Statistics

Lecture 6

Discrete probability distributions

SILESIAN UNIVERSITY
SCHOOL OF BUSINESS ADMINISTRATION IN KARVINA

David Bartl
Statistics
INM/BASTA

Outline of the lecture

- Discrete probability distributions
- Discrete uniform distribution
- Bernoulli Trials
- Binomial distribution
- Poisson distribution
- Some other discrete probability distributions

Experiment — Trial — Random variable

An **experiment** is any physical procedure which can end up with a result from a set of possible outcomes, and the experiment can be repeated (up to) infinitely many times. Each individual repetition of the experiment is called a **trial.**

The final result ω of (each trial of) the experiment is called an **outcome.**

The set Ω of all the outcomes is called the **sample space.** An event E is a set of outcomes (i.e. $E \subseteq \Omega$). The **event space** $\mathcal{F}$ is the collection of all events.

The event space is a σ-algebra, the **probability** $P\colon \mathcal{F} \to \mathbb{R}$ is a non-negative and σ-additive function. The triple $(\Omega, \mathcal{F}, P)$ is a **probability space** and a **random variable** is any measurable function $X\colon \Omega \to \mathbb{R}$.

Random variable — Dataset

To conclude, **the random variable assigns a numerical value to each outcome** of the random experiment.

Now, a **dataset** is a collection of measurements and observations, i.e. it is a collection of data. A **data unit** is an entity of the population under study, and a <u>**data item**</u> or a <u>**variable**</u> is a characteristics of each data unit.

We are considering numerical (quantitative) variables now.

<u>**We assume the hypothesis**</u> that the data items in the dataset are realizations of the random variable, i.e. **the random variable** (via the trials of the random experiment) **generates the data**.

Examples of discrete random variables

- There are 100 rooms in some hotel. The number of rooms booked on the 1st of July is a random variable X whose value is $X \in \{0, 1, 2, \dots, 100\}$.

- The number of customers in a supermarket between 12 and 18 o'clock is a random variable X which (in theory) can attain any non-negative integer value $X \in \{0, 1, 2, 3, \dots\}$.

- The difference between the number of customers in two supermarkets (e.g. Kaufland and Tesco) during a day is a random variable X which (in theory)

Discrete probability distributions

The purpose of this lecture, however, is to present the most important, yet elementary, discrete probability distributions.

We shall present:

- the uniform distribution
- Bernoulli's experiment
- the binomial distribution
- Poisson's distribution

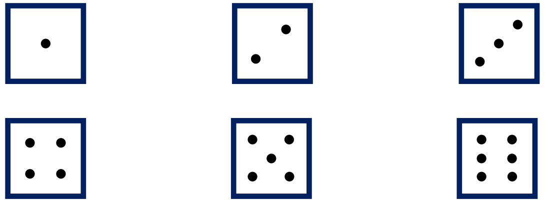
Discrete uniform distribution

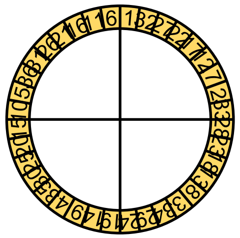
Uniform distribution (discrete)

The discrete uniform distribution relates closely to the classical definition of probability: Considering N distinct outcomes of a random experiment, each of the N outcomes is equally likely to happen.

The classic examples experiments with uniform distribution include:

- tossing a fair coin “heads” or “tails”
- rolling a fair dice
- playing a fair roulette





Uniform distribution (discrete)

Consider a probability space $(\Omega, \mathcal{F}, P)$ where the sample space $\Omega = \{1, 2, \dots, N\}$ is finite, the event space $\mathcal{F} = 2^{\Omega} = \{E : E \subseteq \Omega\}$ consists of all subsets of the sample space Ω, and the probability P is given by its probability mass function p which is

$$p(\omega) = \frac{1}{N} \qquad \text{for every} \quad \omega \in \Omega$$

Then the identity random variable $X : \Omega \to \mathbb{R}$

$$X(\omega) = \omega \qquad \text{for} \quad \omega = 1, 2, \dots, N \in \Omega$$

follows the discrete uniform distribution.

Uniform distribution (discrete)

More generally, let $a, b \in \mathbb{R}$ be real numbers such that $a < b$. Then the random variable $X: \Omega \to \mathbb{R}$ such that

$$X(\omega) = \frac{b-a}{N-1}\omega + \frac{Na-b}{N-1} \qquad \text{for} \quad \omega = 1, 2, \dots, N \in \Omega$$

follows the discrete uniform distribution.

Note that the distribution of the identity random variable $(X(\omega) = \omega)$ above is a special case of this distribution for $a = 1$ and $b = N$.

