Lecture 4 – Basic Statistics

Bulat Ibragimov

bulat@di.ku.dk

Department of Computer Science
University of Copenhagen

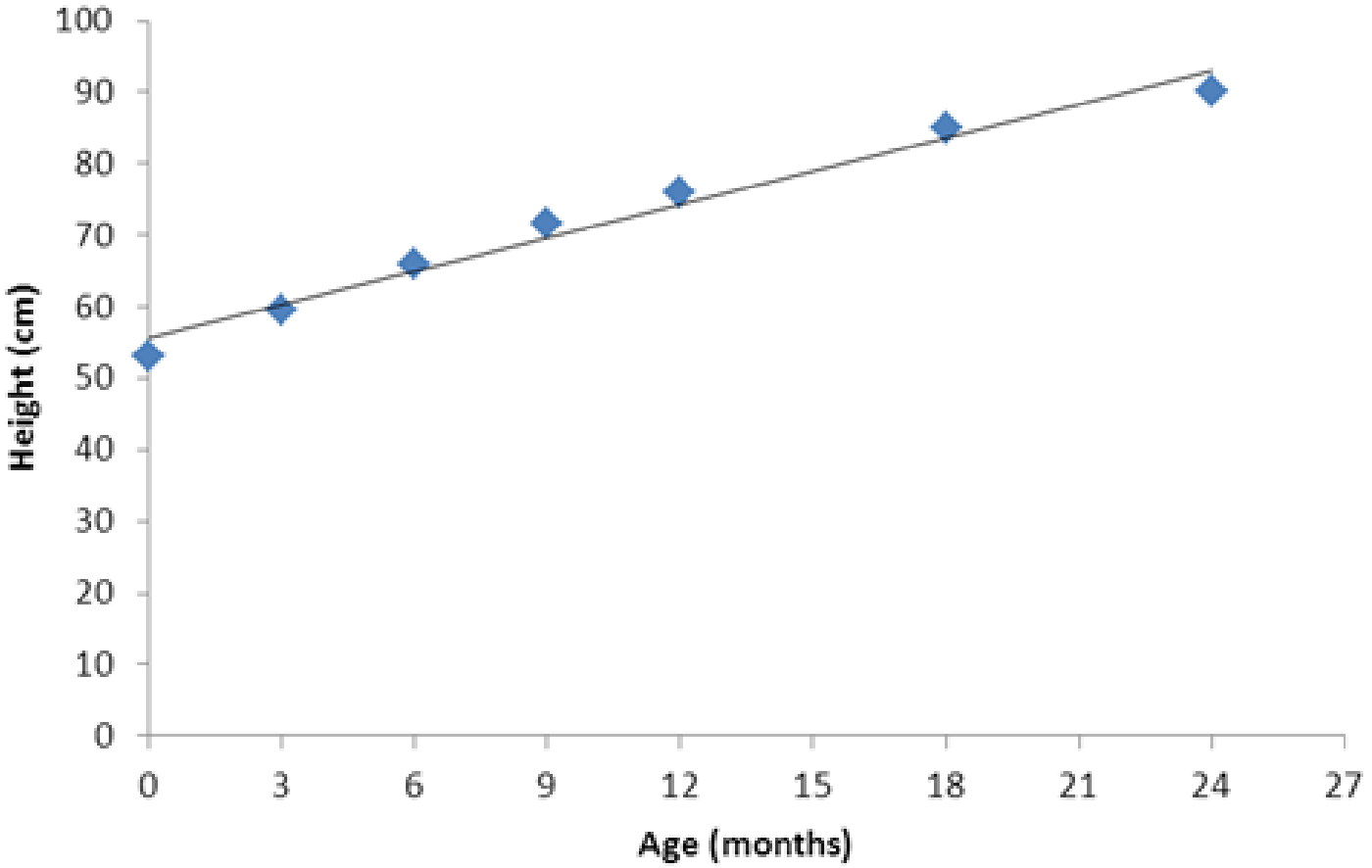
UNIVERSITY OF COPENHAGEN

Lecture X - today

Discrete random variables
Continues random variables
Mean and standard deviation
Bayes' rule
Simple distributions

Random events

- Height vs age in children
- The linear model does not fit perfectly. Why?



