distribution

- **occurrences of rare event within a given time period:**
  - ▶ victims of horse kick in the Prussian army during a year (this was the first statistical example)
  - ▶ number of accidents in a city during a month or a year;
- incoming requests in a system (see also queueing theory):
  - ▶ number of customers in a shop in an hour
  - ▶ number of downloads of a website in an hour
- **in general:** number of events occurring one after each other, with random time gaps

Poisson process: the number of events occurring within a time period of length  $t$  has Poisson distribution with parameter  $c \cdot t$ .

Then the expected value of the number of events during time  $t$ :

# Applications of Poisson distribution

- **occurrences of rare event within a given time period:**
  - ▶ victims of horse kick in the Prussian army during a year (this was the first statistical example)
  - ▶ number of accidents in a city during a month or a year;
- incoming requests in a system (see also queueing theory):
  - ▶ number of customers in a shop in an hour
  - ▶ number of downloads of a website in an hour
- **in general:** number of events occurring one after each other, with random time gaps

Poisson process: the number of events occurring within a time period of length  $t$  has Poisson distribution with parameter  $c \cdot t$ .

Then the expected value of the number of events during time  $t$ :  $c \cdot t$ , that is, it is proportional to the length of the interval.

## Poisson distribution: definition

- out of  $n$  independent experiments each is successful with probability  $p$ , where  $n$  is "large" and  $p$  is "small":  $\lambda = np$  is the expected value of the number of successful experiments;
- the probability that exactly  $k$  experiments are successful, can be approximated with  $\frac{\lambda^k}{k!} e^{-\lambda}$  (based on the calculation and the figure above);
- this is the Poisson distribution, which is often used for modelling the number of **rare events**.

### Definition

Let  $\lambda > 0$ . The random variable  $X$  has **Poisson distribution with parameter  $\lambda$** , if its possible values are:

$$k = 0, 1, 2, \dots, \text{ and } \mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

The **expected value** and **standard deviation** of  $X$  in this case:

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

## Poisson distribution: example

The random variable  $X$  has **Poisson distribution with parameter  $\lambda$** , if its possible values are:

$$k = 0, 1, 2, \dots, \text{ and } \mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

The **expected value** and **standard deviation** of  $X$  in this case:

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

**Example.** Suppose that the number of accidents in a city on a day has Poisson distribution and its **expected value is 3.61**. Then **standard deviation** of the number of accidents on a day:

$$D(X) = \sqrt{3.61} = 1.9.$$

The probability that there will be exactly 5 accidents on a day:

$$\mathbb{P}(X = 5) = \frac{3.61^5}{5!} e^{-5} = 14\%.$$

## Binomial and Poisson distribution in R

- if  $X$  has binomial distribution with order  $n$  and parameter  $p$ , then

$$\mathbb{P}(X = k) : \quad \text{dbinom}(k, \text{size} = n, \text{prob} = p)$$

és

$$\mathbb{P}(X \leq k) : \quad \text{pbinom}(k, \text{size} = n, \text{prob} = p).$$

- In addition,

```
sample<-rbinom(r, size=n, prob=p)
```

constructs a vector "sample", consisting of  $r$  *independent* copies of this random variable.

- if  $X$  has Poisson distribution with parameter  $\lambda$ , then

$$\mathbb{P}(X = k) : \quad \text{dpois}(k, \text{lambda} = \lambda)$$

and

$$\mathbb{P}(X \leq k) : \quad \text{ppois}(k, \text{lambda} = \lambda).$$

- In addition,

```
sample<-rpois(r, lambda= $\lambda$ )
```

constructs a vector "sample", consisting of  $r$  *independent* copies of the Poisson( $\lambda$ ) random variable.

## Example in R

```
> sample=rbinom(100, size=100000, prob=0.0001)
```

```
> hist(sample)
```

```
> dbinom(8, size=100000, prob=0.0001)
```

```
[1] 0.1126013
```

```
> sample=rpois(100, lambda=3.61)
```

```
> hist(sample)
```

```
> mean(sample)
```

```
[1] 3.91
```

```
> dpois(5, lambda=3.61)
```

```
[1] 0.1382139
```

# Hypergeometric distribution

Among  $N = 20$  **members** of a handball team  $M = 9$  **are left-handed**.

There are  $n = 7$  **different** players in the game at the same time.

Suppose that every group of 7 players has the same probability to be chosen.

What is the distribution of the **number of left-handed players,  $X$** ?

# Hypergeometric distribution

Among  $N = 20$  **members** of a handball team  $M = 9$  **are left-handed**.

There are  $n = 7$  **different** players in the game at the same time.

Suppose that every group of 7 players has the same probability to be chosen.

