2 week: Random variables

Random variable denotes a value that depends on the result of some random experiment. Some natural examples of random variables come from gambling and lotteries. There are two main classes of random variables that we will consider in this course. During this week we'll learn discrete random variables that take finite or countable number of values. Discrete random variables can be described by their distribution. We'll consider various discrete distributions, introduce notions of expected value and variance and learn to generate and visualize discrete random variables with Python.

Оглавление:

Random variable in casino

Random variable can be used to model some natural phenomena and it’s the main tool that we’ll use tu study our data. But What is the random variable (RV)?

Casino is the perfect place to perform the RV. Indeed, consider a roulette. There is a wheel with 37 sectors and on each sector has a number. Some sectors are red and some sectors are black. And there is a special sector which is called zero and it’s green.

People can make different bets. For example, one player can bet for a particular number, for ex: 23. And another player can bet, for ex: for even or for red.

- If the ball stops in red sector, she win $100
- Alice bets $100 on red
- If the ball stops in black sector, she loses $100
- In case of zero sector, she loses $50

Teacher don't give answer on : What will be exact payout of Alice after one spinning of the wheel? But you understand how this payout works in probabilistic terms. You can answer different probabilistic questions. For example, what is the probability that Alice wins some positive amount of money? You can find it by calculation.

So Alice payout is an example of random variable**. That is a variable which value depends on the result if some random experiment.** In this case, random experiment is the spinning of the roulette.

Random variable denotes some value that depends on the result of some random experiment. So to defined a RV, we have to describe random experiment in some way.

Of course, random variables are not limited to gambling. (Случайные величины не ограничиваются азартными играми.) We will use them to model a lot if different natural phenomenons. For example:

- How many people will finish this course? → I don't know exactly, but I propose some model that will involve random variable for this number.
- What will be your final grade? ⟶ It's also a RV cause we don't know in advance what will be this value.
- What will be the price of bitcoin tomorrow? ⟶ A lot of people want to know this value, but nobody can. It's an example of a RV as well
- What will be opening weekend box office for particular new film? ⟶ Again, a lot of people interested in this value and nobody knows the exact value before we conduct R, before we perform this film for the 1st weekend and get the actual value.

Examples of random variables

Random variable is a universal model of some quantity that we know with some level of uncertainty. And we eill assume that our data are realization of some random variable

So to study data, we have to study random variable first. We will do it using a very sim;e examples. First, let us consider a random experiment that we discussed previously. Coin tossing. Let us assume that we have a fair coin and we toss it several times.

1. Experiment: toss fair coin n times

Then we can consider different random variables associated with this experiment. For example, we can count how many H we get during this tossing or how many T or smth else.

X - number of H obtained

1.1. n = 1

outcome (all possible)	value of X
H	1
T	0

1.2. n = 2

outcome (all possible)	value of X
HH	2
HT	1
TH	1
TT	0

We can also consider non-fair coin that gives a H with some probability p. This is a very well known binary distribution.

2. Experiment: dice tossing

Y - square of obtained points

outcome (all possible)	Y
1	1
2	4
3	9
4	16
5	25
6	36

3. We have three different coin with different values:
    - $1
    - $0.5
    - $0.25

   Z - summary of two different coin

outcome	Z
$1 and $0.5	1.5
$1 and $0.25	1.25
$0.5 and $0.25	0.75

At first we have to introduce random experiment.

Mathematical definition of random variable

outcome (all possible)	value of X
HH	2
HT	1
TH	1
TT	0

In previous example, we can see that for each outcome we have the corresponding value of X. Mathematically, it means that we defined a **function** on the set of all outcomes. In this case, this function can be described by a phrase "number of H". But from mathematical point of view, this function is described just by this set of numbers (2, 1, 1, 0). And we can introduce different RV by specify different numbers (2, 1, 1, 0).

The only thing that we have to take into account is that **for each outcome, we have to specify exactly one value of the corresponding RV.**

Definition:

Let Omega be sample space of some random experiment.

*sample space mean the space of all outcomes in some random experiment

To define a random experiment, we also have to define some probability. For example, in out simple cases when omega is a finite set of all possible outcomes. Then to define probability, we can define a probability of each outcome. In more complex cases, we have to do it different way.

So we **defined a random variable if we defined a function from Omega to real number.**

X: Omega → R is random variable

w - result of random experiment (w is element of Omega)

X(w) - value of corresponding random variable

X - random variable

If we conduct the same experiment several times, for each time we have its own w(its own outcome) and its own value of random variable.

In this case, we can say that we sample our random variable. And d**uring this sample, we get just a sequence of some numbers.** This sequence is called a **sample** from our random variable.

During our analysis we will assume that our data is sampled from some system of random variables.

Probability distribution and probability mass function (PMF)

we can describe random variable by using definition:

For each outcome defined each value, random variable take for this particular outcome.

However, it can be to complex to do it.

What is the probability that X = 0? We can answer by using table:

outcome (all possible)	value of X	P (fair coin)
HH	2	1/4
HT	1	1/4
TH	1	1/4
TT	0	1/4

- P(X = 0) = P({w belongs Omega | X(w) = 0}) = P({TT}) = 1/4
- P(X = 1) = P({w belongs Omega | X(w) = 1}) = P({HT, TH}) = 2/4
- P(X = 2) = P({w belongs Omega | X(w) = 2}) = P({HH}) = 1/4

*{w belongs Omega | X(w) = 0} = all outcomes of X(w) = 0 holds

We can use the table to answer on different probabilistic questions:

Name is called "Distribution of X"

x	0	1	2
P(X = x)	1/4	2/4	1/4

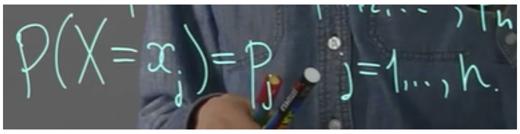
What is the probability that X is > or = 1?

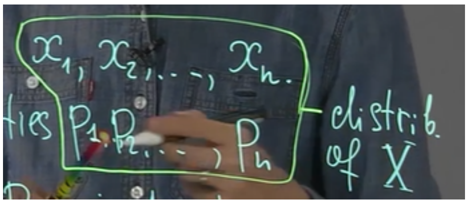
P(X≥1) = P(X=1 U X= 2) = 1/2 + 1/4 = 3/4

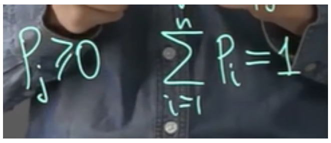
- X=1 and X= 2 are disjoint because when X=1 it could be equal 2.

Definition "Distribution of X"**:**

Len random variable of X takes values x1, x2, .., xn with probability p1, p2, .., pn.





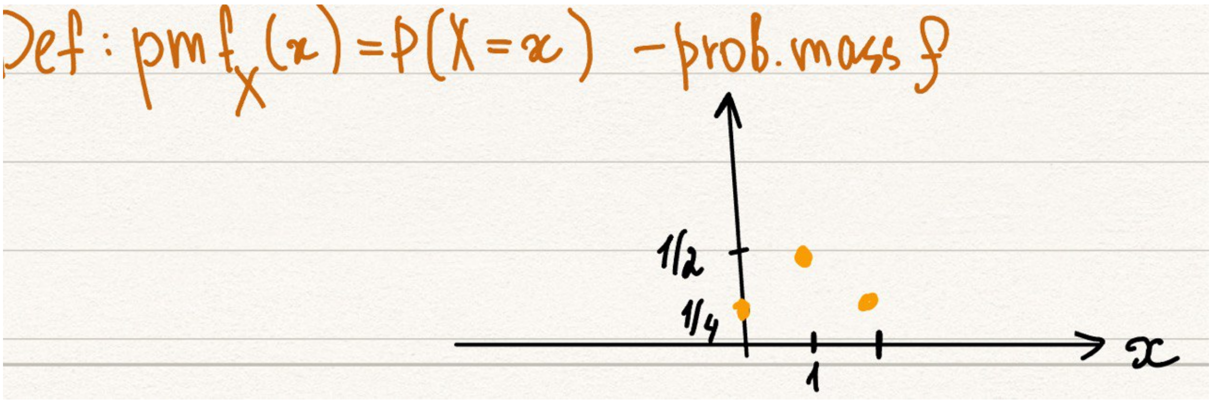


Each of number p is the probability of some event ⟶ p ≥ 0

Definition "Probability Mass Function**":**

For random variable which takes only finite number of values:

pmfX(x) = P(X = x)



Way to visualize probability distributions

Binomial distribution

— is a *distribution of random variable* that counts the number of H in n tossings of a coin that is not necessarily fair.

Experiment: toss non-fair coin n times

P(H) = p

P(T) = 1 - p

X - number of H (random variable)

Let assume that we have a sequence of independent trials. Each trial can result either in success or in failure. If we identify success with H and we say that the probability od success is p, then X is number of successful trials. This is a rather general model for different natural phenomenons.

Example: n = 5

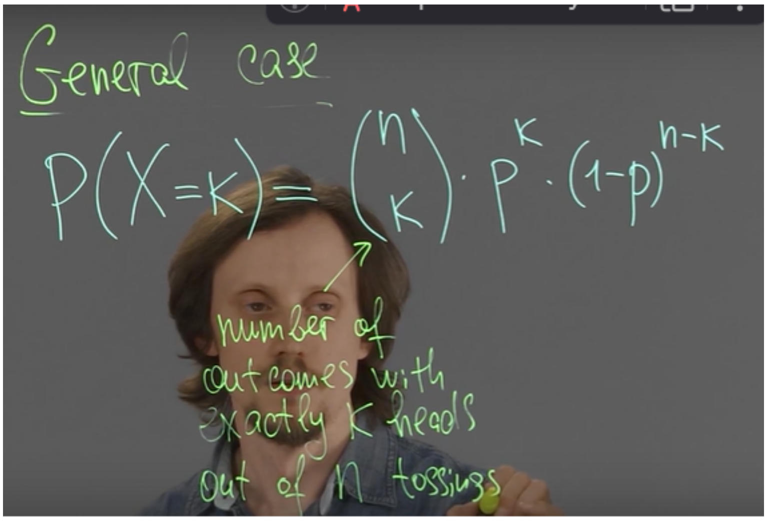
P(X = 0) = P({TTTTT}) = (1-p)^5

P(X=1) = P({HTTTT, THTTT, TTHTT, TTTHT, TTTTH}) = 5 * p^1 * (1 - p)^4

P(X= 2) = P({HHTTT, ...}) = (5 2) * p^2 * (1-p)^3

- use binomial coefficient

General case:



General case

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

number of outcomes with exactly k heads out of n tossings

- все возможные варианты → не важен порядок

k = 0, 1, .., n

If random variable has binomial distribution we will say that this random variable is binomial distributed **with parameters: n and p**

X ~ Bin(n, p)

Application of binomial distribution

Question 1

Biathlonist hits the target in 90% of shots. Find the probability that he hits exactly 3 out of 5 targets. Give the answer up to 4th decimal digit.

