4 week Continuous random variables

4.1 Understand notion of continuous random variable, CDF and PDF, independence, covariance, correlation.

4.2 Understand notion of joint CDF and PDF for systems of random variables.

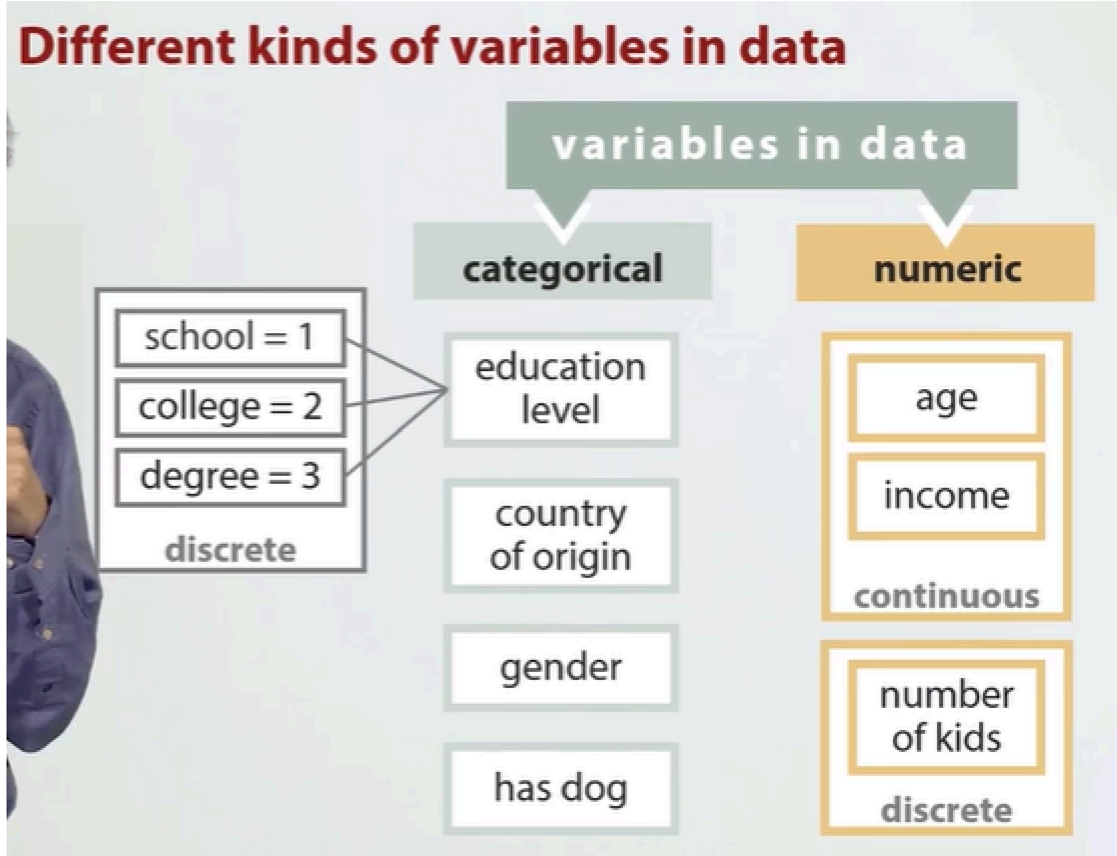
4.3 Distinguish between discrete, continuous and mixed random variables.

4.4 Know definition of expected value and variance of continuous random variable.

Оглавление:

Introduction to the 4th week

In this week we eill study continuous random variables. In contrast with discrete random varibles that can take only some particular values, continuous random variable can take any possible value from some segment or from whole line. Why do we need them? Consider different kinds of variables that we can find in data. Basically, they can be subdivided into two classes: numeric(age, income, number of kids) and categorical(educational level, country of origin, gender, has dog) . In most of cases we have to record this categorical data in numeric. For example, we can say that for educational level we can find numeric records

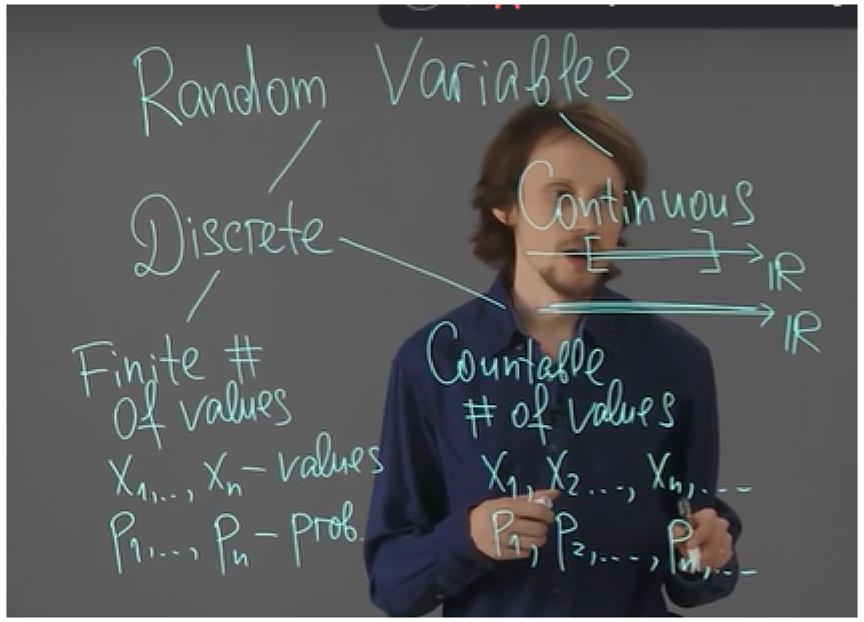


And we use random variable to model this variable. And for this model we usr discrete random variable, because we have only three possible values that can be taken with some probabilities.
However, for income we cannot use discrete random variable to model it because a possible values of income is a some segment and not a particular values.
So we have to introduce continuous random variables to model quantities like income or age

Continuous random variables. Motivation and Example

Random variables can represent different things. For example, some random variables can represent smth like number of Heads in a series of coin tossing. This number can take only finite manu values. However, it's possible that we need random variable that measure smth like temperature, or height, or some other number that can take any values in some segment or at the whole real line. Now, we will discuss R.V. of this second kind.

Random variables thta can take any value from some segment. Bit ti begin, let me recall how we classify random variables.



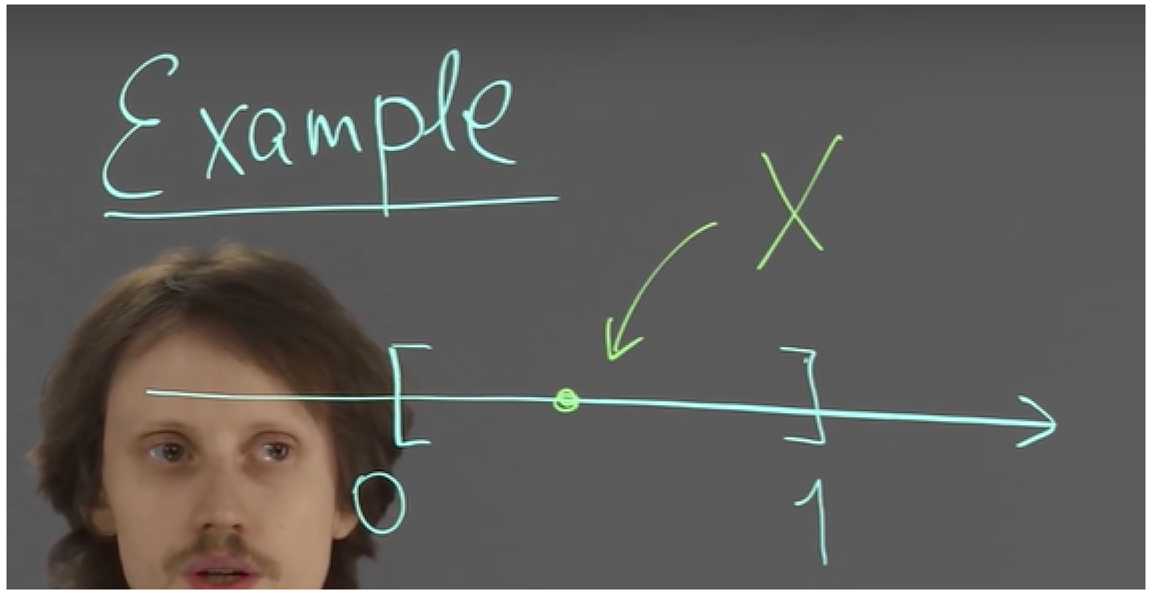
Countable = infinite number of possible values

Continuous means that random variable can take any possible value in some segement or at the whole line.

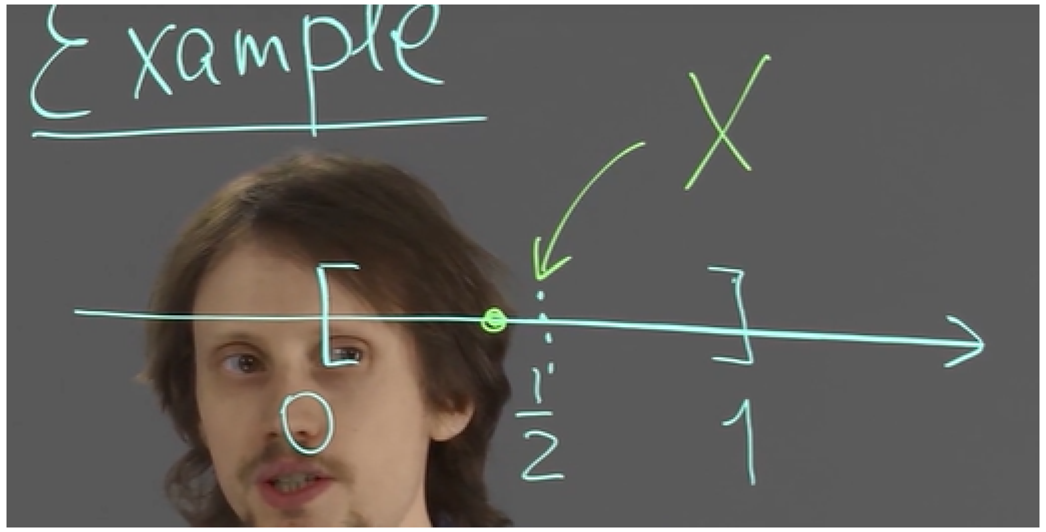
How to define the continuous random varibles?

Example:

Let us assume that we want to pick a random number from 1-0 or, geometrically speaking, we have a segment and we want a pick a randon point on this segment



Let us also assome that we want to shoose this point uniformly over this segment. It means that, for example, the probability for this point to be at the left half of the segment is the same as the probability at the point to be at the righnt half of the segment.There is no preference between left and right parts



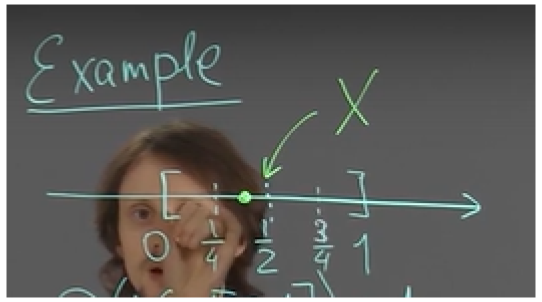
If we said that we need this pont to be inside this segment, then the probability that this point is in this segment equal:

P(X belongs to [0, 1]) = 1

What is the probability that this point in in the left of this segment?

P(X belongs to [0, 1/2]) = P(X belongs to [1/2, 1]) = 1/2

In the same way, we can split our segment into four parts.

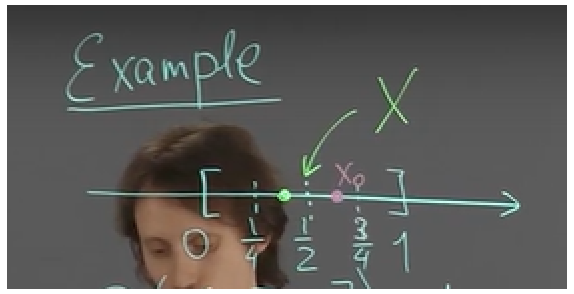


P(X belongs to [0, 1/4]) = P(X belongs to [1/4, 1/2]) = P(X belongs to [1/2, 3/4]) = P(X belongs to [3/4, 1]) = 1/4

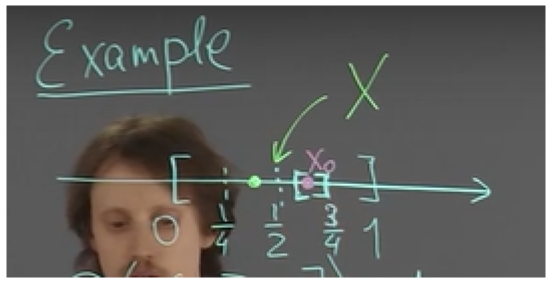
So, we can decide that for any segment, loke from a to b, the probability of this pont to lie inside this segment is equal the length of the segment. The larger segment we consider, the higher probability that our random variable lie inside the segment.

P(X belongs to [a, b]) = b - a

However, we have a problem. Let us ask, what is the probability that our random variable is equal to some predefined point, some point x0



We see that we can take arbitary small segment that contains this point.



P(X = x0) <= P(X belongs to [x0 - E, x0 + E]) <=2E for any E > 0. We see that this probability have to be equal zero.

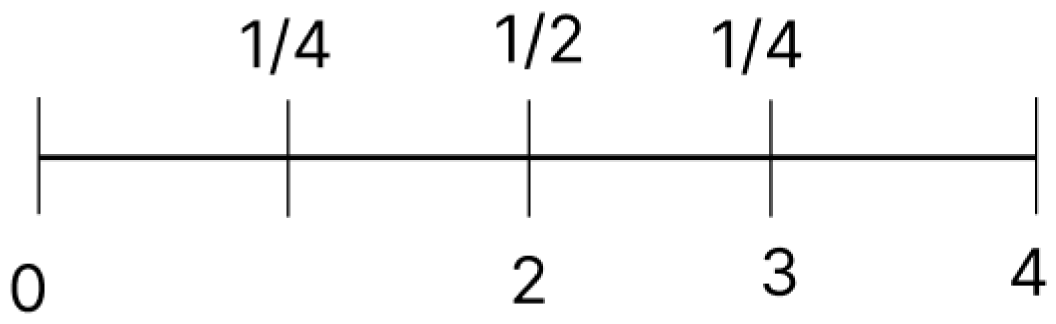
E - epsilon

We see the problem , we described random variables in terms of their possible values and their probabilities. But in case of continuous random variables, we cannot do it like this because the probability that random variable take any particulat value is equal to 0. Now, we see that we cannot describe continuous random variables in terns of the probability of their values as we did it for discrete random variables. However, we can use a different notion of **probability density function to describe continuous random variables.**

Continuous random variable practice

Question 1

A random variable X is uniformly distributed on the segment [1, 5]. Find the probability P(X∈[2,3])
Enter the exact value below with two decimal places or as an irreducible fraction. (e.g. 0.12 or 13/28):



The correct answer is: 1/4

Continuous random variables & independence

Question 1

Let X and Y be independent continuous random variables uniformly distributed on the segment $[0,1]$. Let A be the event $\{|X-Y| \geq 1/2\}$. What's $P(A)$?

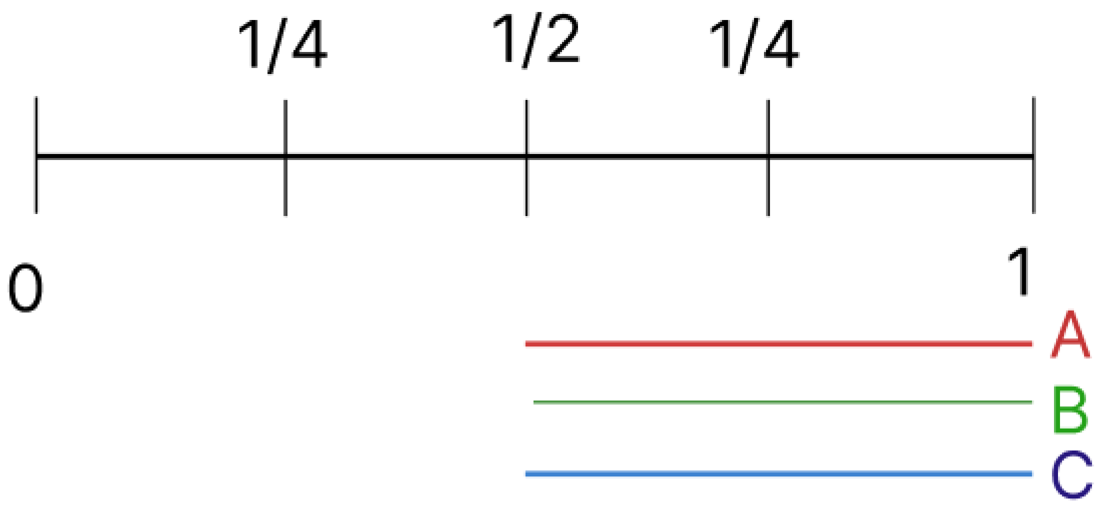
Enter the exact value below with two decimal places or as an irreducible fraction. (e.g. 0.12 or 13/28):

Solution:
P(A belongs to [0, 1/2]) = P(A belongs to [1/2, 1]) = 1/2
1/2 + 1/2 = 1/4

Question 2

Let X and Y be independent continuous random variables uniformly distributed on the segment $[0,1]$. Let A be the event $\{|X-Y| \geq 1/2\}, B = \{|X| > 1/2\}, C = \{|Y| \geq 1/2\}$. Are the events A, B, C pairwise independent?

Solution:



P(A) = P(B) = P(C) = 1/2
P(A ∩ B) = 1/2 *1/2 = 1/4*
P(A ∩ C) = 1/2 1/2 = 1/4
P(B ∩ C) = 1/2 * 1/2 = 1/4
The correct answer is: yes

Question 3

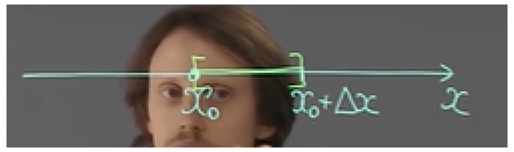
Let X and Y be independent continuous random variables uniformly distributed on the segment $[0, 1]$. Let A be the event $\{|X - Y| \geq 1/2\}, B = \{|X| > 1/2\}, C = \{|Y| \geq 1/2\}$. Are the events A, B, C mutually independent?

P(A ∩ B ∩ C) =1/2 *1/2* 1/2 = 1/4
P(A ∩ B) != P(A ∩ C) != P(B ∩ C) != P(A ∩ B ∩ C) => they aren't mutually independent
The correct answer is: no

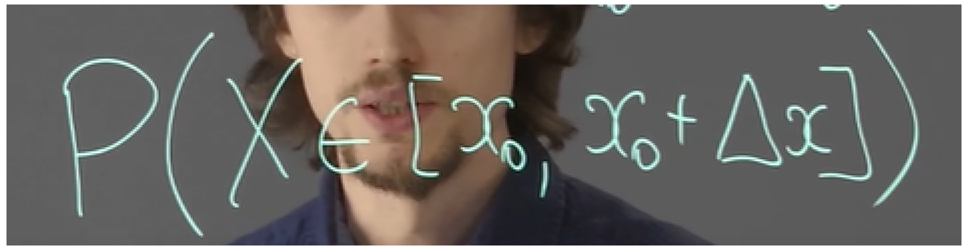
Probability density function (PDF)

As we discussed before, it doesn't make sense to ask about aontinuous random variable: What is the probability that this random variable takes a particular vallue? However, we can ask, what is the probability that our random variable takes a value which is close to some particular number. The answer for this question gives us probability density function.