Random events

Children height depends on may factors:

- Genetics
- Diet
- Health
- etc.

Children height consists of deterministic and random parts

Discrete random variables

The total number of different outcomes is limited

Example:

- Tossing a coin

Outcomes:

- Ω ={Head (H), Tail (T)}

If the coin is fair, the probability of outcomes:

- P(Y = H) = 0.5
- P(Y = T) = 0.5

The sum probability of all outcomes is always one

Continues random variables

The total number of different outcomes is unlimited

Example:

- Throwing a coin on a round table to see how far from the center it will land

Outcomes:

- $\Omega = [0, R]$

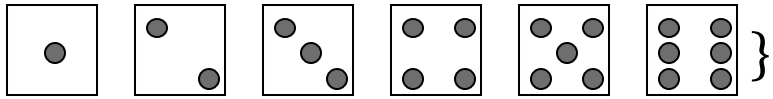
We cannot calculate probability for exact distance, but we can calculate probability for intervals

The probability for the complete interval [0, R] is again one

Adding probabilities

What is the probability of a die landing on x < 4?

Outcomes:

$\Omega = \{$  $\}$

Probability:

$$P(Y < 4) = P(Y = 1) + P(Y = 2) + P(Y = 3) = 1/6 + 1/6 + 1/6 = 0.5$$

Note that events should be **mutually exclusive**!

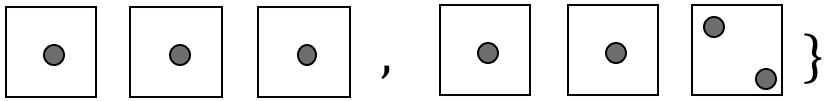
Adding probabilities

What is more likely when you roll three dice:

1) the total is x = 3

2) the total is x = 4

Outcome: 6^3 combinations:

$\Omega = \{$  $\}$

Probability:

$$P(x = 3) = P(A=1, B=1, C=1) = 1/216 = 0.0046$$

$$P(x = 4) = P(A=2, B=1, C=1) + P(A=1, B=2, C=1) + P(A=1, B=1, C=2) = 3/216 = 0.014$$

Conditional probabilities

Example:

- Probability that a person gets university degree is 0.6 (X = 1)
- Probability that a person with university degree gets a well-paid job is 0.7 Y = 1
- Probability that a person without university degree gets a well-paid job is 0.4

Conditional probabilities:

- P(Y = y | X = x) – probability that Y = y happens considering that X = x happened
- P(Y = 0 | X = 0) =
- P(Y = 0 | X = 1) =
- P(Y = 1 | X = 0) =
- P(Y = 1 | X = 1) =

Joint probability

What is the probability that two random persons will get a university degree?

- These events are independent, so the joint probability is multiplication of individual probabilities:
  $$P(Y_1 = 1, Y_2 = 1) = P(Y_1 = 1) \cdot P(Y_2 = 1) = 0.6 \cdot 0.6 = 0.36$$

What is the probability that a person will get a university degree and a well-paid job?

- These events are dependent, we need to use conditional probabilities:
  $$P(Y = 1, X = 1) = P(Y = 1|X = 1) \cdot P(X = 1) = 0.7 \cdot 0.6 = 0.42$$

What are the probabilities of other scenarios for an arbitrary person?

Joint probability

$$P(Y = y, X = x) =$$
$$P(Y = y|X = x) \cdot P(X = x) = P(X = x|Y = y) \cdot P(Y = y)$$

Let's check what are the values of $P(Y = y|X = x)$ and $P(X = x|Y = y)$ for our example with university degrees and incomes?