Биатлонист попадает в цель в 90% бросков. Найти вероятность того, что он поразит ровно 3 мишени из 5. Ответ дайте до 4-й десятичной цифры.

Solution:

X - Biathlonist hits the target

p = 0.9

q = 1-p = 0.1

n = 5

The probability that he hits exactly 3 out of 5 targets:

P(X = 3) = (5 3) * 0.9^3 * 0.1^2 = 0.0729

C = 5! / (3! 2!) = 10

The correct answer is: 0.0729

More exotic distributions

Question 1

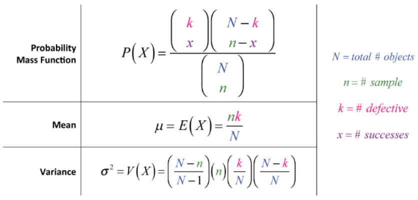
Student knows answers to 75 out of 100 questions in the exam. He picks 3 random questions of the 100 and answers. Three correct answers yield grade A, two correct answers yield grade B, 1 correct answer yields C and 0 yields D.

What's the probability of the student getting **A** grade?

Студент знает ответы на 75 из 100 вопросов экзамена. Он выбирает 3 случайных вопроса из 100 и отвечает. Три правильных ответа дают оценку A, два правильных ответа дают оценку B, 1 правильный ответ дает оценку C, а 0 дает оценку D. Какова вероятность того, что ученик получит оценку «отлично»?

Hint:

This problem leads to hypergeometric distribution. Try to derive the formula for it.



Probability Mass Function	$P(X)=\frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$
Mean	$\mu=E(X)=\frac{nk}{N}$
Variance	$\sigma^2=V(X)=\left(\frac{N-n}{N-1}\right)(n)\left(\frac{k}{N}\right)\left(\frac{N-k}{N}\right)$

$N = total\ \#\ objects$

$n = \#\ sample$

$k = \#\ defective$

$x = \#\ successes$

Solution:

X - student solve 3 random questions correctly

What's the probability of the student getting **A** grade?

Omega = 100 questions

p = 0.75

q = 0.25

P(X = A) = (3 3) * (0.75)^3 * 0.25^0 = 0.41

The correct answer is: 0.41

Question 2

Under the same conditions, what's the probability of the student getting **B** grade.
Какова вероятность того, что при тех же условиях ученик получит оценку B?

Solution:

P(X = B) = (3 2) * (0.75)^2 * 0.25^1 = 0.42

The correct answer is: 0.42

Question 3

Under the same conditions, what's the probability of the student getting **C** grade.
Какова вероятность того, что при тех же условиях ученик получит тройку?

Solution:

P(X =C) = (3 1) * (0.75)^1 * 0.25^2 = 0.13

The correct answer is: 0.13

Question 4

Under the same conditions, what's the probability of the student getting **D** grade (provided that it's the lowest grade for the exam).
Какова вероятность того, что при тех же условиях студент получит оценку D (при условии, что это самая низкая оценка за экзамен).

Solution:

P(X =D) = (3 0) * (0.75)^0 ∗ 0.25^3 = 0.01

The correct answer is: 0.01

Question 5

Under the same circumstances, What's the probability that 2 out of 5 students will get A, 2 will get B and 1 will get C?

Какова вероятность того, что при тех же обстоятельствах 2 из 5 учащихся получат оценку «A», 2 — оценку «B» и 1 — оценку «C»?

Solution:

p = 0.75

q = 0.25

P(X = A) = (5 2) * (0.75)^2 ∗ 0.25^3 = 10 ∗ 0.5625 ∗ 0.015625 = 0.08789062

P(X = B) = (3 2) * (0.75)^2 ∗ 0.25^3 = 3 ∗ 0.5625 ∗ 0.015625= 0.02636719

P(X = C) = (1 1) * (0.75)^1 ∗ 0.25^4 = 0.75 ∗ 0.00390625 = 0.00292969

x	A	B	C
P(X = x)	0.08789062	0.02636719	0.00292969

P(X = A U X = B U X = C) = 0.08789062 + 0.02636719 + 0.00292969 = 0.13

The correct answer is: 0.13

Expected value(математическое ожидание) of random variable. Motivation and definition

Probability distribution gives us a lot of information about random variable. But sometimes we want to answer the following question: **What value a random variable takes on average?**

The simpliest formula doesn’t work here (x1 + … xn)/n because we have to take into account that some values of random variable are taken with large probabilities and other values are taken with low probabilities. It’s natural that those values with large probabilities will contribute more to the average value. The easiest way to think about average of r.v. is to think about average win, average payout in some kind of lottery.

Example: Bob plays lottery

x	5	-1
P(X = x)	0.1	0.9

The question we’re interested in what is an average payout of Bob?

Assume that Bob plays 10 000 times:

- wins: 10 000 ∗ 0.1 = 1 000
- loses: 10 000 ∗ 0.9 = 9 000

Payout:

- wins: 1000 ∗ 5 = 5 000
- loses: 9 000 * (-1) = -9000

Summation: 5 000 - 9 000 = -4 000

Payout (average) per 1 game: -4000 / 10 000 = -0.4

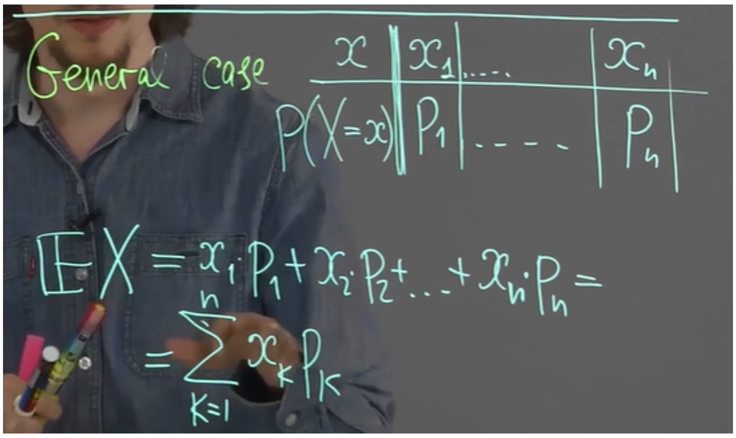
Basically, it means that at every game, Bob loses 49 cents by playing in this lottery

We can this value in a different way:

Expected (average) payout per 1 game:

(10 000 ∗ 0.1 ∗ 5 + 10 000 ∗ 0.9 * (-1)) / 10 000 = 10 000 * (0.1 ∗ 5 + 0.9 * (-1)) / 10 000 = 0.1 ∗ 5 + 0.9 * (-1) = -0.4

Expected value of random variable X:



General case

x	x_1	....	x_n
$P(X=x)$	P_1	....	P_n

$$\mathbb{E}X = x_1 \cdot P_1 + x_2 \cdot P_2 + \ldots + x_n \cdot P_n = \sum_{k=1}^{n} x_k P_k$$

Expected value example and calculation

Remember Alice whose bets $100 on red in roulette game. Let us find her expected payout.

X - Alice’s payout

37 sectors:

- 18 red
- 18 black
- 1 zero

37 sectors	payout	P
18 red	100	18/37
18 black	-100	18/37
1 zero	-50	1/37

EX = 100 ∗ 18/37 - 100 ∗ 18/37 - 50 ∗ 1/37 = -1.35

Now we see that every time when Alice bets $100 on red, she loses on average $1.35. This is not so large of money probably, but it’s important that it’s negative amount. So the more Alive plays this game , the more she lose on average. This is what makes casinos profitable.

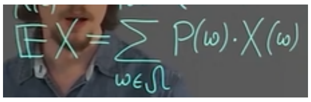
Another way calculate EX:

Omega = {0, 1, 2, 3, .., 36}

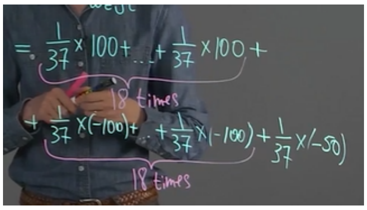
X(0) = -50

X(1) = 100 (1 is red)

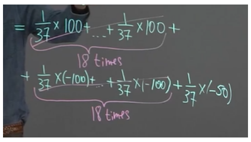
X(2) = - 100 (2 is black)



$$\mathbb{E}X = \sum_{\omega \in \Omega} P(\omega) \cdot X(\omega)$$



$$= \underbrace{\frac{1}{37} \times 100 + \ldots + \frac{1}{37} \times 100}_{18 \text{ times}} + \underbrace{\frac{1}{37} \times (-100) + \ldots + \frac{1}{37} \times (-100)}_{18 \text{ times}} + \frac{1}{37} \times (-50)$$



$$= \underbrace{\frac{1}{37} \times 100 + \ldots + \frac{1}{37} \times 100}_{18 \text{ times}} + \underbrace{\frac{1}{37} \times (-100) + \ldots + \frac{1}{37} \times (-100)}_{18 \text{ times}} + \frac{1}{37} \times (-50)$$

PMF practice

Question 1

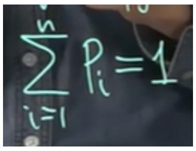
Random variable X is given by its PMF:

X	-2	-1	0	1	3

P	0.1	p	0.4	2p	0.2

Find p. Enter the exact value below with two decimal places or as an ordinary irreducible fraction.

Solution:



0.1 + p + 0.4 + 2p + 0.2 = 1

0.7 + 3p = 1

3p = 0.3

p = 3/10 *1/3 = 1/10 = 0.1

The correct answer is: 1/10

Question 2

Under the same conditions, find EX. Enter the exact value below with two decimal places or as an ordinary irreducible fraction.