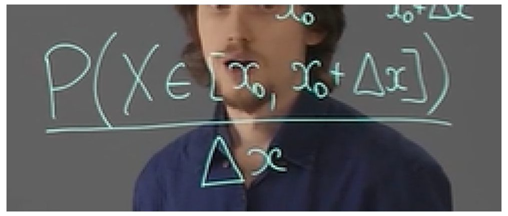
Let us consider some continuous random variable and some particular number x0. How can we ask mathimatically involved about the probability that our random variable takes a value that is close to this x0? We can consider some small segment that is nearer to x0



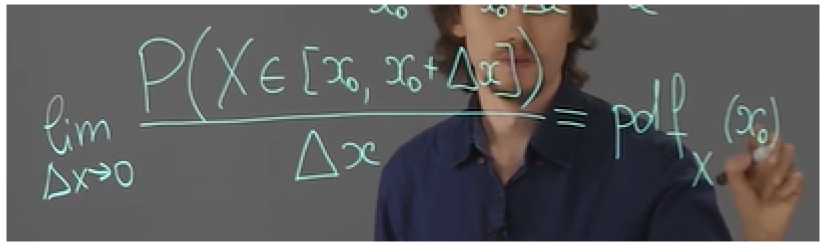
Let us take this segment and ask, **what is the probability that x takes the value in this segment?**



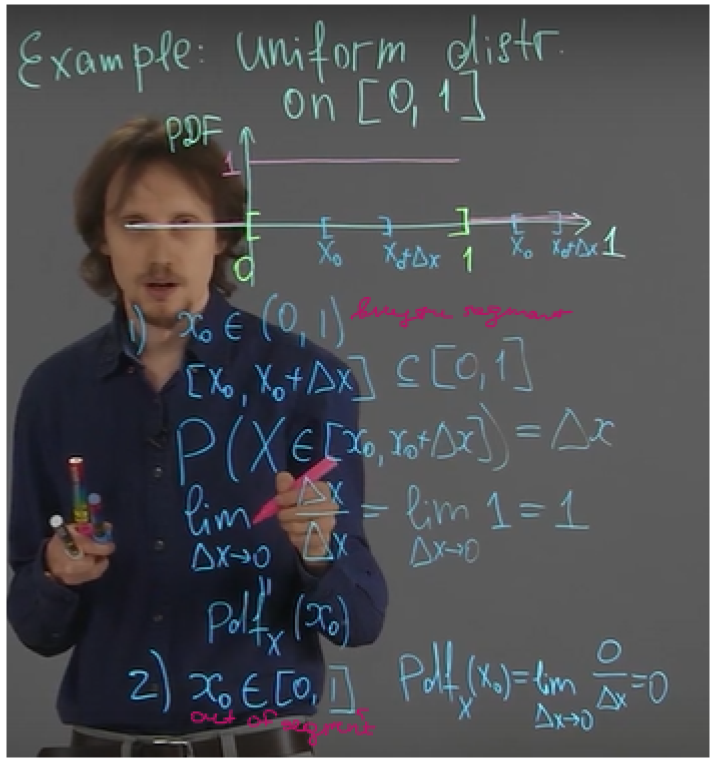
As we discussed before, this probabilty can become very small when the length of this segment becomes small. However, we are interested in small segments(delta x) because we are interested in the small neighborhood of this point x0. So, what can we do? We can devide this probabilty by delta x



Если числитель растёт, то и знаменатель. It is possible that this ration has a limit. If this limit exists, by definition, is equal to the value of pdf of random variable X at point x0.



Example: Now let us consider example that we discussed before. Let us find pdf for a uniforn distribution on segment from 0-1.

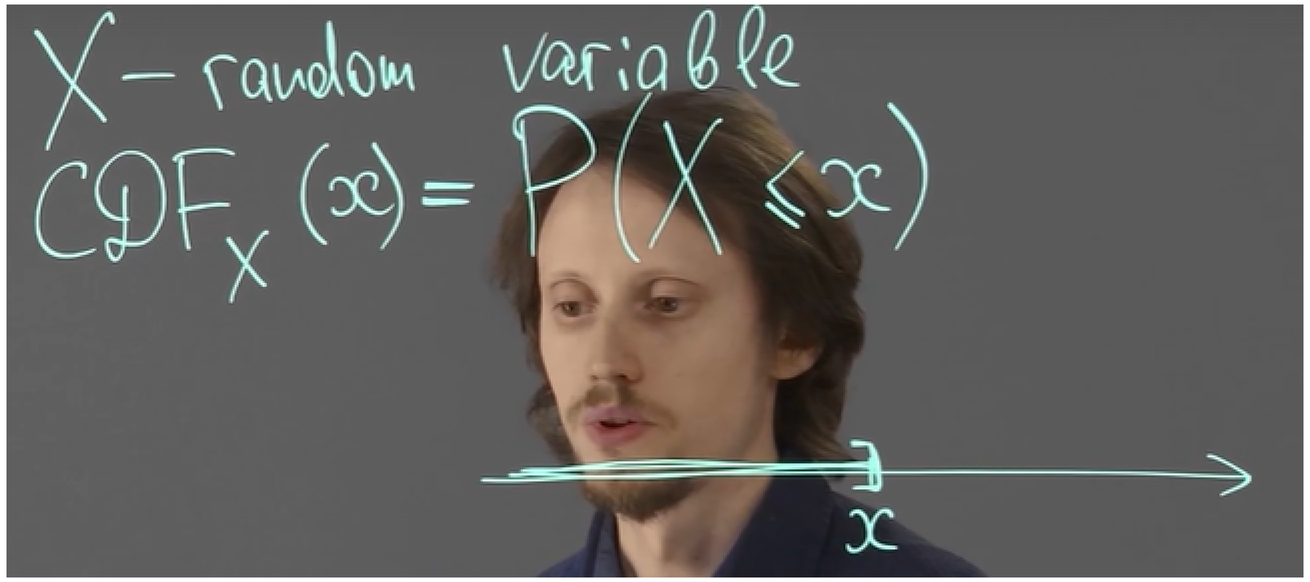


What about points 0 and 1? It appears that pdf is not defined at this point because this limit doesn't exist. But this perfectly normal to have discontinuities of this kind.

So we defined a PDF for uniform distribution. At uniforn distribution, all parts on the segment have the same probability provided that they have the same length because the probabilty is a constant over segment

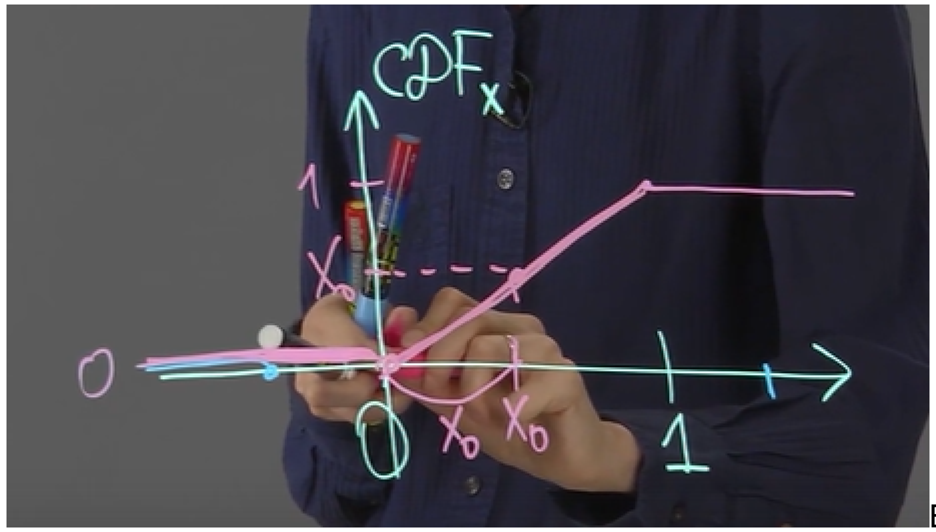
Cumulative distribution function (CDF)

Together with PDF we need yet another one way to descibe continuous random variables. These way is called cumulative distribution function(CDF)

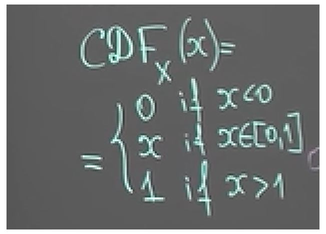


Example: let us consider example of uniform distribution again

Probability that our random variable which is uniformly distributed on the segment [0,1] is equal 0. It means that for all points that are to the left from the zero, the value of CDF of x and this is equal 0. Let us consider some point after 1. What is the probabiity that our random variable uniformly distributed on this segment has value. Of course, if random variable is distributed only on this segment [0,1] then it necessarily to have values that are less than this value (1). So this happens with probability one. It means that for all points that to the righnt from 1, the corresponding CDF is equal to 1.



Basically, we can say that cumulative



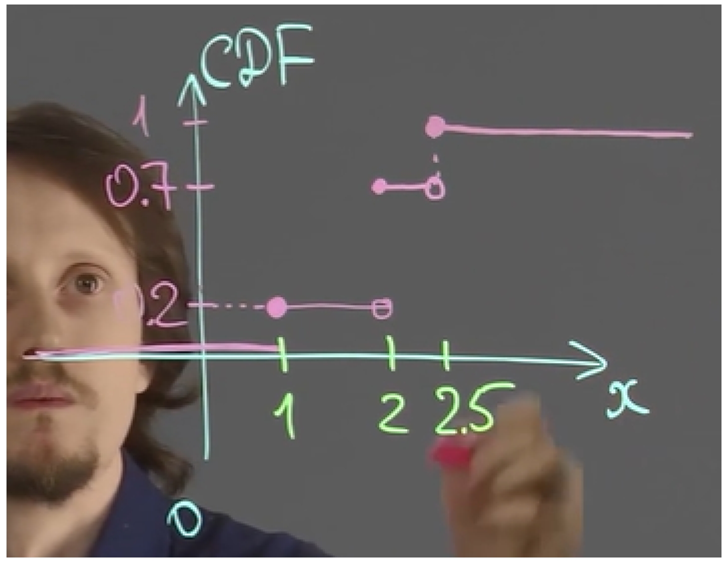
We can defined CDF not only for continuous random variables, but also for discrete random variable on the previous week.

CDF for discrete random variable

X	**1**	**2**	**2.5**
p	0.2	0.5	0.3

We can write CDF for this variable.

Our random variable can take values only 1, 2, 2.5. It means that if we ask what is the probability that x is less than 0.9



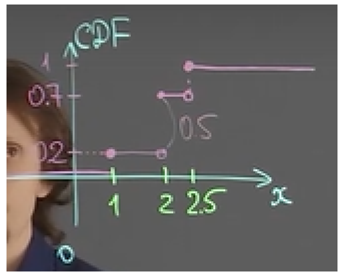
It has several jumps and each jump corresponds to a particular valur that our variable can take. And the size of this jump is equal the probabilty.

P(X <= 0.9) = 0

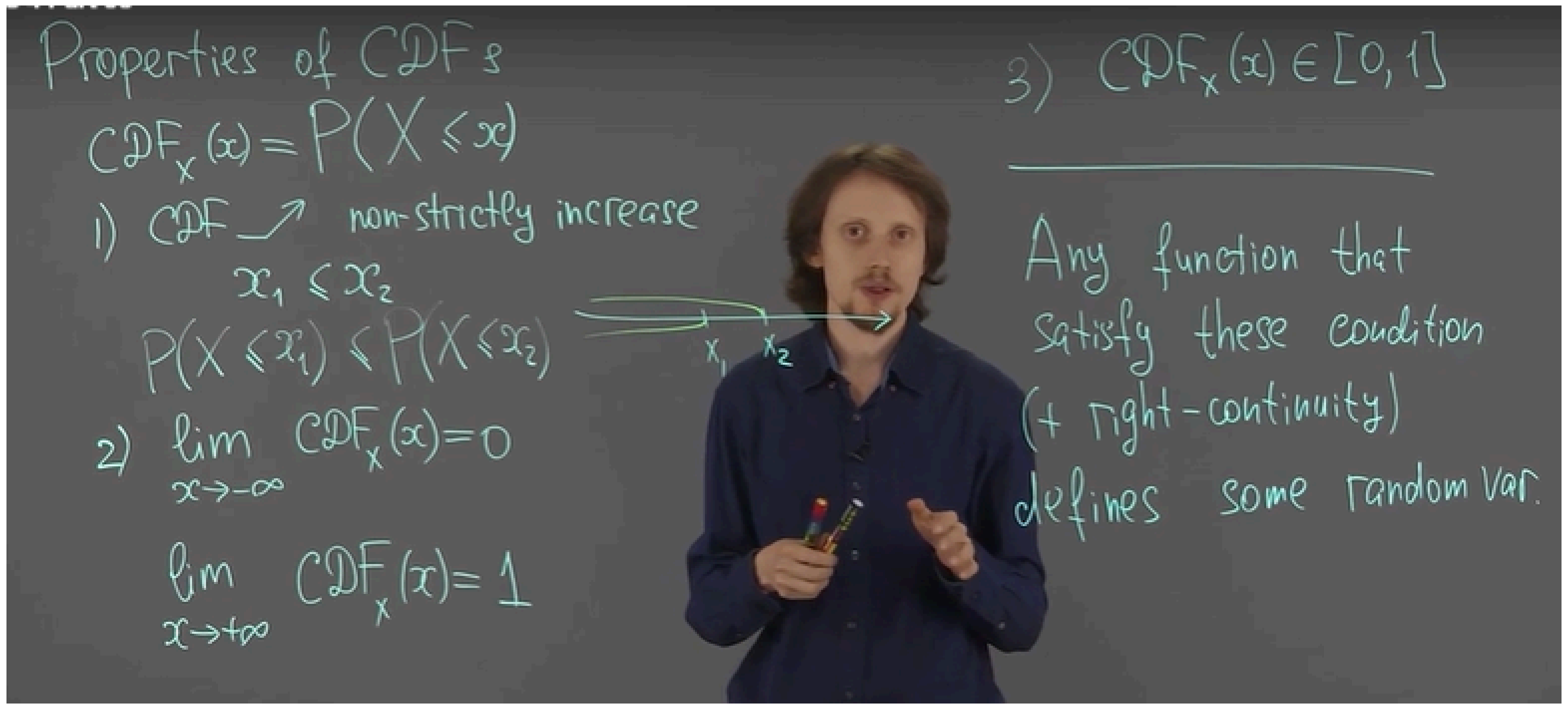
P(X <= 1) = P(X = 1) = 0.2 = CDF(1)

P(X <= 1.5) = P(X = 1) = 0.2 = CDF(1)

P(X <= 2) = P(X = 1) + P(X = 2) = 0.2 + 0.5 = 0.7 = CDF(2)



Properties of CDF



CDF of discrete random variable

Question 1

Distribution of discrete random variable is given by the table:

X	-10	-5	-1	0	1	5	10
P	0.05	0.1	0.2	0.3	0.2	0.1	0.05

Find the value of CDFX(2). Enter the exact value below with two decimal places or as an irreducible fraction. (e.g. 0.12 or 13/28):
Solution:
0.05 + 0.1 + 0.2 + 0.3 + 0.2 = 0.85

Question 2

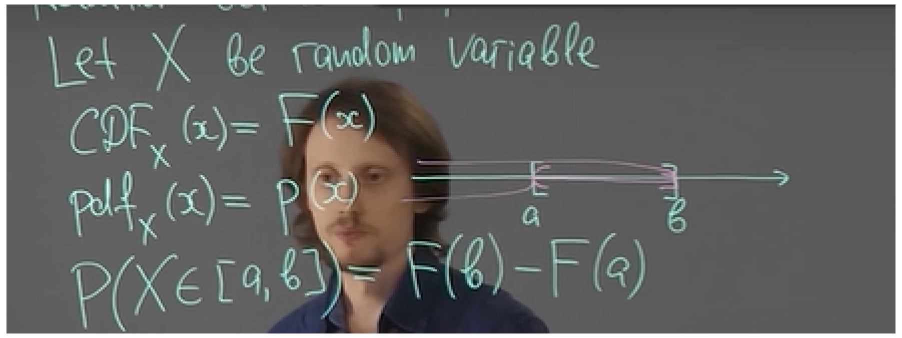
Distribution of discrete random variable is given by the table:

X	-10	-5	-1	0	1	5	10
P	0.05	0.1	0.2	0.3	0.2	0.1	0.05

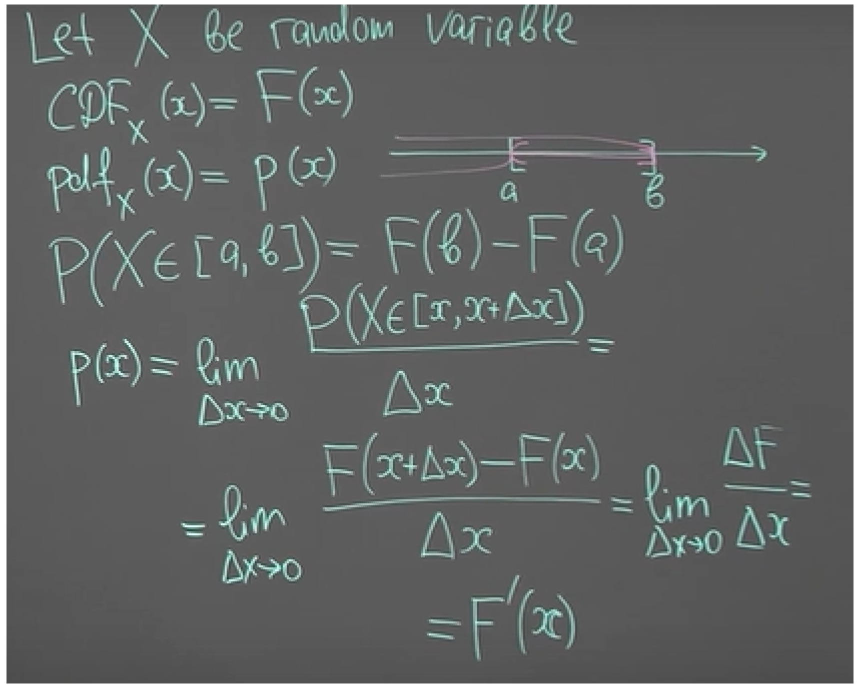
Find the jump of CDFX at point 1. Enter the exact value below with two decimal places or as an irreducible fraction. (e.g. 0.12 or 13/28):
The correct answer is: 1/5 = 0.2

Linking PDF and CDF

A CDF is mathimatically convinient way describe random variable. However, we are more interested in PDF because it's easy intepreture them. So, it's important to discuss the relation between PDF and CDF. This relation is mathimatically very nice)

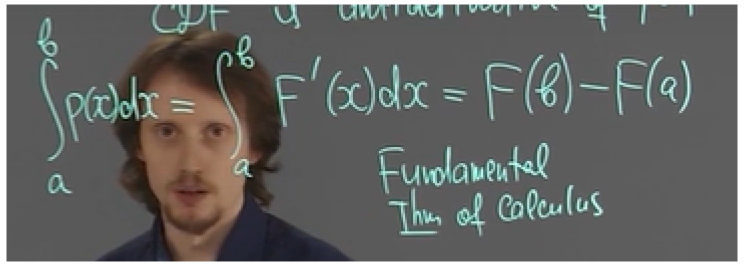


Теоретически, необходимо использовать () скобки, но так как на точке вероятность ровна 0, это можно оставить и так.

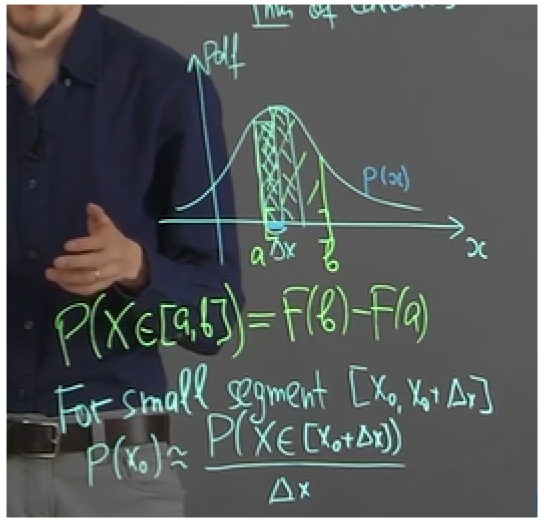


So we can see that

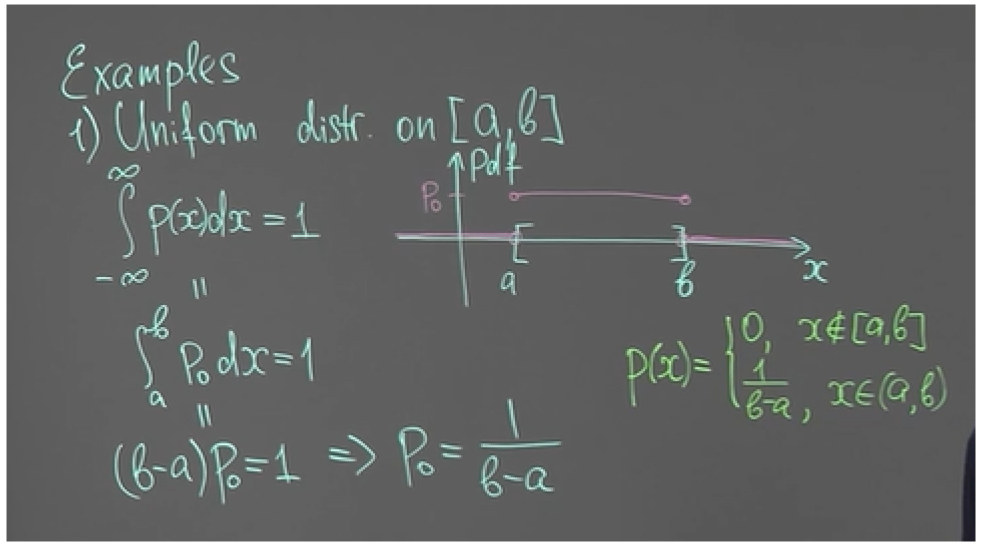
- **pdf is a derivative of CDF**
- **CDF is antiderivative of pdf**



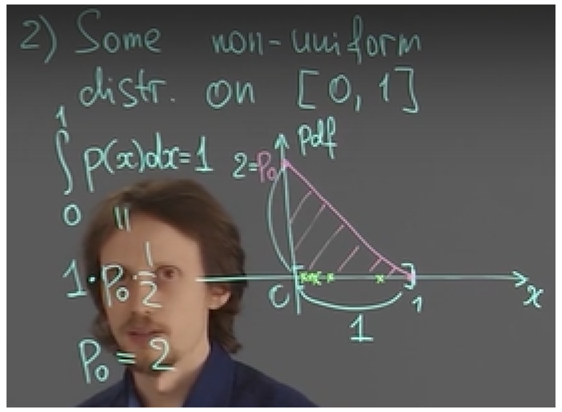
It's more likely to find point somewhere here(yellow).