	Degree	No degree
Well-paid	0.6 · 0.7	0.4 · 0.4
Low-paid	0.6 · 0.3	0.4 · 0.6

	Degree	No degree
Well-paid	0.42	0.16
Low-paid	0.18	0.24

Added, conditional and joint probability: example 1

Input:

- A woman was killed, and her husband is a suspect
- The husband had been abusing the wife
- Defense attorney statement:
  - Only 0.01% of the men who abuse their wives end up murdering them
  - Therefore, the fact that Simpson abused his wife is irrelevant to the case (By irrelevant, he means that the probability of abusiveness importance is very low)
- Why is this a wrong use of conditional probability?

Minecraft speedrun: example 2

Input:

- In 2020, a streamer Dream livestreamed himself speedrunning Minecraft, demonstrating record performance
- He was accused of using a "doctored" version of Minecraft where the probability of important events to happen was increased
- Let's investigate this case

Minecraft speedrun: example 2

Trading with piglins:

- You give piglins gold bars, and they return you some other items
- In speedrun, you are interested in getting pearls from these trades
- The chance of getting a pearl is 20/423 = 0.0473

During his livestreams:

- 262 barters
- 42 pearls traded

Minecraft speedrun: example 2

Getting blaze rods:

- You kill blazes
- From each blaze, you have 50% chance to get a blaze rod

During his livestreams:

- 305 blaze kills
- 211 blaze rods

Minecraft speedrun: example 2

Pearls:

- How many pearls would you expect on average from 262 trades (0.047 probability of getting a pearl)

- 12 pearls (he got 42 pearls)

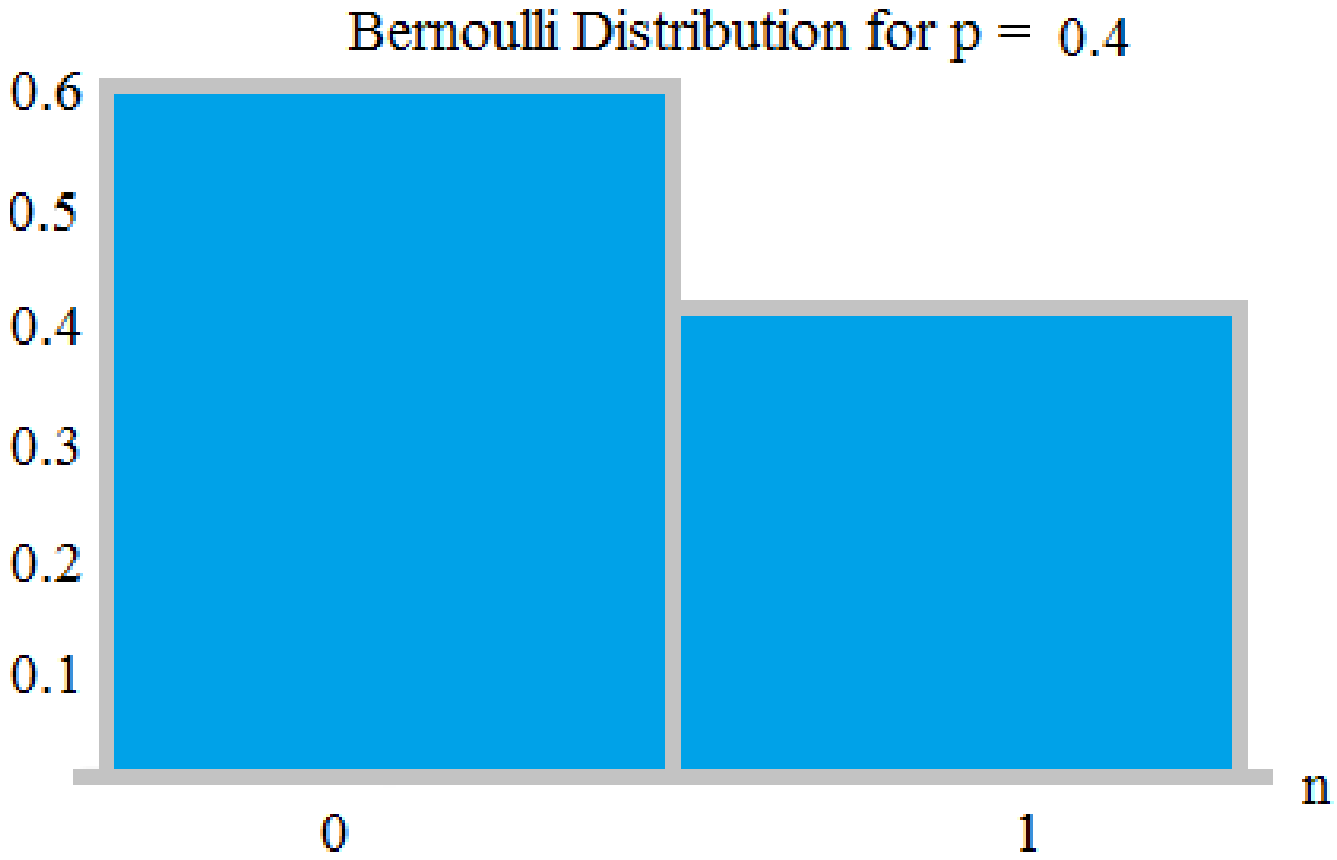
- How many blaze rods would you get from 305 blaze kills (50% probability)

