

# Slides

# Probability Theory

## Chapter 4

### **Ch 4: Discrete distributions**

How to compute probabilities or expectations of (integer) **numbers** of occurring events.

# Main concepts in chapters 1-3

- **Stochastic experiment**
- **Sample space  $S$**
- **Probability measure  $P$**  assigns to  $A$  a probability  $P(A)$ .
- **Probability space  $(S, P)$**
- **Conditional probability:**  $P$  is defined on a part  $B$  of  $S$ :  $(S, P(\cdot | B))$  is a probability space as well.

Ch. 4: **discrete random variables** and their distributions, **expectations and variances**.

**Example:  $X =$  “the number of children of a woman in a population”**

$x$	0	1	2	3	<i>total</i>
$P(X = x)$	0.1	0.4	0.3	0.2	1

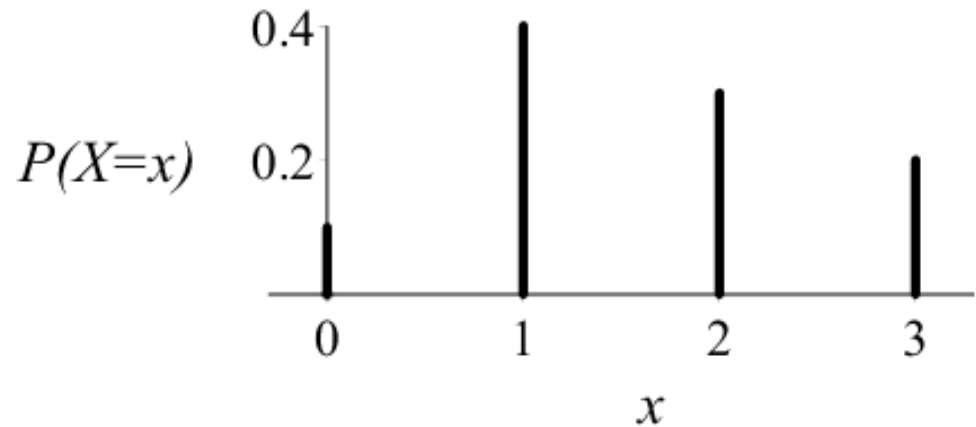
The **graph** of the distribution: a **bar graph** of probabilities.

Verify that:

1.  $P(X = x) \geq 0$
2.  $\sum P(X = x) = 1$

**(1. and 2. should hold for any discrete distribution)**

$X =$  the number of children



**How can you calculate the average number of children?**

Weigh the number of children ( $x$ ) with the probability  $P(X = x)$ :

$$\sum_x x \cdot P(X = x) = 0 \times 0.1 + 1 \times 0.4 + 2 \times 0.3 + 3 \times 0.2 = 1.6 \text{ child/woman}$$

Definition: a **random variable (r.v.)**  $X$  :

a (real) function  $X: S \rightarrow \mathbb{R}$ , for an experiment with sample space  $S$ , that assigns to **each outcome  $s$  a real number  $X(s)$** .

*“A quantitative aspect of the experiment”*

The **range** of  $X$ :  $S_X = \{X(s) \mid s \in S\}$

The r.v.  $X$  is **discrete** if  $S_X$  is **denumerable**,  
so  $S_X = \{x_1, \dots, x_n\}$  or  $S_X = \{x_1, x_2, \dots\}$   
(continuous r.v.: the range is an interval in  $\mathbb{R}$  )

The **event**  $X = x$  is:  $\{s \in S \mid X(s) = x\}$

# The probability function of $X$ :

$$P(X = x), \text{ where } x \in S_X$$

Requirements: 1.  $P(X = x) \geq 0$

$$2. \sum_{x \in S_X} P(X = x) = 1$$

**Distribution of  $X$ :** a list or a formula defining the probabilities  $P(X = x)$  for all  $x$  in  $S_X$

**Expectation or Expected value of  $X$ :**

$$E(X) = \sum_{x \in S_X} x \cdot P(X = x)$$

$E(X)$  is a **measure of centre** of the distribution,  
“the weighted average”