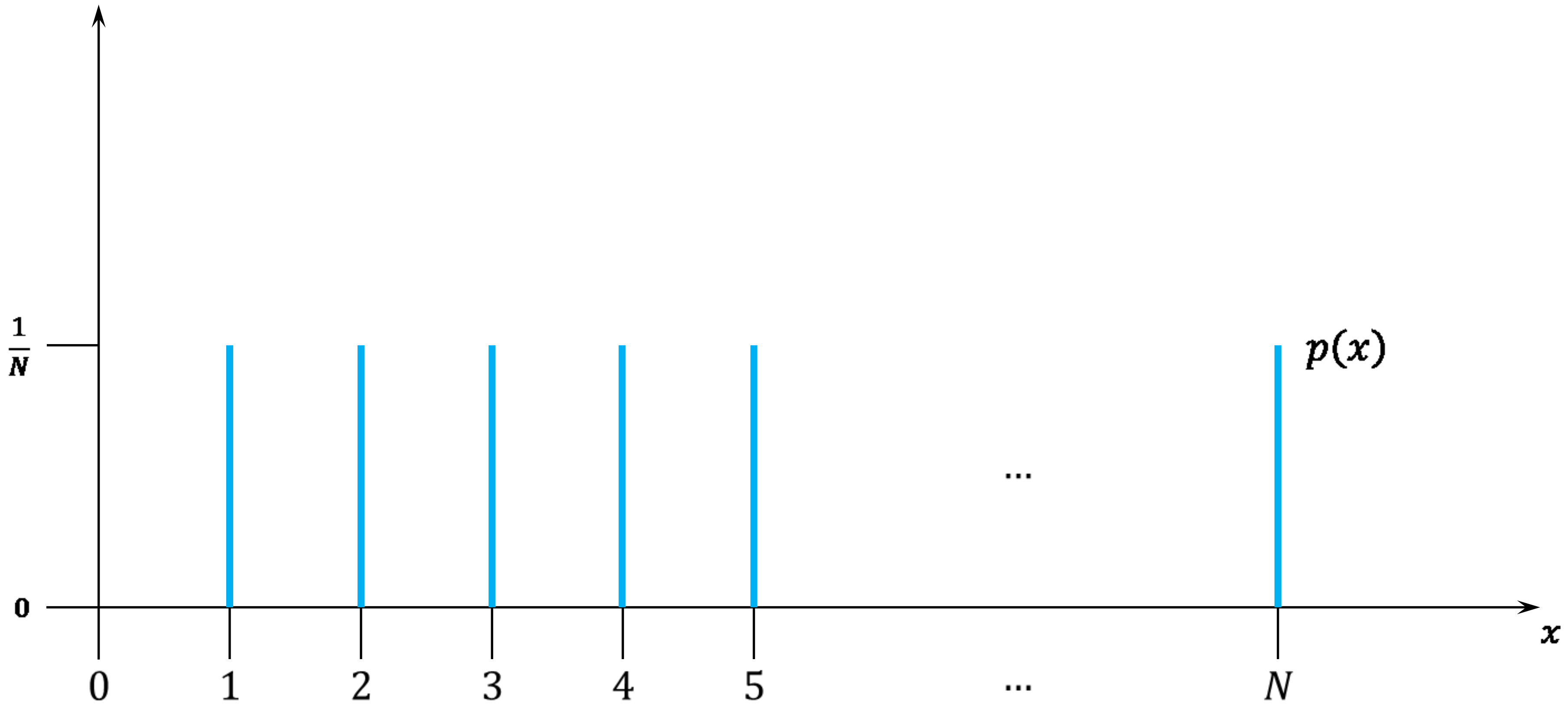
We then say that X is a **discrete uniform random variable** and write

$$X \sim \mathrm{Unif}(N) \qquad \text{or} \qquad X \sim \mathrm{Unif}(N, a, b) \quad \text{in general}$$

Uniform distribution (discrete)

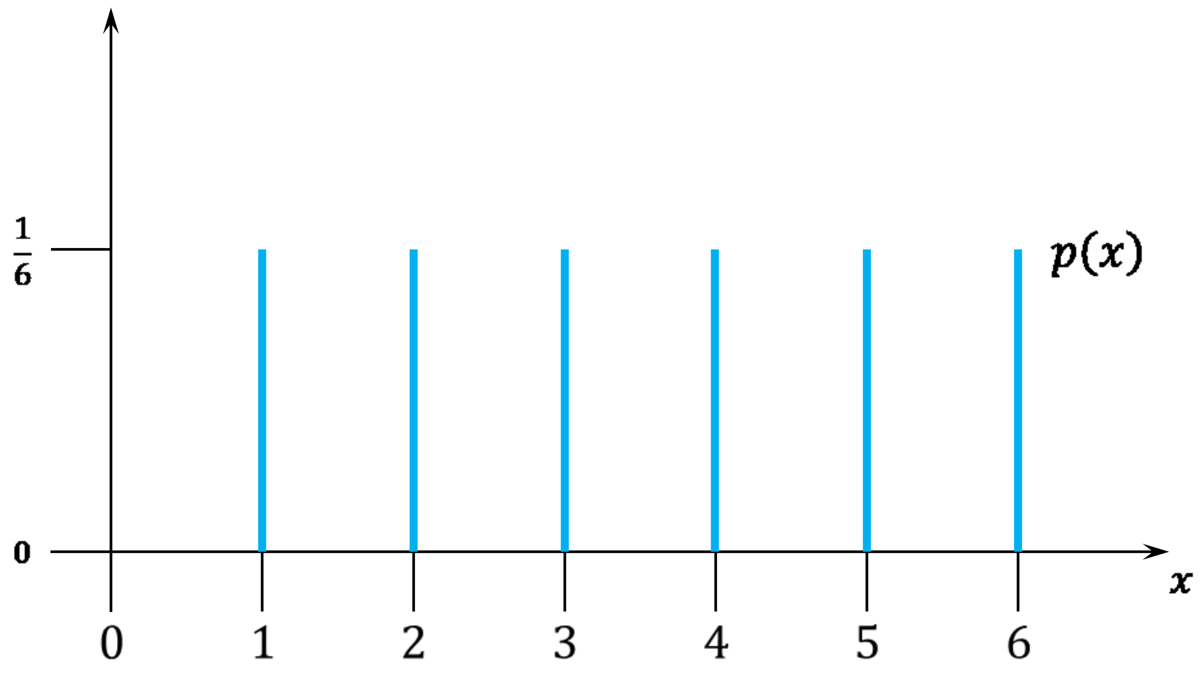
The **graph** of the **probability mass function**

of a discrete uniform random variable $X \sim \mathrm{Unif}(N)$:



Uniform distribution (discrete)

The **graph** of the **probability mass function** of the result of rolling a dice (which is a discrete uniform random variable $X \sim \text{Unif}(6)$):



Uniform distribution (discrete)

Let the random variable X follow the discrete uniform distribution, i.e. $X \sim \mathrm{Unif}(N)$.

