# CENG 222 Probability and Statistics

Discrete Random Variables and Their Distributions

Nesli Erdoğmuş & Burak Galip Aslan



# Contents

# Contents





# Distribution of a random variable

## Main concepts

**DEFINITION 3.1** A random variable is a function of an outcome,

$$\mathbf{X = f(\omega).}$$

In other words, it is a quantity that depends on chance.

The domain of a random variable is the sample space Ω. Its range can be the set of all real numbers R, or only the positive numbers (0,+∞), or the integers Z, or the interval (0, 1), etc., depending on what possible values the random variable can potentially take.

# Distribution of a random variable

**Example 3.1.** Consider an experiment of tossing 3 fair coins and counting the number of heads. Prior to an experiment, its value is not known. All we can say is that $\mathcal{X}$ **has to be an integer between 0 and 3**. Since assuming each value is an event, we can compute probabilities:

P(X=0) = P{three tails} = P{TTT} = (1/2)(1/2)(1/2) = 1/8
P(X=1) = P{HTT} + P{THT} + P{TTH} = (1/8)+(1/8)+(1/8) = 3/8
P(X=2) = P{HHT} + P{HTH} + P{THH} = (1/8)+(1/8)+(1/8) = 3/8
P(X=3) = P{three heads} = P{HHH} = (1/2)(1/2)(1/2) = 1/8



# Distribution of a random variable

**Example 3.1.** Consider an experiment of tossing 3 fair coins and counting the number of heads. Prior to an experiment, its value is not known. All we can say is that $\mathcal{X}$ **has to be an integer between 0 and 3**. Since assuming each value is an event, we can compute probabilities:

| $x$ | $P(\mathcal{X}=x)$ |
|---|---|
| 0 | 1/8 |
| 1 | 3/8 |
| 2 | 3/8 |
| 3 | 1/8 |
| **Total** | 1 |

This table contains everything that is known about random variable X **prior to the experiment**.

# Distribution of a random variable

**DEFINITION 3.2** Collection of all the probabilities related to $\mathcal{X}$ is the distribution of $\mathcal{X}$. The function

$$P(x)=P(\mathcal{X}=x)$$

is the **probability mass function**, or **pmf**.

The **cumulative distribution**, or **cdf** is defined as

$$F(x)=P(\mathcal{X}\le x)=\sum_{y\le x}P(y)$$

The set of possible values of $\mathcal{X}$ is called the **support** of the distribution F .



# Distribution of a random variable

For every outcome ω, the variable $\mathcal{X}$ takes one and only one value $x$. This makes events $\{\mathcal{X}=x\}$ disjoint and exhaustive, and therefore,

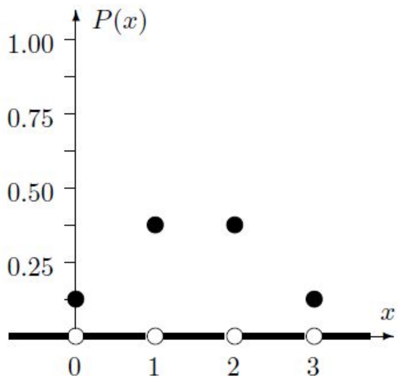
$$\sum_x P(x) = \sum_x P(\mathcal{X}=x) = 1$$

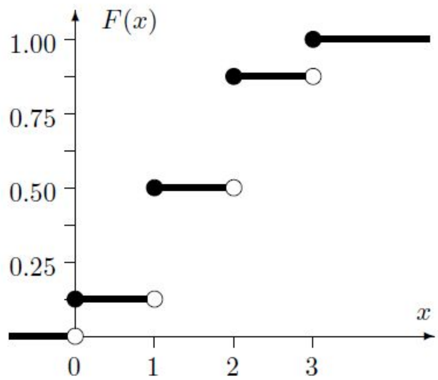
Looking at (3.1), we can conclude that the cdf $F(x)$ is a non-decreasing function of $x$, always between 0 and 1, with

$$\lim_{x\downarrow-\infty}F(x) = 0 \text{ and } \lim_{x\uparrow+\infty}F(x) = 1$$

# Distribution of a random variable

Between any two subsequent values of $\boldsymbol{X}$, F($\boldsymbol{x}$) is constant. It jumps by P($\boldsymbol{x}$) at each possible value $\boldsymbol{x}$ of $\boldsymbol{X}$:






# Distribution of a random variable

Recall that one way to compute the probability of an event is to add probabilities of all the outcomes in it. Hence, for any set A,

$$P\{\boldsymbol{X} \in A) = \sum_{\boldsymbol{x} \in A} P(\boldsymbol{x})$$

When A is an interval, its probability can be computed directly from the cdf F($\boldsymbol{x}$),

$$P\{\boldsymbol{a} < \boldsymbol{X} \leq \boldsymbol{b}\} = F(\boldsymbol{b}) - F(\boldsymbol{a}).$$

**MATLAB DEMO is skipped. We'll discuss computer simulations and histograms in detail later.**

# Distribution of a random variable

**Example 3.3** (Errors in independent modules). A program consists of two modules. The number of errors $X_1$ in the first module has the pmf $P_1(x)$, and the number of errors $X_2$ in the second module has the pmf $P_2(x)$, independently of $X_1$, where

| $x$ | $P_1(x)$ | $P_2(x)$ |
|---|---|---|
| 0 | 0.5 | 0.7 |
| 1 | 0.3 | 0.2 |
| 2 | 0.1 | 0.1 |
| 3 | 0.1 | 0 |

Find the **pmf** and **cdf** of $Y = X_1 + X_2$, the total number of errors.



# Distribution of a random variable

**Example 3.3 - Solution:** $Y = X_1 + X_2$

| $x$ | $P_1(x)$ | $P_2(x)$ |
|---|---|---|
| 0 | 0.5 | 0.7 |
| 1 | 0.3 | 0.2 |
| 2 | 0.1 | 0.1 |
| 3 | 0.1 | 0 |

We break the problem into steps:

- First, determine all possible values of Y,
- then compute the probability of each value.

# Distribution of a random variable

**Example 3.3 - Solution:** $\mathcal{Y} = \mathcal{X}_1 + \mathcal{X}_2$

| $x$ | $P_1(x)$ | $P_2(x)$ |
|---|---|---|
| 0 | 0.5 | 0.7 |
| 1 | 0.3 | 0.2 |
| 2 | 0.1 | 0.1 |
| 3 | 0.1 | 0 |

Clearly, the number of errors $\mathcal{Y}$ is an integer that can be as low as 0 + 0 = 0 and as high as 3 + 2 = 5. Since $P_2(3) = 0$, the second module has at most 2 errors.



# Distribution of a random variable

**Example 3.3 - Solution:** $\mathcal{Y} = \mathcal{X}_1 + \mathcal{X}_2$

| $x$ | $P_1(x)$ | $P_2(x)$ |
|---|---|---|
| 0 | 0.5 | 0.7 |
| 1 | 0.3 | 0.2 |
| 2 | 0.1 | 0.1 |
| 3 | 0.1 | 0 |

$P_{\mathcal{Y}}(0) = P(\mathcal{Y} = 0) = P(\mathcal{X}_1 = \mathcal{X}_2 = 0) = P_1(0)P_2(0) = (0.5)(0.7) = 0.35$

$P_{\mathcal{Y}}(1) = P(\mathcal{Y} =1) = P_1(1)P_2(0)+P_1(0)P_2(1) = (0.3)(0.7)+(0.5)(0.2) = 0.31$

$P_{\mathcal{Y}}(2) = P(\mathcal{Y} =2) = P_1(2)P_2(0)+P_1(1)P_2(1)+P_1(0)P_2(2) = 0.18$

$P_{\mathcal{Y}}(3) = P(\mathcal{Y} =3) = P_1(3)P_2(0)+P_1(2)P_2(1)+P_1(1)P_2(2) = 0.12$

$P_{\mathcal{Y}}(4) = P(\mathcal{Y} =4) = P_1(3)P_2(1)+P_1(2)P_2(2)= 0.03$

$P_{\mathcal{Y}}(5) = P(\mathcal{Y} =5) = P_1(3)P_2(2)= 0.01$

# Distribution of a random variable

**Example 3.3 - Solution:** $y = x_1 + x_2$

| $x$ | $P_1(x)$ | $P_2(x)$ |
|---|---|---|
| 0 | 0.5 | 0.7 |
| 1 | 0.3 | 0.2 |
| 2 | 0.1 | 0.1 |
| 3 | 0.1 | 0 |

| $y$ | $P(y)$ |
|---|---|
| 0 | 0.35 |
| 1 | 0.31 |
| 2 | 0.18 |
| 3 | 0.12 |
| 4 | 0.03 |
| 5 | 0.01 |

Check if the sum of all probabilities is 1.