- 152.5 blaze rods (he got 211 rods)

Minecraft speedrun: Bernoulli distribution

Coin tossing is a good example

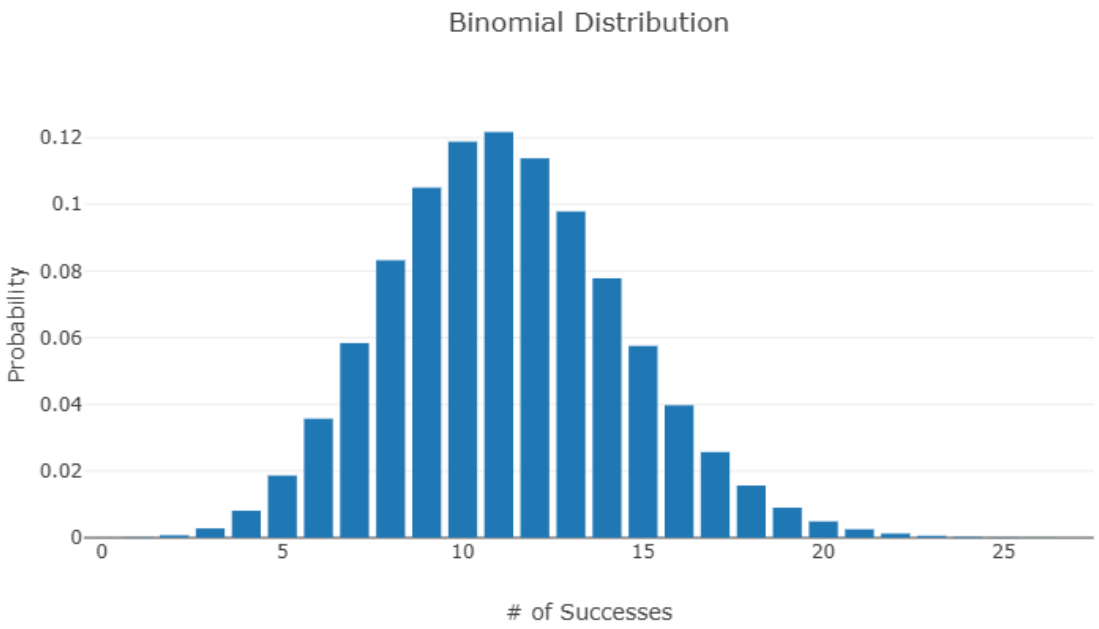
$$P(X= x)= p^x * (1-p)^{1-x}$$



Minecraft speedrun: Binominal distribution

We trade with piglins n times and get k pearls?