Notation:  $E(X) = EX = \mu = \mu_X$

# Expectation or Expected value of a r.v. $X$ :

$$E(X) = \sum_x x \cdot P(X = x)$$

$E(X)$  is a measure of centre:

- Weighted average
- “Population mean”
- “Point of balance” →

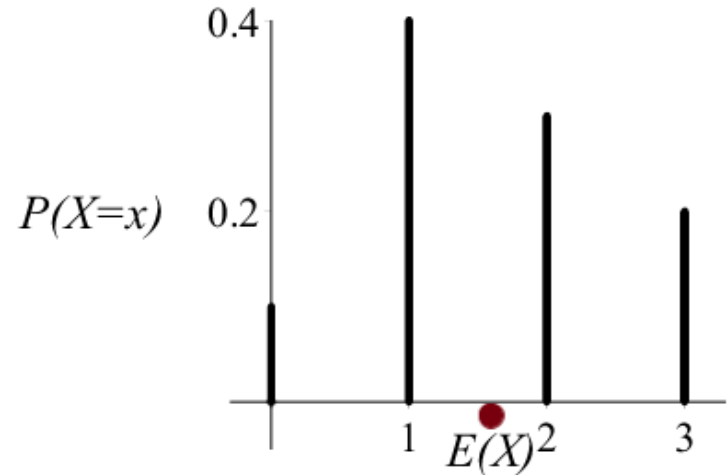
If  $x = \mu$  is the axis of symmetry of  $P(X = x)$ , then  $E(X) = \mu$ .

e.g.:  $X$  = “the result of one roll of a dice”