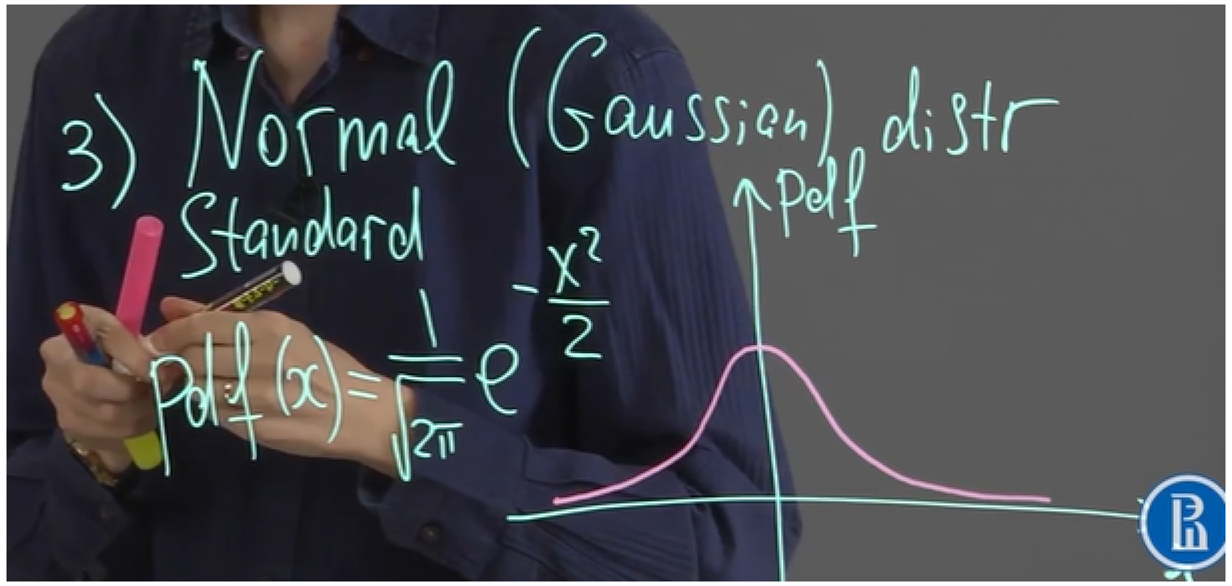
Examples of probability density functions



Точки, которые ближе к 0, будут встречаться чаще, чем точки рядом с 1. Как раз то что показывает нам pdf



Finally, let us discuss an example of random variable that take a value not on the segment but on the whole line.



Большинство точек располагается ближе к 0.

We see that as x increases the exponent becomes very, very small. And we see the same thing on the graph.

PDF and CDF skill test

Question 1

How many correct statements are in this list:

- $PDF_X(x)$ reaches its maximum at $x_0 = \mathbb{E}X$
- $PDF_X(x) \leq 1$ for all x
- if $P(X = x_0) = 0$ then $PDF_X(x_0) = 0$

a. 1

b. 3

c. 2

d. 0 ✔

Question 2

PDF_X is given by function

- 0 for $x < -2$
- $C(4 - x^2)$ for $x \in [-2, 2]$
- 0 for $x > 2$

Find C. Enter the exact value below with two decimal places or as an irreducible fraction. (e.g. 0.12 or 13/28):

Tip: Use that total probability of all values of X is 1

Solution:

$$\int_{-2}^{2}(4-x^2)\,dx = \int 4\,dx - \int x^2\,dx = \left(4x - \frac{x^3}{3}\right)\Big|_{-2}^{2} =$$

$$= \left(4\cdot 2 - \frac{2^3}{3}\right) - \left(4\cdot(-2) - \frac{(-2)^3}{3}\right) = \frac{32}{3} = 10.6 = 10\frac{2}{3}$$

$$\frac{3}{32} \cdot \frac{32}{3} = 1$$

Answer: $3/32 = C$

The correct answer is: 3/32

Question 3

PDF_X is given by function

- 0 for $x < -2$
- $C(4-x^2)$ for $x \in [-2, 2]$
- 0 for $x > 2$

with the value of C defined in the previous task. Find $CDF_X(1)$. Enter the exact value below with two decimal places or as an irreducible fraction. (e.g. 0.12 or 13/28):

Solution:

$$\int_{-2}^{1}(4-x^2) = \left(4x-\frac{x^3}{3}\right)\Bigg|_{-2}^{1} = \left(4-\frac{1}{3}\right)-\left(-8+\frac{8}{3}\right)$$

$$= \frac{11}{3}+\frac{16}{3} = \frac{27}{3} = 9$$

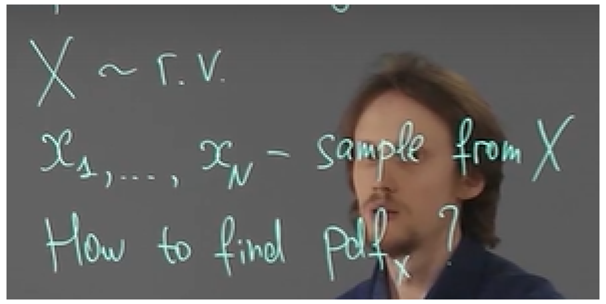
$$\frac{3}{32}\left(4-\frac{1}{3}+8-\frac{8}{3}\right) = \frac{\cancel{3}}{32}\left(\frac{27}{\cancel{3}}\right)$$

Answer 27/32

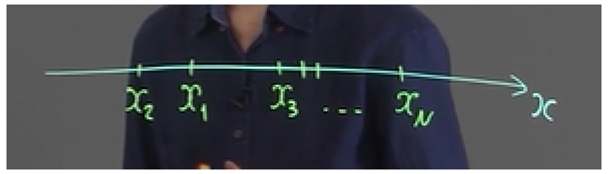
Tip: Use the expression of CDF through PDF that we discussed in lectures and be accurate
The correct answer is: 27/32

Histogram as approximation to a graph of PDF

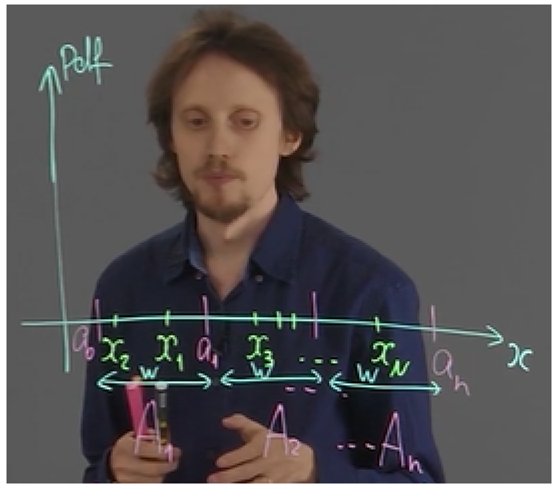
Let us assume that we have a random variable but we don't have access to mathimatical definition. We don't know pdf and CDF. However, we can sample from this random variable. It means that we can take its values several times and this values are independent to each other.
In this case, we can estimate how PDF looks like. It will be done by using histogram (useful tool for data visualization). So let's disscuss relation between pdf and histograms.

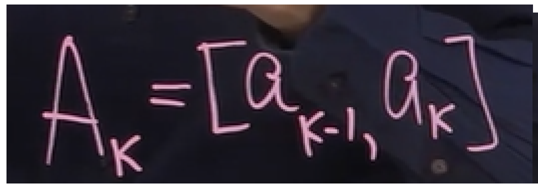


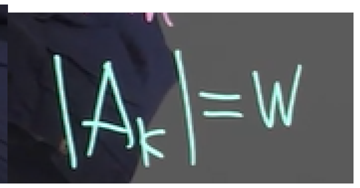
We cannot do it exactly but we can approximate it by using this sample.
The idea is quite simple.



The region with 3 points has larger pdf than region with 1 point. And we can use this idea to draw approximation pdf. To do so, let us devide axis into several segments (theoretically these segments should be small but for example we draw like this)

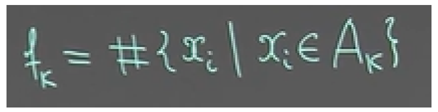




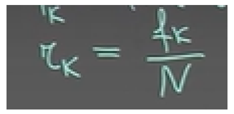


Now we can approximate probability that our random variable take a value on the segment (for example, 2d) by its relative frequence according our sample. What does it mean?

1. Calculate frequences which is number of point on segment k

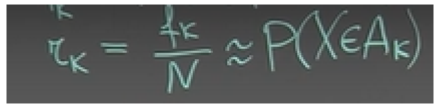
   

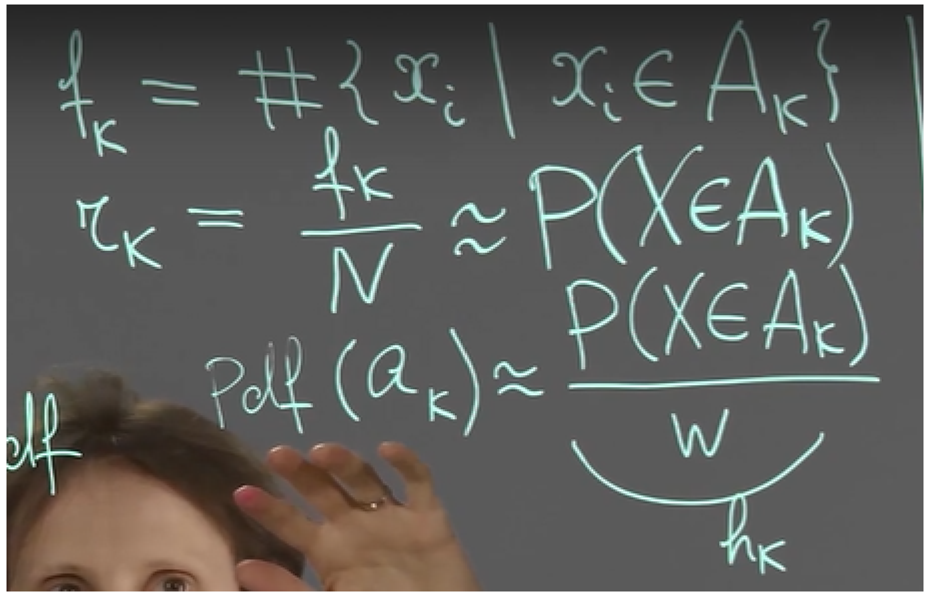
   for our pic: f1 = 2, f2 = 3, f3 =1
2. Find relative frequences

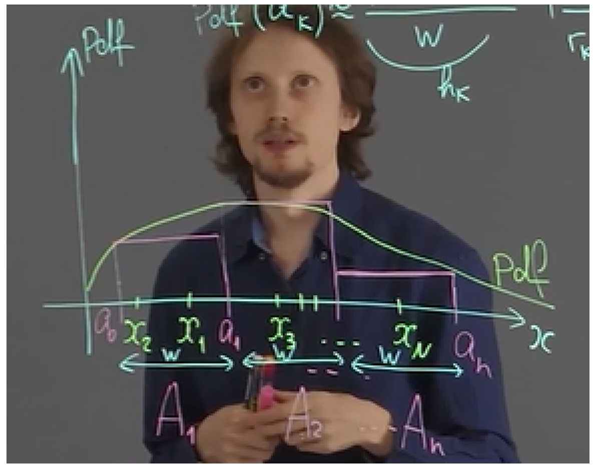
	1	**2**	**3**
fk	2	3	1
rk	2/6	3/6	1/6

As we can see relative frequence is an approximation of probability of the corresponding event.





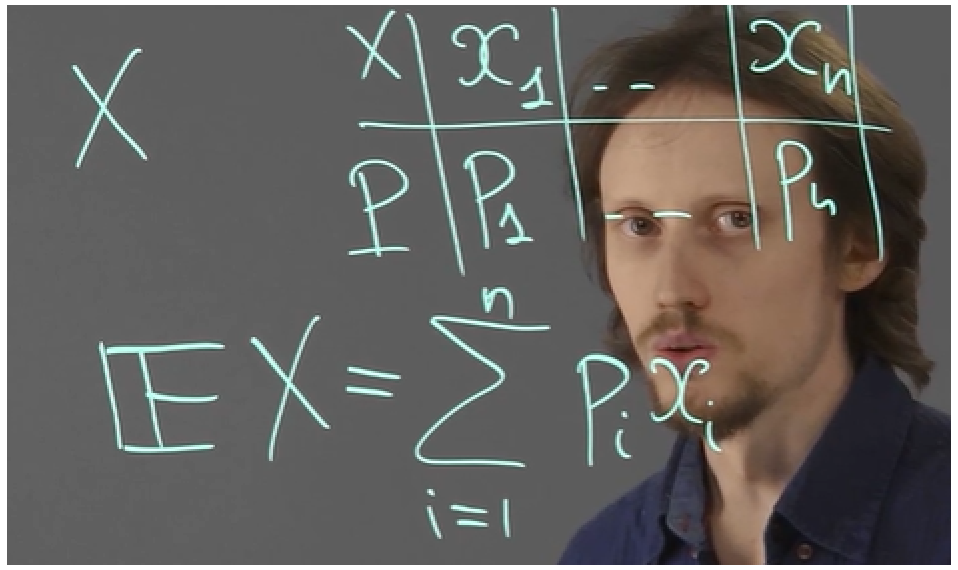
heights of each rectangle is proportional to its number of points that lie on the corresponding segment. This is called histogram. It approximate pdf.



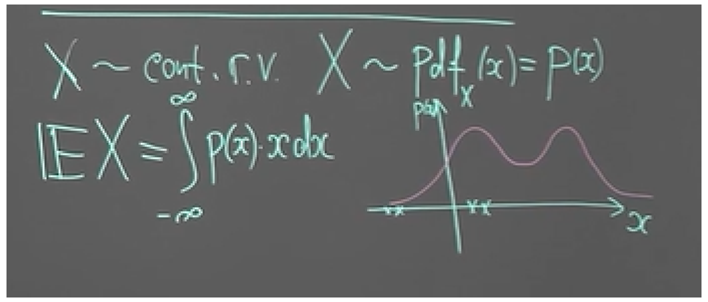
if we

Expected value of continuous random variable

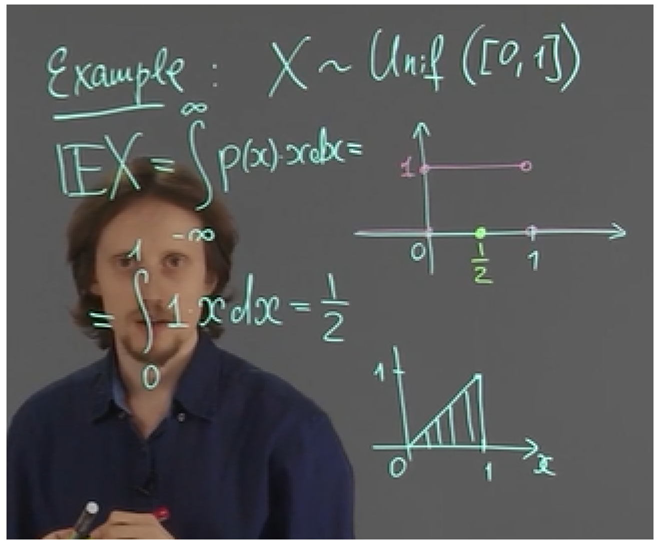
When we discussed discrete random variables we mentioned some different properties of R.V. lile EX, VarX. How do they work with continuous random variables?
Recall for descrete random variable:



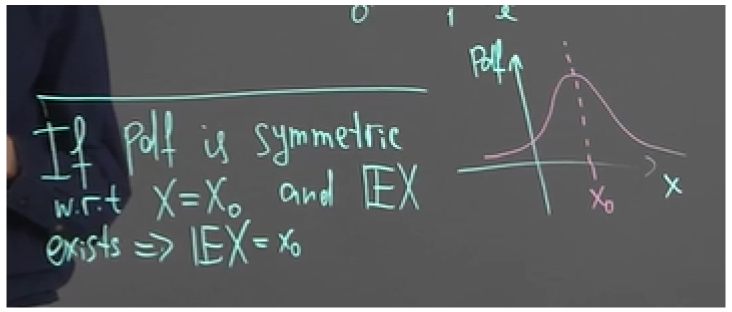
So for continuous R.V.:



Example:



It's quite natural because it's in the middle of the centre. **Actually, we can use the idea that if pdf is summetric where the axis of symmerty has to be equal EX.**



Finding expectation with PDF

Question 1

Find $\mathbb{E}X$ for continuous random variable X with the following PDF:

$2 - 2x$ for $x \in (0, 1)$ and 0 otherwise.

Enter the exact value below with two decimal places or as an irreducible fraction. (e.g. 0.12 or 13/28):

Solution:

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_0^1 x(2-2x) = \int_0^1 2x - 2x^2 =$$

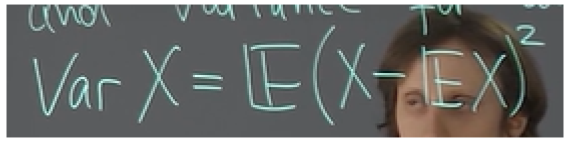
$$= \int_0^1 2x - \int_0^1 2x^2 = \left(\frac{2x^2}{2} - \frac{2\cdot x^3}{3}\right)\Big|_0^1 =$$

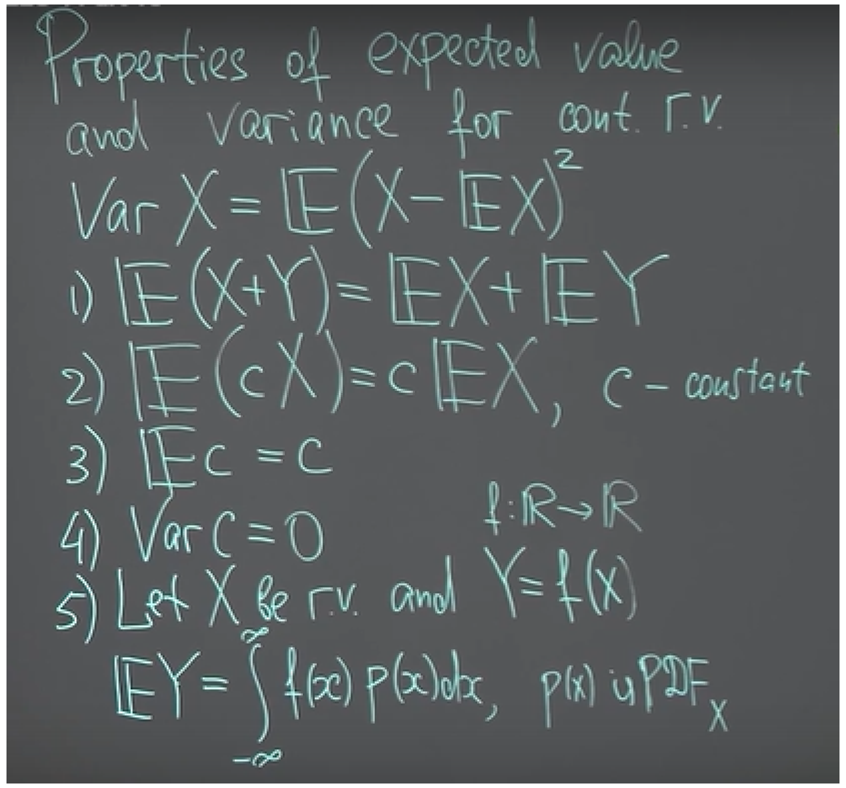
$$= \left(x^2 - \frac{2\cdot x^3}{3}\right)\Big|_0^1 = 1 - \frac{2}{3} = \frac{1}{3}$$

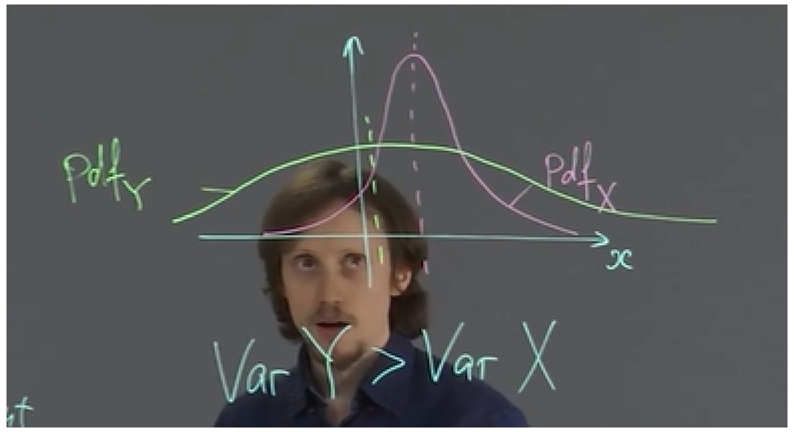
The correct answer is: 1/3

Variance of continuous random variable. Properties of expected value and variance

Properties of expected value and Variance for continous random variables are very similar to the properties of discrete r.v.
The formula of VarX for continuous r.v. is same as for discrete r.v.