Solution:

EX = -2*0.1 + -1*0.1 + 0*0.4 + 1 * 0.2 + 3 *0.2 = -0.2-0.1+0.2+0.6=0.5 = 1/2

The correct answer is: 1/2

Question 3

Find E(X−EX)^2. Enter the exact value below with two decimal places or as an ordinary irreducible fraction.

EX = 0.5 = 1/2

Solution:

X	-2	-1	0	1	3
P	**0.1**	**0.1**	**0.4**	**0.2**	**0.2**
EX = 0.5	0.5	0.5	0.5	0.5	0.5
X−EX	-2.5	-1.5	-0.5	0.5	2.5
(X−EX)^2	6.25	2.25	0.25	0.25	6.25
E(X−EX)^2	0.225	0.25	0.1	0.05	1.25

E(X−EX)^2 = **0.1 *** 6.25 + **0.1 *** 2.25 + **0.4 *** 0.25 + **0.2 *** 0.25 + **0.2 *** 6.25 = 0.625 + 0.225 + 0.1 + 0.05 + 1.25 = 2.25 =9/4

The correct answer is: 9/4

Expectation practice

There are 6 white balls in a box and 4 black balls. One random ball is taken. Find expected value of the indicator of event "the ball is black", i.e. the random variable that equals 1 if the ball is black and 0 otherwise.

В коробке 6 белых шаров и 4 черных шара. Берется один случайный шар. Найти математическое ожидание показателя события «шар черный», т.е. случайной величины, равной 1, если шар черный, и 0 в противном случае.

Solution:

X - шар чёрный

x	1 (шар черный)	0 (шар белый)
P(X = x)	4/10	6/10

EX = 1*4/10 + 0 * 6/10 = 4/10 = 2/5 = 0.4

The correct answer is: 2/5

Expected value exercises

Question 1

What’s the expectation of a random variable that accepts values from -3 to 3 with probabilities set by PMF:

X	-3	-2	-1	0	1	2	3
P(X= x)	1/14	1/7	1/7	2/7	2/7	0	1/14

Solution:

EX = -3 ∗ 1/14 -2* _1/7 -1_ * 1/7 + 0 ∗ 2/7 +1 ∗ 2/7 + 2∗0 + 3 ∗ 1/14 = -3/14 -2/7 - 1/ 7 + 2/7 + 3/14 = -1/7

The correct answer is: -1/7

Question 2

Find the expectation of the product of values on two independently rolled fair dice.

Найдите математическое ожидание произведения значений на двух независимо брошенных игральных костях.

Solution:

1. Define Variables

- Let X be the value of the first die.
- Let Y be the value of the second die.

2. Independence

Since the dice are independent, the expected value of the product is the product of the expected values:
E[XY] = E[X] * E[Y]

3. Expected Value of a Single Die

The expected value of a single fair die is:
E[X] = E[Y] = (1 + 2 + 3 + 4 + 5 + 6) / 6 = 3.5

4. Calculate the Expectation

E[XY] = E[X] * _E[Y] = 3.5_ * 3.5 = **12.25 = 49/4**

The correct answer is: 49/4

Question 3

A player rolls a fair dice and gets "$3 multiplied by the value on the dice" in the case of an even outcome and loses "$4 multiplied by the value on the dice" in the case of an odd outcome. What's his average payout?

Игрок бросает честную игральную кость и получает «3 доллара, умноженные на значение на кубике» в случае четного результата, и теряет «4 доллара, умноженное на значение на кубике» в случае нечетного результата. Какова его средняя зарплата?

Solution:

x	1	2	3	4	5	6
payout	-4	3	-4	3	-4	3
P(X = x)	1/6	1/6	1/6	1/6	1/6	1/6
X = x	-4	6	-12	12	-20	18

EX = 1/6 * (-4 + 6 - 12 + 12 - 20 + 18) = 0

The correct answer is: 0

Question 4

A player is offered to play against croupier who rolls 6 dice simultaneously. If there's a "street" (all 6 digits are yield on different dice), the croupier wins. Otherwise, the player wins. The player is offered to bet 60$ against croupier's 1$ (that is, the player loses $60 if the “street” combination occur and wins $1 otherwise). Will this game be profitable to the player in the long run?

Игроку предлагается сыграть против крупье, который одновременно бросает 6 кубиков. Если есть «улица» (все 6 цифр выпадают на разных кубиках), выигрывает крупье. В противном случае игрок выигрывает. Игроку предлагается поставить 60$ против 1$ крупье (то есть игрок теряет 60$, если выпадает «уличная» комбинация, и выигрывает 1$ в противном случае). Будет ли эта игра прибыльной для игрока в долгосрочной перспективе?

Solution:

1. Вероятность, что выиграет крупье (т.е. выпадут разные цифры на все кубиках)

1/6 ∗ 1/6 ∗ 1/6 ∗ 1/6 ∗ 1/6 ∗ 1/6 = 1/46 656 = 0.00002143

2. Вероятность, что выиграет игрок:

1 - 0.00002143 = 0.9997857

Distribution of X:

x	-60	1
P(X = x)	0.00002143	0.9997857

3. Выигрыш игрока:

0.00002143 * (-60) + 0.9997857 ∗ 1 = -0.0012858 + = 0.9997857

The correct answer is: yes

Expected value as best prediction

One of the basic problems in ML is prediction problem. We have some info, we know smth and we want to use this info to predict some quantity that we don’t know.

Example:

let us consider that we’re Cafe amd we sell pancakes. What we’re interested in is: How maby pancakes should we prepare? It’s determined by the demand. How many pancakes could we sell at a particular day?

Of course, this is a random value, because on different days we have different clients and some of them can be hungry and some of them can be non-hungry and some of them like pancakes and so on.

We have to use random variables to model this demand. What is the best way to predict the value of this random variable?

X - demand of pancakes (random variable)

x̂ - predictions
Let us assume that we choose this prediction once and forever. How can we measure the quality of our prediction?

X - x̂

This difference can be either positive or negative but both is bad for us. For example, if we underestimate the demand, it means if we don’t have enough pancakes, some clients who want pancakes can’t get pancakes and it’s bad for us because lose some money on these clients. On the other hand, if we prepared more pancakes than we can sell, it means that we again lose money.

Better use in some way:

(X - x̂)^2 - **squared error loss**

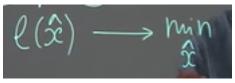
For example, if didn’t prepare enough pancakes , we lose some money because we cannot sell the pancakes to the people who wang them and also these people can become angry an they can give us bad marks on some website sites and so on. So it’s good idea sometimes to consider our penalty as to be square of our error. This is actually a popular loss function in ML which is called squared error.

At different days we have different value of this loss. For example, at one day, our prediction can be good if we are lucky enough.

We want make our prediction are good in average. It means that we have to find EX of loss function.

E(X - x̂)^2 = l(x̂)

This value is not a random variable because we have EX. This is just a number that depends on **x̂.**



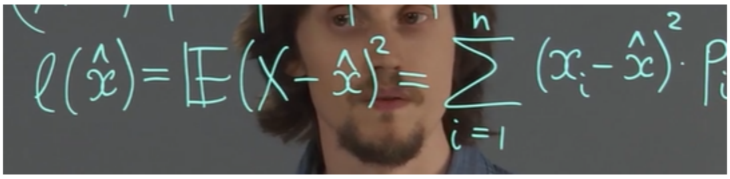
$\ell(\hat{x}) \longrightarrow \min_{\hat{x}}$

How can we find such optimal **x̂?**

So firstly let us consider the distribution of random variable X.

X	x1	x2	...	xn
P(X = x)	p1	p2	...	pn

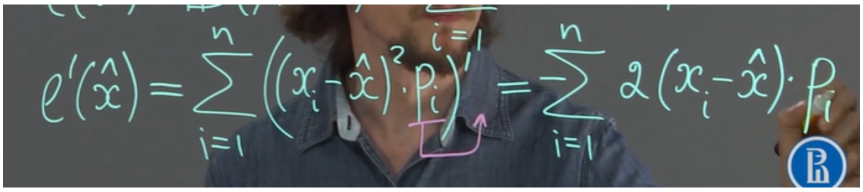
(X - x̂)^2	(x1 - **x̂)^2**	(x2 - **x̂)^2**	...	(xn - **x̂)^2**



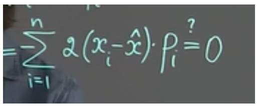
$$l(\hat{x}) = \mathbb{E}(X-\hat{x})^2 = \sum_{i=1}^{n} (x_i-\hat{x})^2 \cdot p_i$$

So we have to minimize **l(x̂)**

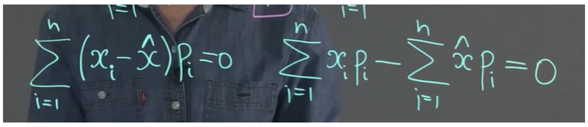
In calculus to find minimum or maximum of sum function,we can use derivative.



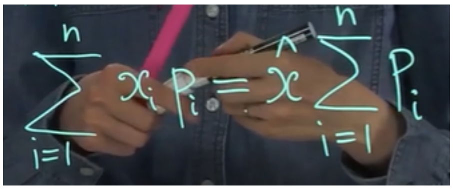
$$l'(\hat{x}) = \sum_{i=1}^{n} \left((x_i-\hat{x})^2 \cdot p_i\right)' = \sum_{i=1}^{n} 2(x_i-\hat{x}) \cdot p_i$$



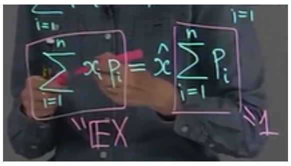
$$\sum_{i=1}^{n} 2(x_i-\hat{x}) \cdot p_i \overset{?}{=} 0$$



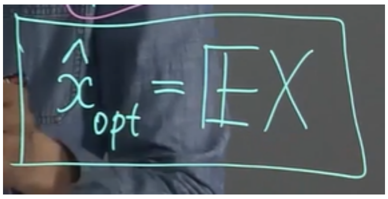
$$\sum_{i=1}^{n} (x_i-\hat{x}) p_i = 0 \qquad \sum_{i=1}^{n} x_i p_i - \sum_{i=1}^{n} \hat{x} p_i = 0$$



$$\sum_{i=1}^{n} x_i p_i = \hat{x} \sum_{i=1}^{n} p_i$$



$$\underbrace{\sum_{i=1}^{n} x_i p_i}_{\mathbb{E}X} = \hat{x} \underbrace{\sum_{i=1}^{n} p_i}_{1}$$



$$\boxed{\hat{x}_{opt} = \mathbb{E}X}$$

💡 **If you want to predict the value of random variable and you have squared error loss than you can choose EX of this variable as your prediction.**

It’s really important result, **if you want to predict the value of random variable and you have squared error loss than you can choose EX of this variable as your prediction.** We will see in the future that if we choose a different loss, we will get a different optimal value. But squared error loss is very popular to good mathematical properties.