$$\mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

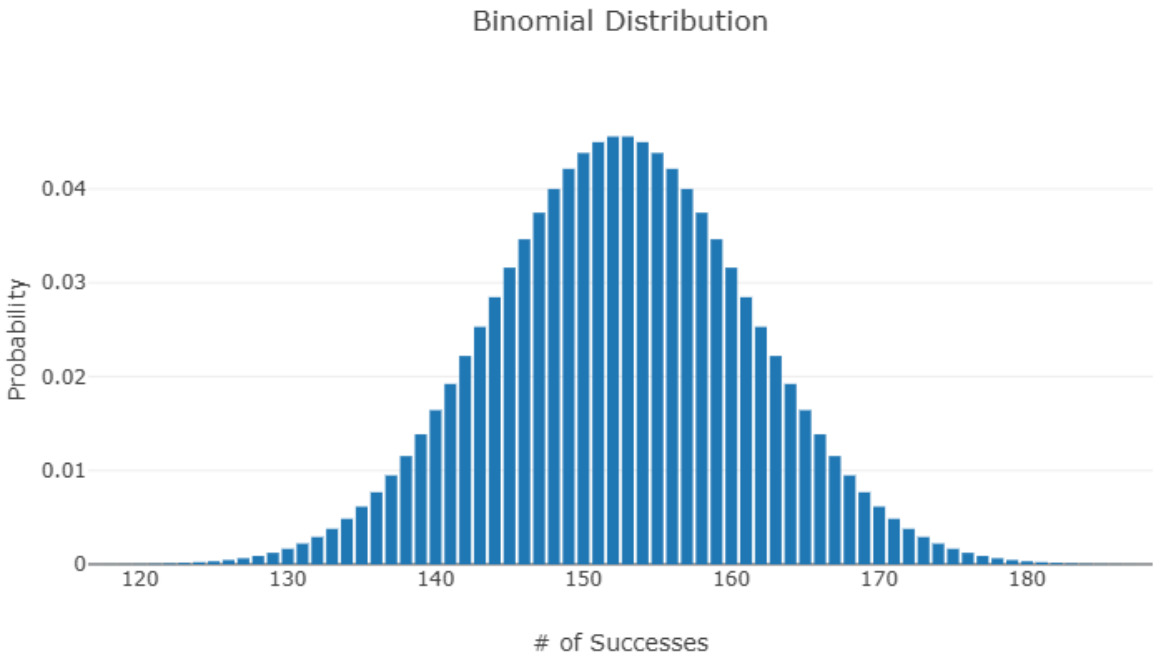


Probability of getting 42 pearls

$4 \cdot 10^{-12}$

Minecraft speedrun : Binomial distribution

For n=305 blaze kills with p=0.5:



Probability of getting 211 rods

$$4.9 \cdot 10^{-12}$$

Minecraft speedrun: Objection 1

We want to check if Dream's extraordinary luck is realistic:

- In his case, we want to check how likely is to be as lucky as he was ***or even luckier***.
- Probability of getting 42 pearls, ***or 43 pearls, or 44 pearls,⋯ or 262 pearls*** in 262 trades.

- $P(X \geq 42) = P(X = 42) + P(X = 43) + \cdots P(X = 262) = 5.6 \cdot 10^{12}$

- Probability of getting 211 rods or more from 305 kills.

- $P(X \geq 211) = P(X = 211) + P(X = 212) \ldots = 8.8 \cdot 10^{12}$

Minecraft speedrun: More objections

The probability of being lucky both with pearls and rods:

$$\left(\mathbf{5.6 \cdot 10^{-12}}\right) \cdot \left(\mathbf{8.8 \cdot 10^{-12}}\right) = \mathbf{4.9 \cdot 10^{-23}}$$

Objections:

- Many people are playing the game, it could have been Mr. Smith or Mr. Xing or Ms Anderson etc instead of Dream.
- Dream plays for many days, today he was lucky, but yesterday he might have been unlucky. But we analyze the day that is "*most accusatory*" for Dream
- Early stoppings

Minecraft speedrun: Resolution

What should be the odds of an event plausible to be done by a human?