CDF, expectation and variance practice

Question 1

CDF of a random variable X is given by table:

0 if $x < 1$;

0.2 if $1 \leq x < 2$;

0.35 if $2 \leq x < 3$;

0.9 if $3 \leq x < 4$;

1 if $4 \leq x$.

Find $P(2 < X \leq 4)$. Enter the exact value below with two decimal places or as an irreducible fraction. (e.g. 0.12 or 13/28):

Solution:

cdf_X	1	2	3	4
P	0.2	0.35	0.9	1

CDF

$(2, 4]$

$$P(a < X \leq b) = F_X(b) - F_X(a)$$

$P(X = 1) = F(1) = 0.2$

$P(X = 2) = F(2) - F(1) = 0.35 - 0.2 = 0.15$

$P(X = 3) = F(3) - F(2) = 0.9 - 0.35 = 0.55$

$P(X = 4) = F(4) - F(3) = 1 - 0.9 = 0.1$

The probability distribution of X is as follows:

$X = x$	1	2	3	4
$P(X = x)$	0.2	0.15	0.55	0.1

$$P(2 < X \leq 4) = P(X = 3) + P(X = 4) = 0.55 + 0.1 = 0.65$$

The correct answer is: 13/20 = 0.65

Question 2

Under the same conditions, find $\mathbb{E}X$. Enter the exact value below with two decimal places or as an irreducible fraction. (e.g. 0.12 or 13/28):

Solution:

$$\mathbb{E}X = \sum p_i x_i = 1\cdot 0.2 + 2\cdot 0.15 + 3\cdot 0.55 + 4\cdot 0.1 =$$

$$= 0.2 + 0.3 + 1.65 + 0.4 = 2.55$$

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

$X = x$	1	2	3	4
$P(X = x)$	0.2	0.15	0.55	0.1
$X - \mathbb{E}X$	−1.55	−0.55	0.45	1.45
$(X-\mathbb{E}X)^2$	2,4	0,3	0,2	2,1

The correct answer is: 51/20 = 2.55

Question 3

Under the same conditions, find $VarX$. Enter the exact value below with two decimal places or as an irreducible fraction. (e.g. 0.12 or 13/28):

Solution:

$\mathrm{Var}X = \mathbb{E}((X-\mathbb{E}X)^2) = 0.2 \cdot 2{,}4 + 0.15 \cdot 0.3 + 0.55 \cdot 0.2 + 0.1 \cdot 2.1 =$

$= 0.48 + 0{,}045 + 0.11 + 0.21 = 0.845$

The correct answer is: 0.84

CDF, expectation and variance of continuous random variables skill test

Question 1

Let X be a continuous random variable without atoms, i.e. it takes no values with positive probability. Let F be its CDF given by the formula:

$C \cdot \frac{\log(1-x)}{x}$ if $x < 0$;

1 if $0 \leq x$.

Find C:

Solution:

By assumption, F is continuous at 0. The right limit is 1, while the left limit is

$$\lim_{y \to 0^-} \frac{C \log(1-y)}{y} = C \lim_{y \to 0^-} \frac{\log(1-y)}{y} = C \cdot (-1) = -C.$$

Therefore, $-C = 1$, which gives $C = -1$.

Т.к. CDF значит $C \cdot \frac{\log(1-x)}{x}$ уже результат интегрирования с pdf.

$$\boxed{\int_{-\infty}^{+\infty} pdf_x(x)dx = 1} \Rightarrow C \cdot \frac{\log(1-x)}{x} = 1 \Rightarrow C \cdot (-1) = 1 \Rightarrow C = -1$$

т.к. $-1 \cdot (-1) = 1$

$\Rightarrow \lim_{x \to -\infty} \left(C \cdot \frac{\log(1-x)}{x} \right) \Rightarrow C \cdot \lim_{x \to \infty} \left(\frac{\log(1-x)}{x} \right)$. Т.к. $x < 0$ мы сможем взять, например, $x = -1 \Rightarrow \frac{\log(1-(-1))}{-1} = \frac{\log(2)}{-1} \Rightarrow$

$$\lim_{x \to -\infty} \left(\frac{\log(2)}{-1} \right) = -1$$

The correct answer is: -1

Question 2

Under the same conditions, find probability of X having value in the segment [−1,1] . Give the answer up to 2nd decimal digit:

Solution:

$$\left(\frac{\log(1-x)}{x}\right)\Bigg|_{-1}^{1} = \frac{\log(0)}{1} \leftarrow \text{не определён} \quad - \frac{\log(2)}{-1}$$

$$= \frac{\log(2)}{1} = \log(2) \approx 0{,}301$$

Question 3

CDF of a continuous random variable X is given by the formula:

0 if $x < 0$;

Cx if $0 \leq x < 10$;

1 if $10 \leq x$.

Find C. Enter the exact value below with two decimal places or as an irreducible fraction. (e.g. 0.12 or 13/28):

Solution:

CDF_x	0	10
p	cx	1

CDF уже интегрирован $\Rightarrow$ сразу подставим:

$$(cx)\Big|_0^{10} = c\,10$$

$$\Rightarrow c = \frac{1}{10}, \quad \text{т.к. } \frac{1}{10}\cdot 10 = 1$$

$$\boxed{\int_{-\infty}^{+\infty} pdf_x(x)\,dx = 1}$$

Question 4

Under the same conditions, find $P(X = 2)$

Solution:

?

Question 5

Under the same conditions, find $P(X < 3)$.

Enter the exact value below with two decimal places or as an irreducible fraction. (e.g. 0.12 or 13/28):

Solution:

$$\mathbb{E}X = \left(\int_0^{10} x\right) \cdot \mathrm{pdf}_x = \left(\int_0^{10} x\right) \cdot \frac{1}{10} = \frac{x^2}{2} \cdot \frac{1}{10} = \frac{x^2}{20}\Big|_0^{10}$$

$$= \frac{100}{20} = 5$$

$$\mathrm{Var}\,X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{10}{3} - 5^2 = \frac{10}{3} - \frac{25 \cdot 3}{3} =$$

$$=$$

$$\mathbb{E}X^2 = \left(\int_0^{10} x^2\right) \cdot \mathrm{cdf}_x = \frac{x^3}{3} \cdot \frac{1}{10} = \frac{x^3}{30} = \frac{100}{30} =$$

$$= \frac{10}{3}$$

Question 6

Under the same conditions, find EX:
Recall how Expectation is calculated via CDF

$$EX = \left(\int_0^{10} x\right) \cdot pdf_x = \left(\int_0^{10} x\right) \cdot \frac{1}{10} = \frac{x^2}{2} \cdot \frac{1}{10} = \frac{x^2}{20}\Big|_0^{10}$$

$$= \frac{100}{20} = 5$$

The correct answer is: 5

Question 7

Under the same conditions, find $VarX$. Enter the exact value below with two decimal places or as an irreducible fraction. (e.g. 0.12 or 13/28):

Recall how Variance is calculated via CDF
?
The correct answer is: 25/3

PDF practice

Question 1

PDF of a continuous random variable X is given by formula:

0 if $x < 0$;

$x/32$ if $0 \leq x \leq 8$;

0 if $8 < x$.

Find the value of CDF of X in x = 4.

Enter the exact value below with two decimal places or as an irreducible fraction. (e.g. 0.12 or 13/28):

Solution:

pdf

x	0	8
p	0	$x/32$

cdf_x	0	8
p	0	$\frac{1}{4}$

$$CDF_x = \int_0^8 \frac{x}{32} = \frac{1}{32}\left(\frac{x^2}{2}\right)\Big|_0^8 =$$

$$= \frac{\cancel{16}^{8}}{32\cdot\cancel{2}} = \frac{1}{4}$$

Answer: $\frac{1}{4}$

The correct answer is: 1/4

Question 2

Under the same conditions, find EX. Enter the exact value below with two decimal places or as an irreducible fraction. (e.g. 0.12 or 13/28):

Solution:

$$\mathbb{E}X = \int_0^8 x \cdot pdf_X(x)\,dx =$$

$$\int_0^8 x \cdot \frac{x}{32} = \int_0^8 \frac{x^2}{32} = \frac{1}{32}\int_0^8 x^2 =$$

$$= \frac{1}{32}\cdot\left(\frac{X^3}{3}\right) = \frac{8^3}{3\cdot 32} = \frac{5\cancel{12}}{3\cdot 3\cancel{2}} = \frac{16}{3}$$

The correct answer is: 16/3

Question 3

Under the same conditions, find VarX. Enter the exact value below with two decimal places or as an irreducible fraction. (e.g. 0.12 or 13/28):

Solution:

cdf_{X}	0	8
p	0	$\frac{1}{4}$

$$=$$

$$\mathbb{E}X^2 = \int_0^8 x^2 \cdot \frac{x}{32} = \frac{1}{32}\int_0^8 x^3 = \frac{1}{32}\cdot\left(\frac{x^4}{4}\right)\Big|_0^8 =$$

$$= \frac{8^4}{32\cdot 4} = \frac{4096}{32\cdot 4} = \frac{4096}{128} = 32$$

$$Var(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = 32 - \left(\frac{16}{3}\right)^2 = \frac{32}{9}.$$

The correct answer is: 32/9

Finding variance with PDF

Question 1

Find $varX$ for continuous random variable X with the following PDF:

$2 - 2x$ for $x \in (0, 1)$ and 0 otherwise.

Enter the exact value below with two decimal places or as an irreducible fraction. (e.g. 0.12 or 13/28):

Solution:

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_0^1 x(2-2x) = \int_0^1 2x - 2x^2 =$$

$$= \int_0^1 2x - \int_0^1 2x^2 = \left(\frac{2x^2}{2} - \frac{2\cdot x^3}{3}\right)\Big|_0^1 =$$

$$= \left(x^2 - \frac{2\cdot x^3}{3}\right)\Big|_0^1 = 1 - \frac{2}{3} = \frac{1}{3}$$

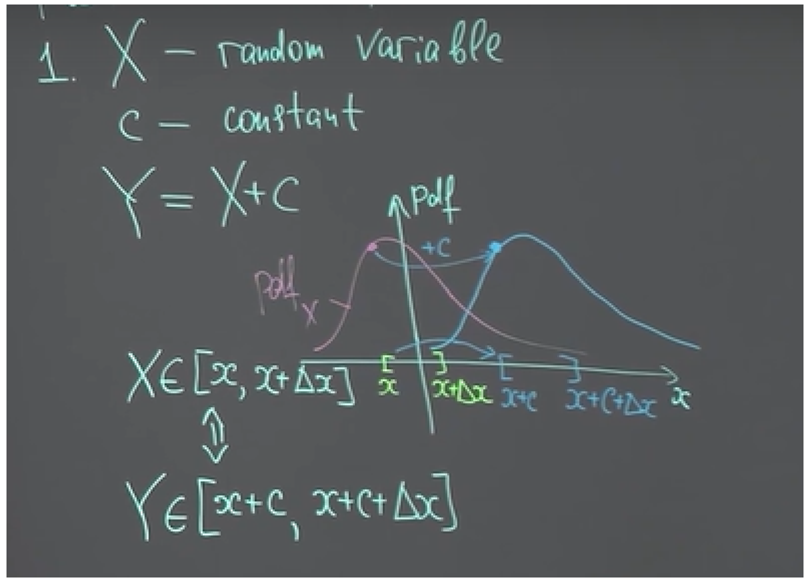
VarX = E(X - EX)^2 = ?

The correct answer is: 1/18

Transformations of continuous random variables and their PDFs

We often need to transform random variables, for example, by adding a constant or multipying random variable by a constant. So it's useful to understand how PDF changes when we do such a transform.

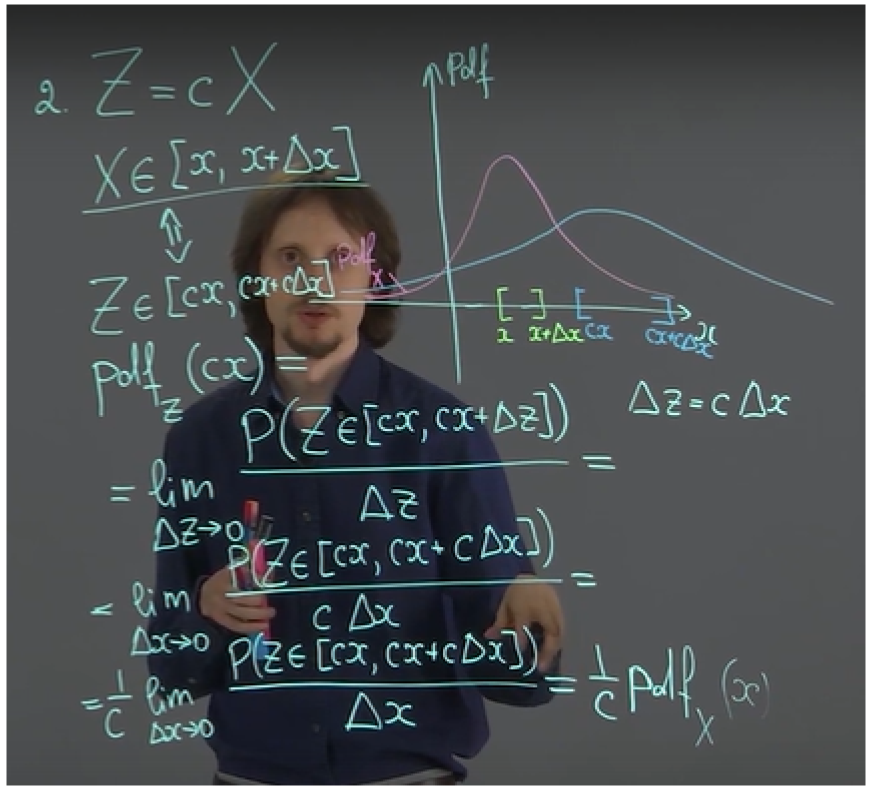
1. Adding a constant



If these two events happened simultaneously, their probability equals. However, the probability of event of getting our random variable X in yellow segment is involved in the definition of the value of pdf on the red point on the red line pdf X. In the same way, probability of event of getting our random variable Y is involved in definition pdf of the value of pdf on the blue point. So the red point was shifted to the right by adding a constant C such a whole graph.

2. Multipying random variable by a constant

- firstly, stretch curve by multiplying by C
- squeeze(сжать) graph vertically by 1/C



The fact that we have to compensate expanding in horizontal direction by contraction in vertical direction preserves the probability of getting somewhere, or in other sense, the area under blue curve to be equal 1 as for red curve.

Expectation of a function of random variable

Question 1

Find $\mathbb{E}(\pi \cos \pi x)$ for X uniformly distributed on [0,1].
Enter the value below:

Solution:

Find the integral with respect to x.
$\sin(\pi x) + C$

Tap to view steps...

$E(\pi \cos \pi x) \Rightarrow$ continuous r.v. $\Rightarrow E(x) = \int_a^b x \cdot f(x)\,dx$

$$f(x) = \begin{cases} 1, & \text{for } 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

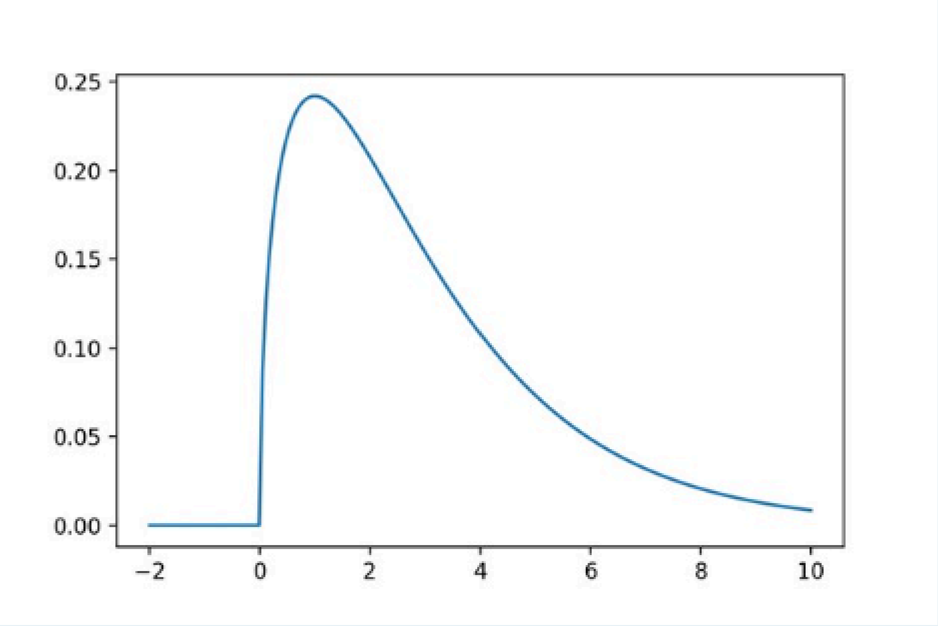
$$\int_0^1 (\pi \cos \pi x) \cdot 1\,dx = \sin \pi - \sin 0 = 0$$

The correct answer is: 0

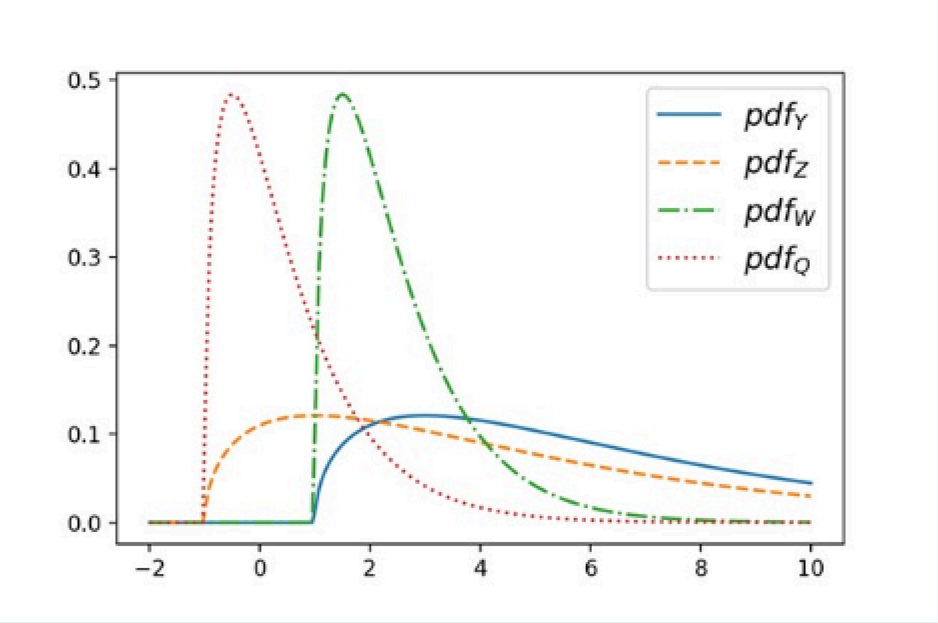
PDF skill test

Question 1

Continuous random variable X has PDF represented by the graph:



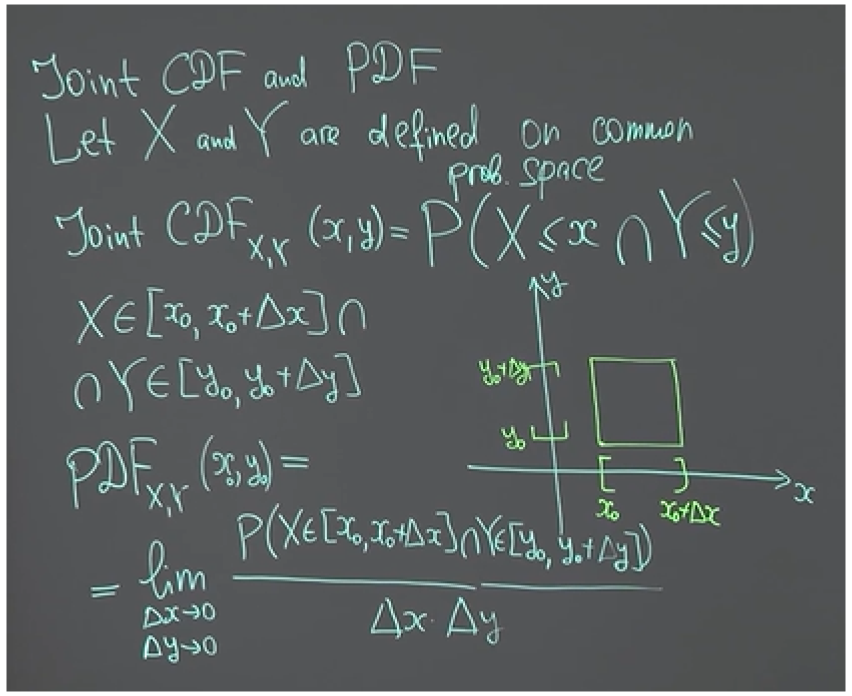
Which of the random variables represented by their PDF graphs below is 2X-1?