Variance(Дисперсия) of random variable. Motivation and definition

The measure how far our random variable deviates from the EX we will use Variance.

Example:

Return to Bob whose plays lottery. Bets $1 and can win $5.

x	5	-1
P(X = x)	0.1	0.9

x - EX	5.4	-0.6
(x - EX)^2	29.16	0.36

EX = -0.4

How far the values x from its EX?

x - EX

We see that x - EX again is a random variable.

We're interested in how far X from its EX **on average**

E(x - EX) = 5.4 *(0.1) + 0.9*(-0.6) = 0.54 - 0.54 = 0

We can solve problem with sign which we can get in EX by square:

E((x - EX)^2) = VarX

E((x - EX)^2) = 29.16 ∗ 0.1 + 0.36 ∗ 0.9 = 3.24

This number show the average deviation of the value x from its squares EX.

This thing is actually the same as we discussed before when we were discussing a square error.

Definition:

X - random variable

Variance of X(VarX) = E((X - EX)^2)

Variance skill test

Question 1

What's the variance of a random variable that takes values from -3 to 3 with probabilities set by PMF:

X	-3	-2	-1	0	1	2	3
P	1/14	1/7	1/7	2/7	2/7	0	1/14

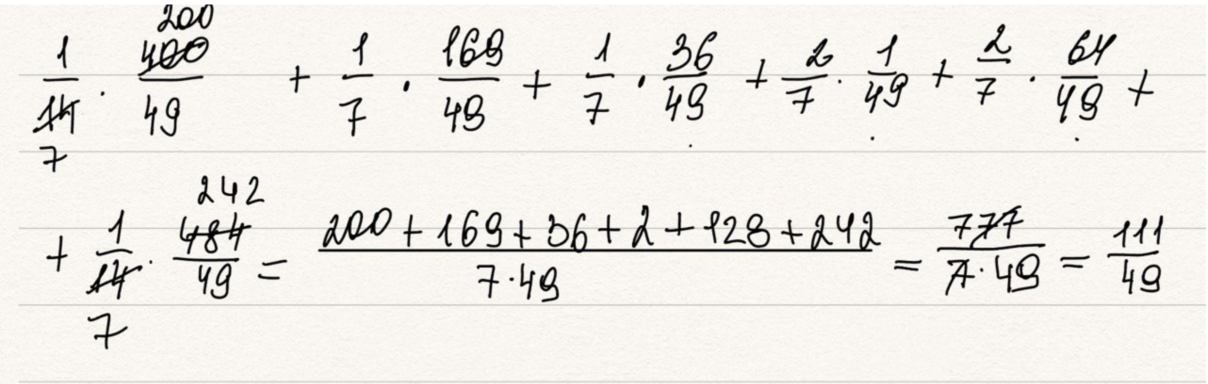
Solution:

X(VarX) = E((X - EX)^2)

EX = -3/14 -2/7 -1/7 +2/7 + 3/14 = -1/7

X	-3	-2	-1	0	1	2	3
P(X = x)	1/14	1/7	1/7	2/7	2/7	0	1/14
EX	-1/7	-1/7	-1/7	-1/7	-1/7	-1/7	-1/7
X - EX	-20/7	-13/7	-6/7	1/7	8/7	15/7	22/7
(X - EX)^2	400/49	169/49	36/49	1/49	64/49	225/49	484/49

E((X - EX)^2) = 1/14 ∗ 400/49 + 1/7 ∗ 169/49 + 1/7 ∗ 36/49 + 2/7 ∗ 1/49 + 2/7 ∗ 64/49 + 1/14 ∗ 484/49 =



$$\frac{1}{\underset{7}{\cancel{14}}}\cdot\frac{\overset{200}{\cancel{400}}}{49}+\frac{1}{7}\cdot\frac{169}{49}+\frac{1}{7}\cdot\frac{36}{49}+\frac{2}{7}\cdot\frac{1}{49}+\frac{2}{7}\cdot\frac{64}{49}+$$

$$+\frac{1}{\underset{7}{\cancel{14}}}\cdot\frac{\overset{242}{\cancel{484}}}{49}=\frac{200+169+36+2+128+242}{7\cdot 49}=\frac{\cancel{777}}{\cancel{7}\cdot 49}=\frac{111}{49}$$

The correct answer is: 111/49

Question 2

Find variance of the value on a fair dice rolled one time.

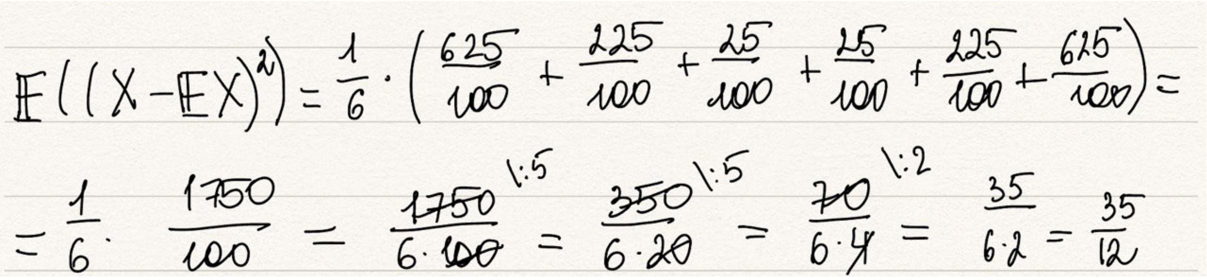
Solution:

X(VarX) = E((X - EX)^2)

EX = 1/6 + 2/6 + 3/6 + 4/6 + 5/6 + 6/6 = 21/6 = 3, 5

x	1	2	3	4	5	6
P(X = x)	1/6	1/6	1/6	1/6	1/6	1/6
EX	**21/6**	**21/6**	**21/6**	**21/6**	**21/6**	**21/6**
X - EX	-2.5	-1.5	-0.5	0.5	1.5	2.5
(X - EX)^2	6.25 = 625/100	2.25 = 225/100	0.25 = 25/100	25/100	225/100	625/100

E((X - EX)^2)



$$E((X-EX)^2) = \frac{1}{6}\cdot\left(\frac{625}{100}+\frac{225}{100}+\frac{25}{100}+\frac{25}{100}+\frac{225}{100}+\frac{625}{100}\right) =$$

$$= \frac{1}{6}\cdot\frac{1750}{100} = \frac{\cancel{1750}}{6\cdot\cancel{100}}\Big\backslash :5 = \frac{\cancel{350}}{6\cdot\cancel{20}}\Big\backslash :5 = \frac{\cancel{70}}{6\cdot\cancel{4}}\Big\backslash :2 = \frac{35}{6\cdot 2} = \frac{35}{12}$$

The correct answer is: 35/12

More skill tests on Expectation and Variance (не решила)

Question 1

There are 100 lottery tickets in a box. 10 of them win $1 each, 5 win 5$ each, 3 win 10$ each and 2 win $100 each. Player takes one ticket from the box. What's expectation of the amount he wins? Enter the exact value below (e.g., 13/28 or 0.12):

В коробке 100 лотерейных билетов. 10 из них выигрывают по 1 доллару каждый, 5 выигрывают по 5 долларов каждый, 3 выигрывают по 10 долларов каждый и 2 выигрывают по 100 долларов каждый. Игрок достает из коробки один билет. Каковы ожидания суммы, которую он выиграет?

Solution:

n = 100

X	1	5	10	100
P(X = x)	10/100	5/100	3/100	2/100

EX = 10/100 + 25/100 + 30/100 + 200/100 = 265/100 = 53/20 = 2.65

The correct answer is: 53/20

Question 2

Under the same conditions lottery organizer charges $3 for an attempt to pick a ticket. After each attempt ticket returns to the box. What's expectation of the organizer's profit after one attempt? Enter the exact value below (e.g., 13/28 or 0.12):

При тех же условиях за попытку подобрать билет организатор лотереи взимает 3 доллара. После каждой попытки билет возвращается в коробку. Какова ожидаемая прибыль организатора после одной попытки?

Solution:

3 - 2.65 = 0.35 = 7/20

The correct answer is: 7/20

Question 3

There are 6 white balls in a box and 4 black balls. Two random balls are taken. Find expected value of the number of black balls among them. Enter the exact value below (e.g., 13/28 or 0.12):

В коробке 6 белых шаров и 4 черных шара. Берутся два случайных шара. Найдите математическое ожидание количества черных шаров среди них.

Solution:

X	2
P(X = x)	6/10

EX black= 2 ∗ 6/10 = 12/10 = 4/5

The correct answer is: 4/5

Question 4

Discrete random variable takes only 2 values x and y, x<y.

P(X=x)=0.2, EX=2.6,VarX=0.64

Find P(X=y).

Solution:

X	x	y
P(X = x)	0.2	?

0.2 + y = 1

P(X = y) = 0.8 = 8/10 = 4/5

The correct answer is: 4/5

Question 5

Under the same conditions find x . Enter the value below:

Solution:

The correct answer is: 1

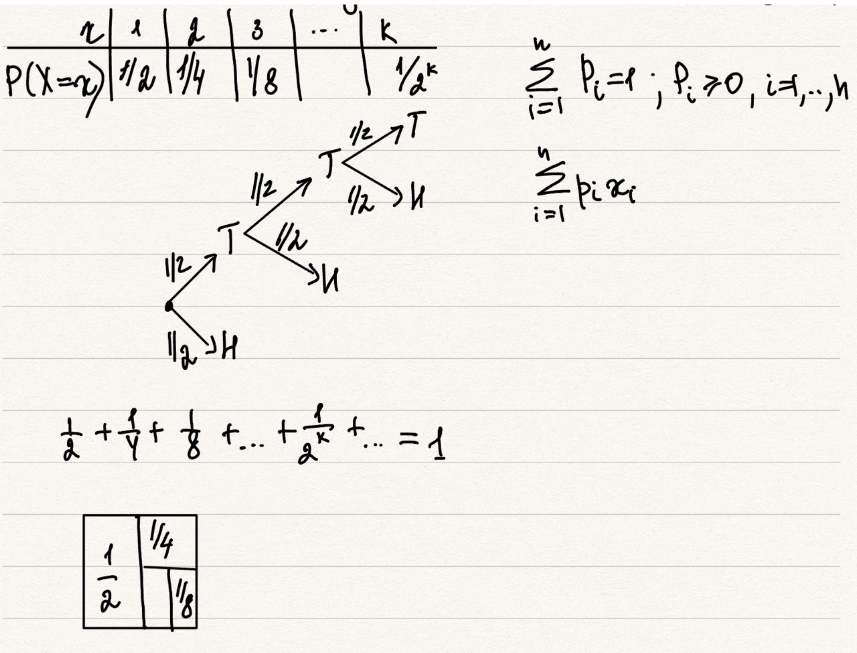
Discrete random variables with infinite number of values

So far, we can see that there are random variable with a finite number of possible values. However, it’s possible to deak with some phenomena in which the corresponding variable can take infinite number of values.