**Simply obtaining the sum as 1 does not guarantee that we made no mistake in our solution. However, if this equality is not satisfied, we have a mistake for sure.**



# Distribution of a random variable

**Example 3.3 - Solution:** $y = x_1 + x_2$

| $x$ | $P_1(x)$ | $P_2(x)$ |
|---|---|---|
| 0 | 0.5 | 0.7 |
| 1 | 0.3 | 0.2 |
| 2 | 0.1 | 0.1 |
| 3 | 0.1 | 0 |

The cumulative function can be computed as:

| $y$ | $P(y)$ | $F(y)$ |
|---|---|---|
| 0 | 0.35 | $P(y=0) = 0.35$ |
| 1 | 0.31 | $F(y=0)+P(y=1) = 0.66$ |
| 2 | 0.18 | $F(y=1)+P(y=2) = 0.84$ |
| 3 | 0.12 | $F(y=2)+P(y=3) = 0.96$ |
| 4 | 0.03 | $F(y=3)+P(y=4) = 0.99$ |
| 5 | 0.01 | $F(y=4)+P(y=5) = 1.00$ |

# Types of random variables

So far, we are dealing with discrete random variables. These are variables whose range is finite or countable. In particular, it means that their values can be listed, or arranged in a sequence.

Examples include the number of jobs submitted to a printer, the number of errors, the number of error-free modules, the number of failed components, and so on. **Discrete variables don’t have to be integers**. For example, the proportion of defective components in a lot of 100 can be 0, 1/100, 2/100, ..., 99/100, or 1. This variable assumes 101 different values, so it is discrete, although not an integer.



# Types of random variables

On the contrary, **continuous random variables assume a whole interval of values**. This could be a bounded interval (a, b), or an unbounded interval such as (a,+∞), (−∞, b), or (−∞,+∞). Sometimes, it may be a union of several such intervals. Intervals are uncountable, therefore, all values of a random variable cannot be listed in this case.

Examples of continuous variables include various times (software installation time, code execution time, connection time, waiting time, lifetime), also physical variables like weight, height, voltage, temperature, distance, the number of miles per gallon, etc. We shall discuss continuous random variables in detail in Chapter 4.

# Types of random variables

**Example 3.4.** For comparison, observe that a **long jump** is formally a **continuous** random variable because an athlete can jump any distance within some range. **Results of a high jump**, however, are **discrete** because the bar can only be placed on a finite number of heights.



# Contents

# Distribution of a random vector

Often we deal with several random variables simultaneously. We may look at the size of a RAM and the speed of a CPU, the price of a computer and its capacity, temperature and humidity, technical and artistic performance, etc.



# Distribution of a random vector

Often we deal with several random variables simultaneously. We may look at the size of a RAM and the speed of a CPU, the price of a computer and its capacity, temperature and humidity, technical and artistic performance, etc.

**Joint distribution and marginal distributions**

**DEFINITION 3.3** If $\mathcal{X}$ and $\mathcal{Y}$ are random variables, then the pair $(\mathcal{X}, \mathcal{Y})$ is a **random vector**. Its distribution is called the **joint distribution** of $\mathcal{X}$ and $\mathcal{Y}$. Individual distributions of $\mathcal{X}$ and $\mathcal{Y}$ are then called the **marginal distributions**.

# Distribution of a random vector

Similarly to a single variable, the joint distribution of a vector is a collection of probabilities for a vector $(\mathcal{X},\mathcal{Y})$ to take a value $(x,y)$.

Recall that two vectors are equal, $(\mathcal{X},\mathcal{Y}) = (x,y)$, if $\mathcal{X} = x$ and $\mathcal{Y} = y$. This "and" means the intersection, therefore, the joint probability mass function of X and Y is

$$P(x,y) = P\left((\mathcal{X},\mathcal{Y}) = (x,y)\right) = P(\mathcal{X} = x \cap \mathcal{Y} = y).$$

Again, $\{(\mathcal{X},\mathcal{Y}) = (x,y)\}$ are exhaustive and mutually exclusive events for different pairs (x, y), therefore,

$$\sum_x \sum_y P(x,y) = 1$$



# Distribution of a random vector

The joint distribution of $(\mathcal{X}, \mathcal{Y})$ carries the complete information about the behavior of this random vector.

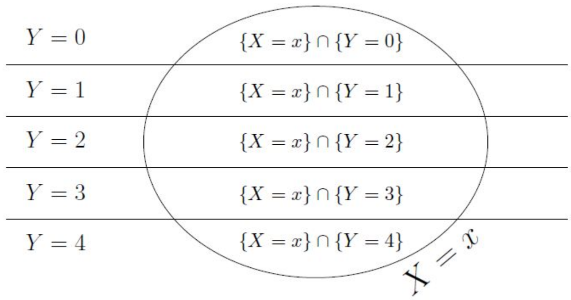
In particular, the marginal probability mass functions of $\mathcal{X}$ and $\mathcal{Y}$ can be obtained from the joint pmf by the **Addition Rule**.

$$P_{\mathcal{X}}(x) = P(\mathcal{X}=x) = \sum_y P_{(\mathcal{X},\mathcal{Y})}(x,y)$$

$$P_{\mathcal{Y}}(y) = P(\mathcal{Y}=y) = \sum_x P_{(\mathcal{X},\mathcal{Y})}(x,y)$$

# Distribution of a random vector

The **Addition Rule** is illustrated in the following figure:




Events $\{\mathcal{Y} = y\}$ for different values of $y$ partition the sample space Ω. Hence, their intersections with $\{\mathcal{X} = x\}$ partition the event $\{\mathcal{X} = x\}$ into mutually exclusive parts.



# Distribution of a random vector

By the rule for the union of mutually exclusive events,

| | |
|---|---|
| **Probability of a union** | $\boldsymbol{P}\{A \cup B\} = \boldsymbol{P}\{A\} + \boldsymbol{P}\{B\} - \boldsymbol{P}\{A \cap B\}$<br>For mutually exclusive events,<br>$\boldsymbol{P}\{A \cup B\} = \boldsymbol{P}\{A\} + \boldsymbol{P}\{B\}$ |

their probabilities should be added. These probabilities are precisely $P_{(\mathcal{X},\mathcal{Y})}(x,y)$.

# Distribution of a random vector

**Independence of random variables**

$\mathcal{X}$ and $\mathcal{Y}$ are independent if $P_{(\mathcal{X},\mathcal{Y})}(x, y) = P_{\mathcal{X}}(x)\, P_{\mathcal{Y}}(y)$ for all values of $x$ and $y$.

- This means, events $\{\mathcal{X} = x\}$ and $\{\mathcal{Y} = y\}$ are independent for all $x$ and $y$; in other words, variables $\mathcal{X}$ and $\mathcal{Y}$ take their values independently of each other.
- In problems, to show independence of $\mathcal{X}$ and $\mathcal{Y}$, we have to check whether the joint pmf factors into the product of marginal pmfs for all pairs $x$ and $y$.
- To prove dependence, we only need to present one counter example, a pair $(x, y)$ with $P(x, y) \neq P_{\mathcal{X}}(x)P_{\mathcal{Y}}(y)$.



# Distribution of a random vector

**Example 3.6.** A program consists of two modules. The number of errors, $\mathcal{X}$, in the first module and the number of errors, $\mathcal{Y}$, in the second module have the joint distribution,

P(0,0) = P(0,1) = P(1,0) = 0.2, P(1,1) = P(1,2) = P(1,3) = 0.1,

P(0,2) = P(0,3) = 0.05.

**a.** Find the marginal distributions of $\mathcal{X}$ and $\mathcal{Y}$.
**b.** Find the probability of no errors in the first module.
**c.** Find the distribution of the total number of errors in the program.
**d.** Find out if errors in the two modules occur independently.

# Distribution of a random vector

**Example 3.6.** Let's summarize everything in a table and

**(a)** compute the marginal probabilities for $\mathcal{X}$ and $\mathcal{Y}$:

| $P_{(X,Y)}(x,y)$ | | $y$ = 0 | 1 | 2 | 3 | $P_X(x)$ |
|---|---|---|---|---|---|---|
| $x$ | 0 | 0.20 | 0.20 | 0.05 | 0.05 | 0.50 |
| | 1 | 0.20 | 0.10 | 0.10 | 0.10 | 0.50 |
| $P_Y(y)$ | | 0.40 | 0.30 | 0.15 | 0.15 | 1.00 |

**(b)** the probability of no errors in the first module, $P_{\mathcal{X}}(0) = 0.50$.



# Distribution of a random vector

**Example 3.6.** In order to compute **(c)** the distribution of the total number of errors in the program, we define a new random variable $\mathcal{Z} = \mathcal{X} + \mathcal{Y}$ which is the total number of errors.