The correct answer is: Z

Joint CDF and PDF. Level charts. Marginal PDF

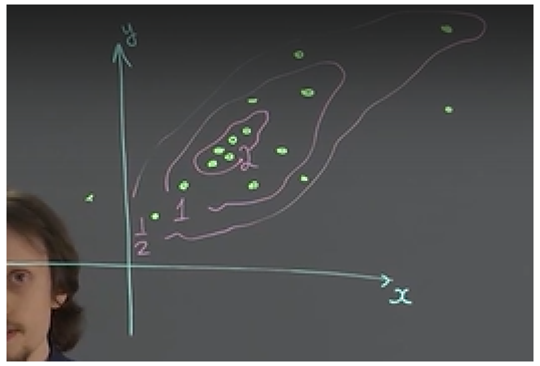
If we have two random variables and we are interested in iteraction between them, we have to consider joint distribution and joint pdf.



So, the points itself lies in small rectangle.

When delta X or delta Y becomes smaller this rectangle becomes smaller, and the probability becomes smaller correspondingly. So we are interested in relation between probabilities X and Y. And the area of rectangle is deltaX * deltaY.

The function of two variables can be visualized either by 3-dimensional graph or by its level curves. Let us use level curves to draw how pdf can looks like.

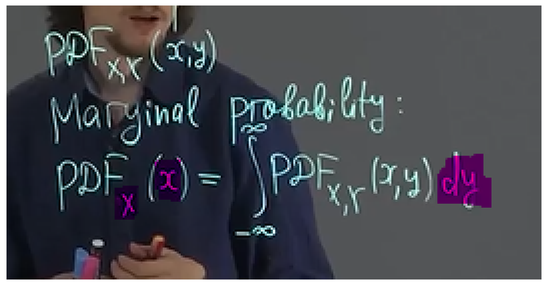


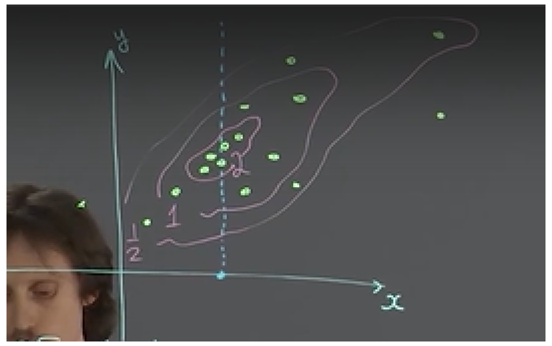
It's possible that in this area 2 values of pdf are large like greater than 2. In this area, they are between 1 and 2. And in this area, there are between 1 and 1/2. And the last area there're less than 1/2. It means that it's probable to have a point that is in the deep circle then in similar rectangle somewhere further to the left.

So pdf gives us information about how likely is to have a values of our random variables that together gives us a point near some particular point. If we will sample from the pair of random variables, which pdf looks like above pic, we will get more often points in the center circle, a little bit less often points in the 2d circle from the center, even less points on 3d circle and almost no points after 3d circle.

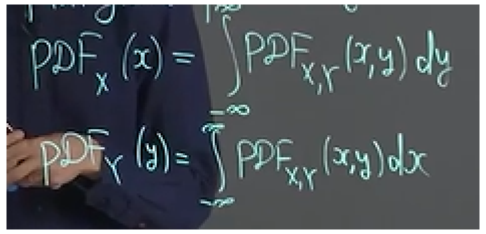
So this is a pin which we cann obtain when generate values from the pair of from random variables that we had chosen.

Now, like in discrete case, we can consider **marginal distribution**. So, if we have a pdf of a pair of twp variables, How can we get pdf of one variable, for example, x. In discrete case, we had to consider a sum **over all possible** values of the other variable, in this case **y**. In continuous case, we have to replace summation by integration.



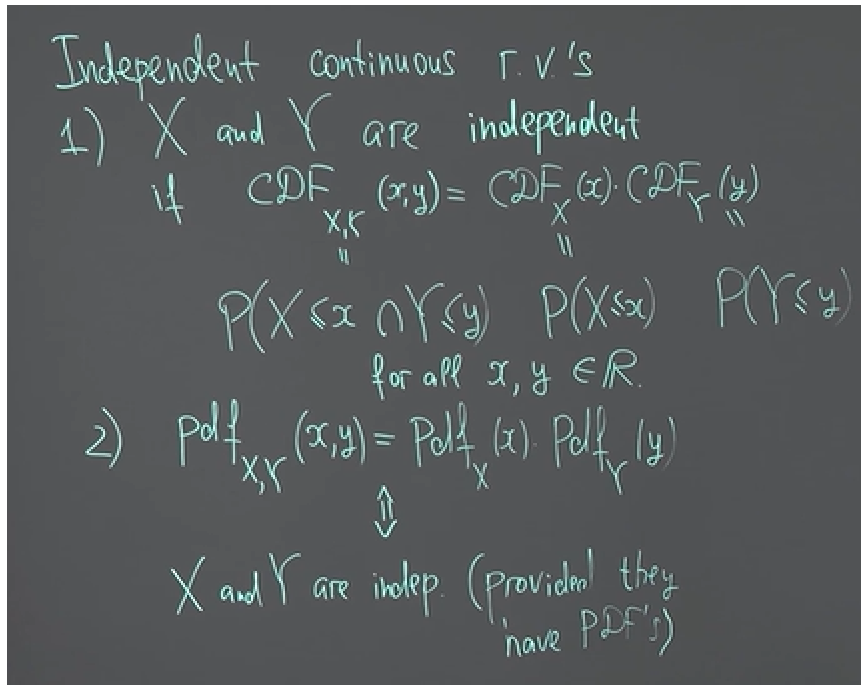


Geometrically that means that we fix some of X and we want to find pdf of random variable X at this point. In this case, we have to draw a straight line, and other this straigth line our joint probability is a function of one variable y. The integral of this funciton gives you the value of pdfX at the point x. When you're changing the point, you change blue line and get different values. This is a very close to what we did in discrete random variables. As we said, we just only have to change probabilities to pdf and summation by integration. In the same way we can define a pdf for y.

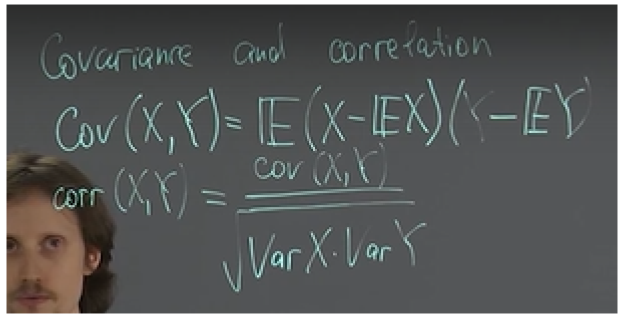


Independence, covariance and correlation of continuous random variables

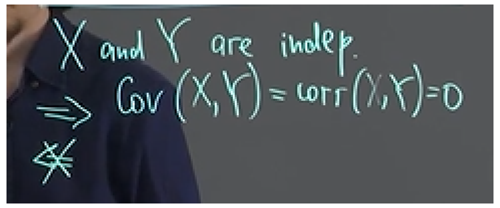
Important case is when two random variables are independent to each other. It means as in discrete case that knoweledge of the value of one of these variables gives you new information about possible values, and the probabilities of another variable. We can state it mathimatically in different ways.
First of all we can discuss it in terms of cdf(s).



So let's closely discuss covariance and corellation which is close to independence:
Just like in case of disccrete random variables:



Like in discrete random variables, if events are independent, their correlation and covariance will be equal 0.



Let us discuss now what does it mean that correlation between X and Y equals to 0 or 1?

It immediately follows that if two random variables are non correlated meaning that their covarance equals to 0. Then Variance of sum equald to sum of Variances.

Variance of sum of Gaussian random variables

Question 1

> Find the variance of the sum of two independent random variables with Gaussian distribution $\mathcal{N}(0, \sigma^2)$ with $\sigma = 2$. Enter the integer number:

Check out what σ means in definition of Gaussian distribution and use elementary property of variance

Solution:

Var(X, Y) = Var(X) + Var(Y) + 2Cov(X, Y)

σ(standard deviation)=2 => σ^2 (Variance) = 4

X and Y are independent => 2Cov(X, Y) = 0 => Var(X, Y) = Var(X) + Var(Y) = 4 + 4 = 8

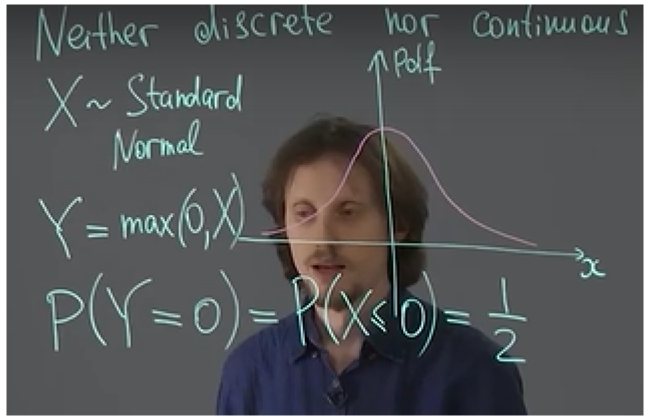
The correct answer is: 8

Mixed random variables. Example

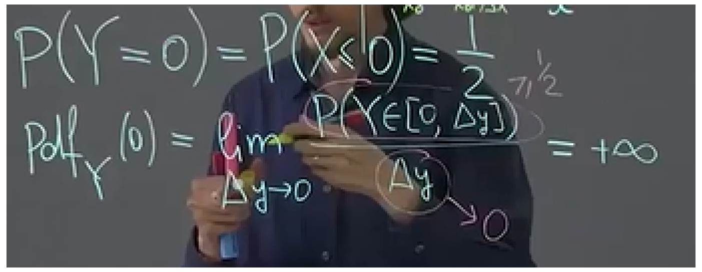
We discussed discrete random variables that can take only finite or countable number of values with some nonzero probabilies and continuous random variables that have a pdf. However, it's possible to consider random variables that are neither discrete non-continuous. Sometimes these random variables can be useful for our statistical modeling.

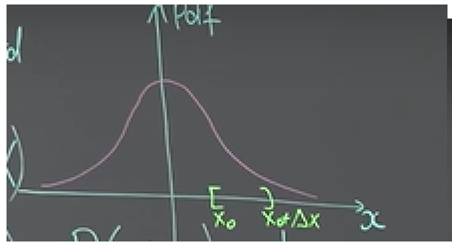
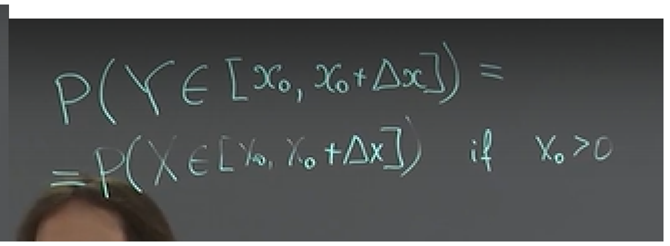
We begin with continuous random variable, which is distributed according to a normal distribution.

pdf is summetric



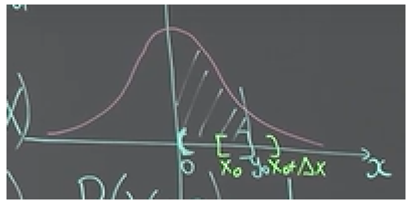
This means that random variable Y cannot have pdf defined at point zero.

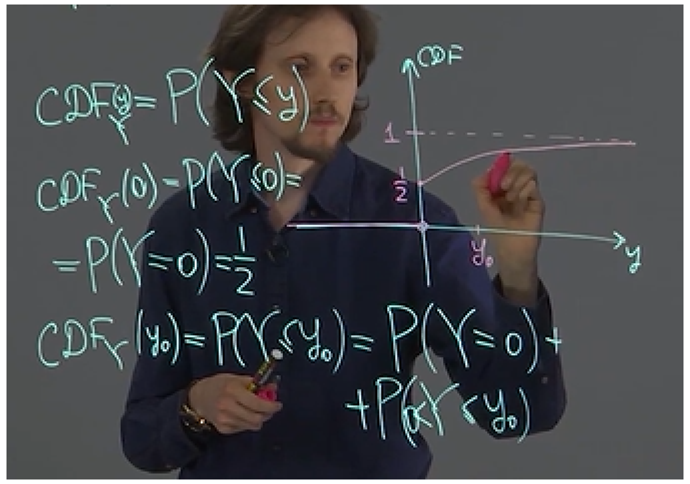


It means that for this part of the pic, we can define pdf for Y and it will the same as for X. So it will be a half of the standard normal distribution, the right half.

We see that, for example, probability of any point at the right side from zero are greater than 0 is zero as it goes for pdf. So we have a kind of hybrid random variable that on one hand behaves like discrete random variable, and on the other hand for different values, it behaves like a continuous random variable. We cannot defin pdf for random variable Y, however we can defne CDF.





As we can see we can define a little bit strange random variable by definding its cdf.

Distinguishing random variables

Question 1

An observer observes cars passing by her on a highway and builds random model of this process. In the model that she builds, what can be said about the following random variable:

The time moment of first car passing by her

- this random variable is discrete
- this random variable is mixed (neither discrete not continuous)
- this random variable is continuous ✔

Question 2

An observer observes cars passing by her on a highway and builds random model of this process. In the model that she builds, what can be said about the following random variable:

Number of cars that bypassed the observer within one hour

- this random variable is mixed (neither discrete not continuous)
- this random variable is discrete ✔
- this random variable is continuous

Question 3

90% of clients of an insurance company didn't report any insurance cases in 2019. Other 10% reported cases and got payouts. Considering insurance payout on a reported case being some continuous random variable, what can be said about the following random variable:

Payout value of the insurance company to a client in 2019

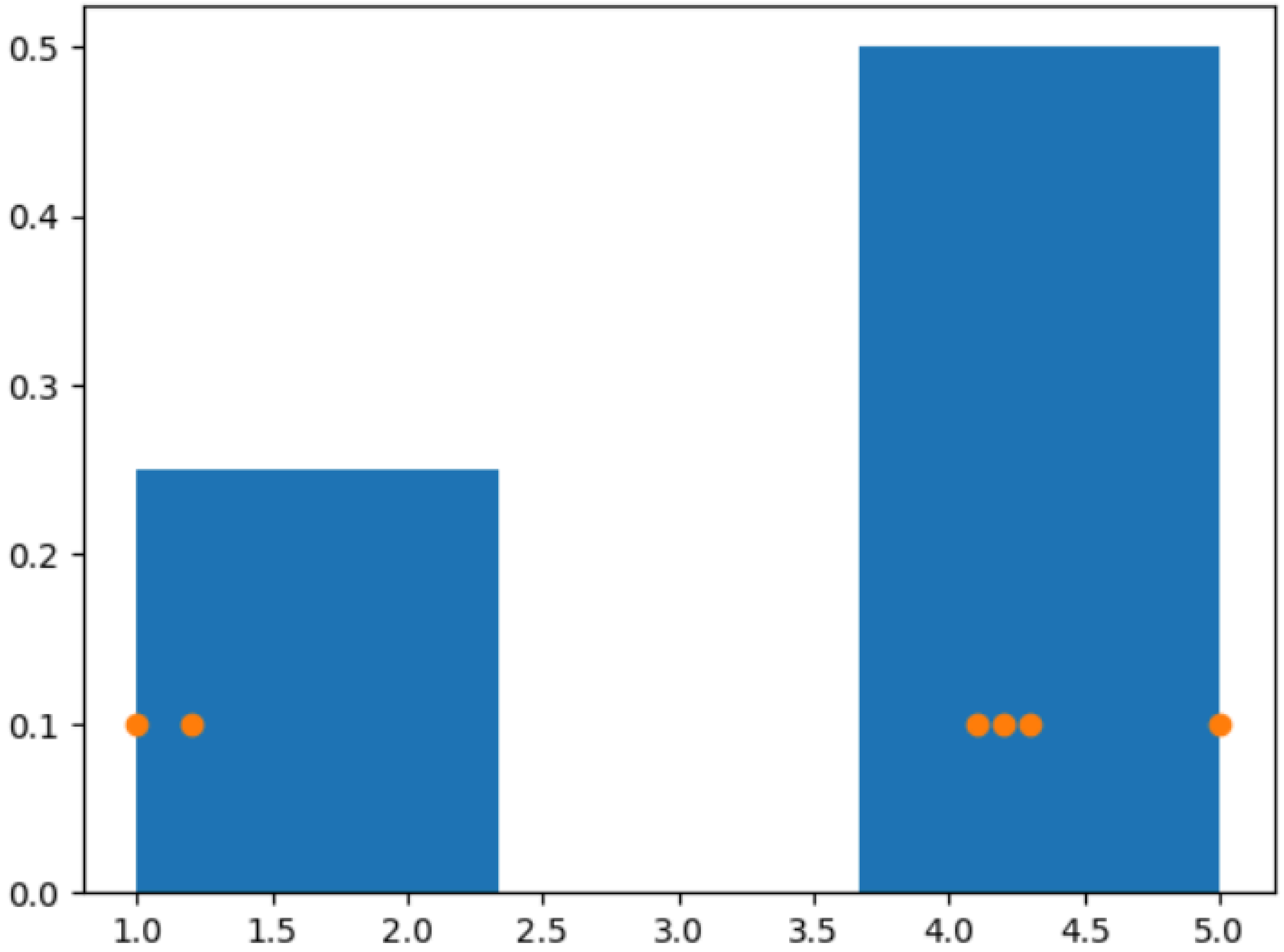
- this random variable is mixed (neither discrete not continuous) ✔
- this random variable is discrete
- this random variable is continuous

Когда чего-то половину нет, то функция не определена, но cdf существует. Это как правило neither discrete not continuous