Experiment: toss a fair coin until 1st head

X - number of tossings until 1st H (including H) (random variable)

x	1	2	3		k
P (X= x)	1/2	1/4	1/8		1/2^k



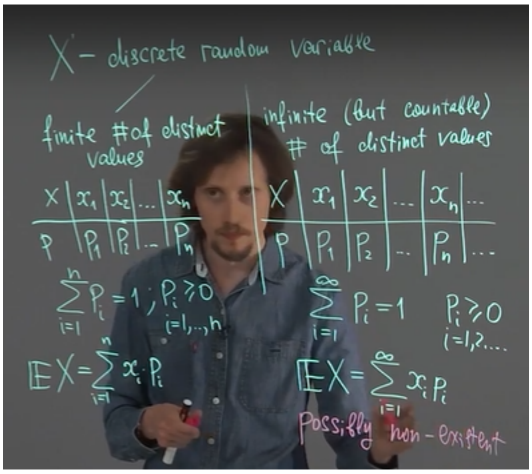
x	1	2	3	...	k
$P(X=x)$	1/2	1/4	1/8		$1/2^k$

$$\sum_{i=1}^{n} p_i = 1 \; ; \; p_i \geq 0, \; i=1,..,n$$

$$\sum_{i=1}^{n} p_i x_i$$

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots + \frac{1}{2^k} + \ldots = 1$$

- геометрическая прогрессия



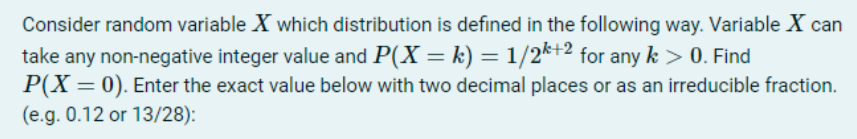
X – discrete random variable
finite # of distinct values
infinite (but countable) # of distinct values
X | x1 | x2 | ... | xn
P | P1 | P2 | ... | Pn
$\sum_{i=1}^{n} P_i = 1$; $P_i \geq 0$ $i=1,..,n$
$EX = \sum_{i=1}^{n} x_i P_i$
X | x1 | x2 | ... | xn | ...
P | P1 | P2 | ... | Pn | ...
$\sum_{i=1}^{\infty} P_i = 1$ $P_i \geq 0$ $i=1,2,...$
$EX = \sum_{i=1}^{\infty} x_i P_i$
Possibly non-existent

infinite but countable sets of sequences

> Know that in contrast with finite sums, infinite sum can be non existent ans EX.

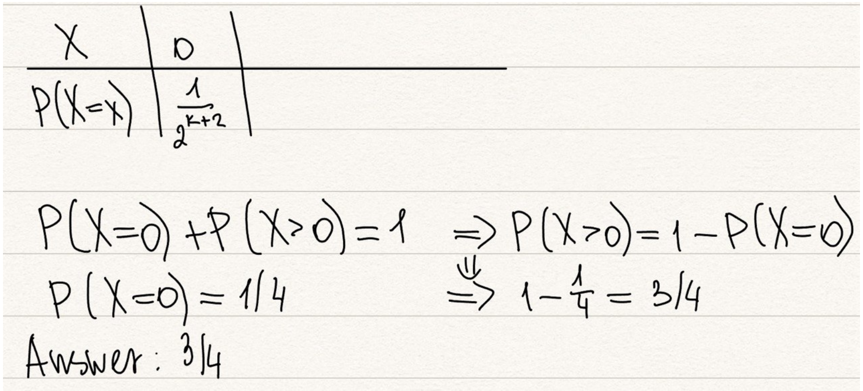
Random variables and geometric series

Question 1



Consider random variable X which distribution is defined in the following way. Variable X can take any non-negative integer value and $P(X=k) = 1/2^{k+2}$ for any $k > 0$. Find $P(X=0)$. Enter the exact value below with two decimal places or as an irreducible fraction. (e.g. 0.12 or 13/28):

Solution:



X	0
P(X=x)	$\frac{1}{2^{k+2}}$

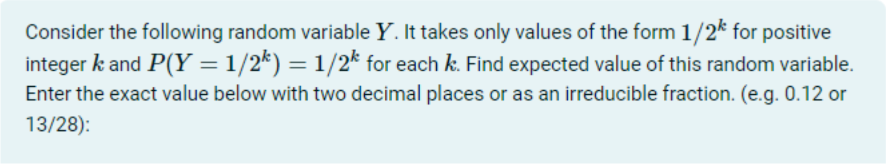
$P(X=0)+P(X>0)=1 \Rightarrow P(X>0)=1-P(X=0)$

$P(X=0)=1/4 \quad \overset{\Downarrow}{\Rightarrow} 1-\frac{1}{4}=3/4$

Answer: 3/4

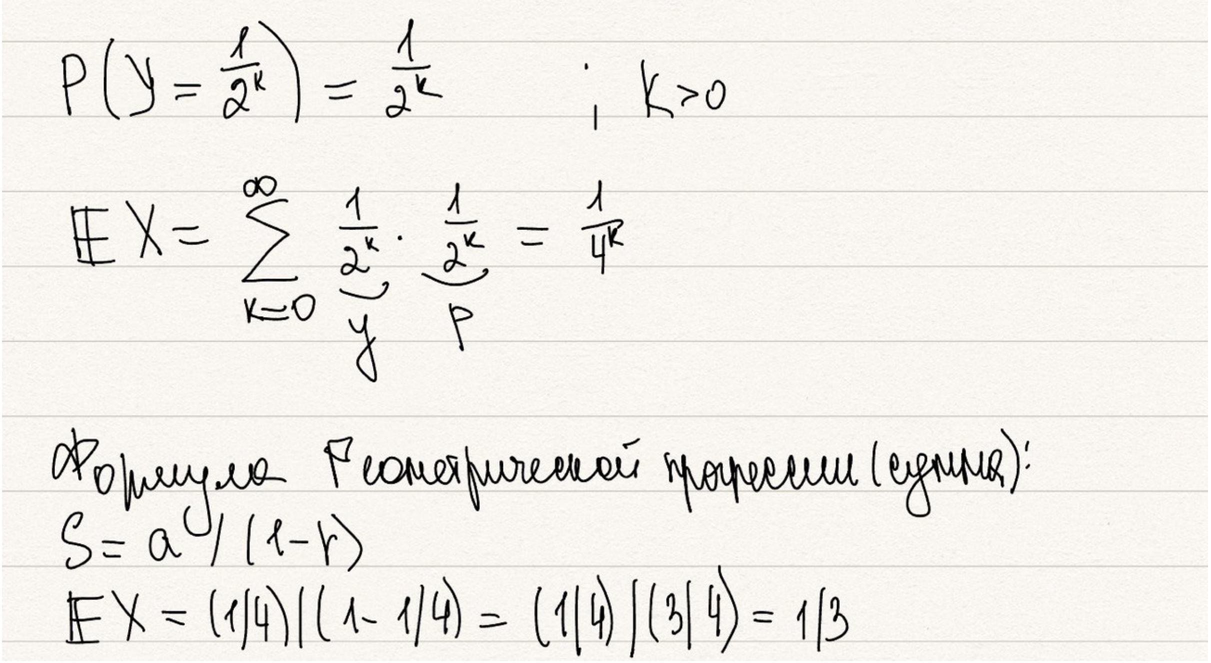
The correct answer is: 3/4

Question 2



Consider the following random variable Y. It takes only values of the form $1/2^k$ for positive integer k and $P(Y=1/2^k)=1/2^k$ for each k. Find expected value of this random variable. Enter the exact value below with two decimal places or as an irreducible fraction. (e.g. 0.12 or 13/28):

Solution:



$P\left(Y=\frac{1}{2^k}\right)=\frac{1}{2^k} \quad ; \; k>0$

$$\mathbb{E}X=\sum_{k=0}^{\infty}\underbrace{\frac{1}{2^k}}_{y}\cdot\underbrace{\frac{1}{2^k}}_{P}=\frac{1}{4^k}$$

Формула геометрической прогрессии (суммы):