<u>Calculate as an exercise:</u>

- Mean value: $\mu = \mathrm{E}[X] = \frac{N+1}{2}$
- Variance: $\sigma^2 = \mathrm{Var}(X) = \frac{N^2-1}{12}$
- Mode: $\hat{X}$ is any element $\omega = 1, 2, \dots, N \in \Omega$
- Median: $\tilde{X} = \frac{N+1}{2}$

An application: the German Tank Problem

A sample of k elements $x_1, x_2, \dots, x_k$ is selected without repetition out of the set $\Omega = \{1, 2, \dots, N\}$. That is, $x_1 \in \Omega$, $x_2 \in \Omega \setminus \{x_1\}$, etc., $x_k \in \Omega \setminus \{x_1, \dots, x_{k-1}\}$. When selecting an element x_{i+1} each element of the set $\Omega \setminus \{x_1 \dots, x_i\}$ is equally probable to be selected (i.e. the uniform distribution is assumed).

The number N exists, but is not known, however.

The goal: Based on the sample $x_1, x_2, \dots, x_k$, find an estimate of the number N.

— let $m = \max\{x_1, x_2, \dots, x_k\}$ be the maximum in the sample

— the estimate is $N^{\approx} = \frac{k+1}{k} m - 1 = m + \frac{m}{k} - 1$

Binomial distribution

- Bernoulli Trials
- Binomial distribution

Bernoulli Trials

A Bernoulli trial is a random experiment with exactly two possible outcomes: "success" and "failure". Each repetition of the experiment is called a trial.

The probability p of the success is given in advance and is the same in each trial. (In particular, the probability of the success in a trial does not depend on the results in the previous trials.) The probability of the failure is then $q = 1 - p$.

We have $p, q \geq 0$ and $p + q = 1$.

Examples of Bernoulli Trials

Tossing a coin:

- outcomes: “heads” (success) or “tails” (failure)
- probabilities: $p = \frac{1}{2}$ and $q = \frac{1}{2}$

Rolling a dice:

- outcomes: “6” (success) or “not 6” (failure)
- probabilities: $p = \frac{1}{6}$ and $q = \frac{5}{6}$

Mathematical model of the Bernoulli Trial

To model the Bernoulli trial mathematically, consider a probability space $(\Omega, \mathcal{F}, P)$ with $\Omega = \{0, 1\}$ (where "1" and "0" means "success" and "failure", respectively), with $\mathcal{F} = \{\emptyset, \{0\}, \{1\}, \Omega\}$ and with the probability P such that $P(\emptyset) = 0$, $P(\Omega) = 1$, and

$$P(\{1\}) = p$$

$$P(\{0\}) = q$$

Binomial distribution

Assume that the Bernoulli trial, with the probability of the success being p, is repeated n times where n is a given natural number.

Let X be the random variable whose value is the number of successes in the series of the n Bernoulli trials.

Obviously, the range of the random variable X is the set $\{0, 1, 2, 3, \dots, n\}$.

The probability that X attains the value k is

$$\binom{n}{k} p^k q^{n-k} \qquad \text{for} \quad k = 0, 1, 2, 3, \dots, n$$

Binomial distribution

Formally, let p be the probability of the success in the Bernoulli trial, let $q = 1 - p$, and let n be the given natural number. Consider the probability space $(\Omega, \mathcal{F}, P)$ where $\Omega = \{0,1\}^n$ is the collection of all ordered n-tuples of 0's and 1's, moreover $\mathcal{F} = 2^\Omega$, and the probability mass function $\bar{p}$ of the probability P is such that

$$\bar{p}(\omega) = \binom{n}{k} p^k q^{n-k} \qquad \text{for every } \omega \in \Omega \text{ such that}$$

there are exactly k 1's in the ω

for $k = 0, 1, 2, 3, \ldots, n$

Binomial distribution

Considering the just introduced probability space $(\Omega, \mathcal{F}, P)$ with $\Omega = \{0,1\}^n$ etc., the random variable $X: \Omega \to \mathbb{R}$ such that

$$X(\omega) = \text{the number of } 1\text{'s in the } \omega \qquad \text{for every} \quad \omega \in \Omega$$