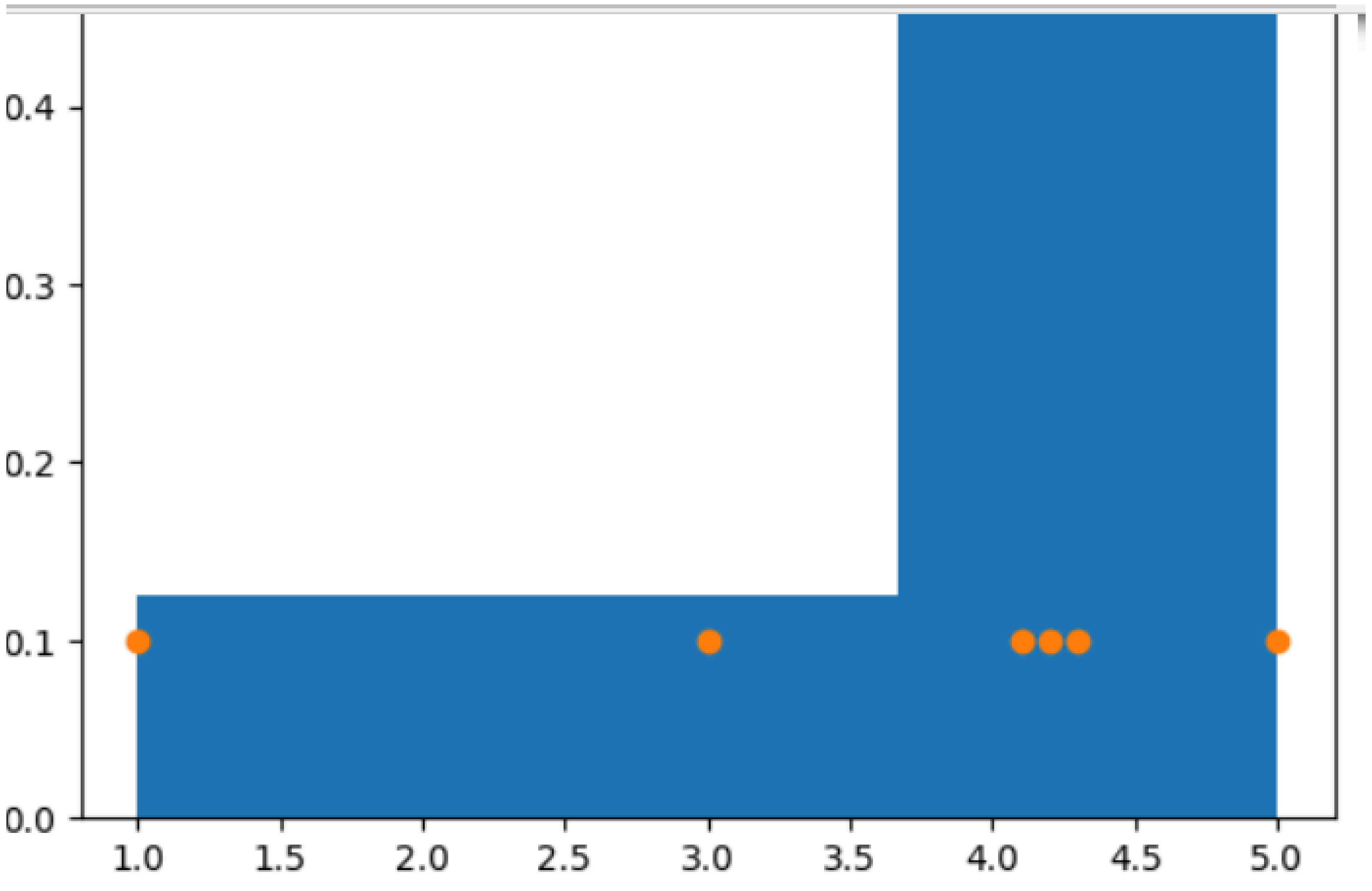
Generating and visualizing continuous random variables with Python

Previously we discussed how to generate discrete random variables like bimonial or Poisson distribution. Now, let us generate some continuous random variables like uniform. Again, we can use special functions from libraries.

```
from numpy. random import uniform

#segment low, where start; high, where end
#каждый раз разное число
uniform(low= -2, high = 2)
#0.3240354027864836



uniform(low= -2, high = 2, size = 10)
#array([-0.01868147,  0.72588419,  0.42288175, -1.73623413, -0.33741349,
        1.03193569,  0.16426092,  1.21748572, -0.49316448, -1.29554088])


import matplotlib.pyplot as plt
%matplotlib inline

data = [1, 1.2, 4.1, 4.2, 4.3, 5]

plt.hist(data, bins = 3, density = True)
#bins - кол-во столбиков
#density = True переделываем картинку так, чтобы размер рисунка был 1X1
plt.plot(data, [0.1] * len(data), 'o')
```

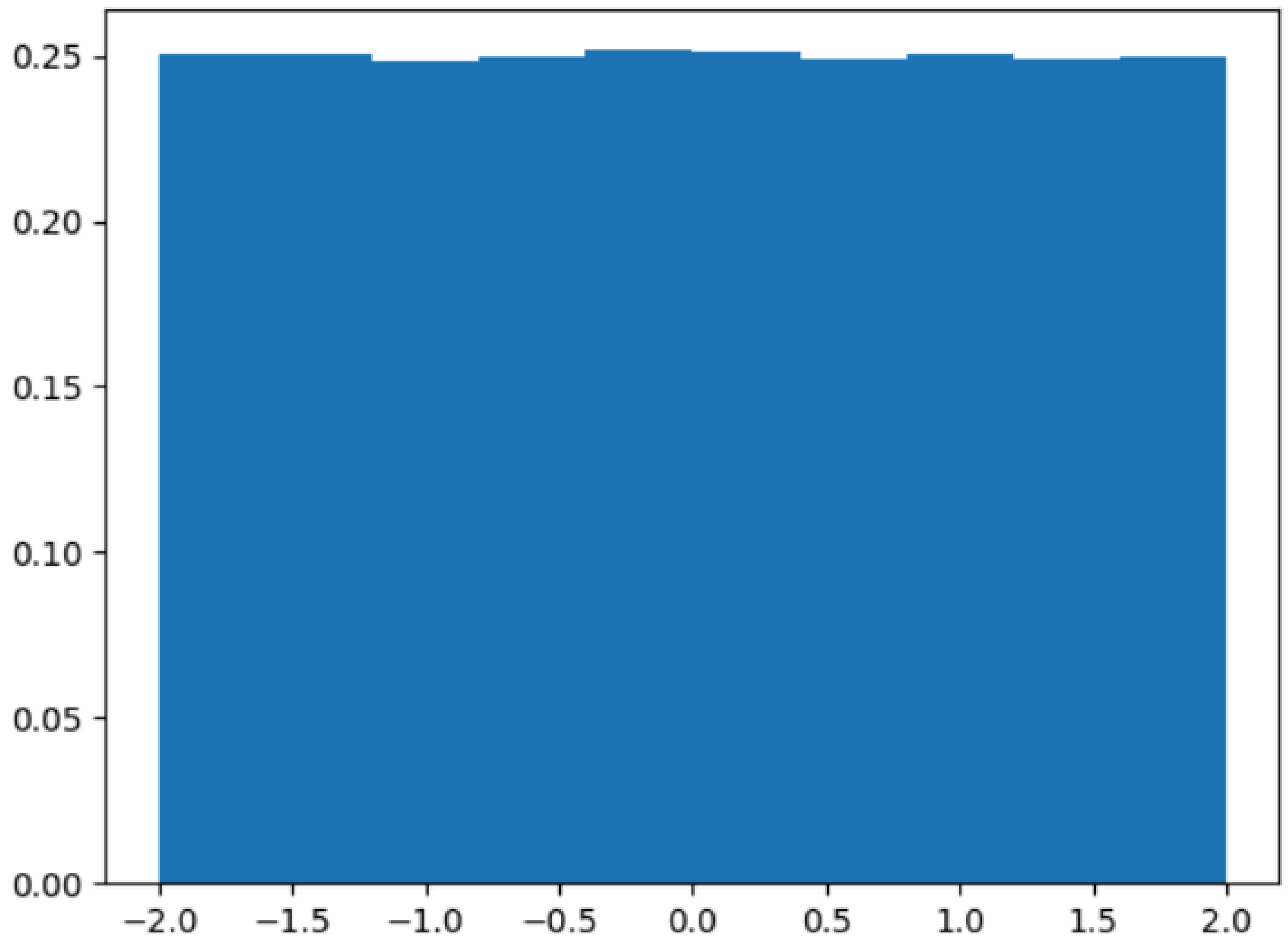



```
data = [1, 3, 4.1, 4.2, 4.3, 5]
plt.hist(data, bins = 3, density = True)
#bins - кол-во столбиков
#density = True переделываем картинку так, чтобы размер рисунка был 1X1
plt.plot(data, [0.1] * len(data), 'o')
```



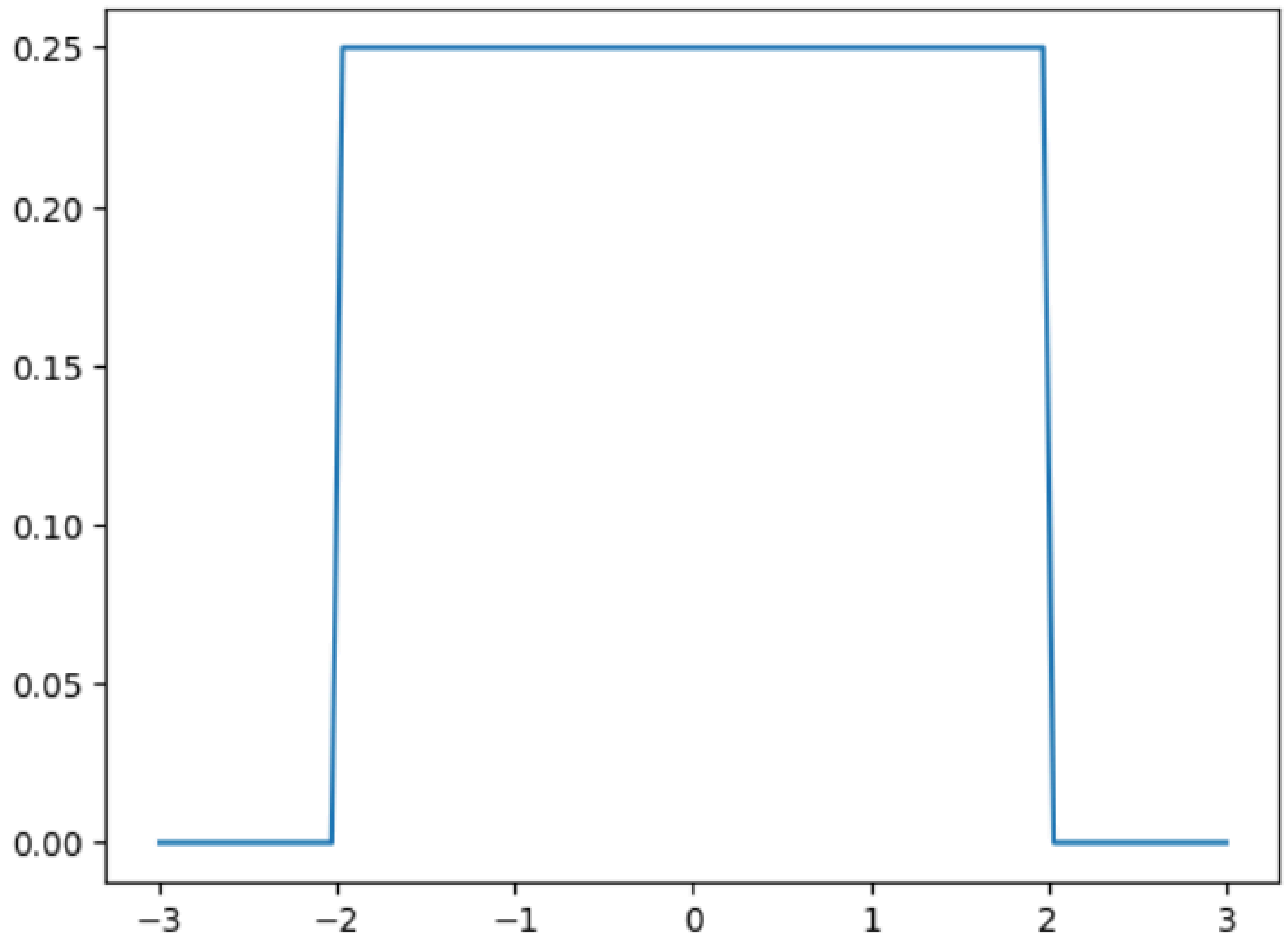
we have 2 rectangles with the same hight 1.

```
sample = uniform(low= -2, high = 2, size = 1000000)
plt.hist(sample, density = True)
```



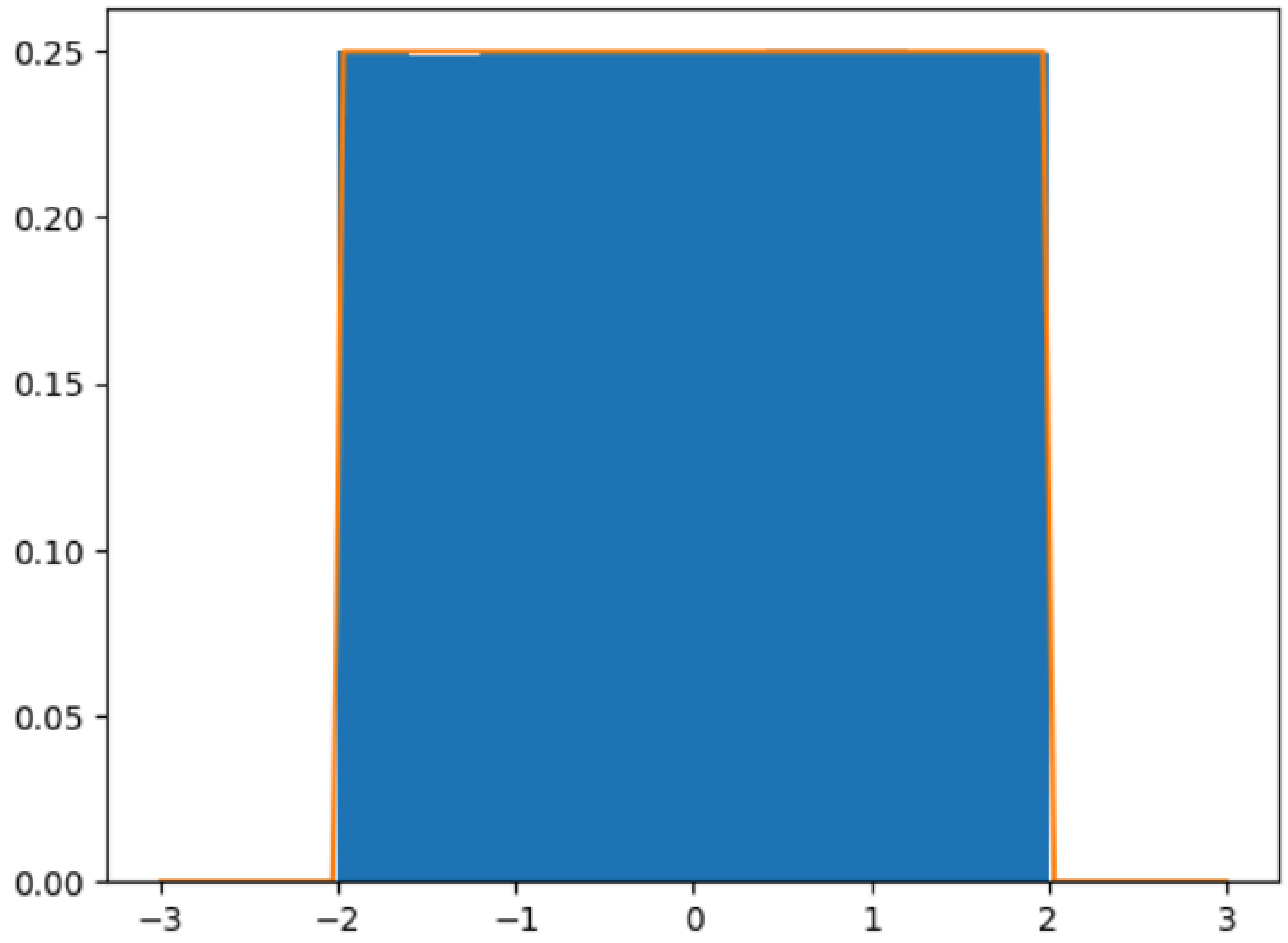
From this hostogram we see that the probability of getting a random variable on every segment is mpre or less the same provided that the length of these segments(rectangles) is the same. It works with size-> infinite

```
import scipy.stats

X = scipy.stats.uniform(-2, 4) #left end point, lenght of the segment
X.pdf(1.2) #0.25

import numpy as np

x = np.linspace(-3, 3, 100)
#left point, right point, size. create a list for pic
```



Совмещаем ручной вариант с вариантом библиотечным

```
sample = uniform(low= −2, high = 2, size = 1000000)
plt.hist(sample, density = True)
plt.plot(x, X.pdf(x))
```



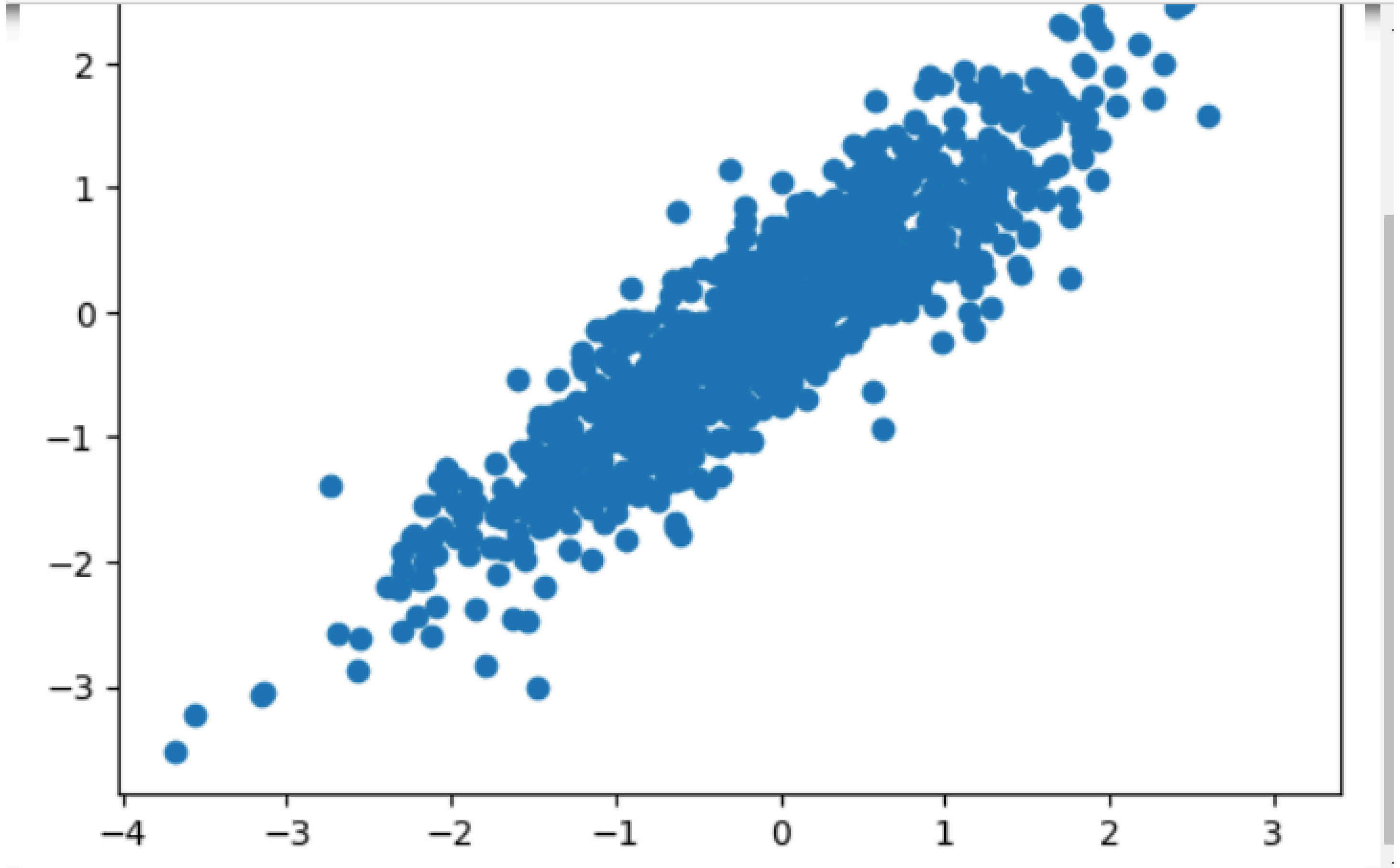
Now we see that our histogram actually approximates the corresponding pdf. And the larger sample size => the better approximation is.

Generating correlated random variables with Python

When you sample from some random variable using Python, these samples are independent to each other. But it's also possible to generate depend random variables, for example, correlated normal random variables. This can be dome using special function NumPy random multivariate normal. Let us use this funciton to see how correlation affects the joined distribution of random variables.