We see that Z can be as small as 0 and as large as 4.

$P_{\mathcal{Z}}(0) = P(\mathcal{Z} = 0) = P(\mathcal{X} = 0 \cap \mathcal{Y} = 0) = P_{(\mathcal{X},\mathcal{Y})}(0, 0) = 0.20$

$P_{\mathcal{Z}}(1) = P(\mathcal{Z} = 1) = P(\mathcal{X} = 0 \cap \mathcal{Y} = 1) + P(\mathcal{X} = 1 \cap \mathcal{Y} = 0)$

$P_{\mathcal{Z}}(1) = P_{(\mathcal{X},\mathcal{Y})}(0, 1) + P_{(\mathcal{X},\mathcal{Y})}(1, 0) = 0.20 + 0.20 = 0.40$

…

# Distribution of a random vector

**Example 3.6.** Now, let's (d) find out if errors in the two modules occur independently.

| $P_{(X,Y)}(x,y)$ | | $y$ | | | | $P_X(x)$ |
|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | |
| $x$ | 0 | 0.20 | 0.20 | 0.05 | 0.05 | 0.50 |
| | 1 | 0.20 | 0.10 | 0.10 | 0.10 | 0.50 |
| $P_Y(y)$ | | 0.40 | 0.30 | 0.15 | 0.15 | 1.00 |

To decide on the independence of $\mathcal{X}$ and $\mathcal{Y}$, check if their joint pmf factors into a product of marginal pmfs:



# Distribution of a random vector

**Example 3.6.** Now, let's (d) find out if errors in the two modules occur independently.

| $P_{(X,Y)}(x,y)$ | | $y$ | | | | $P_X(x)$ |
|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | |
| $x$ | 0 | 0.20 | 0.20 | 0.05 | 0.05 | 0.50 |
| | 1 | 0.20 | 0.10 | 0.10 | 0.10 | 0.50 |
| $P_Y(y)$ | | 0.40 | 0.30 | 0.15 | 0.15 | 1.00 |

To decide on the independence of $\mathcal{X}$ and $\mathcal{Y}$, check if their joint pmf factors into a product of marginal pmfs:

$P(X,Y)(0, 0) = 0.20 = P_{\mathcal{X}}(0)P_{\mathcal{Y}}(0) = (0.5)(0.4)$ ✔

## Distribution of a random vector

**Example 3.6.** Now, let's (d) find out if errors in the two modules occur independently.

| $P_{(X,Y)}(x,y)$ | | $y$ = 0 | 1 | 2 | 3 | $P_X(x)$ |
|---|---|---|---|---|---|---|
| $x$ | 0 | 0.20 | 0.20 | 0.05 | 0.05 | 0.50 |
| | 1 | 0.20 | 0.10 | 0.10 | 0.10 | 0.50 |
| $P_Y(y)$ | | 0.40 | 0.30 | 0.15 | 0.15 | 1.00 |

To decide on the independence of $\mathcal{X}$ and $\mathcal{Y}$, check if their joint pmf factors into a product of marginal pmfs:

P(X,Y)(0, 0) = 0.20 = $P_{\mathcal{X}}(0)P_{\mathcal{Y}}(0)$ = (0.5)(0.4) ✔

P(X,Y)(0, 1) = 0.20 ≠ $P_{\mathcal{X}}(0)P_{\mathcal{Y}}(1)$ = (0.5)(0.3) ✘



## Distribution of a random vector

**Example 3.6.** Now, let's (d) find out if errors in the two modules occur independently.

| $P_{(X,Y)}(x,y)$ | | $y$ = 0 | 1 | 2 | 3 | $P_X(x)$ |
|---|---|---|---|---|---|---|
| $x$ | 0 | 0.20 | 0.20 | 0.05 | 0.05 | 0.50 |
| | 1 | 0.20 | 0.10 | 0.10 | 0.10 | 0.50 |
| $P_Y(y)$ | | 0.40 | 0.30 | 0.15 | 0.15 | 1.00 |

To decide on the independence of $\mathcal{X}$ and $\mathcal{Y}$, check if their joint pmf factors into a product of marginal pmfs:

P(X,Y)(0, 0) = 0.20 = $P_{\mathcal{X}}(0)P_{\mathcal{Y}}(0)$ = (0.5)(0.4) ✔

P(X,Y)(0, 1) = 0.20 ≠ $P_{\mathcal{X}}(0)P_{\mathcal{Y}}(1)$ = (0.5)(0.3) ✘

No need to check further!

# Contents





# Expectation and variance

The distribution of a random variable or a random vector, the full collection of related probabilities, contains <u>the entire information about its behavior</u>. This detailed information can be summarized in a few vital characteristics describing the average value, the most likely value of a random variable, its spread, variability, etc.

The most commonly used are the **expectation, variance, standard deviation, covariance, and correlation**. Also rather popular and useful are the *mode, moments, quantiles, and interquartile range* that we'll discuss in the coming weeks.

# Expectation

**DEFINITION 3.5** **Expectation** or **expected value** of a random variable $\mathcal{X}$ is its mean, the average value.

We know that $\mathcal{X}$ can take different values with different probabilities. For this reason, its average value is not just the average of all its values. Rather, it is a **weighted average**.



# Expectation

**Example 3.7.** Consider a variable that takes values 0 and 1 with probabilities P(0) = P(1) = 0.5. That is,

$$\mathcal{X} = \begin{cases} 0 & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2 \end{cases}$$

Observing this variable many times, we shall see $\mathcal{X}$=0 about 50% of times and $\mathcal{X}$=1 about 50% of times. The average value of $\mathcal{X}$ will then be close to 0.5, so it is reasonable to have $\mathbf{E}(\mathcal{X}) = 0.5$.

# Expectation

**Example 3.8.** Suppose that P(0) = 0.75 and P(1) = 0.25. Then, in a long run, $X$ is equal 1 only 1/4 of times, otherwise it equals 0. Suppose we earn $1 every time we see X = 1.

On the average, we earn $1 every four times, or $0.25 per each observation. Therefore, in this case $\mathbf{E}(X)$ = 0.25.



# Expectation

Here is a physical model for these two examples:

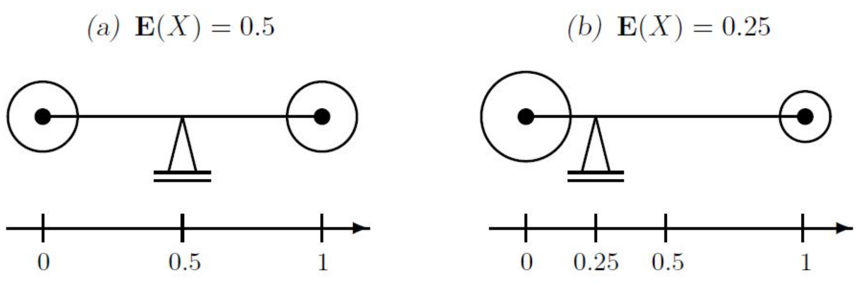



FIGURE 3.4: *Expectation as a center of gravity.*

# Expectation

This formula returns the center of gravity for a system with masses $P(\boldsymbol{x})$ allocated at points $\boldsymbol{x}$. Expected value is often denoted by a Greek letter **μ**.

$$\mu = \mathbf{E}(\mathcal{X}) = \sum_{\boldsymbol{x}} \boldsymbol{x} P(\boldsymbol{x})$$

In a certain sense, expectation is the best forecast of $\mathcal{X}$. The variable itself is random. It takes different values with different probabilities $P(\boldsymbol{x})$. At the same time, it has just one expectation $\mathbf{E}(\mathcal{X})$ which is non-random.



# Expectation

**Example:** Consider the experiment of rolling a dice n times. If we denote the outcomes as $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, the average will be $(\boldsymbol{x}_1+\boldsymbol{x}_2+\ldots+\boldsymbol{x}_n)/n$. The sum $(\boldsymbol{x}_1+\boldsymbol{x}_2+\ldots+\boldsymbol{x}_n)$ can also be calculated by grouping same numbers as $(\boldsymbol{x}_1+\boldsymbol{x}_2+\ldots+\boldsymbol{x}_n)/n = (\sum_{k=1:6} k \cdot N_k)/n$ where $N_k$ is the number of k's in the outcome.

$$\frac{\boldsymbol{x}_1+\boldsymbol{x}_2+\ldots+\boldsymbol{x}_n}{n} = \frac{\sum_{k=1:6} k \cdot N_k}{n} = \sum_{k=1:6} k \cdot (N_k/n)$$

As $n \to \infty$, $(N_k/n) \to P(k)$, so the average becomes $\sum_{k=1:6} k \cdot P(k) = \mathbf{3.5}$.