$S=a/(1-r)$

$\mathbb{E}X=(1/4)/(1-1/4)=(1/4)/(3/4)=1/3$

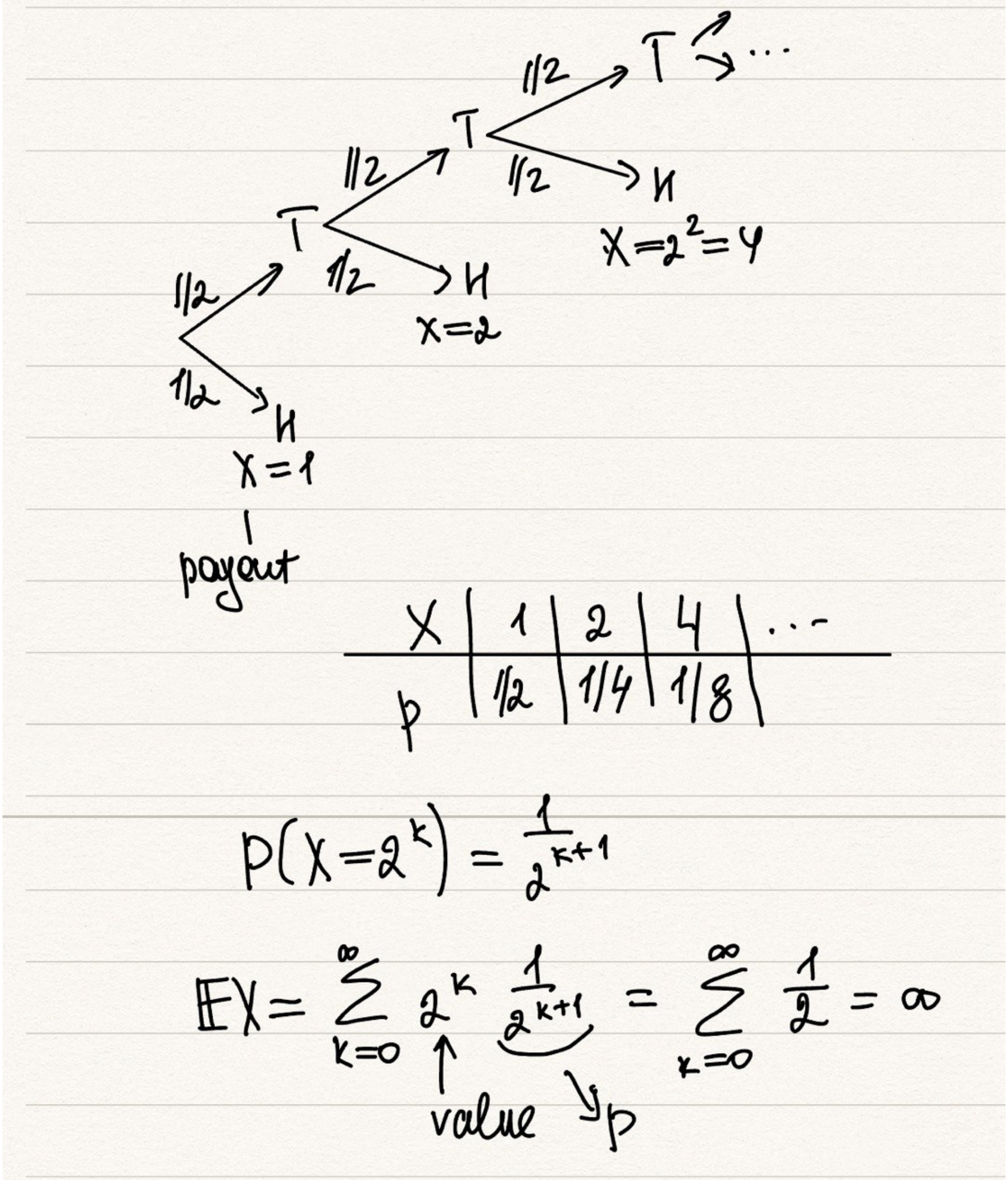
The correct answer is: 1/3

Saint Petersburg Paradox. Example of infinite expected value

Game:

I have a fair coin, I toss this coin. If I have H, I give you $1. If I have T, I toss it again. And if I have H, I give you $2. If I have T, I toss it again. Now, if I have H, I give you $4 and so on.

How much money do you want to give me fo opportunity to play this game with me?



X	1	2	4	...
p	1/2	1/4	1/8	

$$P(X=2^k)=\frac{1}{2^{k+1}}$$

$$\mathbb{E}X=\sum_{k=0}^{\infty} \underset{\substack{\uparrow\\ \text{value}}}{2^k} \underbrace{\frac{1}{2^{k+1}}}_{p} = \sum_{k=0}^{\infty} \frac{1}{2} = \infty$$

This expected value is you expected payout.

It’s obvious that You will play this game with me if the price of the ticket is lower tham your expected payout.

So you see that youe expected payout is infinity large. You’ll probably want to pay any amount of money to be able to play this game with me. However, the ral persons don;t behave in this way, in experiments people do not want to pay large amounts of money to play this game.

This is interesting example which doesn’t have expected value.

Geometric and Poisson distributions

Let us discuss a couple of distributions of **random variable with infinite number of possible values** that can be useful in practice.

1. **Geometric distribution**

Experiment: toss non-fair coin until 1st H appear

X - number of tossings (including H)

p(H) = p

TTTTT…TH

#T = k-1

P(X = k) = ((1 - p)^(k-1)) * p, where k is positive integer number (1, 2, ..)

EX = 1/p

VarX = (1- p)/p^2

2. **Poisson distribution**

- is distribution which can take any no negative integer from 0 to infinity.

It can be used to model a value that is equal to number of, for example visitors of your shop during some period of time (ex: during the day).

It’s possible that during some period of time you have zero visitors or theoretically it can be any positive number.

Another example, you’re call center and you want to know *the number of calling during some period of time.* Again, you don’t know in advance how many people will call you during some period of time.

$$\lambda > 0$$

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

For example, there exists so called Poisson regression.

Geometric distribution practice

Question 1

Find Expectation of Geometric distribution with probability of success equal to 0.1. Enter the value:

Hint:

Use the definition of Expectation and calculate the sum emerging series as the derivative of the sum of geometric sequence.

Solution:

EX = 1/p = 1/0.1 = 10

The correct answer is: 10

Question 2

Find Variance of Geometric distribution with probability of success equal to 0.1. Enter the value:

Hint:

Find EX2 by definition. Calculate the sum emerging series as the derivative of the sum of the series emerging in previous task.

Solution:

VarX = (1- p)/p^2 = 0.9 / 0.1^2 = 0.9 / 0.01 = 90

The correct answer is: 90

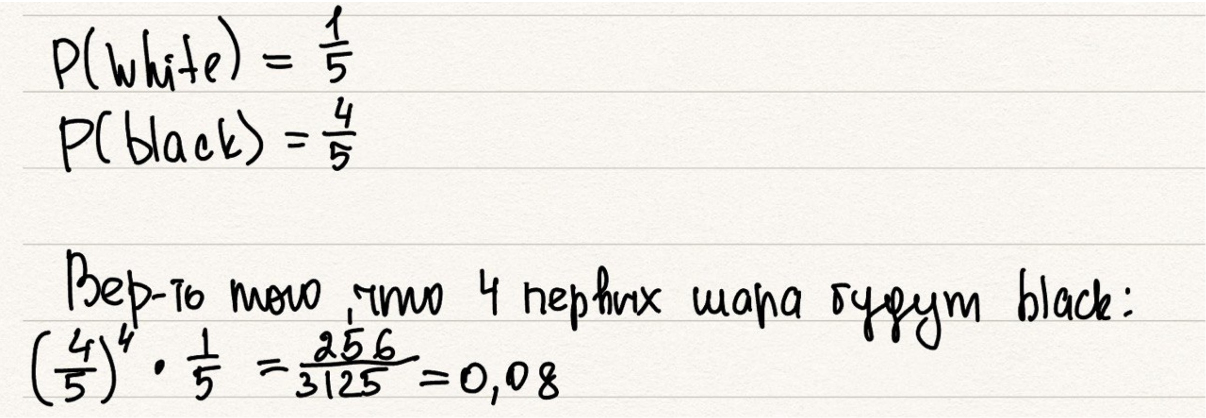
Geometric distribution skill test

Question 1

There are 4 black balls and 1 white ball in a box. The balls are taken out until the white ball occurs. Black balls return to the box. Find the probability that the experiment stops after exactly 5 tries. Enter the exact value below with two decimal places or as an irreducible fraction. (e.g. 0.12 or 13/28):

В коробке 4 черных шара и 1 белый шар. Шарики вынимаются до тех пор, пока не появится белый шарик. Черные шары возвращаются в коробку. Найти вероятность того, что эксперимент остановится ровно после 5 попыток. Введите ниже точное значение с двумя знаками после запятой или в виде несократимой дроби. (например, 0,12 или 13/28):

Solution:



$P(\text{white}) = \frac{1}{5}$

$P(\text{black}) = \frac{4}{5}$

Вер-ть того, что 4 первых шара будут black:

$\left(\frac{4}{5}\right)^4 \cdot \frac{1}{5} = \frac{256}{3125} = 0{,}08$

The correct answer is: 256/3125

Generating discrete random variables with Python

```
from numpy.random import choice

choice([1, 2, 4], p=[0.2, 0.5, 0.3]) #2
```

Здесь функция будет давать разное число каждый раз при запуске, но чаще всего будет выходить 2, так как её вероятность появления больше чем у других чисел

Напишем функцию частоты встречания элементов:

```
def count_frequencies(data, relative=False):
    counter = {}
    for element in data:
        if element not in counter:
        # get this element fpr the 1st time
            counter[element] = 1
        else:
            counter[element] += 1
    if relative:
        for element in counter:
            counter[element] = counter[element] / len(data)
    return counter

count_frequencies([1, 2, 2, 1]) #{1: 2, 2: 2}
#relative=True --> делим на кол-во элеметонов
count_frequencies([1, 2, 2, 1], relative=True) #{1: 0.5, 2: 0.5}
```

Увеличим кол-во срабатывания функции choice:

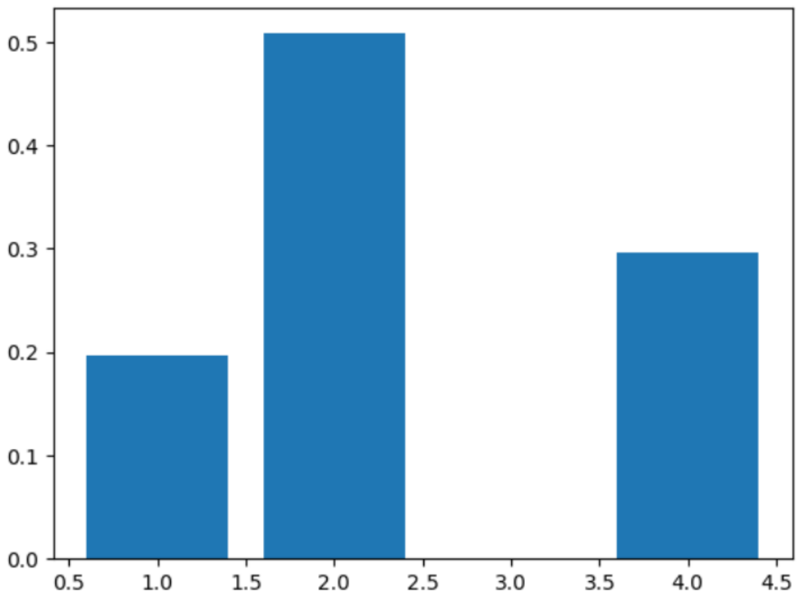
```
sample = [choice([1, 2, 4], p=[0.2, 0.5, 0.3]) for _ in range(1000)] #list

count_frequencies(sample) #{2: 501, 4: 292, 1: 207}
count_frequencies(sample, relative=True) #{2: 0.501, 4: 0.292, 1: 0.207}
```