$$P(X = x) = \frac{1}{6}, x = 1, \dots, 6: P(X=x)$$

$$\mu_X = 3.5$$

$X =$  the number of children



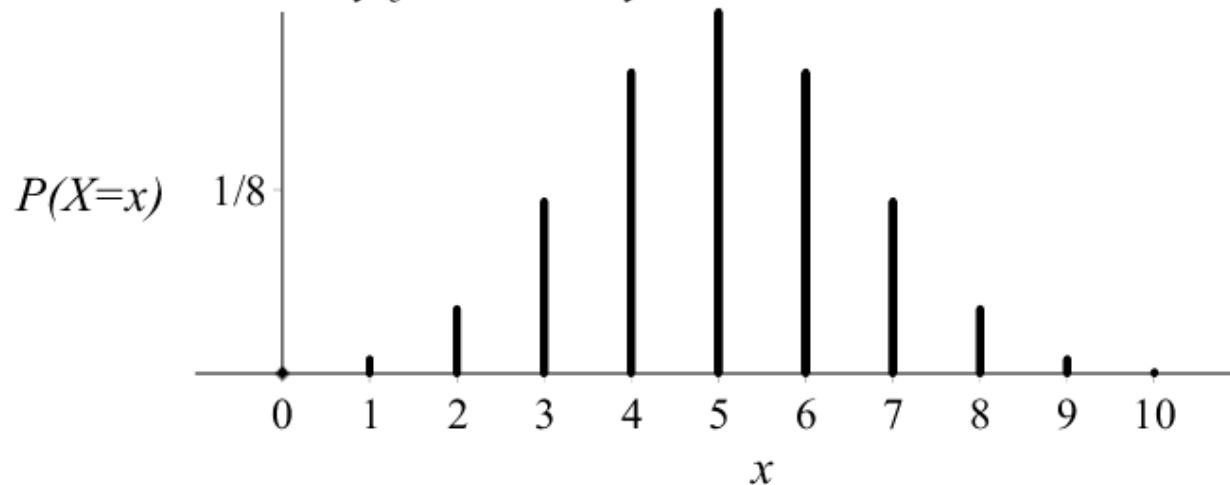
$X =$  Result of a dice roll



# Properties of $E(X)$ , measure of the centre

1. If the probability function of  $X$  is symmetric about  $x = c$ , then  $E(X) = c$

*Probability function symmetric about  $x = 5$*



$$2. E g(X) = \sum_{x \text{ in } S_X} g(x) P(X = x)$$

$$3. E(aX + b) = aE(X) + b \quad (a \text{ and } b: \text{ real numbers})$$

$$4. E[ag(X) + bh(X)] = aEg(X) + bEh(X)$$

# Measure of variation: $\mathit{var}(X)$

The moments:  $E(X^k)$  is the  $k^{\text{th}}$  moment van  $X$

$k = 1$ :  $\mu_X = E(X^1)$  is the first moment

$k = 2$ :  $E(X^2)$  is het 2<sup>nd</sup> moment:  $E(X^2) \neq (EX)^2$

**Measures of variation** of a r.v.  $X$ :

1. The variance of  $X$ :  $\mathit{var}(X) = E(X - \mu_X)^2$

Notation:  $\mathit{var}(X)$  or  $\sigma^2$  (or  $\sigma_X^2$ )

2. The standard deviation  $\sigma_X = \sqrt{\mathit{var}(X)}$



# Properties of $\text{var}(X)$ and $\sigma_X$

1.  $\text{var}(X) \geq 0$  and  $\sigma_X \geq 0$

2.  $\text{var}(X) = E(X^2) - (EX)^2$

*“the calculation formula”*

Alternative notation:  $\sigma^2 = E(X^2) - \mu^2$

3. If  $\text{var}(X) > 0$ , then  $E(X^2) > (EX)^2$ .

If  $\text{var}(X) = 0$ , then  $P(X = EX) = 1$ :

In the latter case  $X$  attains only one value:

$X$  has a **degenerate distribution**.

4.  $\text{var}(aX + b) = a^2 \cdot \text{var}(X)$  and  $\sigma_{aX+b} = |a| \cdot \sigma_X$

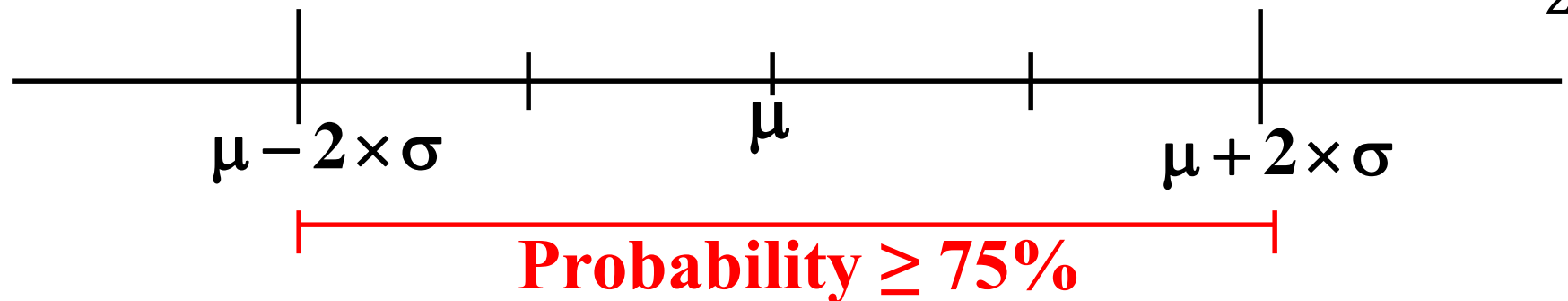
# The meaning of the standard deviation $\sigma$ (1)

For every distribution Chebyshev's inequality holds:

$$\text{for all } c \text{ we have: } P(|X - \mu_X| \geq c) \leq \frac{\text{var}(X)}{c^2}$$