# Expectation of a function

Often we are interested in another variable, $\mathcal{Y}$, that is a function of $\mathcal{X}$. For example, downloading time depends on the connection speed, profit of a computer store depends on the number of computers sold, and bonus of its manager depends on this profit.

Expectation of $\mathcal{Y} = g(\mathcal{X})$ is computed by a similar formula:

$$\mathbf{E}\{g(\mathcal{X})\} = \sum_{x} g(x)P(x)$$



# Expectation of a function

**Remark:**

- If g is a one-to-one function, then Y takes each value $y = g(x)$ with probability $P(x)$, and the formula for $\mathbf{E}(\mathcal{Y})$ can be applied directly.
- If g is **not** one-to-one, then some values of $g(x)$ will be repeated in the sum. However, they are still multiplied by the corresponding probabilities. In the addition, these probabilities are also added, thus each value of $g(x)$ is still multiplied by the probability $P_Y(g(x))$.

# Properties of expectations

For any random variables $\mathcal{X}$ and $\mathcal{Y}$ and any non-random numbers a, b, and c, we have:

- $\mathbf{E}(a\mathcal{X} + b\mathcal{Y} + c) = a\mathbf{E}(\mathcal{X}) + b\mathbf{E}(\mathcal{Y}) + c$
- In particular:
  - $\mathbf{E}(\mathcal{X} + \mathcal{Y}) = \mathbf{E}(\mathcal{X}) + \mathbf{E}(\mathcal{Y})$
  - $\mathbf{E}(a\mathcal{X}) = a\mathbf{E}(\mathcal{X})$
  - $\mathbf{E}(c) = c$
- For independent $\mathcal{X}$ and $\mathcal{Y}$, $\mathbf{E}(\mathcal{X}\mathcal{Y}) = \mathbf{E}(\mathcal{X})\mathbf{E}(\mathcal{Y})$



# Properties of expectations

**Back to Example 3.6:**

| $P_{(X,Y)}(x,y)$ | | $y$ = 0 | 1 | 2 | 3 | $P_X(x)$ |
|---|---|---|---|---|---|---|
| $x$ | 0 | 0.20 | 0.20 | 0.05 | 0.05 | 0.50 |
| | 1 | 0.20 | 0.10 | 0.10 | 0.10 | 0.50 |
| $P_Y(y)$ | | 0.40 | 0.30 | 0.15 | 0.15 | 1.00 |

- $\mathbf{E}(\mathcal{X}) = (0)(0.5) + (1)(0.5) = 0.5$
- $\mathbf{E}(\mathcal{Y}) = (0)(0.4) + (1)(0.3) + (2)(0.15) + (3)(0.15) = 1.05$,
- $\mathbf{E}(\mathcal{X} + \mathcal{Y}) = 0.5 + 1.05 = \mathbf{1.65}$.

**Remark:** The program will never have 1.65 errors, because the number of errors is always integer. Then, should we round 1.65 to 2 errors? **Absolutely not**, it would be a mistake. Although both $\mathcal{X}$ and $\mathcal{Y}$ are integers, their expectations, or average values, do not have to be integers at all.

# Variance and standard deviation

Expectation shows where the average value of a random variable is located, or where the variable is *expected* to be, plus or minus some error.

- How large could this "error" be?
- How much can a variable vary around its expectation?



# Variance and standard deviation

**Example 3.10.** Consider two users. One receives either 48 or 52 e-mail messages per day, with a 50-50% chance of each. The other receives either 0 or 100 e-mails, also with a 50-50% chance. What is a common feature of these two distributions, and how are they different?

We see that both users receive the same average number of emails:

$$\mathbf{E}(\mathcal{X}) = \mathbf{E}(\mathcal{Y}) = 50.$$

## Variance and standard deviation

**Example 3.10.** Consider two users. One receives either 48 or 52 e-mail messages per day, with a 50-50% chance of each. The other receives either 0 or 100 e-mails, also with a 50-50% chance. What is a common feature of these two distributions, and how are they different?

However, in the first case, the actual number of emails is always close to 50, whereas it always differs from it by 50 in the second case. The first random variable, $\mathcal{X}$, is more stable; it has **low variability**. The second variable, $\mathcal{Y}$, has **high variability**.



## Variance and standard deviation

This example shows that variability of a random variable is measured by its distance from the mean μ = **E**($\mathcal{X}$). In its turn, this distance is random too, and therefore, cannot serve as a characteristic of a distribution. It remains to square it and take the expectation of the result.

**DEFINITION 3.6 Variance** of a random variable is defined as the expected squared deviation from the mean. For discrete random variables, variance is

$$\sigma^2 = \mathrm{Var}(\mathcal{X}) = \mathbf{E}(\mathcal{X} - \mathbf{E}\mathcal{X})^2 = \sum_{x}(x - \mu)^2 P(x)$$

# Variance and standard deviation

**Remark**

Notice that if the distance to the mean is not squared, then the result is always $\mu - \mu = 0$ bearing no information about the distribution of $X$.

According to this definition, **variance is always non-negative**. Further, it equals 0 only if $x = \mu$ for all values of $x$, i.e., when $X$ is constantly equal to $\mu$. Certainly, a constant (non-random) variable has zero variability.

- Variance can also be computed as $\text{Var}(X) = \mathbf{E}(X^2) - \mu^2$.
  Can you prove this?




# Variance and standard deviation

**DEFINITION 3.7** Standard deviation is a square root of variance,

$$\sigma = \text{Std}(X) = \sqrt{\text{Var}(X)}$$

Continuing the Greek-letter tradition, variance is often denoted by $\sigma^2$. Then, standard deviation is $\sigma$.

- If $X$ is measured in some units, then its mean $\mu$ has the same measurement unit as $X$.
- Variance $\sigma^2$ is measured in squared units, and therefore, it cannot be compared with $X$ or $\mu$.
- Standard deviation $\sigma$ is measured in the same units as $X$. This is why we have another measure of variability, $\sigma$.

# Covariance and correlation

Expectation, variance, and standard deviation characterize the **distribution** of a single random variable. Now we introduce measures of **association** of two random variables.

**DEFINITION 3.8**

**Covariance** $\sigma_{XY} = \mathrm{Cov}(\mathcal{X}, \mathcal{Y})$ is defined as

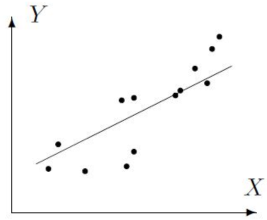
$$\mathrm{Cov}(\mathcal{X}, \mathcal{Y}) = \mathbf{E}\{(\mathcal{X} - \mathbf{E}\mathcal{X})(\mathcal{Y} - \mathbf{E}\mathcal{Y})\}$$
$$= \mathbf{E}(\mathcal{X}\mathcal{Y}) - \mathbf{E}(\mathcal{X})\mathbf{E}(\mathcal{Y})$$

It summarizes interrelation of two random variables.



# Covariance and correlation

If $\mathrm{Cov}(\mathcal{X}, \mathcal{Y}) > 0$, then positive deviations $(\mathcal{X} - \mathbf{E}\mathcal{X})$ are more likely to be multiplied by positive $(\mathcal{Y} - \mathbf{E}\mathcal{Y})$, and negative $(\mathcal{X} - \mathbf{E}\mathcal{X})$ are more likely to be multiplied by negative $(\mathcal{Y} - \mathbf{E}\mathcal{Y})$. In short, large $\mathcal{X}$ imply large $\mathcal{Y}$, and small $\mathcal{X}$ imply small $\mathcal{Y}$. These variables are **positively correlated**,

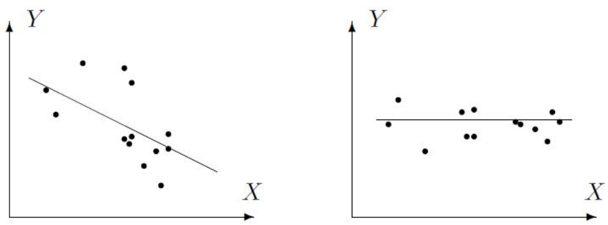



(a) $\mathrm{Cov}(X, Y) > 0$

# Covariance and correlation

Conversely, Cov($\mathcal{X}$, $\mathcal{Y}$) < 0 means that large $\mathcal{X}$ generally correspond to small $\mathcal{Y}$ and small $\mathcal{X}$ correspond to large $\mathcal{Y}$. These variables are **negatively correlated**.

If Cov($\mathcal{X}$, $\mathcal{Y}$) = 0, we say that $\mathcal{X}$ and $\mathcal{Y}$ are **uncorrelated**.