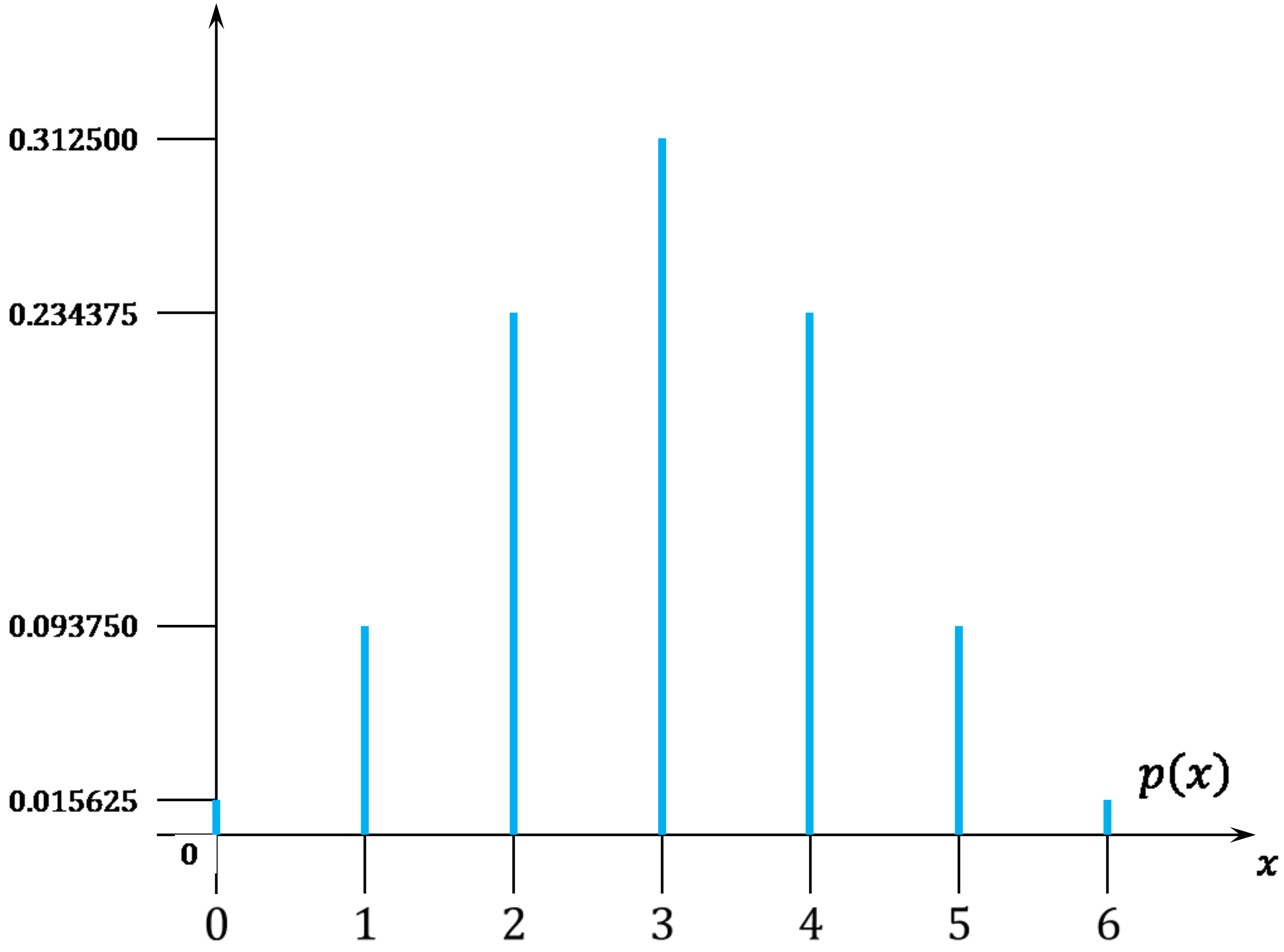
follows the binomial distribution.