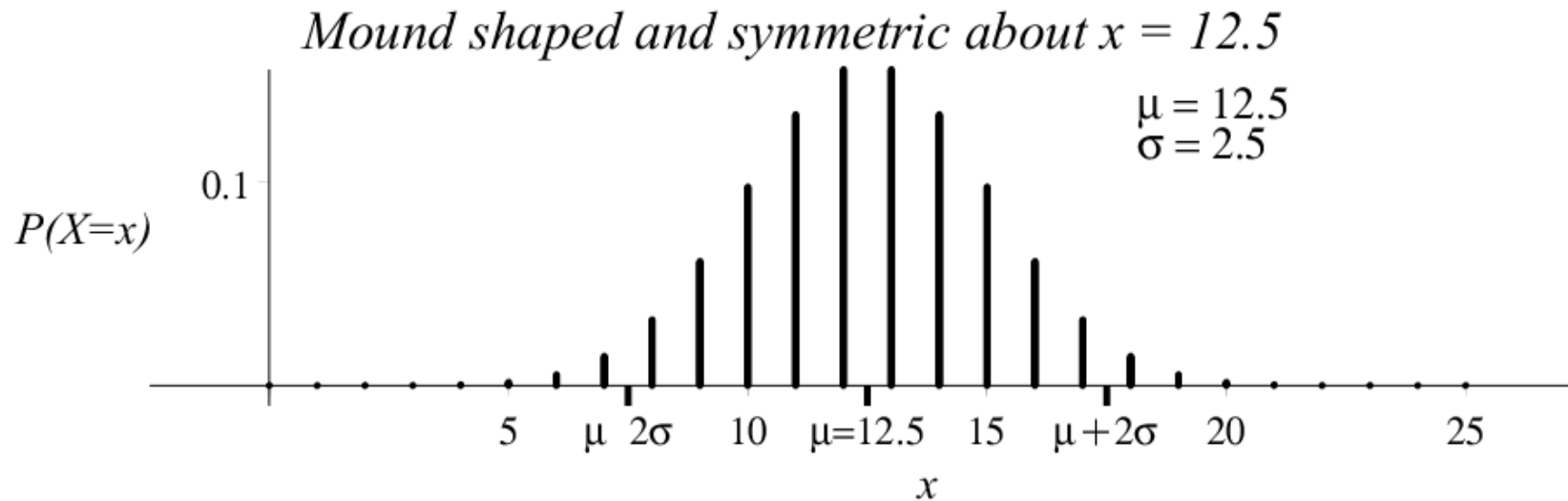
If we choose  $c = k \cdot \sigma$  it follows that  $X$  attains values in  $(\mu - k\sigma, \mu + k\sigma)$  with probability at least  $1 - \frac{1}{k^2}$

**$k = 2$** :  $X$  is in  $(\mu - 2\sigma, \mu + 2\sigma)$  with prob.  $\geq 1 - \frac{1}{2^2}$ .



**$k = 3$** :  $X$  is in  $(\mu - 3\sigma, \mu + 3\sigma)$  with prob. at least  $\frac{8}{9}$

# The meaning of the standard deviation $\sigma$ (2)



The “empirical rule” for  $\sigma$  holds only for (approximately) “mound shaped” distributions. Then the following probabilities hold **approximately**:

Interval	<i>Empirical rule:</i> Proportion within interval	<i>Chebyshev`s rule</i>
$[\mu - \sigma, \mu + \sigma]$	$\approx 68\%$	$\geq 0\%$
$[\mu - 2\sigma, \mu + 2\sigma]$	$\approx 95\%$	$\geq 75\%$
$[\mu - 3\sigma, \mu + 3\sigma]$	$\approx 99.7\%$	$\geq 89\%$

# The binomial distribution – 2 examples

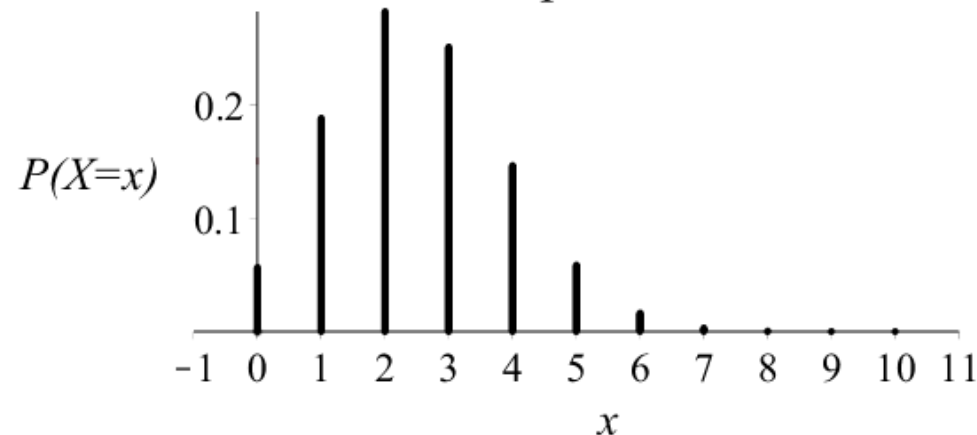
Applicable if we have  $n$  (ind.) Bernoulli trials with success rate  $p$