(b) $\mathrm{Cov}(X,Y) < 0$ (c) $\mathrm{Cov}(X,Y) = 0$




# Covariance and correlation

**DEFINITION 3.9 Correlation** coefficient between variables $\mathcal{X}$ and $\mathcal{Y}$ is defined as

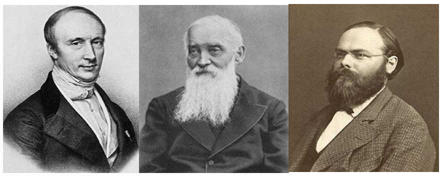
$$\rho = \frac{\mathrm{Cov}(\mathcal{X},\mathcal{Y})}{\mathrm{Std}\,\mathcal{X}\,\mathrm{Std}\,\mathcal{Y}}$$

Correlation coefficient is a rescaled, normalized covariance.

Covariance has a measurement unit. It is measured in units of $\mathcal{X}$ multiplied by units of $\mathcal{Y}$. Hence, it is not clear from Cov($\mathcal{X}$, $\mathcal{Y}$) whether $\mathcal{X}$ and $\mathcal{Y}$ are strongly or weakly correlated.

Correlation coefficient compares Cov($\mathcal{X}$, $\mathcal{Y}$) with the magnitude of $\mathcal{X}$ and $\mathcal{Y}$ and, as a result, it is **dimensionless**.

# Covariance and correlation



How do we interpret the value of ρ? What possible values can it take?

As a special case of famous *Cauchy(-Bunyakovsky)-Schwarz inequality*,

$$-1 \le \rho \le 1,$$

where $|\rho| = 1$ is possible only when all values of $\mathcal{X}$ and $\mathcal{Y}$ lie on a straight line. Further, values of ρ near 1 indicate strong positive correlation, values near (−1) show strong negative correlation, and values near 0 show weak correlation or no correlation.



# Properties of variances and covariances

For any random variables $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$, and $\mathcal{W}$ and any non-random numbers *a*, *b*, *c* and *d*:

- $\mathrm{Var}(a\mathcal{X} + b\mathcal{Y} + c) = a^2\mathrm{Var}(\mathcal{X}) + b^2\mathrm{Var}(\mathcal{Y}) + 2ab\mathrm{Cov}(\mathcal{X}, \mathcal{Y})$
- $\mathrm{Cov}(a\mathcal{X} + b\mathcal{Y}, c\mathcal{Z} + d\mathcal{W})$

  $= ac\mathrm{Cov}(\mathcal{X},\mathcal{Z}) + ad\mathrm{Cov}(\mathcal{X},\mathcal{W}) + bc\mathrm{Cov}(\mathcal{Y},\mathcal{Z}) + bd\mathrm{Cov}(\mathcal{Y},\mathcal{W})$
- $\mathrm{Cov}(\mathcal{X}, \mathcal{Y}) = \mathrm{Cov}(\mathcal{Y}, \mathcal{X})$
- $\rho(\mathcal{X}, \mathcal{Y}) = \rho(\mathcal{Y}, \mathcal{X})$

# Properties of variances and covariances

For any random variables $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$, and $\mathcal{W}$ and any non-random numbers *a*, *b*, *c* and *d*, in particular:

- $\mathrm{Var}(a\mathcal{X} + b) = a^2\mathrm{Var}(\mathcal{X})$ (Can you prove this?)
- $\mathrm{Cov}(a\mathcal{X} + b, c\mathcal{Y}+d) = ac\mathrm{Cov}(\mathcal{X},\mathcal{Y})$
- $\rho(a\mathcal{X} + b, c\mathcal{Y}+d) = \rho(\mathcal{X}, \mathcal{Y})$
- For independent $\mathcal{X}$ and $\mathcal{Y}$:
  - $\mathrm{Cov}(\mathcal{X}, \mathcal{Y}) = 0$
  - $\mathrm{Var}(\mathcal{X} + \mathcal{Y}) = \mathrm{Var}(\mathcal{X}) + \mathrm{Var}(\mathcal{Y})$



# Properties of variances and covariances

- We see that **independent variables are always uncorrelated**. The reverse is not always true. There exist some variables that are **uncorrelated but not independent**.
- **Adding a constant** does not affect the variables' **variance** or **covariance**. It shifts the whole distribution of $\mathcal{X}$ without changing its variability or degree of dependence of another variable.
- The **correlation** coefficient does not change even **when multiplied** by a constant because it is recomputed to the unit scale, due to Std($\mathcal{X}$) and Std($\mathcal{Y}$) in the denominator.

# Covariance and correlation

| $P_{(X,Y)}(x,y)$ | | $y$ = 0 | 1 | 2 | 3 | $P_X(x)$ |
|---|---|---|---|---|---|---|
| $x$ | 0 | 0.20 | 0.20 | 0.05 | 0.05 | 0.50 |
| | 1 | 0.20 | 0.10 | 0.10 | 0.10 | 0.50 |
| $P_Y(y)$ | | 0.40 | 0.30 | 0.15 | 0.15 | 1.00 |

| $x$ | $P_X(x)$ | $xP_X(x)$ | $x-\mathbf{E}X$ | $(x-\mathbf{E}X)^2P_X(x)$ |
|---|---|---|---|---|
| 0 | 0.5 | 0 | $-0.5$ | 0.125 |
| 1 | 0.5 | 0.5 | 0.5 | 0.125 |
| | $\mu_X = 0.5$ | | $\sigma_X^2 = 0.25$ | |

| $y$ | $P_Y(y)$ | $yP_Y(y)$ | $y^2$ | $y^2P_Y(y)$ |
|---|---|---|---|---|
| 0 | 0.4 | 0 | 0 | 0 |
| 1 | 0.3 | 0.3 | 1 | 0.3 |
| 2 | 0.15 | 0.3 | 4 | 0.6 |
| 3 | 0.15 | 0.45 | 9 | 1.35 |
| | $\mu_Y = 1.05$ | | $\mathbf{E}(Y^2) = 2.25$ | |

**Example 3.11.** Result: Var($\mathcal{X}$) = 0.25,

Var($\mathcal{Y}$) = 2.25 − $1.05^2$ = 1.1475,

Std($\mathcal{X}$) = $\sqrt{0.25}$ = 0.5, and Std($\mathcal{Y}$) = $\sqrt{1.1475}$ = 1.0712. Also,

$$\mathbf{E}(XY) = \sum_x \sum_y xyP(x,y) = (1)(1)(0.1) + (1)(2)(0.1) + (1)(3)(0.1) = 0.6$$

(the other five terms in this sum are 0). Therefore,

Cov($\mathcal{X}$, $\mathcal{Y}$) = E($\mathcal{XY}$) − E($\mathcal{X}$)E($\mathcal{Y}$) = 0.6 − (0.5)(1.05) = 0.075 and

$$\rho = \frac{\mathrm{Cov}(X,Y)}{(\mathrm{Std}X)(\mathrm{Std}Y)} = \frac{0.075}{(0.5)(1.0712)} = 0.1400.$$

Thus, the numbers of errors in two modules are **positively** and **not very strongly** correlated.



# Properties of expectations

**Back to Example 3.6:**

| $P_{(X,Y)}(x,y)$ | | $y$ = 0 | 1 | 2 | 3 | $P_X(x)$ |
|---|---|---|---|---|---|---|
| $x$ | 0 | 0.20 | 0.20 | 0.05 | 0.05 | 0.50 |
| | 1 | 0.20 | 0.10 | 0.10 | 0.10 | 0.50 |
| $P_Y(y)$ | | 0.40 | 0.30 | 0.15 | 0.15 | 1.00 |

- Var($\mathcal{X}$) = $(0-0.5)^2$*0.5+$(1-0.5)^2$*0.5 = 0.25
- Var($\mathcal{Y}$) = $\mathbf{E}(\mathcal{Y}^2)$ - $\mathbf{E}(\mathcal{Y})^2$ = 2.25 − $1.05^2$ = 1.1475
- $\mathbf{E}(\mathcal{XY}) = \sum_x\sum_y xyP(x,y)$ = (1)(1)(0.1) + (1)(2)(0.1) + (1)(3)(0.1) = 0.6
- Cov($\mathcal{X}$, $\mathcal{Y}$) = $\mathbf{E}(\mathcal{XY})$ - $\mathbf{E}(\mathcal{X})\mathbf{E}(\mathcal{Y})$ = 0.6 - (0.5)(1.05) = 0.075
- ρ = Cov($\mathcal{X}$, $\mathcal{Y}$) / (Std$\mathcal{X}$)(Std$\mathcal{Y}$) = 0.075 / (0.5)(1.0712) = 0.1400

# Covariance and correlation

| | | | |
|---|---|---|---|
| Notation | $\mu$ or $\mathbf{E}(X)$ | = | expectation |
| | $\sigma_X^2$ or $\mathrm{Var}(X)$ | = | variance |
| | $\sigma_X$ or $\mathrm{Std}(X)$ | = | standard deviation |
| | $\sigma_{XY}$ or $\mathrm{Cov}(X,Y)$ | = | covariance |
| | $\rho_{XY}$ | = | correlation coefficient |



# Chebyshev's inequality

Knowing just the expectation and variance, one can find the range of values most likely taken by this variable.