We then say that X is a **discrete binomial random variable** and write

$$X \sim \mathrm{Bi}(n, p)$$

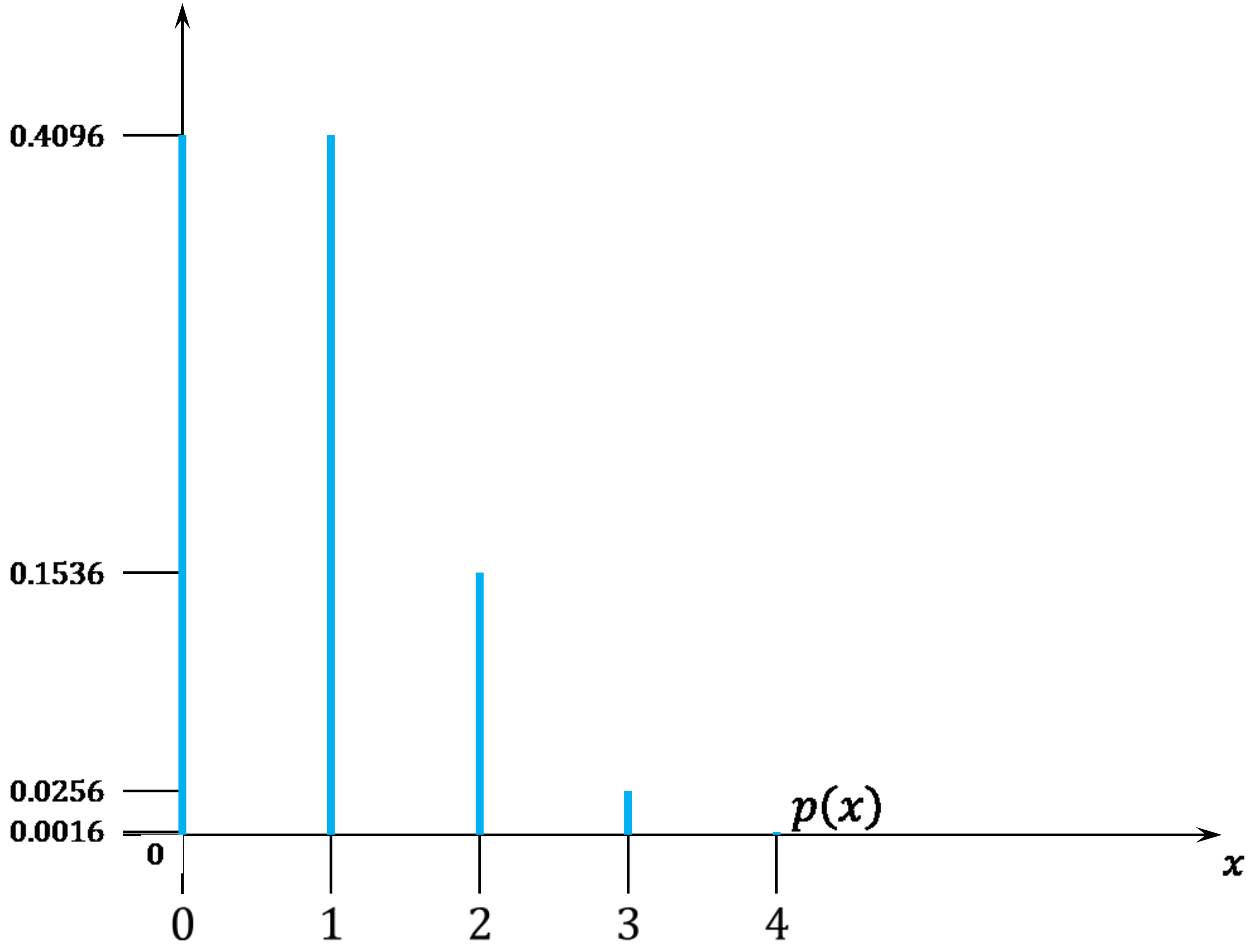
Binomial distribution

The graph of the probability mass function of an $X \sim \mathrm{Bi}(6, 0.5)$:



Binomial distribution

The **graph** of the **probability mass function** of an $X \sim \mathrm{Bi}(4, 0.2)$:



Binomial distribution

Let the random variable $X \sim \mathrm{Bi}(n, p)$.

<u>Calculate as an exercise:</u>

- Mean value: $\mu = \mathrm{E}[X] = np$
- Variance: $\sigma^2 = \mathrm{Var}(X) = np(1-p)$
- Mode: $\hat{X}$

Binomial distribution in Excel

In Excel, use the functions:

=BINOM.DIST$(n; k; p; \mathrm{TRUE})$	to get the value of the cumulative distribution function of the random variable $X \sim \mathrm{Bi}(n, p)$
=BINOM.DIST$(n; k; p; \mathrm{FALSE})$	to get the value of the probability mass function of the random variable $X \sim \mathrm{Bi}(n, p)$
=BINOM.INV$(n; p; \alpha)$	to get the quantile of the random variable $X \sim \mathrm{Bi}(n; p)$ $(0 < \alpha \leq 1)$

Binomial distribution in Excel

In Excel, use the functions:

=BINOM.DIST.RANGE($n; p; k_1; k_2$)

to get the probability that the result of the random variable $X \sim \mathrm{Bi}(n, p)$ is between k_1 and k_2

=BINOMDIST()

the same as =BINOM.DIST(), deprecated

Poisson distribution

Poisson distribution

There are some events, such as

— customers coming to a shop during one hour (between 10:00 and 11:00, say)

— telephone calls incoming during one hour (between 10:00 and 11:00, say)

— requests incoming to a server during one minute (between 10:00 and 10:01)

— meteorites of diameter ≥ 1 meter hitting the Earth during a year

— decay events from a radioactive source

that (as we suppose) have some properties in common.

Poisson distribution

Suppose that a random event occurs repeatedly and satisfies the following assumptions:

- the event can occur at any time
- the average number of occurrences of the event during an interval of time of a fixed length is constant; the number does not depend on the beginning of the interval, and does not depend on the number of occurrences of the event before the beginning of the time interval
- the average number of occurrences of the event during an interval of time is proportional to the length of the interval
- …

Poisson distribution

Suppose that a random event occurs repeatedly and satisfies the following assumptions:

- …
- if the length of the interval is very small, then there is no more than one occurrence of the event in the interval;

  in other words, denoting by $p_t^{\geq 2}$ the probability that the event occurs at least two times during a time interval of length $t > 0$, it holds $p_t^{\geq 2} \times t \longrightarrow 0$ as $t \to 0$

Poisson distribution

Now, consider a time interval of length $t > 0$, where the length t is fixed.

Divide the interval into n subintervals of the length t/n.

Let $p^1_{t/n}$ and $p^{\geq 2}_{t/n}$ denote the probability that the event occurs exactly once and at least two times, respectively, during the interval of the length t/n.

Since we have $p^{\geq 2}_{t/n} \times t/n \longrightarrow 0$ as $n \to \infty$ by our assumptions, it follows that

$$p^1_{t/n} \times n \longrightarrow \lambda \qquad \text{as} \quad n \to \infty$$

where λ is the average number of occurrences of the event during any interval of the given length $t > 0$.

Poisson distribution

Choose a non-negative natural number $k \in \mathbb{N}_0$. Consider a sufficiently large natural number $n \in \mathbb{N}$ so that $p^{\geq 2}_{t/n} \times t/n$ is near zero (and $n \geq k$).

Now, letting the probability of the success be $p^1_{t/n}$, repeat the Bernoulli trial n times. Then the probability that the success occurs exactly k times is

$$\binom{n}{k}\left(p^1_{t/n}\right)^k\left(1 - p^1_{t/n}\right)^{n-k}$$

which (approximately) is also the probability that the event occurs exactly k times during the time interval of the given length $t > 0$.

As $n \to \infty$, the above probability converges to the (exact) probability that the event occurs exactly k times during the time interval of the given length $t > 0$.

Poisson's Theorem

Under the above assumptions it holds that

$$\binom{n}{k}\left(p^1_{t/n}\right)^k\left(1-p^1_{t/n}\right)^{n-k} \longrightarrow \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda} \qquad \text{as} \quad n \to \infty$$

which (approximately) is also the probability that the event occurs exactly k times during the time interval of the given length $t > 0$.