- 100 years = 3,153,600,000 seconds
- Number of humans that live at the same time <10,000,000,000
- If every human alive would try something every second, the total number of attempts will be less than $3 \cdot 10^{19}$
- This number is way smaller than Dream's odd of $4.9 \cdot 10^{-23}$

Bayes' rule

From joint probability formulation:

$$P(Y = y, X = 1) = P(Y = y|X = x) \cdot P(X = x) = P(X = x|Y = y) \cdot P(Y = y)$$

We can get Bayes' rule:

$$P(X = x|Y = y) = \frac{P(Y = y|X = x) \cdot P(X = x)}{P(Y = y)}$$

Bayes' rule

Returning to our example of university degrees and success (probability of university degree, on condition of well-paid job):

$$P(X = 1|Y = 1) = \frac{P(Y = 1|X = 1) \cdot P(X = 1)}{P(Y = 1)}$$

	Degree	No degree
Well-paid	0.6 · 0.7	0.4 · 0.4
Low-paid	0.6 · 0.3	0.4 · 0.6

	Degree	No degree
Well-paid	0.42	0.16
Low-paid	0.18	0.24

Bayes' rule: example

Input:

The probability of a certain medical test being positive is 90% if a patient has disease D. The 1% of the population have the disease, and the test records a false positive 5% of the time. If a random person receives a positive test, what is the probability of D for him?

Bayes' rule: example

Input:

The probability of a certain medical test being positive is 90% if a patient has disease D. The 1% of the population have the disease, and the test records a false positive 5% of the time. If a random person receives a positive test, what is the probability of D for him?

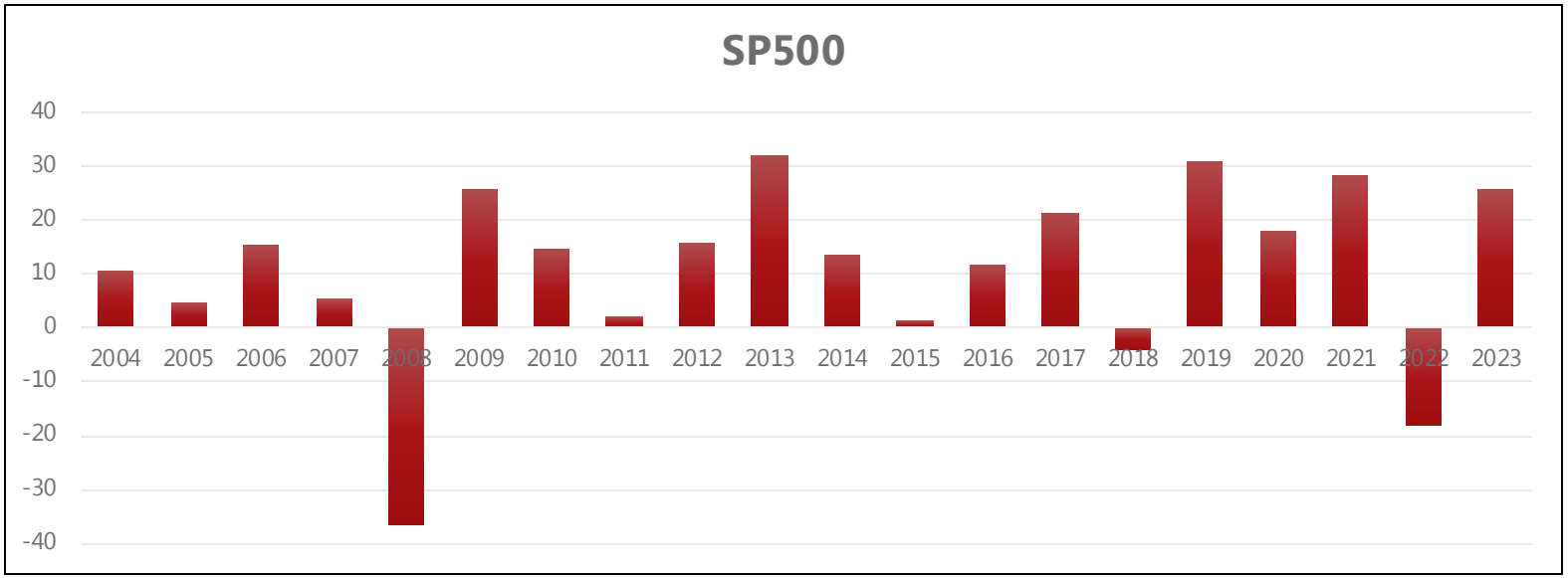
$$P(D|+) = \frac{P(+|D) \cdot P(D)}{P(+)}$$