Russian mathematician Pafnuty Chebyshev (1821–1894) showed that any random variable $X$ with expectation $\mu = \mathbf{E}(X)$ and variance $\sigma^2 = \mathrm{Var}(X)$ belongs to the interval $\mu \pm \varepsilon = [\mu - \varepsilon, \mu + \varepsilon]$ with probability of at least $1 - (\sigma/\varepsilon)^2$.

$$P(|X-\mu| > \varepsilon) \leq (\sigma/\varepsilon)^2$$

(Can you prove this?)

# Chebyshev's inequality

**Example 3.12**. Suppose the number of errors in a new software has expectation μ = 20 and a standard deviation of 2. According to Chebyshev's inequality, there are more than 30 errors with probability:

$P(X > 30) \leq P\,(|X\text{-}20| > 10\} \leq (2/10)^2 = 0.04$.

However, if the standard deviation is 5 instead of 2, then the probability of more than 30 errors can only be bounded by $(5/10)^2 = 0.25$.



# Contents

# Bernoulli distribution


The simplest random variable (excluding non-random ones!) takes just two possible values. Call them 0 and 1.

**DEFINITION 3.10** A random variable with two possible values, 0 and 1, is called a **Bernoulli variable**, its distribution is **Bernoulli distribution**, and any experiment with a binary outcome is called a **Bernoulli trial**.

Good or defective components, parts that pass or fail tests, transmitted or lost signals, working or malfunctioning hardware, benign or malicious attachments, sites that contain or do not contain a keyword, girls and boys, heads and tails, and so on, are examples of Bernoulli trials.



# Bernoulli distribution

All these experiments fit the same Bernoulli model, where we shall use generic names for the two outcomes: "***successes***" and "***failures***." These are nothing but commonly used generic names; in fact, successes do not have to be good, and failures do not have to be bad.

If $P(1) = p$ is the probability of a success, then $P(0) = q = 1 - p$ is the probability of a failure.

# Bernoulli distribution

We can compute the expectation and variance as:

$\mathbf{E}(\mathcal{X}) = \sum_{x} x P(x) = (0)(1\text{-}p) + (1)(p) = p$

$\mathrm{Var}(\mathcal{X}) = \sum_{x} (x\text{-}p)^2 P(x) = (0\text{-}p)^2(1\text{-}p) + (1\text{-}p)^2(p) = (1\text{-}p)(p^2+p\text{-}p^2)$

$= (1\text{-}p)p = qp$

Another approach to compute $\mathrm{Var}(\mathcal{X})$?

- There is a whole family of Bernoulli distributions, indexed by a parameter p. Every p between 0 and 1 defines another Bernoulli distribution.



# Bernoulli distribution

We can compute the expectation and variance as:

$\mathbf{E}(\mathcal{X}) = \sum_{x} x P(x) = (0)(1\text{-}p) + (1)(p) = p$

$\mathrm{Var}(\mathcal{X}) = \sum_{x} (x\text{-}p)^2 P(x) = (0\text{-}p)^2(1\text{-}p) + (1\text{-}p)^2(p) = (1\text{-}p)(p^2+p\text{-}p^2)$

$= (1\text{-}p)p = qp$

Another approach to compute $\mathrm{Var}(\mathcal{X})$?

- The distribution with p = 0.5 carries the highest level of uncertainty because the variance is maximized.
- Distributions with lower or higher p have lower variances.
- For p=0 and p=1, the variance is 0.

# Binomial distribution

Consider a sequence of independent Bernoulli trials and count the number of successes in it. This may be the number of defective computers in a shipment, the number of updated files in a folder, the number of girls in a family, etc.

**DEFINITION 3.11** A variable described as the number of successes in a sequence of independent Bernoulli trials has **Binomial distribution**. Its parameters are *n*, the number of trials, and *p*, the probability of success.

**Remark:** "Binomial" can be translated as "two numbers," bi meaning "two" and nom meaning "a number".



# Binomial distribution

Binomial probability mass function is:

$$P(x) = P(X = x) = C(n,x)p^{x}q^{n-x},\ x = 0,1,\ldots,n$$

which is the probability of exactly *x* successes in *n* trials.

- $p^{x}$ is the probability of *x* successes, probabilities being multiplied due to independence of trials.
- $q^{n-x}$ is the probability of the remaining $(n - x)$ trials being failures.
- $C(n,x)$ is the number of elements of the sample space Ω that form the event $\{X = x\}$. This is the number of possible orderings of $x$ successes and $(n - x)$ failures among n trials.

# Binomial distribution

Due to a somewhat complicated form of (3.9), practitioners use a table of Binomial distribution, Table A2. Its entries are values of the Binomial cdf F(*x*). If we need a pmf instead of a cdf, we can always compute it from the table as P($\boldsymbol{x}$) = F($\boldsymbol{x}$) - F($\boldsymbol{x}$-1)

$$F(x) = \boldsymbol{P}\{X \leq x\} = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}$$

| $n$ | $x$ | $p$ | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | .050 | .100 | .150 | .200 | .250 | .300 | .350 | .400 | .450 | .500 | .550 | .600 | .650 | .700 | .750 | .800 | .850 | .900 | .950 |
| 1 | 0 | .950 | .900 | .850 | .800 | .750 | .700 | .650 | .600 | .550 | .500 | .450 | .400 | .350 | .300 | .250 | .200 | .150 | .100 | .050 |
| 2 | 0 | .903 | .810 | .723 | .640 | .563 | .490 | .423 | .360 | .303 | .250 | .203 | .160 | .123 | .090 | .063 | .040 | .023 | .010 | .003 |
| | 1 | .998 | .990 | .978 | .960 | .938 | .910 | .878 | .840 | .798 | .750 | .698 | .640 | .578 | .510 | .438 | .360 | .278 | .190 | .098 |
| 3 | 0 | .857 | .729 | .614 | .512 | .422 | .343 | .275 | .216 | .166 | .125 | .091 | .064 | .043 | .027 | .016 | .008 | .003 | .001 | .000 |
| | 1 | .993 | .972 | .939 | .896 | .844 | .784 | .718 | .648 | .575 | .500 | .425 | .352 | .282 | .216 | .156 | .104 | .061 | .028 | .007 |
| | 2 | 1.0 | .999 | .997 | .992 | .984 | .973 | .957 | .936 | .909 | .875 | .834 | .784 | .725 | .657 | .578 | .488 | .386 | .271 | .143 |



# Binomial distribution

**Example 3.16.** As part of a business strategy, randomly selected 20% of new internet service subscribers receive a special promotion from the provider. A group of 10 neighbors signs for the service. What is the probability that at least 4 of them get a special promotion?

# Binomial distribution

**Example 3.16.** As part of a business strategy, randomly selected 20% of new internet service subscribers receive a special promotion from the provider. A group of 10 neighbors signs for the service. What is the probability that at least 4 of them get a special promotion?

**Solution.** We need to find the probability P $\{X \geq 4\}$, where $X$ is the number of people, out of 10, who receive a special promotion. This is the number of successes in 10 Bernoulli trials, therefore, $X$ has Binomial distribution with parameters $n$ = 10 and $p$ = 0.2.



# Binomial distribution

**Example 3.16.** As part of a business strategy, randomly selected 20% of new internet service subscribers receive a special promotion from the provider. A group of 10 neighbors signs for the service. What is the probability that at least 4 of them get a special promotion?