To see that, notice that, as $p^1_{t/n} \times n \longrightarrow \lambda$, we have $p^1_{t/n} \approx \frac{\lambda}{n}$.

Then $\binom{n}{k}\left(p^1_{t/n}\right)^k \approx \binom{n}{k}\left(\frac{\lambda}{n}\right)^k = \frac{n(n-1)(n-2)\ldots(n-k+1)}{k!}\frac{\lambda^k}{n^k} \longrightarrow \frac{\lambda^k}{k!}$

and $\left(1-p^1_{t/n}\right)^{n-k} \approx \left(1-\frac{\lambda}{n}\right)^{n-k} \longrightarrow \mathrm{e}^{-\lambda}$.

Poisson distribution

Consider a probability space $(\Omega, \mathcal{F}, P)$ with the sample space $\Omega = \mathbb{N}_0 =$ $= \{0, 1, 2, 3, 4, 5, \dots\}$, the event space $\mathcal{F} = 2^\Omega = \{E : E \subseteq \Omega\}$ consisting of all subsets of the sample space Ω, and the probability P given by its probability mass function p which is

$$p(k) = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda} \qquad \text{for every} \quad k \in \Omega$$

where $\lambda > 0$ is a fixed parameter (the average number of the events occurring during a time interval of the given length).

Poisson distribution

Having the above probability space $(\Omega, \mathcal{F}, P)$ with $\Omega = \mathbb{N}_0 = \{0, 1, 2, 3, \ldots\}$ etc. and

$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda} \qquad \text{for every} \quad k = 0, 1, 2, 3, \ldots \in \Omega$$

then the identity random variable $X: \Omega \to \mathbb{R}$

$$X(k) = k \qquad \text{for} \quad k = 0, 1, 2, 3 \ldots \in \Omega$$

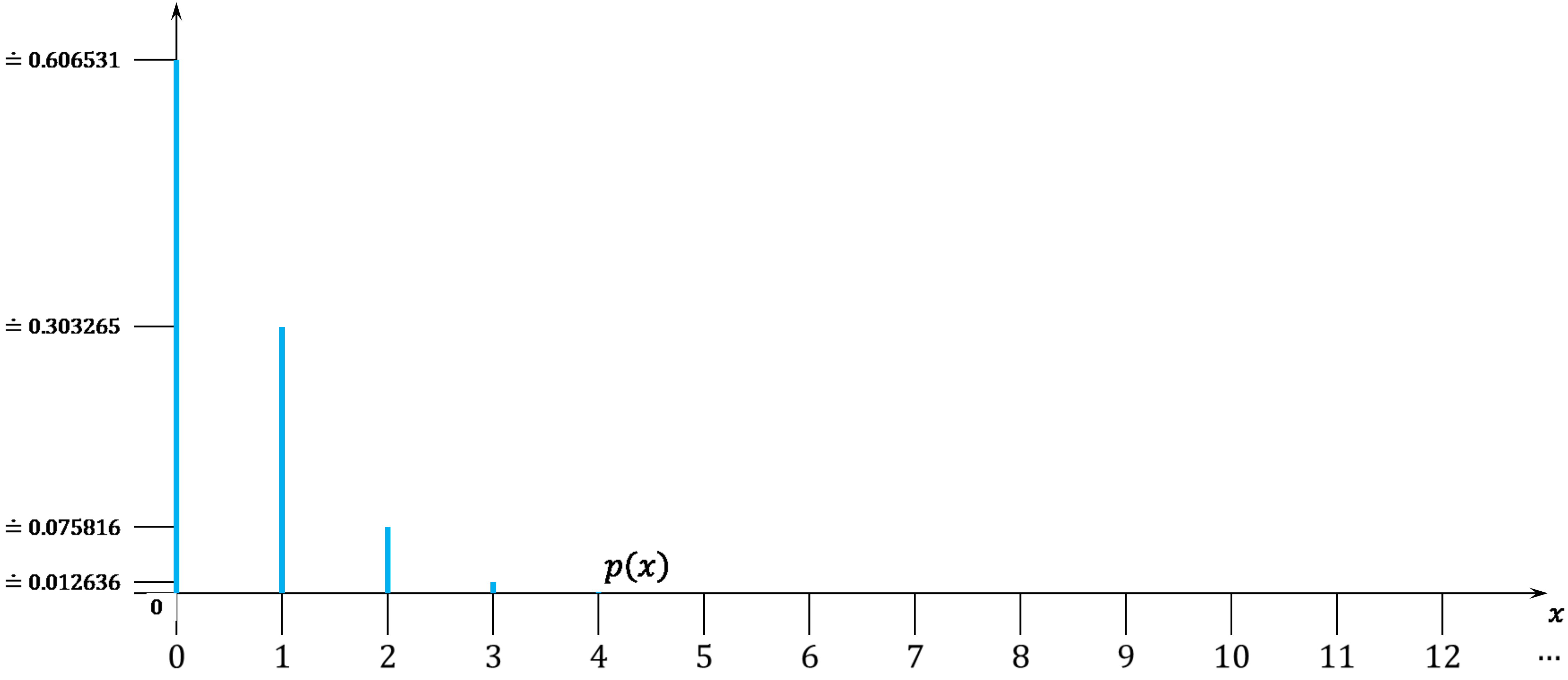
follows the Poisson distribution.