What is the distribution of the **number of left-handed players,  $X$** ?

Number of groups of 7 players:

choosing  $k$  left-handed players:

choosing  $7 - k$  right-handed players:

Number of good outcomes:

# Hypergeometric distribution

Among  $N = 20$  **members** of a handball team  $M = 9$  **are left-handed**.

There are  $n = 7$  **different** players in the game at the same time.

Suppose that every group of 7 players has the same probability to be chosen.

What is the distribution of the **number of left-handed players,  $X$** ?

Number of groups of 7 players:  $\binom{20}{7}$

choosing  $k$  left-handed players:  $\binom{9}{k}$

choosing  $7 - k$  right-handed players:  $\binom{11}{7-k}$

Number of good outcomes:  $\binom{9}{k} \cdot \binom{11}{7-k}$  ←

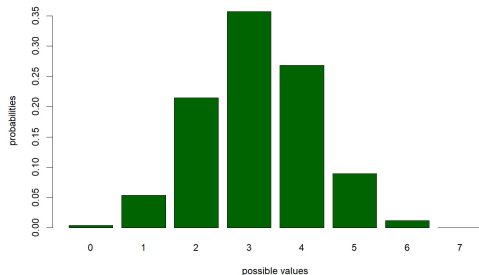
**multiplication:**  
every choice of the left-handed works with every choice of the right-handed

**division:** each group is chosen with the same probability

$$\mathbb{P}(X = k) = \frac{\binom{9}{k} \cdot \binom{11}{7-k}}{\binom{20}{7}}$$

sampling without replacement

# Hypergeometric distribution



The number of left-handed players in the team has hypergeometric distribution, with parameters  $N = 20$ ,  $M = 9$ ,  $n = 7$

x-axis:  $k$ , height of columns:  $\mathbb{P}(X = k) = \frac{\binom{9}{k} \cdot \binom{11}{7-k}}{\binom{20}{7}}$ .

# Hypergeometric distribution

Let  $N, M, n$  be positive integers such that  $1 \leq n \leq M \leq N$ . The random variable  $X$  has **hypergeometric distribution**, if

$$\mathbb{P}(X = k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} \quad (k = 0, 1, \dots, n).$$

- **sampling without replacement**:  $N$  balls, among which  $M$  are red, we draw  $n$  times without replacement and consider the number of chosen red balls

## Expected value and standard deviation of hypergeometric distribution

If  $X$  has hypergeometric distribution with parameters  $M, N, n$ , that is,

$$\mathbb{P}(X = k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} \quad (k = 0, 1, \dots, n),$$

then

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

For example, if among  $N = 20$  players,  $M = 9$  are left-handed, and we choose  $n = 7$  different players randomly, then, the number of left-handed players in the sample,  $X$  has **expected value** and **standard deviation**:

$$\mathbb{E}(X) = \frac{9}{20} \cdot 7 = 3.15; \quad D(X) = \sqrt{7 \cdot \frac{9}{20} \cdot \left(1 - \frac{9}{20}\right) \cdot \frac{13}{19}} = 1.09.$$

# Geometric distribution

In a survey, everyone answers one particular question with probability 0.2 independently. Let  $Y$  denote the number of people that we have to ask until we find someone who answers this question.

$$\mathbb{P}(Y = 1) = \mathbb{P}(\text{first person answers}) = 0.2;$$

$$\mathbb{P}(Y = 2) = \mathbb{P}(\text{first does not answers, second one does}) =$$

# Geometric distribution

In a survey, everyone answers one particular question with probability 0.2 independently. Let  $Y$  denote the number of people that we have to ask until we find someone who answers this question.

$$\mathbb{P}(Y = 1) = \mathbb{P}(\text{first person answers}) = 0.2;$$

$$\mathbb{P}(Y = 2) = \mathbb{P}(\text{first does not answers, second one does}) = 0.8 \cdot 0.2;$$

$$\mathbb{P}(Y = 3) = \mathbb{P}(\text{first two do not answer, third one does}) =$$

# Geometric distribution

In a survey, everyone answers one particular question with probability 0.2 independently. Let  $Y$  denote the number of people that we have to ask until we find someone who answers this question.

$$\mathbb{P}(Y = 1) = \mathbb{P}(\text{first person answers}) = 0.2;$$

$$\mathbb{P}(Y = 2) = \mathbb{P}(\text{first does not answers, second one does}) = 0.8 \cdot 0.2;$$

$$\mathbb{P}(Y = 3) = \mathbb{P}(\text{first two do not answer, third one does}) = 0.8^2 \cdot 0.2;$$

$$\mathbb{P}(Y = k) = \mathbb{P}(\text{first } k - 1 \text{ do not answer, the } k\text{th one does}) =$$

# Geometric distribution

In a survey, everyone answers one particular question with probability 0.2 independently. Let  $Y$  denote the number of people that we have to ask until we find someone who answers this question.