**Solution.** $P(X \geq 4) = 1 - F(3)$

$= 1 - 0.879$

$= 0.121$

| $n$ | $x$ | .050 | .100 | .150 | .200 | .250 | .300 |
|---|---|---|---|---|---|---|---|
| 10 | 0 | .599 | .349 | .197 | .107 | .056 | .028 |
| | 1 | .914 | .736 | .544 | .376 | .244 | .149 |
| | 2 | .988 | .930 | .820 | .678 | .526 | .383 |
| | 3 | .999 | .987 | .950 | .879 | .776 | .650 |
| | 4 | 1.0 | .998 | .990 | .967 | .922 | .850 |
| | 5 | 1.0 | 1.0 | .999 | .994 | .980 | .953 |
| | 6 | 1.0 | 1.0 | 1.0 | .999 | .996 | .989 |
| | 7 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | .998 |
| | 8 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| | 9 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |

# Binomial distribution

Computing the expected value directly by its equation results in a complicated formula,

$$E(X) = \sum_{x=0\ldots n} x\,C(n,x)p^{x}q^{n-x} = \ldots ?$$

A shortcut can be obtained from the following important property: Each Bernoulli trial is associated with a Bernoulli variable that equals 1 if the trial results in a success and 0 in case of a failure. Then, a sum of these variables is the overall number of successes. Thus, any Binomial variable *X* can be represented as a sum of independent Bernoulli variables.

$$X = X_1 + \ldots + X_n$$



# Binomial distribution

$$X = X_1 + \ldots + X_n$$

Using the addition properties of expectation and variance, we can compute:

$$\mathbf{E}(X) = \mathbf{E}(X_1 + \ldots + X_n) = \mathbf{E}(X_1) + \ldots + \mathbf{E}(X_n) = p+\ldots+p = np$$

$$\mathrm{Var}(X) = \mathrm{Var}(X_1 + \ldots + X_n) = \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n) = npq$$

# Binomial distribution

**Example 3.17.** An exciting computer game is released. Sixty percent of players complete all the levels. Thirty percent of them will then buy an advanced version of the game. Among 15 users, what is the expected number of people who will buy the advanced version? What is the probability that at least two people will buy it?



# Binomial distribution

**Solution.** Let X be the number of people (successes), among the mentioned 15 users (trials), who will buy the advanced version of the game. It has Binomial distribution with $n$ = 15 trials and the probability of success p = P{buy advanced}

p = P{buy advanced | complete all levels} P{complete all levels}

p = (0.30)(0.60) = 0.18

$\mathbf{E}(X) = np = (15)(0.18) = 2.7$

$P(x) = P(X = x) = C(n,x)p^x q^{n-x}$, $x = 0,1,\ldots,n$

$P(X \geq 2) = 1 - P(0) - P(1) = 1 - (1-p)^n - np(1-p)^{n-1} = 0.7813$

# Bernoulli and Binomial distributions

**Bernoulli distribution**

$$
\begin{array}{lcl}
p & = & \text{probability of success} \\
P(x) & = & \begin{cases} q = 1 - p & \text{if} \quad x = 0 \\ \quad p & \text{if} \quad x = 1 \end{cases} \\
\mathbf{E}(X) & = & p \\
\mathrm{Var}(X) & = & pq
\end{array}
$$

**Binomial distribution**

$$
\begin{array}{lcl}
n & = & \text{number of trials} \\
p & = & \text{probability of success} \\
P(x) & = & \binom{n}{x} p^x q^{n-x} \\
\mathbf{E}(X) & = & np \\
\mathrm{Var}(X) & = & npq
\end{array}
$$



# Geometric distribution

Again, consider a sequence of independent Bernoulli trials. Each trial results in a “success” or a “failure.”

**DEFINITION 3.12** The number of Bernoulli trials needed to get the first success has **Geometric distribution**.

**Example 3.18.** A search engine goes through a list of sites looking for a given key phrase. Suppose the search terminates as soon as the key phrase is found. The number of sites visited is Geometric.

# Geometric distribution

- Geometric random variables can take any integer value from 1 to infinity, because one needs at least 1 trial to have the first success.
- The number of trials needed is not limited by any specific number. (For example, there is no guarantee that among the first 10 coin tosses there will be at least one head.)
- The only parameter is p, the probability of a "success."

Geometric probability mass function has the form:

$P(\boldsymbol{x})$ = P{the 1st success occurs on the $\boldsymbol{x}$-th trial} = $(1-p)^{\boldsymbol{x}-1}p$, where $\boldsymbol{x} = 1, 2, \ldots$



# Geometric distribution

Geometric probability mass function has the form:

$P(\boldsymbol{x})$ = P{the 1st success occurs on the $\boldsymbol{x}$-th trial} = $(1-p)^{\boldsymbol{x}-1}p$, where $\boldsymbol{x} = 1, 2, \ldots$

- $(1-p)^{\boldsymbol{x}-1}$ is the probability of $(x - 1)$ failures
- f is the probability of one success in the $\boldsymbol{x}$-th trial.

This is the first time we see an **unbounded** random variable, that is, with no upper bound.

# Geometric distribution

Geometric distribution has expectation $\mu = 1/p$ and variance $\sigma^2 = (1 - p)/p^2$.

**Geometric distribution**

$$
\begin{aligned}
p &= \text{probability of success} \\
P(x) &= (1-p)^{x-1}p, \quad x = 1, 2, \ldots \\
\mathbf{E}(X) &= \frac{1}{p} \\
\mathrm{Var}(X) &= \frac{1-p}{p^2}
\end{aligned}
$$



# Geometric distribution

**Example 3.20** (St. Petersburg Paradox). This paradox was noticed by a Swiss mathematician Daniel Bernoulli (1700–1782), a nephew of Jacob. It describes a gambling strategy that enables one to win any desired amount of money with probability one.

Consider a game that can be played any number of times:

- Rounds are independent.
- Each time your winning probability is p.
- For each round, you bet some amount *x*. In case of a success, you win *x*. If you lose a round, you lose *x*.

# Geometric distribution

**Example 3.20** Consider a game that can be played any number of times:

- Rounds are independent.
- Each time your winning probability is p.
- For each round, you bet some amount *x*. In case of a success, you win *x*. If you lose a round, you lose *x*.

The strategy is simple:

- Your initial bet is the amount that you like to win eventually.
- If you win a round, stop. Otherwise, double your bet and continue.



# Geometric distribution

Let the desired profit be \$100. The game will progress as follows:

| Round | Bet | Balance... | |
|---|---|---|---|
| | | ... if lose | ... if win |
| 1 | 100 | $-100$ | $+100$ and stop |
| 2 | 200 | $-300$ | $+100$ and stop |
| 3 | 400 | $-700$ | $+100$ and stop |
| ... | ... | ... | ... |

Sooner or later, the game will stop, and at this moment, your balance will be \$100. Guaranteed!

However, this is not what D. Bernoulli called a paradox.

# Geometric distribution

- How many rounds should be played?

  Since each round is a Bernoulli trial, the number of them, $\mathcal{X}$, until the first win is a Geometric random variable with parameter *p*.

- Is the game endless?

  No, on the average, it will last $\mathbf{E}(\mathcal{X}) = 1/p$ rounds. In a fair game with $p = 1/2$, one will need 2 rounds, on the average, to win the desired amount. If $p = 0.2$, i.e., one win in five rounds, then on the average, one stops after $1/p = 5$ rounds.

  This is not a paradox yet.



# Geometric distribution

- How much money does one need to have in order to be able to follow this strategy?

  Let $\mathcal{Y}$ be the amount of the last bet. According to the strategy, $\mathcal{Y} = 100 \cdot 2^{\mathcal{X}-1}$. It is a discrete random variable with:

  $$E(\mathcal{Y}) = \sum_{x}(100 \cdot 2^{x-1})P_{\mathcal{X}}(x) = 100\sum_{x=1:\infty}2^{x-1}(1-p)^{x-1}p$$

  $$E(\mathcal{Y}) = 100p\sum_{x=1:\infty}(2(1-p))^{x-1}$$

  $$E(\mathcal{Y}) = \begin{cases} \dfrac{100p}{2(1-p)} & \text{if } p > 1/2 \\ +\infty & \text{if } p \leq 1/2 \end{cases}$$

# Geometric distribution

- How much money does one need to have in order to be able to follow this strategy?

  Let $\mathcal{Y}$ be the amount of the last bet. According to the strategy, $\mathcal{Y} = 100 \cdot 2^{\mathcal{X}-1}$. It is a discrete random variable with:

  $E(\mathcal{Y}) = \sum_{x}(100 \cdot 2^{x-1})P_{\mathcal{X}}(x) = 100\sum_{x=1:\infty} 2^{x-1}(1-p)^{x-1}p$

  $E(\mathcal{Y}) = 100p\sum_{x=1:\infty}(2(1-p))^{x-1}$

  $$E(\mathcal{Y}) = \begin{cases} \dfrac{100p}{2(1-p)} & \text{if } p > 1/2 \\ +\infty & \text{if } p \leq 1/2 \end{cases}$$

  This is the St. Petersburg Paradox! A random variable that is always finite has an infinite expectation!



# Poisson distribution

- This distribution is related to a concept of *rare events*, or Poissonian events.
- It means that two such events are extremely unlikely to occur simultaneously or within a very short period of time.
- Arrivals of jobs, telephone calls, e-mail messages, traffic accidents, network blackouts, virus attacks, errors in software, floods, and earthquakes are examples of rare events.