Example:  $X =$  “the number of correct random answers to 10 MC-items”:

here  $n = 10$  and  $p = \frac{1}{4}$

$$P(X = x) = \binom{10}{x} \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{10-x}, \quad x = 0, \dots, 10$$

$n=10$  and  $p=0.25$



$$E(X) = 10 \cdot \frac{1}{4} = 2.5$$

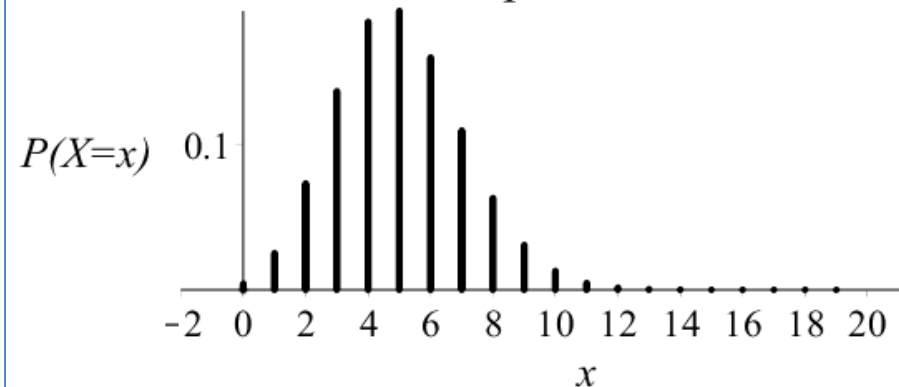
Example:

$X =$  “The number of sixes in 30 rolls of a dice.”

So  $n = 30$  and  $p = \frac{1}{6}$

$$P(X = x) = \binom{30}{x} \left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{30-x}, \quad x = 0, \dots, 30$$

$n = 30$  and  $p = 0.167$



$$E(X) = 30 \cdot \frac{1}{6} = 5$$

# The binomial distribution: $X \sim B(n, p)$

$X$  = “the number of successes in  $n$  Bernoulli trials”

Example: count the number of red balls if we draw, at random and with replacement, 8 balls from a box with 10 red and 15 white balls.