Если увеличить кол-во итераций, то мы приблизимся к изначально заданным вероятностям в функции choice:

```
sample = [choice([1, 2, 4], p=[0.2, 0.5, 0.3]) for _ in range(10000)]

count_frequencies(sample) #{2: 5015, 4: 3042, 1: 1943}
count_frequencies(sample, relative=True) #{2: 0.5015, 4: 0.3042, 1: 0.1943}
```

```
#можно сразу задать нудное кол-во в переменное size и получить numpy.array
sample = choice([1, 2, 4], size = 10000, p = [0.2, 0.5, 0.3])
#array([2, 2, 2, ..., 2, 2, 2])

# Нарисуем PMF
import matplotlib.pyplot as plt
%matplotlib inline

freqs = count_frequencies(sample, relative=True)
plt.bar(list(freqs.keys()), list(freqs.values()))
```

```
<BarContainer object of 3 artists>
```



Numpy, scipy and matplotlib for generation and visualization of common distributions

```
from numpy.random import binomial

#каждый раз при запуске меняет значение
binomial(10, 0.3) #n -number of tossings, p - prob. of success #2


binomial(10, 0.3, size = 100)
```

```
array([3, 2, 3, 2, 1, 4, 3, 3, 0, 2, 2, 2, 4, 5, 1, 5, 5, 1, 4, 3, 3, 2,
       2, 3, 1, 1, 3, 4, 4, 4, 1, 3, 3, 4, 1, 3, 6, 2, 2, 2, 3, 3, 1, 4,
       4, 4, 4, 5, 2, 2, 4, 5, 3, 2, 0, 2, 6, 4, 1, 4, 2, 5, 2, 3, 3, 3,
       4, 4, 5, 2, 0, 3, 3, 5, 2, 4, 3, 2, 4, 4, 4, 1, 1, 4, 4, 3, 3, 4,
       3, 2, 1, 2, 4, 6, 2, 2, 1, 0, 4, 5])


def count_frequencies(data, relative=False):
    counter = {}
    for element in data:
        if element not in counter:
        # get this element fpr the 1st time
            counter[element] = 1
        else:
            counter[element] += 1
    if relative:
        for element in counter:
            counter[element] = counter[element] / len(data)
    return counter


sample = binomial(10, 0.3, size = 100)
count_frequencies(sample, relative=True)
#{2: 0.26, 1: 0.11, 4: 0.2, 5: 0.14, 3: 0.24, 7: 0.01, 6: 0.03, 0: 0.01}


from scipy.stats import binom
X = binom(10, 0.3)
X.pmf(1) #find probability at point
#0.12106082099999992
```

Мы можем видеть, что вероятность X.pmf(1)= 0.12106082099999992 достаточно близка к тому, что мы получили, используя функцию count_frequencies(sample, relative=True) , где 1: 0.11. Но чтобы результат был лучше, увеличим размер выборки.

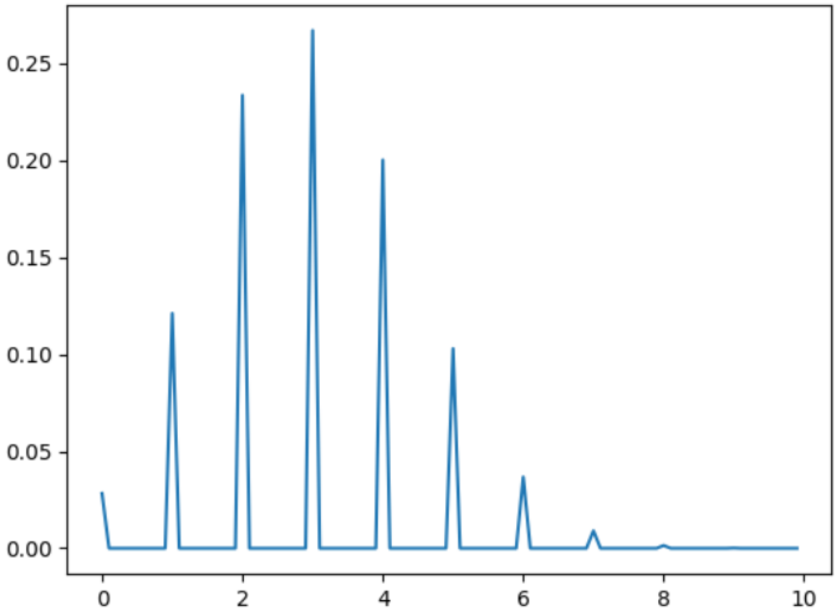
```
sample = binomial(10, 0.3, size = 1000)
count_frequencies(sample, relative=True)
{3: 0.253,
 5: 0.109,
 4: 0.217,
 0: 0.022,
 2: 0.222,
 1: 0.124,
 8: 0.006,
 7: 0.006,
 6: 0.041}
```

Теперь вероятности приблизительно одинаковые.

Давайте визиализируем:

```
import matplotlib.pyplot as plt
%matplotlib inline
import numpy as np
```

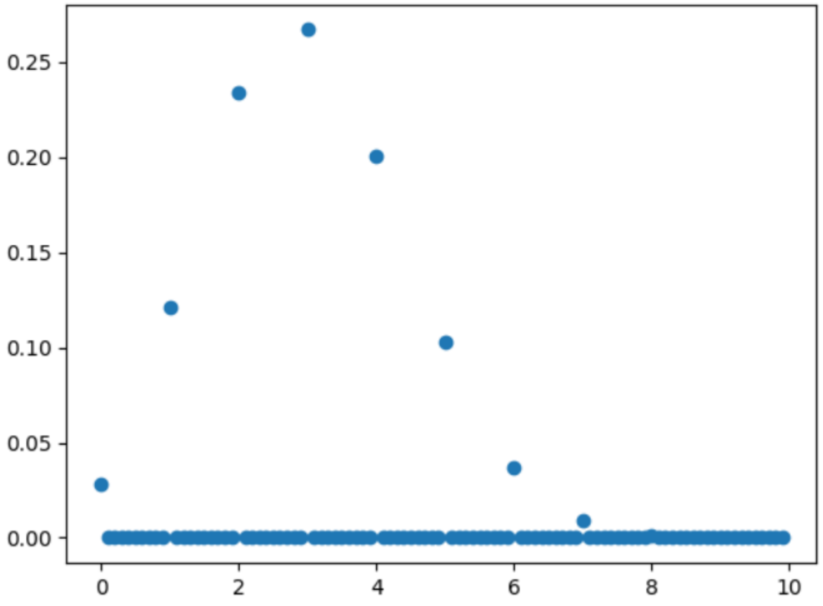
```
x = np.arange(0, 10, 0.1)
plt.plot(x, X.pmf(x))
```



Можно видеть что это дискретные величины, а непрерывные, поэтому при каких-то нецелых значениях не существует такой вероятности появления их.

```
plt.plot(x, X.pmf(x), 'o')
```

```
[<matplotlib.lines.Line2D at 0x1cea7b39d80>]
```

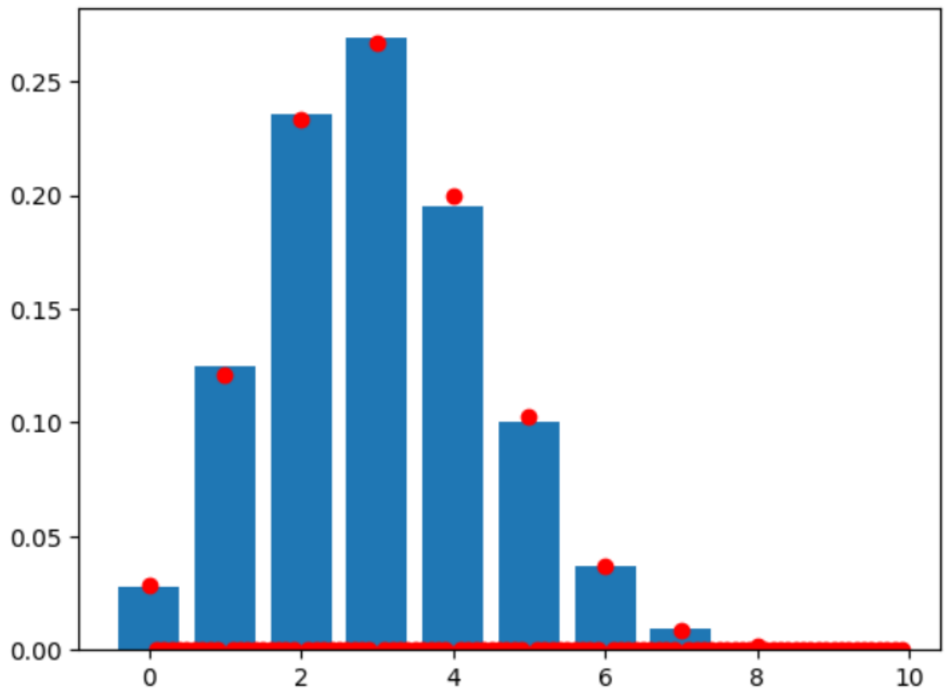


Также мы можем наложить график нашей написанной функции и график X.pmf(x). Можно увидеть, что вероятности очень близки:

```
sample = binomial(10, 0.3, size = 10000)
freqs = count_frequencies(sample, relative=True)
```

```
plt.plot(x, X.pmf(x), 'o', color = 'red')
plt.bar(list(freqs.keys()), list(freqs.values()))
```