**DEFINITION 3.14** The number of rare events occurring within a fixed period of time has **Poisson distribution**.

# Poisson distribution

**Poisson distribution**

$$\begin{aligned} \lambda &= \text{frequency, average number of events} \\ P(x) &= e^{-\lambda}\frac{\lambda^x}{x!},\ x = 0, 1, 2, \ldots \\ \mathbf{E}(X) &= \lambda \\ \mathrm{Var}(X) &= \lambda \end{aligned}$$

Poisson variable can take any nonnegative integer value because there may be no rare events within the chosen period, on one end, and the possible number of events is not limited, on the other end.

Poisson distribution has one parameter, λ>0, which is the **average number** of the considered rare events.



# Poisson distribution

Values of its cdf are given in the following table for different λ:

$$F(x) = \boldsymbol{P}\{X \le x\} = \sum_{k=0}^{x} \frac{e^{-\lambda}\lambda^k}{k!}$$

| $x$ | $\lambda$ | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 3.5 | 4.0 | 4.5 | 5.0 | 5.5 | 6.0 | 6.5 | 7.0 | 7.5 | 8.0 | 8.5 | 9.0 | 9.5 | 10.0 | 10.5 |
| 0 | .030 | .018 | .011 | .007 | .004 | .002 | .002 | .001 | .001 | .000 | .000 | .000 | .000 | .000 | .000 |
| 1 | .136 | .092 | .061 | .040 | .027 | .017 | .011 | .007 | .005 | .003 | .002 | .001 | .001 | .000 | .000 |
| 2 | .321 | .238 | .174 | .125 | .088 | .062 | .043 | .030 | .020 | .014 | .009 | .006 | .004 | .003 | .002 |
| 3 | .537 | .433 | .342 | .265 | .202 | .151 | .112 | .082 | .059 | .042 | .030 | .021 | .015 | .010 | .007 |
| 4 | .725 | .629 | .532 | .440 | .358 | .285 | .224 | .173 | .132 | .100 | .074 | .055 | .040 | .029 | .021 |
| 5 | .858 | .785 | .703 | .616 | .529 | .446 | .369 | .301 | .241 | .191 | .150 | .116 | .089 | .067 | .050 |
| 6 | .935 | .889 | .831 | .762 | .686 | .606 | .527 | .450 | .378 | .313 | .256 | .207 | .165 | .130 | .102 |
| 7 | .973 | .949 | .913 | .867 | .809 | .744 | .673 | .599 | .525 | .453 | .386 | .324 | .269 | .220 | .179 |
| 8 | .990 | .979 | .960 | .932 | .894 | .847 | .792 | .729 | .662 | .593 | .523 | .456 | .392 | .333 | .279 |
| 9 | .997 | .992 | .983 | .968 | .946 | .916 | .877 | .830 | .776 | .717 | .653 | .587 | .522 | .458 | .397 |
| 10 | .999 | .997 | .993 | .986 | .975 | .957 | .933 | .901 | .862 | .816 | .763 | .706 | .645 | .583 | .521 |

# Poisson distribution

**Example 3.22** (New accounts). Customers of an internet service provider initiate new accounts at the average rate of 10 accounts per day.

(a) What is the probability that more than 8 new accounts will be initiated today?

**Solution.** (a) New account initiations qualify as rare events because no two customers open accounts simultaneously. Then the number $X$ of today's new accounts has Poisson distribution with parameter $\lambda = 10$.

$P(X > 8) = 1 - F_X(8)$



# Poisson distribution

Values of its cdf are given in the Table A3 for different $\lambda$ and $x$:

$$F(x) = \boldsymbol{P}\{X \leq x\} = \sum_{k=0}^{x} \frac{e^{-\lambda}\lambda^k}{k!}$$

| $x$ | $\lambda$ | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 3.5 | 4.0 | 4.5 | 5.0 | 5.5 | 6.0 | 6.5 | 7.0 | 7.5 | 8.0 | 8.5 | 9.0 | 9.5 | 10.0 | 10.5 |
| 0 | .030 | .018 | .011 | .007 | .004 | .002 | .002 | .001 | .001 | .000 | .000 | .000 | .000 | .000 | .000 |
| 1 | .136 | .092 | .061 | .040 | .027 | .017 | .011 | .007 | .005 | .003 | .002 | .001 | .001 | .000 | .000 |
| 2 | .321 | .238 | .174 | .125 | .088 | .062 | .043 | .030 | .020 | .014 | .009 | .006 | .004 | .003 | .002 |
| 3 | .537 | .433 | .342 | .265 | .202 | .151 | .112 | .082 | .059 | .042 | .030 | .021 | .015 | .010 | .007 |
| 4 | .725 | .629 | .532 | .440 | .358 | .285 | .224 | .173 | .132 | .100 | .074 | .055 | .040 | .029 | .021 |
| 5 | .858 | .785 | .703 | .616 | .529 | .446 | .369 | .301 | .241 | .191 | .150 | .116 | .089 | .067 | .050 |
| 6 | .935 | .889 | .831 | .762 | .686 | .606 | .527 | .450 | .378 | .313 | .256 | .207 | .165 | .130 | .102 |
| 7 | .973 | .949 | .913 | .867 | .809 | .744 | .673 | .599 | .525 | .453 | .386 | .324 | .269 | .220 | .179 |
| 8 | .990 | .979 | .960 | .932 | .894 | .847 | .792 | .729 | .662 | .593 | .523 | .456 | .392 | .333 | .279 |
| 9 | .997 | .992 | .983 | .968 | .946 | .916 | .877 | .830 | .776 | .717 | .653 | .587 | .522 | .458 | .397 |
| 10 | .999 | .997 | .993 | .986 | .975 | .957 | .933 | .901 | .862 | .816 | .763 | .706 | .645 | .583 | .521 |

# Poisson distribution

**Example 3.22** (New accounts). Customers of an internet service provider initiate new accounts at the average rate of 10 accounts per day.

(a) What is the probability that more than 8 new accounts will be initiated today?

**Solution.** (a) New account initiations qualify as rare events because no two customers open accounts simultaneously. Then the number $X$ of today's new accounts has Poisson distribution with parameter λ = 10.

$P(X > 8) = 1 - F_X(8) = 1 - 0.333 = 0.667$



# Poisson distribution

**Example 3.22** (New accounts). Customers of an internet service provider initiate new accounts at the average rate of 10 accounts per day.

(b) What is the probability that more than 16 accounts will be initiated within 2 days?

**Solution.** (b) The number of accounts, $Y$, opened within 2 days does **not** equal $2X$. Rather, $Y$ is another Poisson random variable whose parameter equals 20.

The parameter is the average number of rare events, which, over the period of two days, doubles the one-day average.

# Poisson distribution

**Example 3.22** (New accounts). Customers of an internet service provider initiate new accounts at the average rate of 10 accounts per day.

(b) What is the probability that more than 16 accounts will be initiated within 2 days?

**Solution.** (b) The number of accounts, $\mathcal{Y}$, opened within 2 days does **not** equal $2\mathcal{X}$. Rather, $\mathcal{Y}$ is another Poisson random variable whose parameter equals 20.

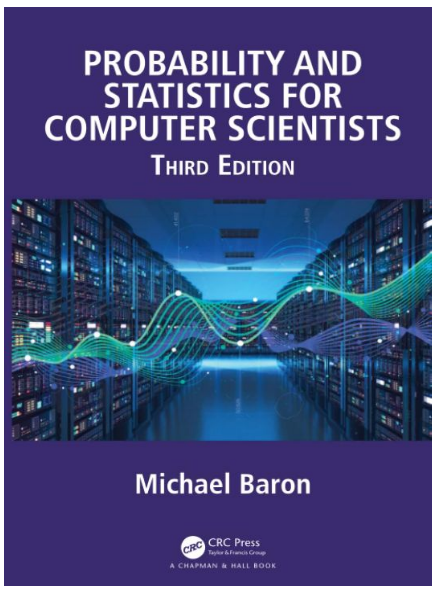
Again, using Table A3: $P(\mathcal{Y} > 16) = 1 - F_{\mathcal{Y}}(16) = 1 - 0.221 = 0.779$



# References

WIKIPEDIA

Google images



# Appendix


THE GREEK ALPHABET
Α α
Alpha
Β β
Beta
Γ γ
Gamma
Δ δ
Delta
Ε ε
Epsilon
Ζ ζ
Zeta
Η η
Eta
Θ θ
Theta
Ι ι
Iota
Κ κ
Kappa
Λ λ
Lambda
Μ μ
Mu
Ν ν
Nu
Ξ ξ
Xi
Ο ο
Omicron
Π π
Pi
Ρ ρ
Rho
Σ σ ς
Sigma
Τ τ
Tau
Υ υ
Upsilon
Φ φ
Phi
Χ χ
Chi
Ψ ψ
Psi
Ω ω
Omega