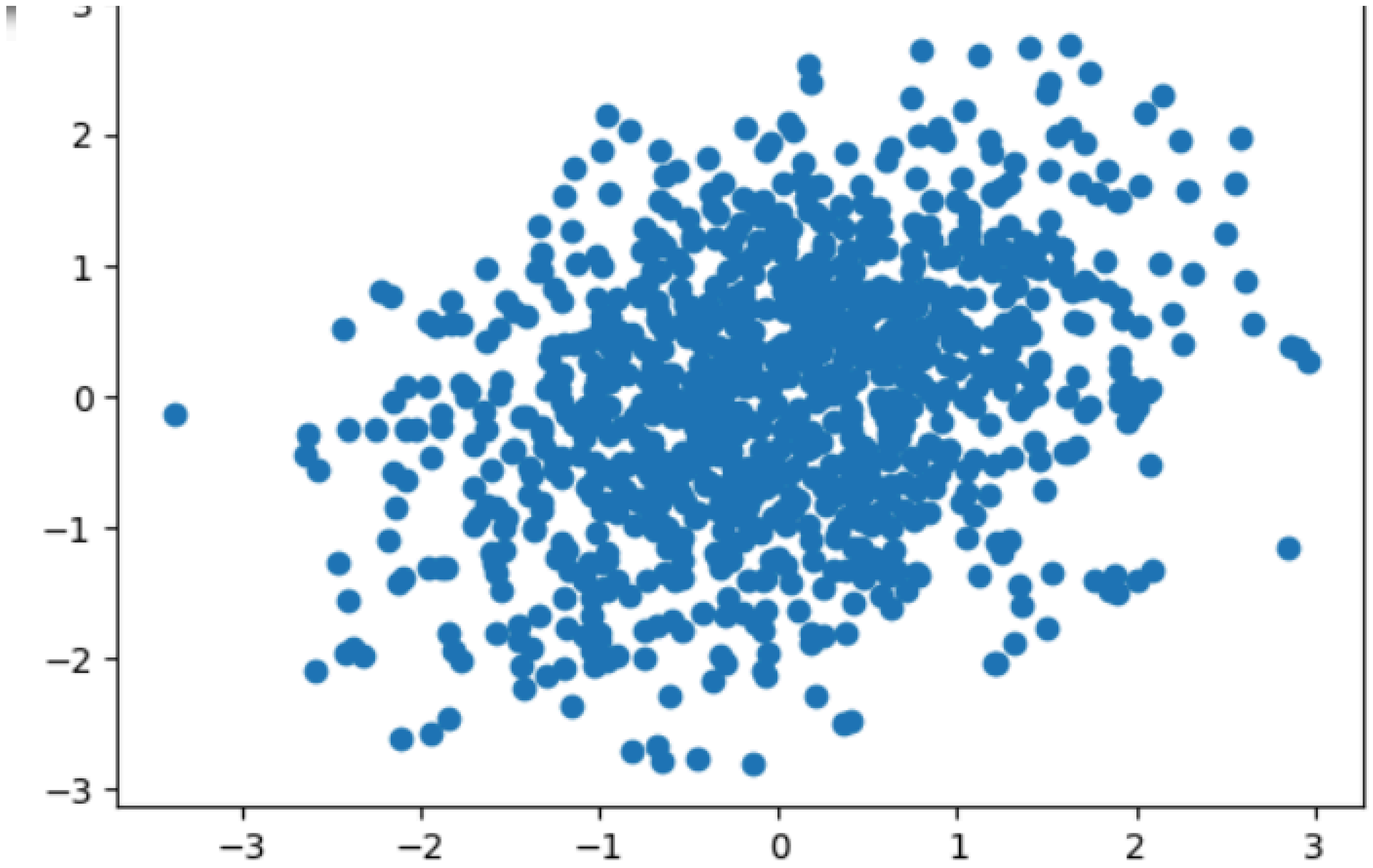
```
from numpy.random import multivariate_normal
import matplotlib.pyplot as plt
%matplotlib inline
import numpy as np

var1 = 1 #Variance
var2 = 1 #Variance
cov = 0.9
cov_matrix = [[var1, cov],
             [cov, var2]]
data = multivariate_normal([0,0], cov_matrix, size =1000)
#[0,0] - expected values
#It creates 2-dimensional matrix where 1st column - 1st r.v., 2d column - 2d r.v.
```

```
plt.scatter(data[:, 0], data[:, 1]) #1st r.v., 2d r.v.
```

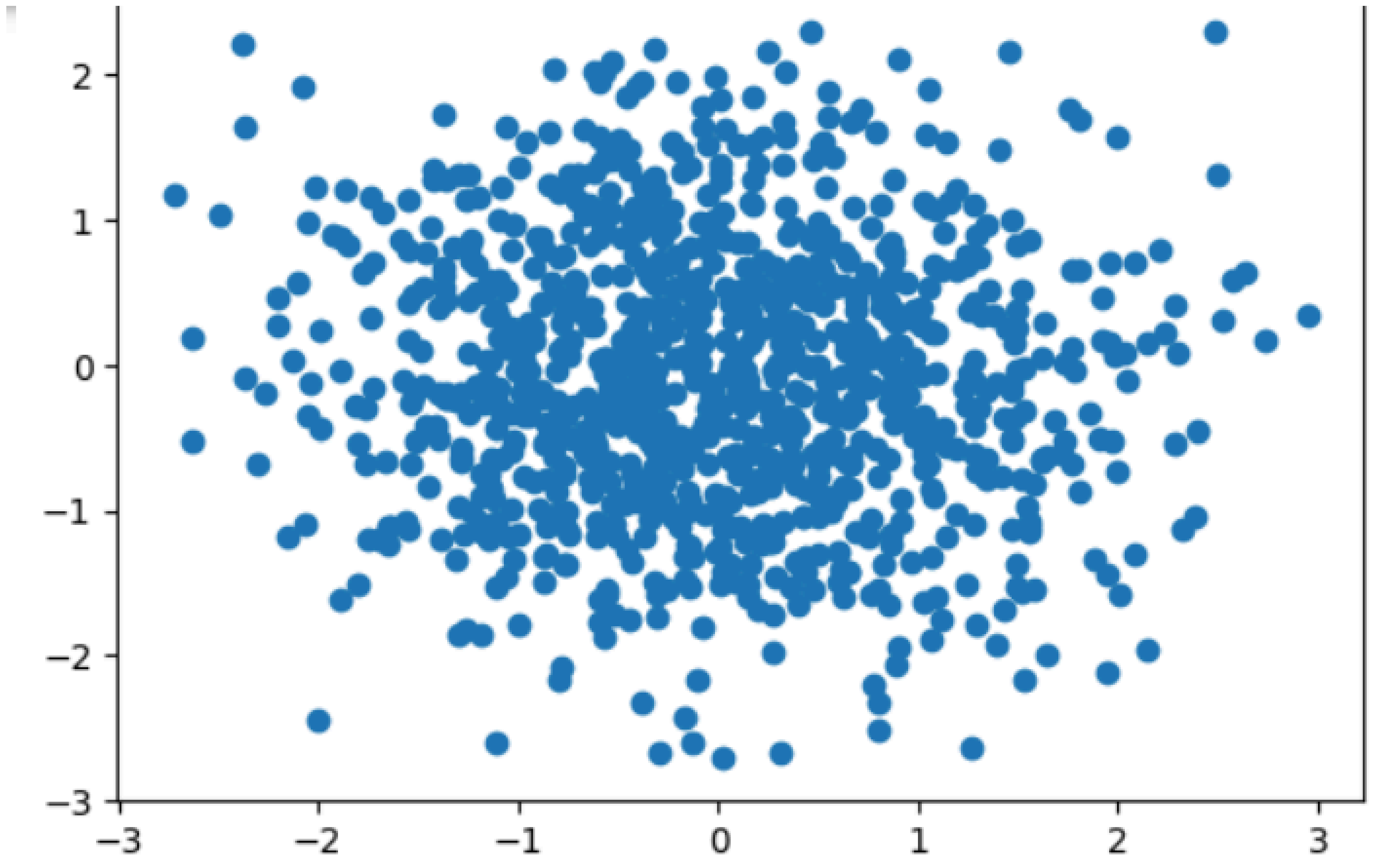


We can see a clear correlation between our random variables.

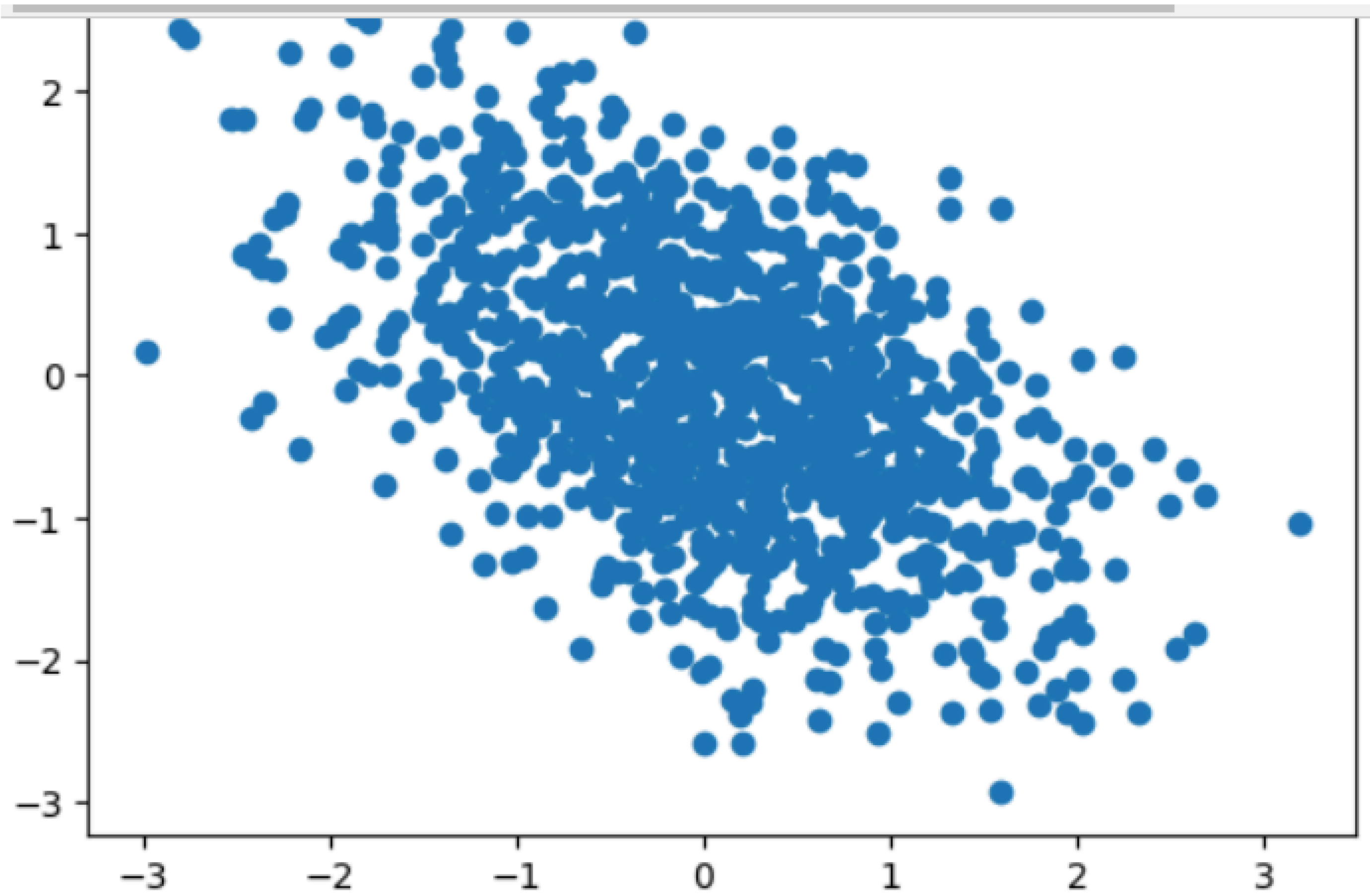
```
var1 = 1 #Variance
var2 = 1 #Variance
cov = 0.3
cov_matrix = [[var1, cov],
              [cov, var2]]
data = multivariate_normal([0,0], cov_matrix, size =1000) #[0,0] - expected values
plt.scatter(data[:, 0], data[:, 1]) #1st r.v., 2d r.v.
```



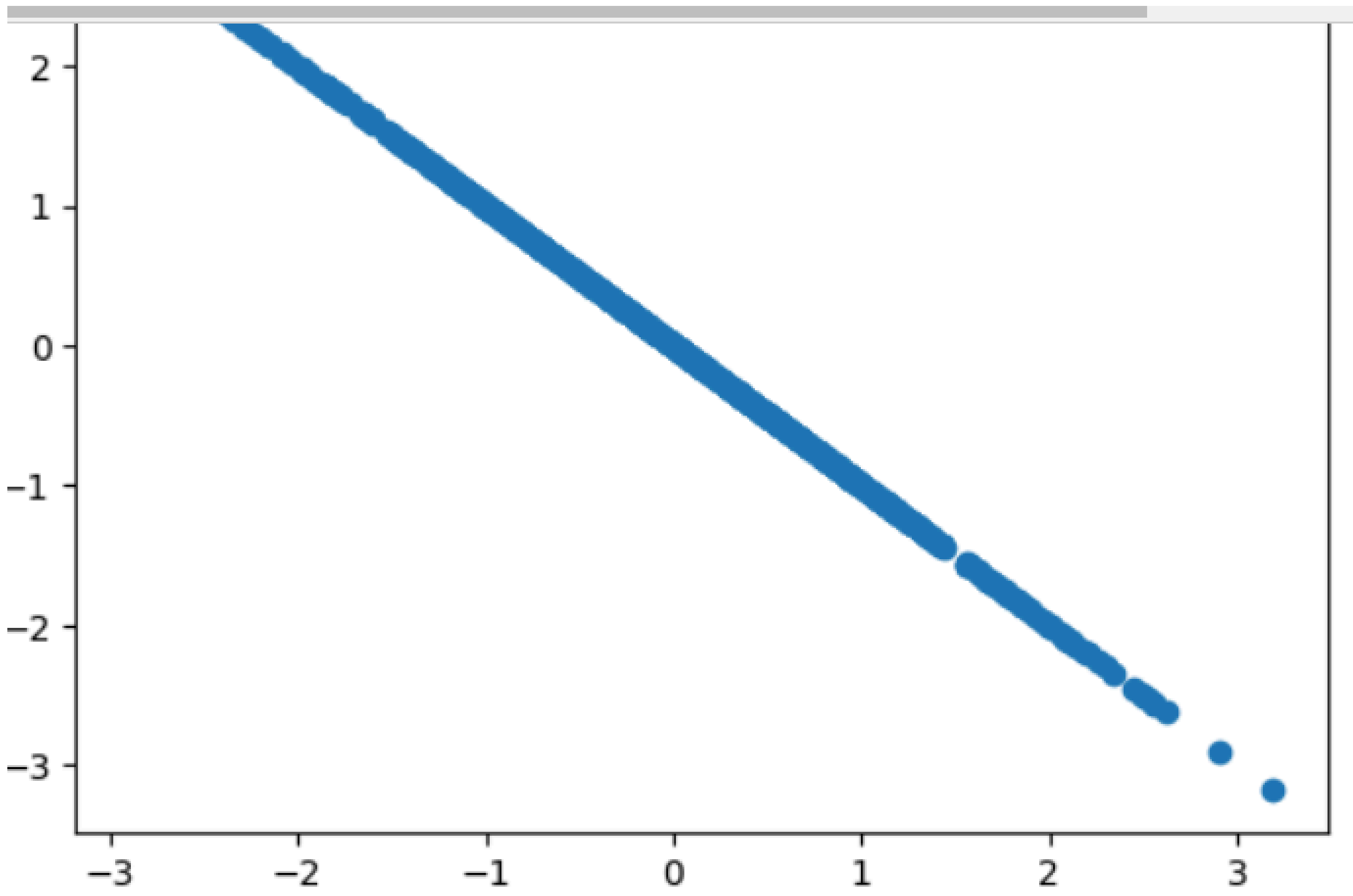
```
var1 = 1 #Variance
var2 = 1 #Variance
cov = 0 #no correlation
cov_matrix = [[var1, cov],
             [cov, var2]]
data = multivariate_normal([0,0], cov_matrix, size =1000) #[0,0] - expected values
plt.scatter(data[:, 0], data[:, 1]) #1st r.v., 2d r.v.
```



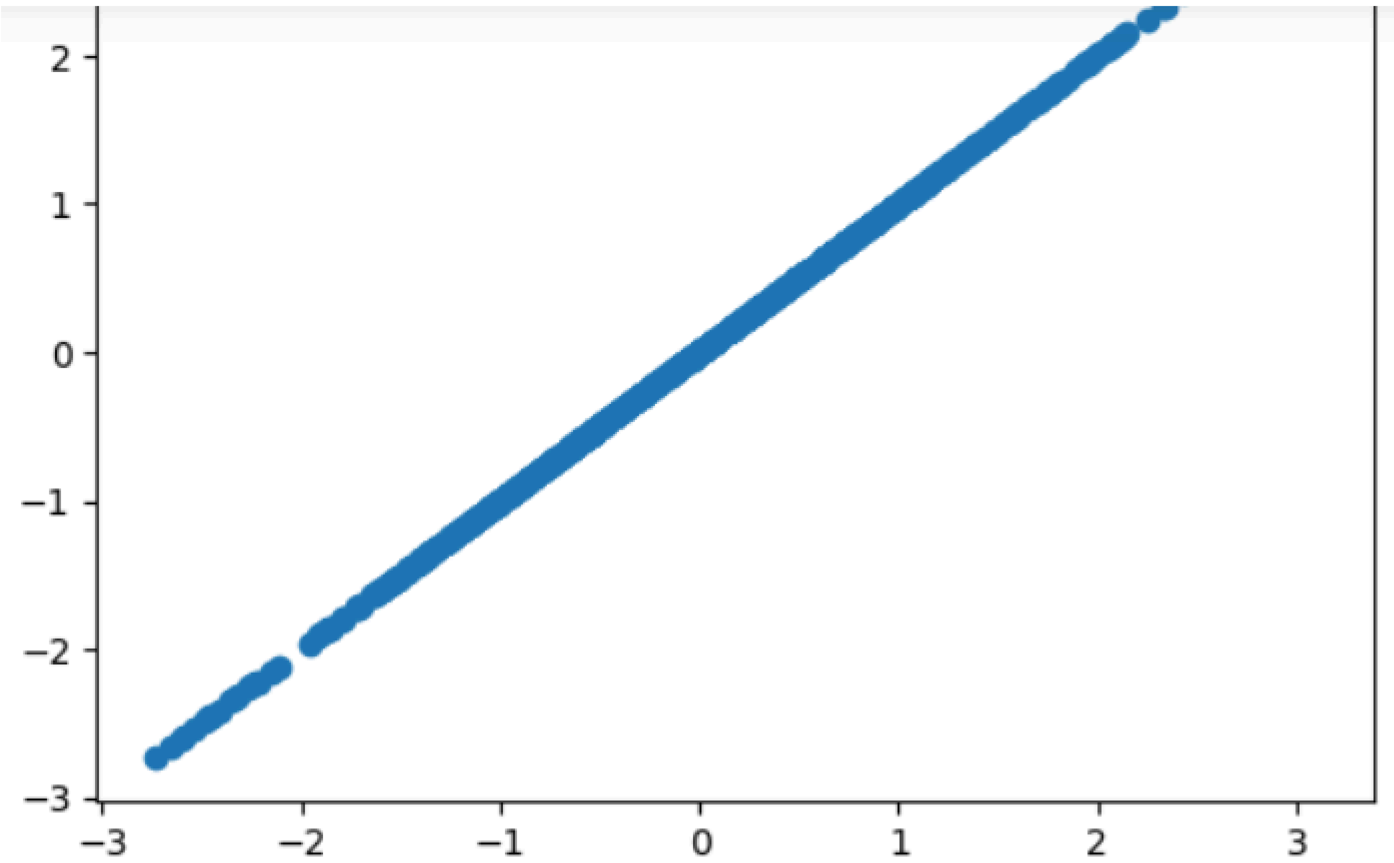

```
var1 = 1 #Variance
var2 = 1 #Variance
cov = -0.5 #no correlation
cov_matrix = [[var1, cov],
              [cov, var2]]
data = multivariate_normal([0,0], cov_matrix, size =1000) #[0,0] - expected values
plt.scatter(data[:, 0], data[:, 1]) #1st r.v., 2d r.v.
```