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

Short notation:  $X$  has a  $B(n, p)$ -distribution

or:  $X \sim B(n, p)$

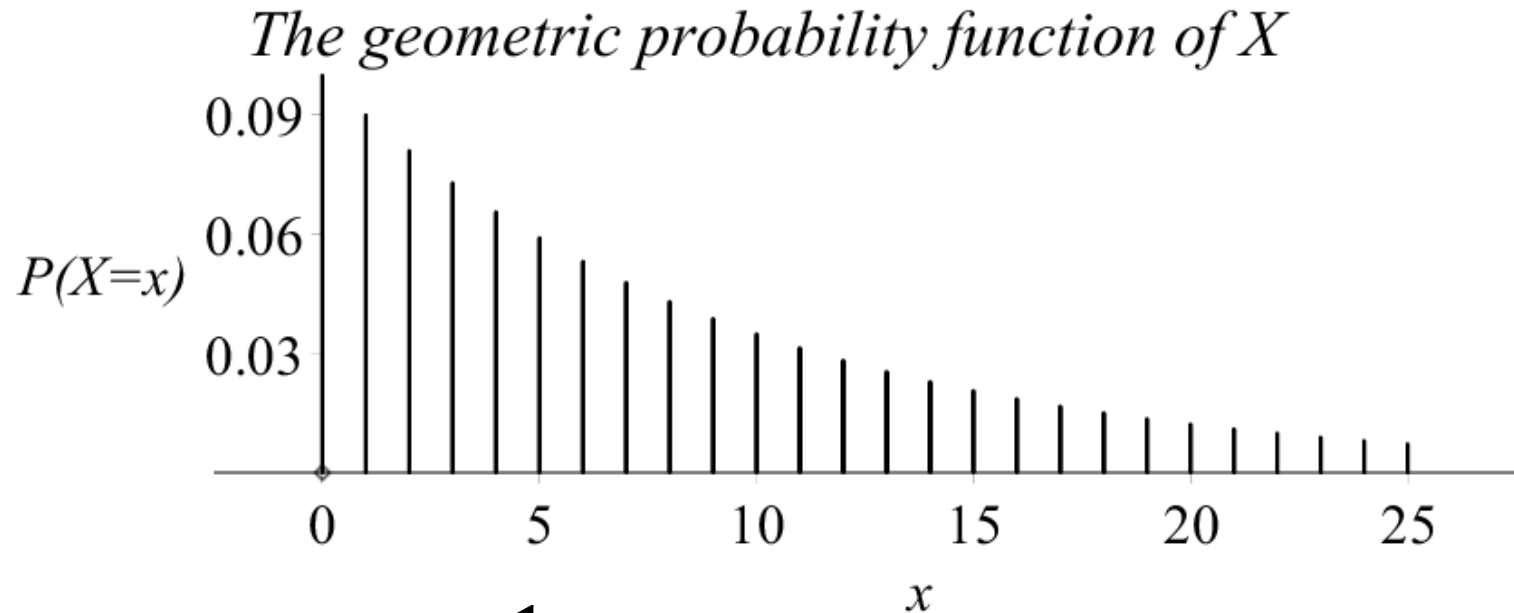
Properties:

$$E(X) = np \quad \text{and} \quad \text{var}(X) = np(1 - p)$$

# Example geometric distribution

Suppose that 10% of the passing cars are Mercedes.  
 $X$  = “the number of the first passing Mercedes”

$$P(X = k) = 0.9^{k-1} \cdot 0.1, \text{ with } k = 1, 2, 3, \dots$$



We have:  $E(X) = \frac{1}{0.1} = 10$

and  $P(X > 10) = 0.9^{10} \approx 34.9\%$

# The geometric distribution

$X =$  “the number of Bernoulli trials until the 1<sup>st</sup> success occurs”

Example:  $X$  is the required number of rolls of a dice to obtain a “6”.

$$P(X = k) = (1 - p)^{k-1}p, \quad k = 1, 2, 3, \dots$$

Properties:

1.  $P(X > k) = (1 - p)^k$ , for  $k = 0, 1, 2, \dots$

2.  $E(X) = \frac{1}{p}$  and  $var(X) = \frac{1-p}{p^2}$

# Example hypergeometric formula: the lotto.

6 draws *without* replacement from balls 1 to 45

Compute the  
Probability of  
4 correct  
(and 2 wrong)

	<b>correct</b>	wrong	total
Population:	<b>6</b>	39	45
	↓	↓	↓
Sample:	<b>4</b>	2	<b>6</b>

$X$  = “the number of draws with one of my 6 choices”

$$P(X = 4) = \frac{\binom{6}{4} \binom{39}{2}}{\binom{45}{6}} \approx 0.14\%$$

$$\text{and } E(X) = 6 \times \frac{6}{45} = 0.8$$



# The hypergeometric distribution

$X$  = “The number of red balls if we draw  $n$  balls, at random and without replacement, from a box with  $R$  red and  $N - R$  white balls”.

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

Properties (success probability is  $p = \frac{R}{N}$ ):

1.  $E(X) = np$  and  $var(X) = np(1 - p)$
2. For large  $N$  ( $N \geq 5n^2$ )  $X$  has an **approximate**  
 **$B(n, p)$ -distribution** with  $p = \frac{R}{N}$

# Examples for the Poisson distribution

- $X =$  “**the number of car accidents** on a busy junction per week”.  $X$  varies weekly, but from police records we know that the average number is 2.5 per week”:  
 $\mu = 2.5$  is the expected number of car accidents per week.
- $X$ , **the number of phone calls** at a helpdesk in 10 minutes time interval, can be modelled as a Poisson distributed variables, where  $\mu$  is the so called **intensity** of the calls.
  - For half an hour the expected number of calls is  $3\mu$ .
- $X =$  “**The number of heart attacks on a day** in a town with 100 000 inhabitants”. If the expected number is known, e.g.  $\mu = 2.5$ , we can use the Poisson distribution to compute probabilities. But we could also model the number  $X$  of heart attacks  $B\left(100000, \frac{2.5}{100000}\right)$ -distribution.