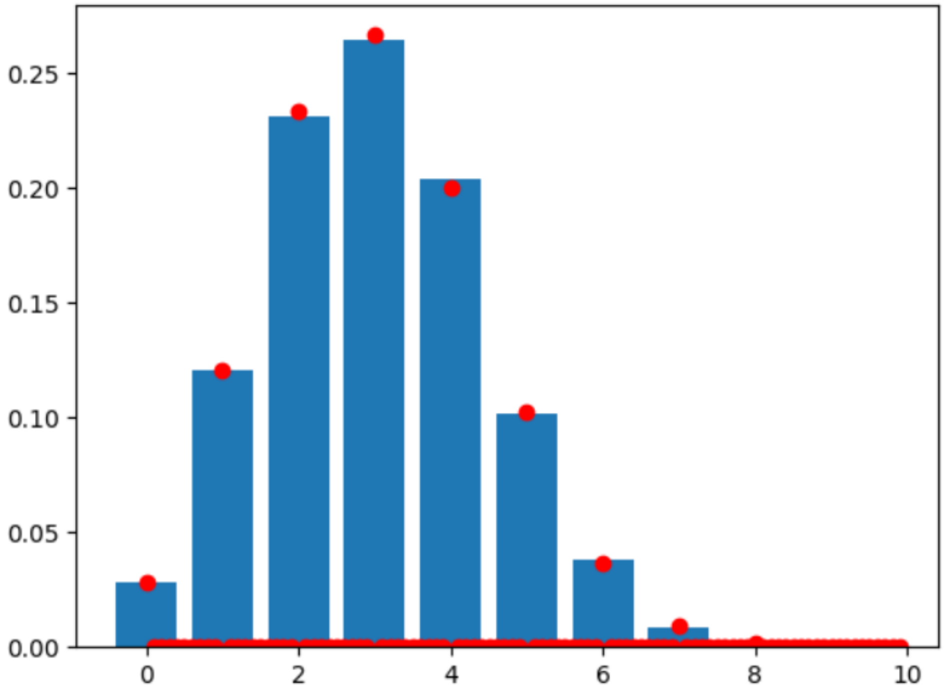
```
<BarContainer object of 9 artists>
```



Как же значения n и p в функции:

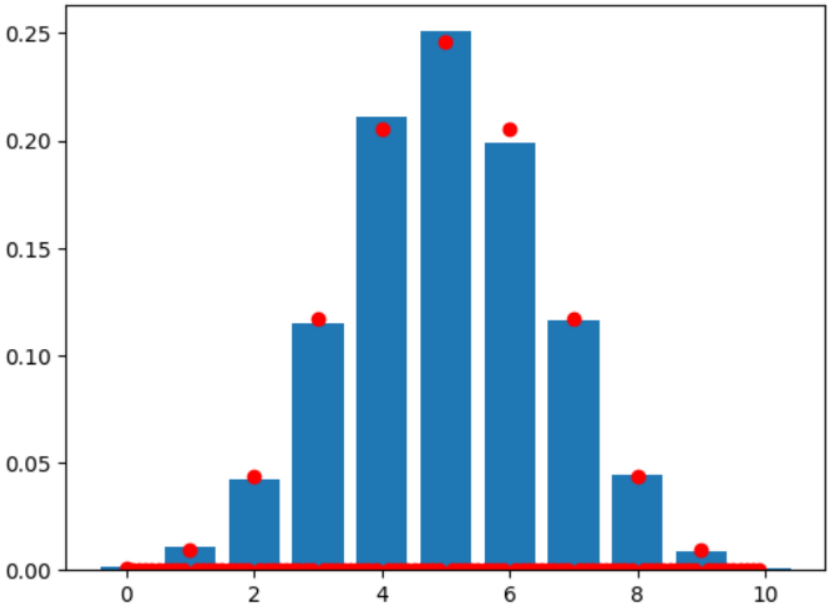
```
n = 10 #number of trials
p = 0.3 #prob. of success
X = binom(n, p)
sample = binomial(n, p, size = 10000)
freqs = count_frequencies(sample, relative=True)
plt.plot(x, X.pmf(x), 'o', color = 'red')
plt.bar(list(freqs.keys()), list(freqs.values()))
```

```
<BarContainer object of 10 artists>
```

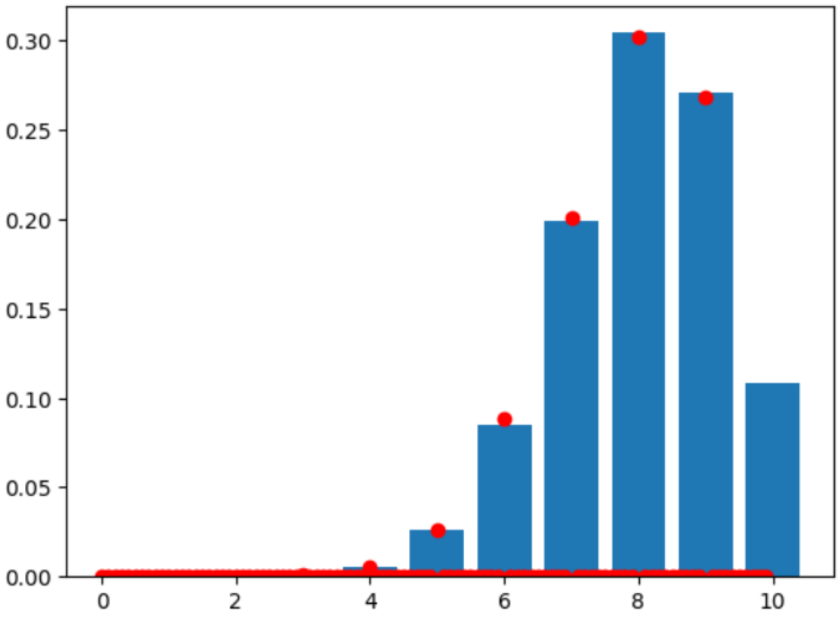


Изменим p:

```
n = 10 #number of trials
p = 0.5 #prob. of success
X = binom(n, p)
sample = binomial(n, p, size = 10000)
freqs = count_frequencies(sample, relative=True)
plt.plot(x, X.pmf(x), 'o', color = 'red')
plt.bar(list(freqs.keys()), list(freqs.values()))
```



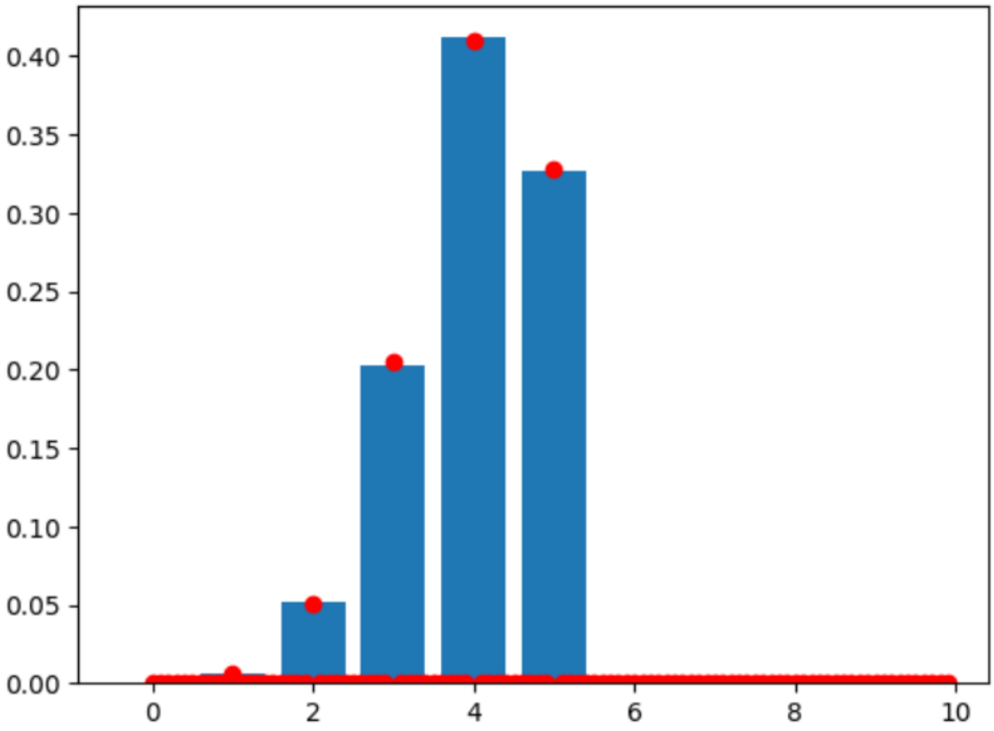
```
n = 10 #number of trials
p = 0.8 #prob. of success
X = binom(n, p)
sample = binomial(n, p, size = 10000)
freqs = count_frequencies(sample, relative=True)
plt.plot(x, X.pmf(x), 'o', color = 'red')
plt.bar(list(freqs.keys()), list(freqs.values()))
```



You see that when you increase probability of success, you will get much more cases when the corresponding value is closer to 10.

We can also change n - number of trials:

```
n = 5 #number of trials
p = 0.8 #prob. of success
X = binom(n, p)
sample = binomial(n, p, size = 10000)
freqs = count_frequencies(sample, relative=True)
plt.plot(x, X.pmf(x), 'o', color = 'red')
plt.bar(list(freqs.keys()), list(freqs.values()))
```



Когда мы увеличиваем кол-во n, то становится лучше.

Python skill test

Question 1

Assume that number of calls that some call center receives during one minute is Poisson random variable with parameter lambda=2. Use Python to find probability that number of calls is larger than 5. Enter the answer with first 5 digits after the decimal point.

Hint: Use scipy.stats.poisson object.

Solution:

```
from scipy.stats import poisson

# Parameter lambda
lambd = 2

# Probability of more than 5 calls
probability = poisson.sf(5, lambd)

# Print the result with 5 digits after the decimal point
print(f"{probability:.5f}")
```

The correct answer is: 0.01656

Question 2

This problem continues previous one. Assume now that one operator can handle one call in one minute. If call is not handled, it's missed. How many operators should I hire to be sure that probability to miss a call during one minute is not larger than 0.05? Of course I want to minimize number of operators hired.

Hint: Denote number of calls by X (it's a random variable) and number of operators hired by q (it's integer number). If X>q, then call is lost. You need to find minimal value qq such that P(X>q)≤0.05. You can re-state this question in terms of CDF values, then answer it by investigating of CDF function in Python. However, there is more efficient way to do it: scipy.stats random variables have .ppf method that calculates percent point function (also known as quantile function) that is inverse function for CDF. For any value pp it finds a minimal value q such that CDF(q)≥p.

Solution:

```
from scipy.stats import poisson

# Parameter lambda
lambd = 2

# Probability threshold for missing a call
threshold = 0.05

# Find the minimal number of operators using the percent point function (ppf)
q = int(poisson.ppf(threshold, lambd)) + 1

# Print the result
print(f"Number of operators to hire: {q}")
```

The correct answer is: 5