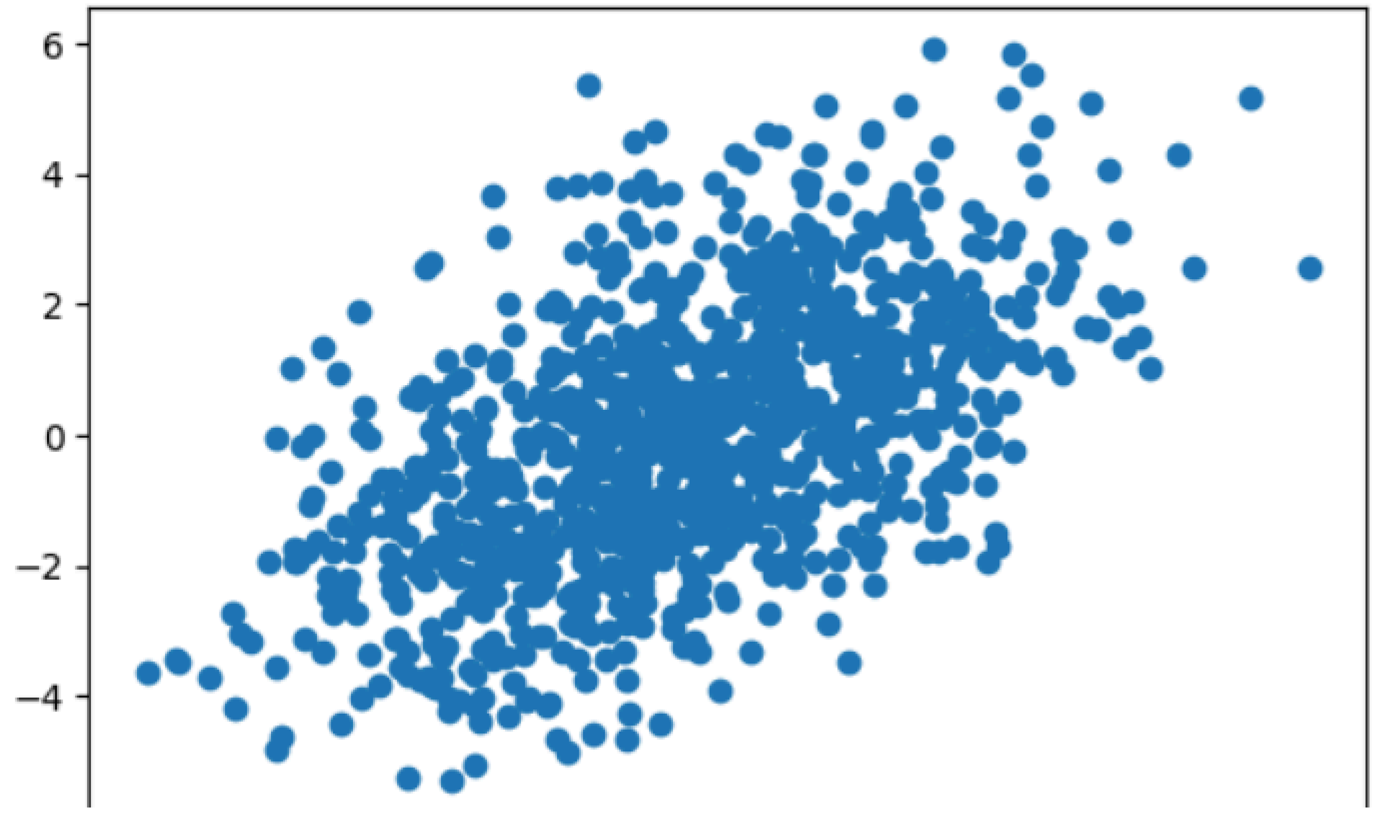
```
var1 = 1 #Variance
var2 = 1 #Variance
cov = -1 #no correlation
cov_matrix = [[var1, cov],
             [cov, var2]]
data = multivariate_normal([0,0], cov_matrix, size =1000) #[0,0] - expected values
plt.scatter(data[:, 0], data[:, 1]) #1st r.v., 2d r.v.
```




```
var1 = 1 #Variance
var2 = 1 #Variance
cov = 1 #no correlation
cov_matrix = [[var1, cov],
             [cov, var2]]
data = multivariate_normal([0,0], cov_matrix, size =1000) #[0,0] - expected values
plt.scatter(data[:, 0], data[:, 1]) #1st r.v., 2d r.v.
```

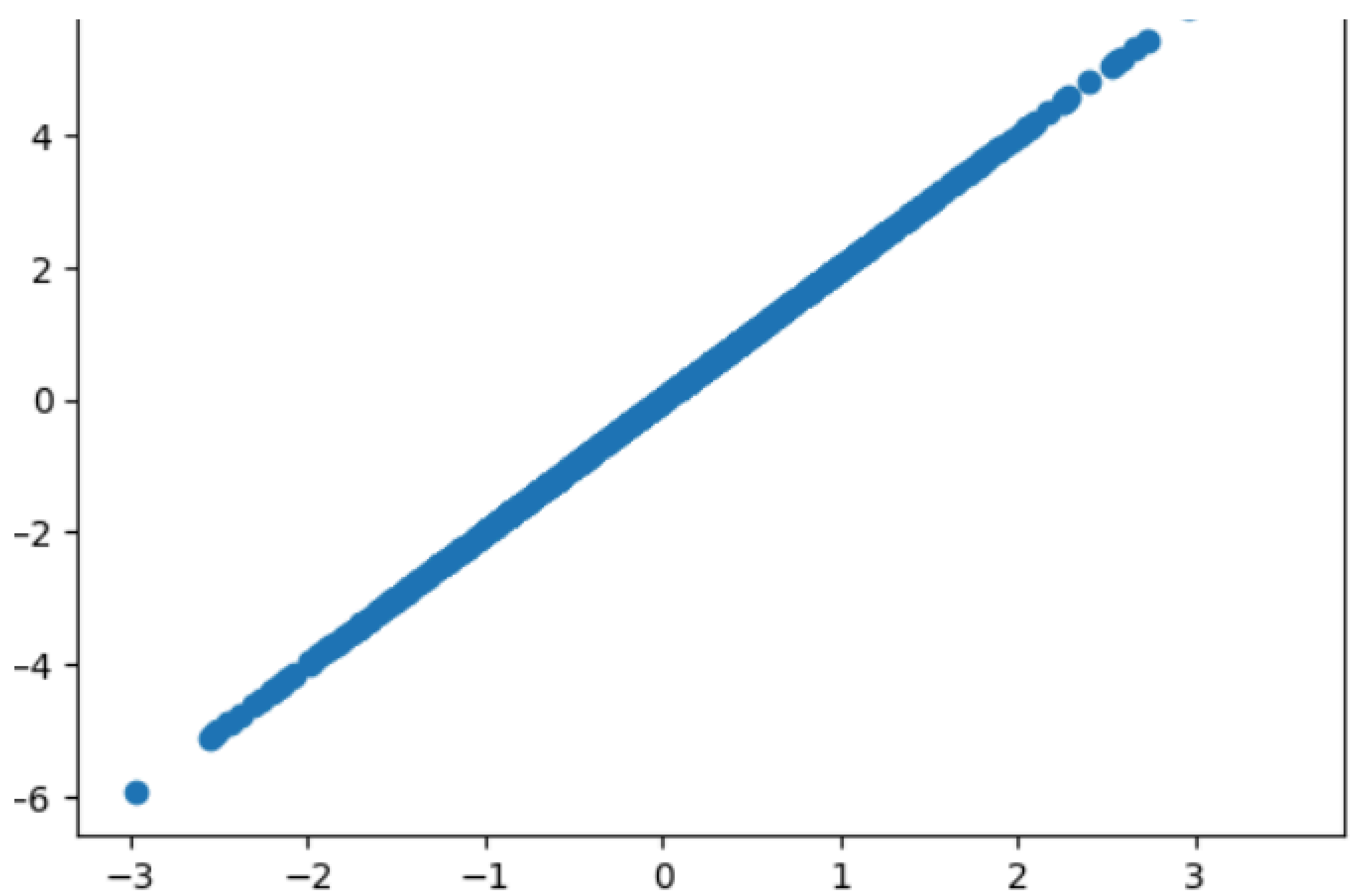



```
var1 = 1 #Variance
var2 = 4 #Variance
cov = 1
cov_matrix = [[var1, cov],
             [cov, var2]]
data = multivariate_normal([0,0], cov_matrix, size =1000) #[0,0] - expected values
plt.scatter(data[:, 0], data[:, 1]) #1st r.v., 2d r.v.
```



The range of the 2d variable increase on the y-axis on the pic. But not give us perfect correlation.

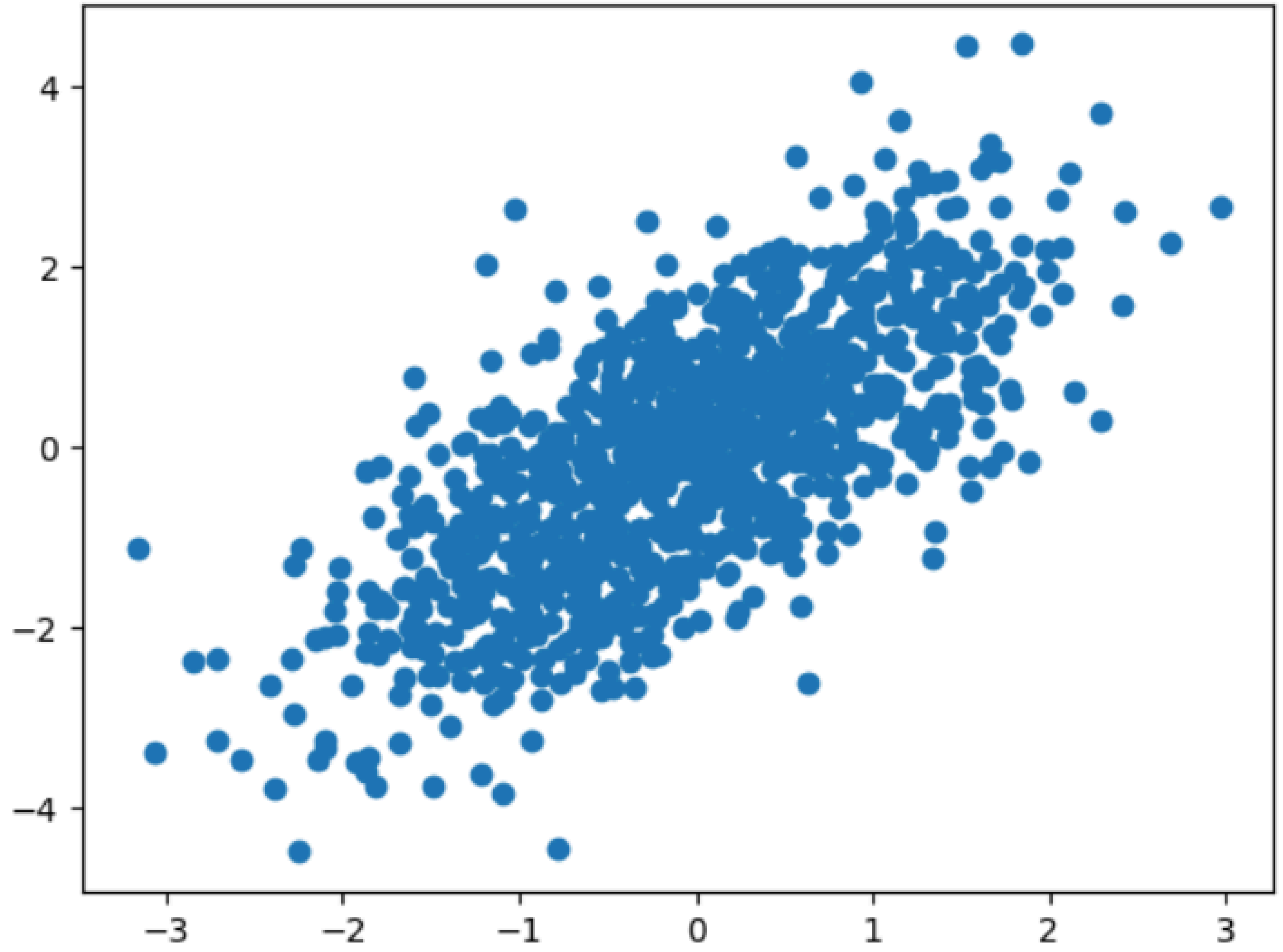
```
var1 = 1 #Variance
var2 = 4 #Variance
cov = 2
cov_matrix = [[var1, cov],
             [cov, var2]]
data = multivariate_normal([0,0], cov_matrix, size =1000) #[0,0] - expected values
plt.scatter(data[:, 0], data[:, 1]) #1st r.v., 2d r.v.
```



And again correlation = 1

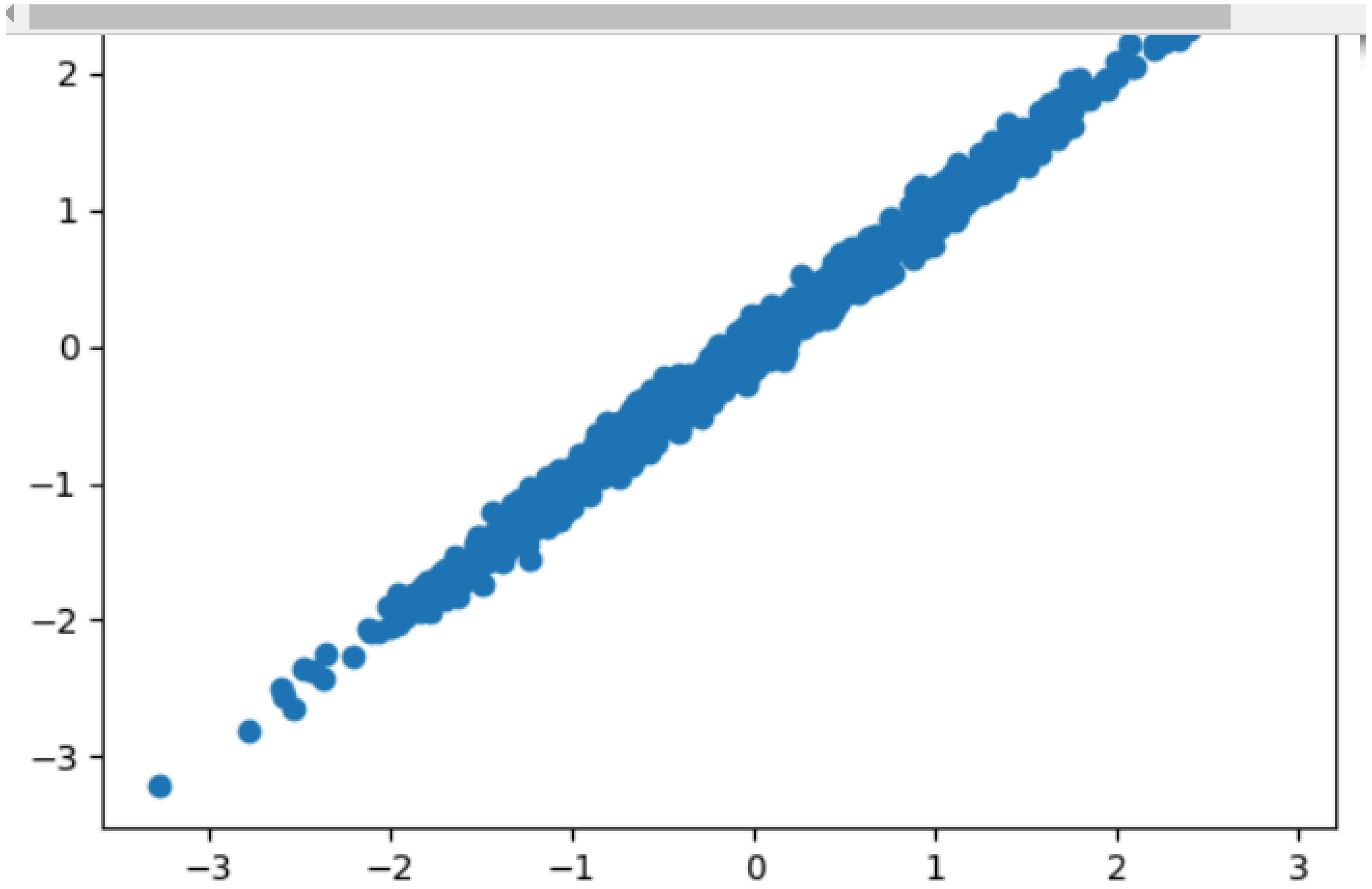
How can we generate some correlated random variables without using funtion from liabrary?

```
N = 1000
x = np.random.normal(size=N) #use standard normal distribution for simplicity
eps = np.random.normal(size=N)
y = x + eps
plt.scatter(x, y)
```

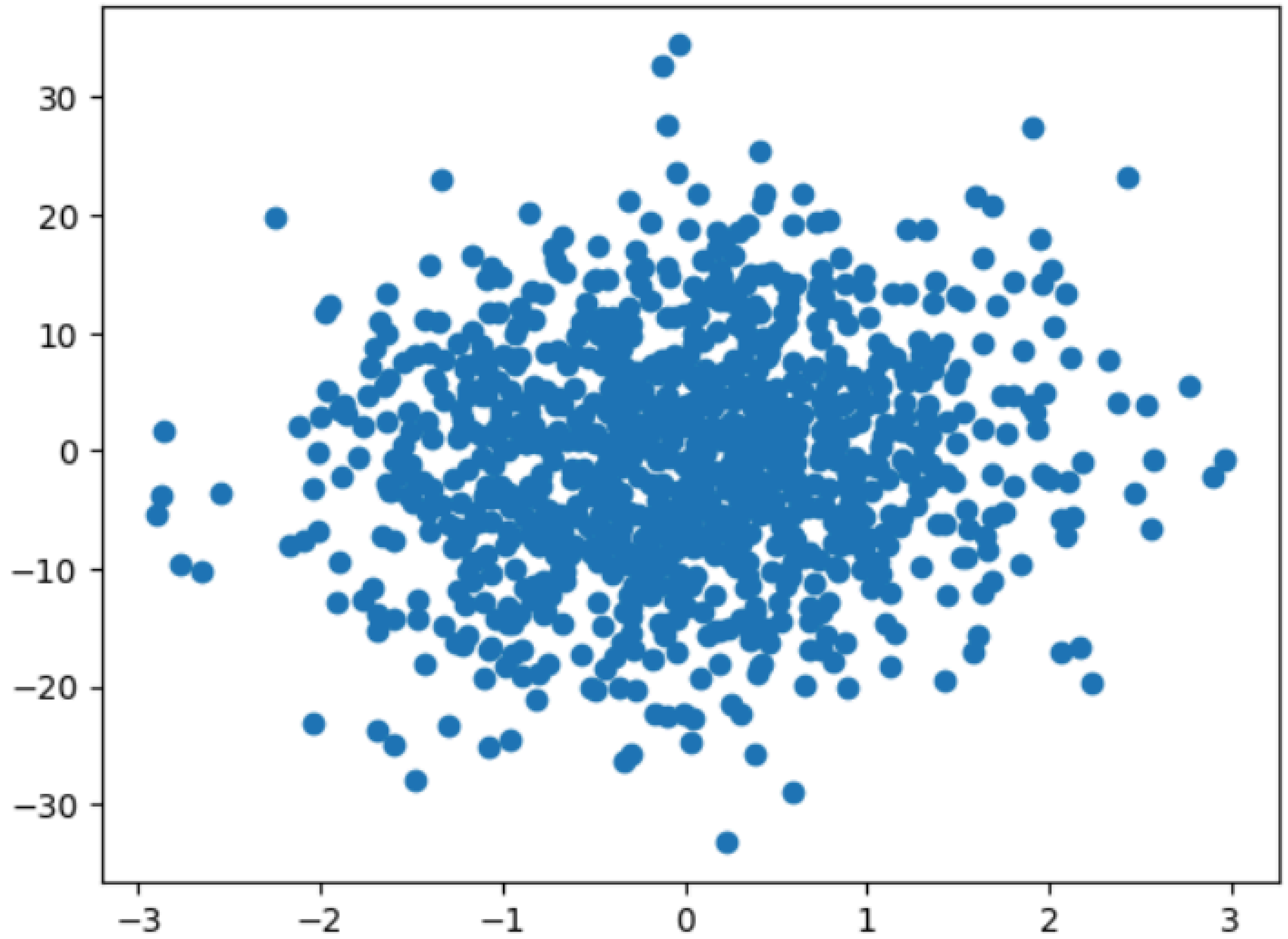


We can change correlation coeficient by multiplying by coefficient eps.

```
N = 1000
x = np.random.normal(size=N) #use standard normal distribution for simplicity
eps = np.random.normal(size=N)
y = x + 0.1 * eps
plt.scatter(x, y)
```




```
N = 1000
x = np.random.normal(size=N) #use standard normal distribution for simplicity
eps = np.random.normal(size=N)
y = x + 10 * eps
plt.scatter(x, y)
```


30
20
10
0
−10
−20
−30
−3
−2
−1
0
1
2
3