# The Poisson distribution

$X$  = “the number of “rare” events in a given area and/or interval of time”.

Example: the number of customers entering a system within a 10 min. time interval: on average  $\mu = 2.5$

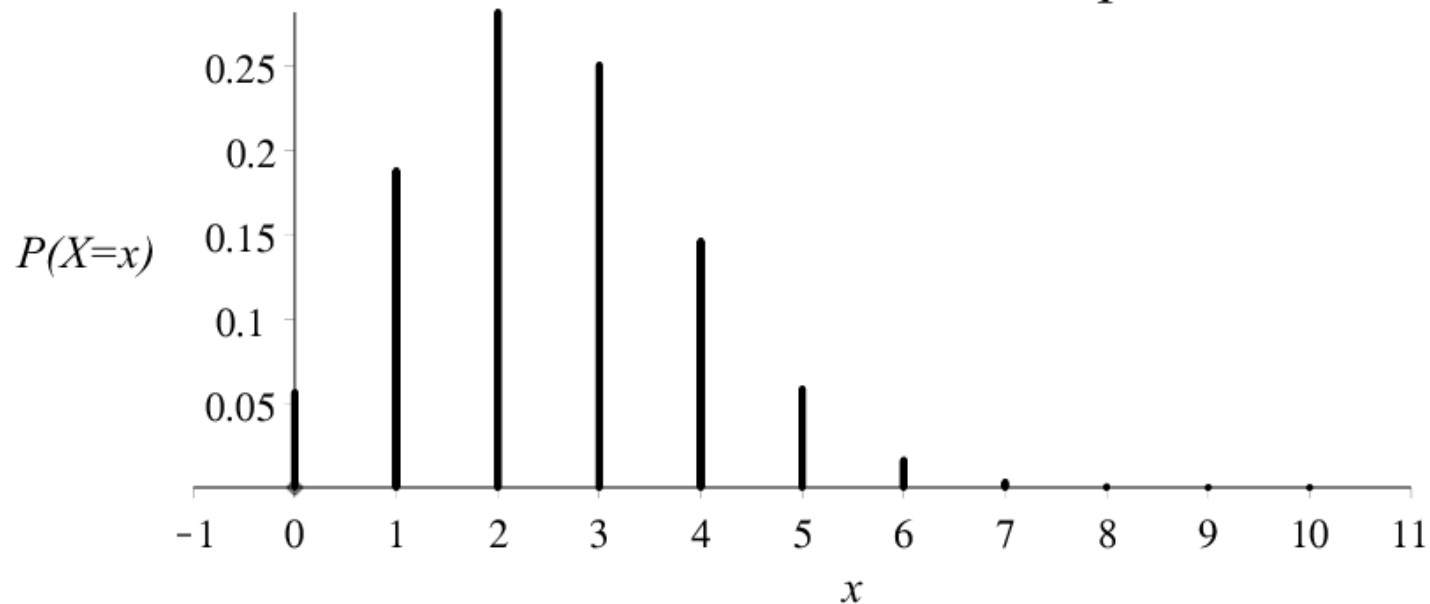
$$P(X = k) = \frac{\mu^k e^{-\mu}}{k!}, \quad k = 0, 1, \dots$$

Properties:

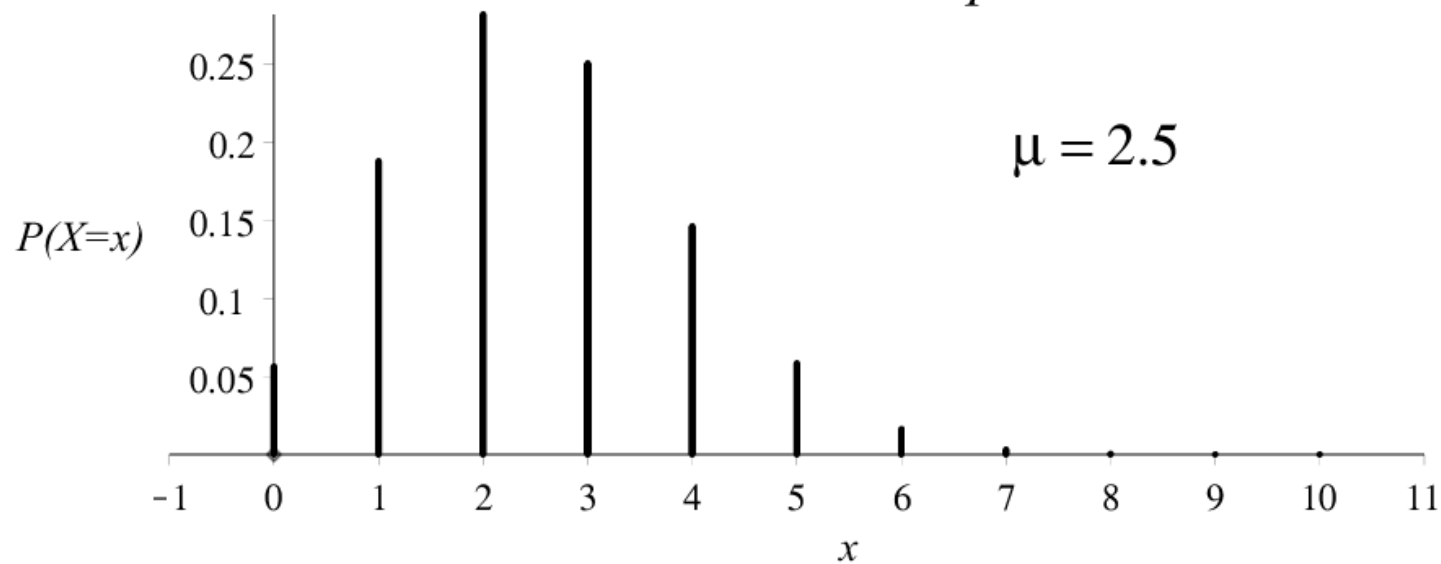
1.  $E(X) = \mu$  and  $var(X) = \mu$
2. If  $X$  has a  $B(n, p)$ -distribution with **large  $n$  and small  $p$** , then  $X$  is approximately **Poisson** with  $\mu = np$ .

**Rule of thumb:**  $n \geq 25$  and  $np < 10$  or  $n(1 - p) < 10$

*Binomial distribution with  $n = 10$  and  $p = 0.25$*



*Poisson distribution with  $\mu = np = 10 * 0.25$*



# Families of discrete distributions

Name + parameter(s)	Probability function	$E(X)$	$Var(X)$
<b>Homogeneous</b> on $1, \dots, n$	$P(X = k) = \frac{1}{n}, k = 1, \dots, n$	$\frac{n + 1}{2}$	Compute using $var(X)$ $= E(X^2) - \mu^2$
<b>Binomial</b> $B(n, p)$	$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k},$ $k = 0, 1, \dots, n$	$np$	$np(1 - p)$
<b>Geometric</b> $(p)$	$P(X = k) = (1 - p)^{k-1} p,$ $k = 1, 2, 3, \dots$	$\frac{1}{p}$	$\frac{1 - p}{p^2}$
<b>Poisson</b> $(\mu)$	$P(X = k) = \frac{\mu^k e^{-\mu}}{k!}, k = 0, 1, \dots$	$\mu$	$\mu$
<b>Hyper-geometric</b> $(R, N, n)$	$P(X = k) = \frac{\binom{R}{k} \binom{N-R}{n-k}}{\binom{N}{n}},$ $k = 0, 1, \dots, n$	$n \cdot \frac{R}{N}$	$n \frac{R}{N} \left(1 - \frac{R}{N}\right) \cdot \frac{N-n}{N-1}$