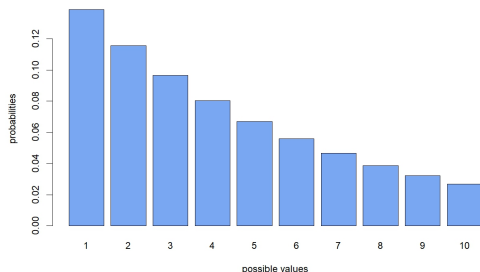
$$\mathbb{P}(Y = 1) = \mathbb{P}(\text{first person answers}) = 0.2;$$

$$\mathbb{P}(Y = 2) = \mathbb{P}(\text{first does not answer, second one does}) = 0.8 \cdot 0.2;$$

$$\mathbb{P}(Y = 3) = \mathbb{P}(\text{first two do not answer, third one does}) = 0.8^2 \cdot 0.2;$$

$$\mathbb{P}(Y = k) = \mathbb{P}(\text{first } k - 1 \text{ do not answer, the } k\text{th one does}) = 0.8^{k-1} \cdot 0.2.$$

## Example: geometric distribution



Distribution of the first 6 with a fair die: geometric distribution with  $p = 1/6$ , until  $k = 10$

# Geometric distribution

- independent experiments;
- each of them is successful with probability  $p$  independently;
- $Y$ : which one is the first successful experiment.

# Geometric distribution

- independent experiments;
- each of them is successful with probability  $p$  independently;
- $Y$ : which one is the first successful experiment.

## Definition

The random variable  $Y$  has **geometric distribution** with parameter  $p$ , if its possible values are:

$$1, 2, 3 \dots$$

and for every integer  $1 \leq k$  we have

$$\mathbb{P}(Y = k) = (1 - p)^{k-1} p.$$

( $0 < p < 1$ .) Notation:  $\text{Geo}(p)$ . Other name: Pascal distribution.

Since  $\sum_{k=1}^{\infty} (1 - p)^{k-1} p = \frac{1}{1 - (1 - p)} \cdot p = 1$ , this is indeed a probability distribution; the probability that the experiment is never successful is 0.

# Expected value and standard deviation of geometric distribution

## Proposition

If the random variable  $X$  has geometric distribution with parameter  $p$ , that is,

$$\mathbb{P}(X = k) = (1 - p)^{k-1}p \quad (k = 1, 2, \dots),$$

then

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

## Expected value and standard deviation of geometric distribution

### Proposition

If the random variable  $X$  has geometric distribution with parameter  $p$ , that is,

$$\mathbb{P}(X = k) = (1 - p)^{k-1}p \quad (k = 1, 2, \dots),$$

then

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

**Example.** Suppose that a given political party is supported by each citizen with probability  $p = 0.06$  independently. Let  $X$  be the number of people that we have to ask to find someone who supports this party. Then

$$\mathbb{E}(X) = \frac{1}{0.06} = 16.67; \quad D(X) = \sqrt{\frac{0.94}{0.06^2}} = 16.16.$$

## Expected value and standard deviation of common discrete distributions

- 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

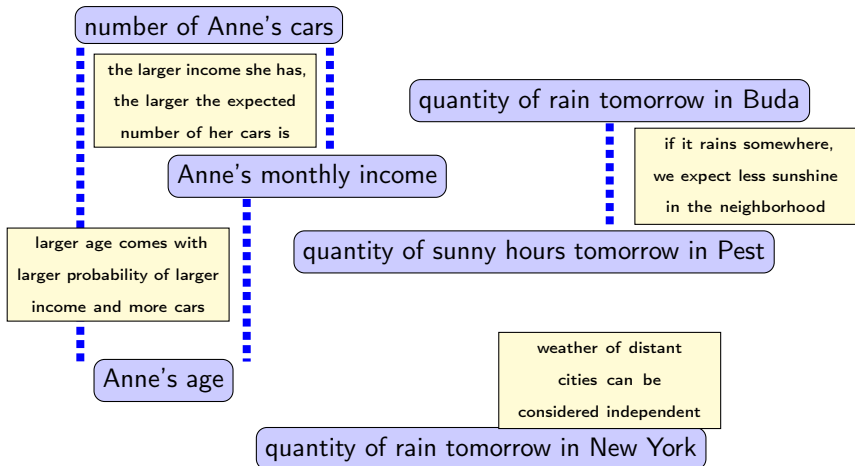
quantity of sunny hours tomorrow in Pest

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.**