We then say that X is a **Poisson random variable** and write

$$X \sim \mathrm{Po}(\lambda)$$

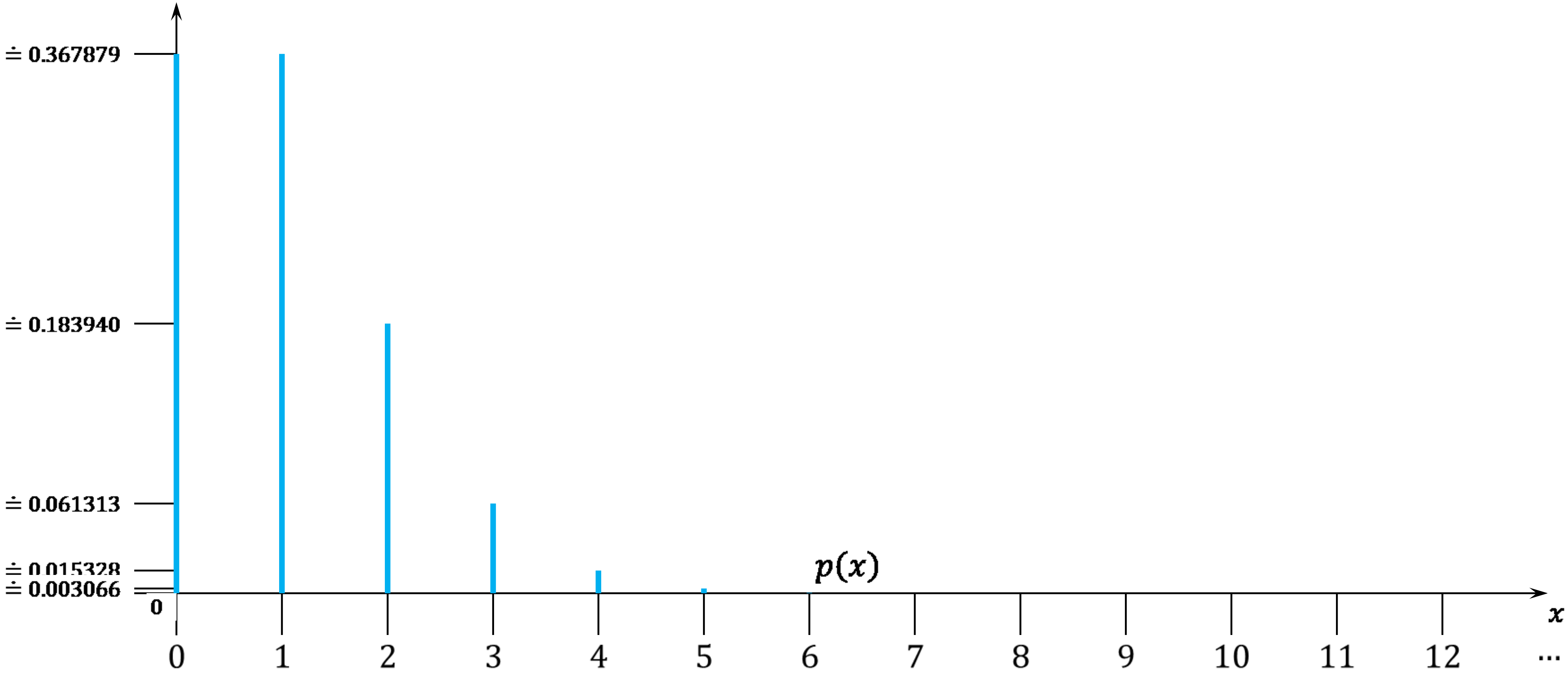
Poisson distribution

The graph of the probability mass function of an $X \sim \mathrm{Po}(0.5)$:



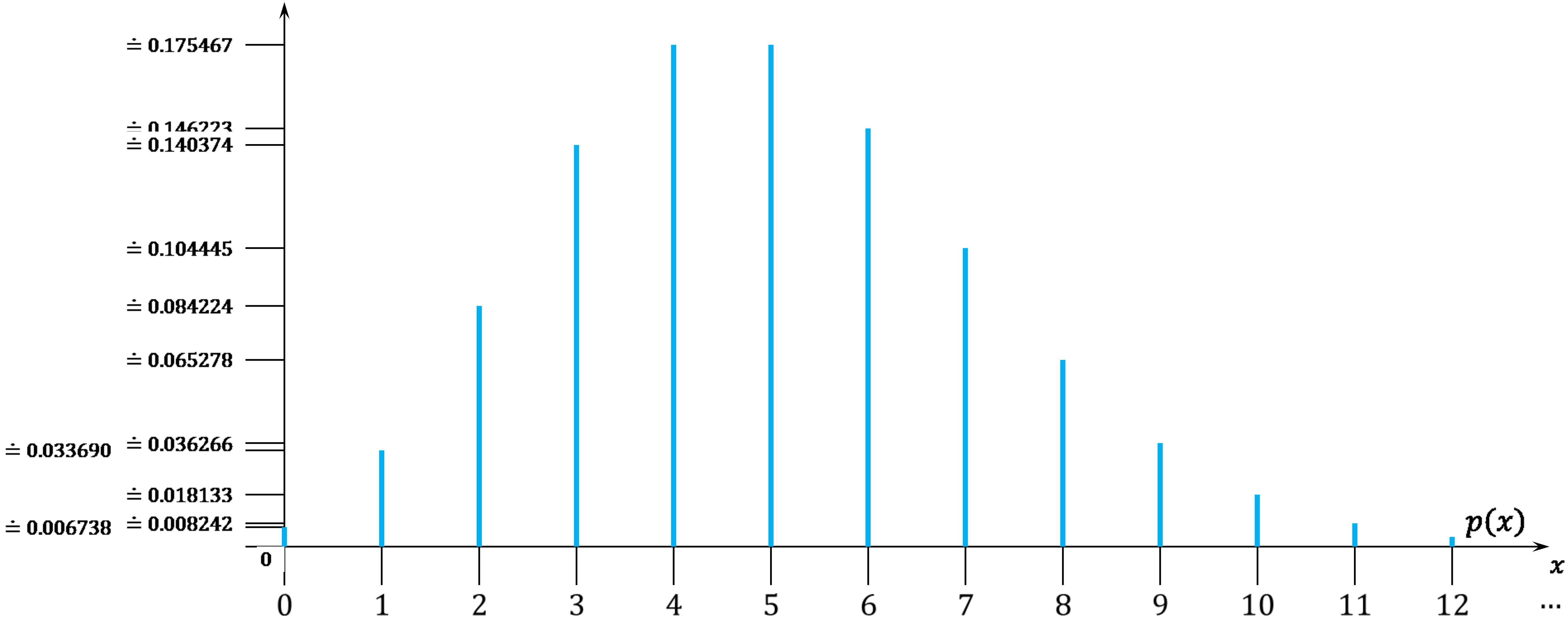
Poisson distribution

The graph of the probability mass function of an $X \sim \mathrm{Po}(1)$:



Poisson distribution

The graph of the probability mass function of an $X \sim \text{Po}(5)$:



Poisson distribution

Let the random variable $X \sim \mathrm{Po}(\lambda)$.

<u>Calculate as an exercise:</u>

- Mean value: $\mu = \mathrm{E}[X] = \lambda$
- Variance: $\sigma^2 = \mathrm{Var}(X) = \lambda$
- Mode: $\hat{X}$

Poisson distribution in Excel

In Excel, use the functions:

=POISSON.DIST$(n; \lambda; \mathrm{TRUE})$ to get the value of the cumulative distribution function of the random variable $X \sim \mathrm{Po}(\lambda)$

=POISSON.DIST$(n; \lambda; \mathrm{FALSE})$ to get the value of the probability mass function of the random variable $X \sim \mathrm{Po}(\lambda)$

=POISSONDIST() the same as =POISSON.DIST(), deprecated

Poisson distribution: Examples

- The number of telephone calls received by a call centre per hour.
- The number of customers coming to the shop per hour.
- The number of radioactive decay events per second from a radioactive source.
- The number of clicks per second of a Geiger-Müller counter.
- The number of defaults per year in risk modelling.
- The number of some failures / accidents / … per year.

Some other discrete probability distributions

- Negative binomial distribution
- Lady tasting the tea
- Hypergeometric distribution

Negative binomial distribution

Assume that the Bernoulli trial, with the probability of the success being p, is repeated until we encounter n successes where n is a given natural number. Let X be the random variable whose value is the number of failures until we encountered n successes in the series of the Bernoulli trials.

The probability that X attains the value k is

$$\binom{k+n-1}{k} p^n (1-p)^k \qquad \text{for} \quad k = 0, 1, 2, 3, \dots$$

Then X is a negative binomial random variable and we write

$$X \sim \mathrm{NB}(n, p)$$

Negative binomial distribution

Let the random variable $X \sim \mathrm{NB}(n, p)$.

<u>Calculate as an exercise:</u>

- Mean value: $\mu = \mathrm{E}[X] = \frac{n(1-p)}{p}$
- Variance: $\sigma^2 = \mathrm{Var}(X) = \frac{n(1-p)}{p^2}$
- Mode: $\hat{X} = \begin{cases} \left\lfloor \frac{(n-1)(1-p)}{p} \right\rfloor, & n = 2, 3, 4, \dots \\ 0, & n = 1 \end{cases}$

Hypergeometric distribution

A Lady Tasting the Tea: We prepare a cup of tea with milk.

There are two ways to prepare the cup:

- pour the tea into the cup first and then add the milk,
- pour the milk into the cup first and then add the tea.

A lady says that she can recognize by the taste of the tea

how the cup was prepared.

Hypergeometric distribution

A Lady Tasting the Tea: We prepare 8 cups of tea with milk.

We prepare:

- 4 cups so that we pour the tea first and then add the milk,
- 4 cups so that we pour the milk first and then add the tea.

We are to choose 4 cups out of the 8 cups. What is the probability – if one selects the cups randomly – that we choose the 4 cups where the tea was first correctly? (I.e., we correctly recognize whether the tea was first in the cup?)

— By the way, the Lady recognized the cups correctly.

Hypergeometric distribution

In general, we have a total of N objects, out of which K objects have some specific feature ("success", say).

Out of the population of the N objects, we are selecting a sample of n objects (without replacement). What is the probability that there are exactly k object with the feature in the sample?

The probability is

$$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Hypergeometric distribution

Consider the probability space $(\Omega, \mathcal{F}, P)$ where the sample space is

$$\Omega = \{\max(0, K + n - N), \ldots, \min(n, K)\}$$

the event space is $\mathcal{F} = 2^{\Omega}$ and the probability P is given by its probability mass function

$$p(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \qquad \text{for} \quad k \in \Omega$$

Then the identity random variable $X: \Omega \to \mathbb{R}$

$$X(k) = k \qquad \text{for} \quad k \in \Omega$$

follows the hypergeometric distribution.

Hypergeometric distribution

Let the random variable X follow the hypergeometric distribution with N, K, n.

<u>Calculate as an exercise:</u>

- Mean value: $\mu = \mathrm{E}[X] = n\frac{K}{N}$
- Variance: $\sigma^2 = \mathrm{Var}(X) = n\frac{K}{N}\left(1 - \frac{K}{N}\right)\frac{N-n}{N-1}$

The binomial coefficient in Excel

In Excel, use the function:

=COMBIN$(n; k)$ to get the value of the binomial coefficient

$$\binom{n}{k}$$