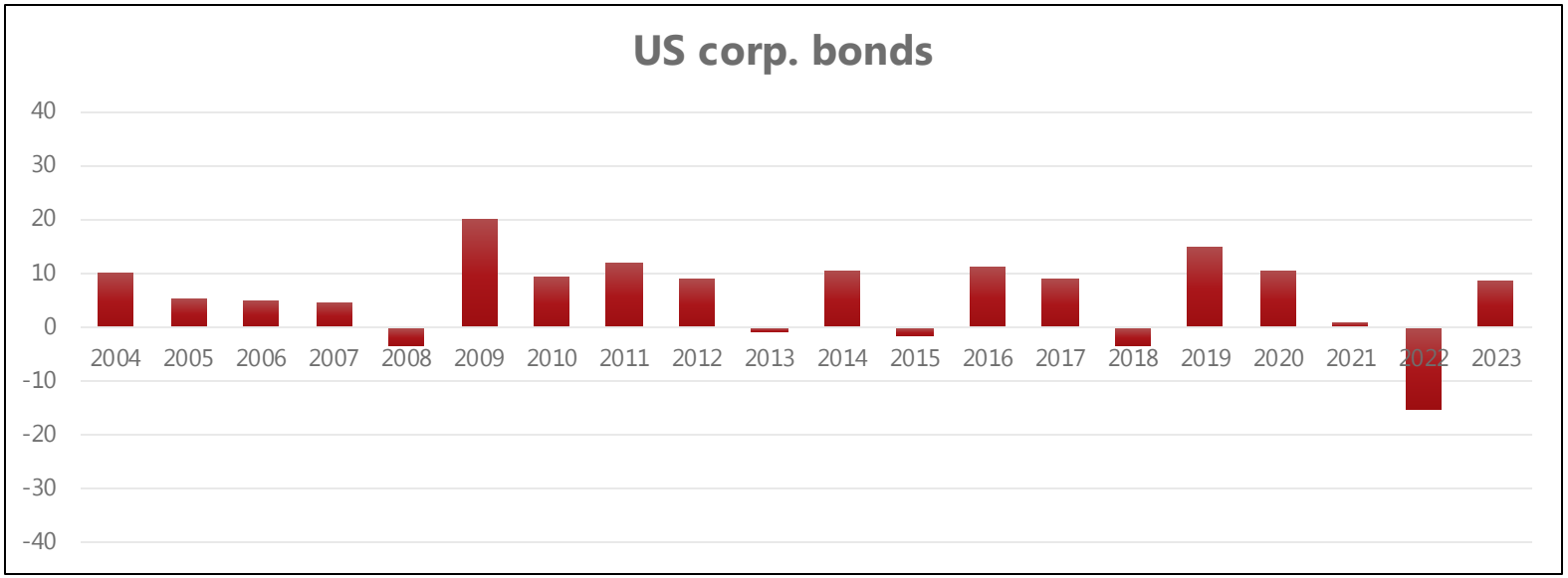
$P(D) =$ $P(+|D) =$

$P(+) = P(+|D) \cdot P(D) + P(+|no\ D) \cdot P(no\ D) =$

$P(D|+) = \frac{P(+|D) \cdot P(D)}{P(+)} =$

Statistical measures





Mean

Yearly return SP500 = [10.7, 4.8, 15.6, 5.5, -36.6, ⋯, 18, 28.5, -18, 26.1]

Yearly return bonds = [10.4, 5.3, 5.2, 4.8, -3.5, ⋯, 10.6, 0.9, -15.1, 8.7]

Mean value for SP500:

$$\mathbf{E}_{P(x)}\{X\} = \sum_x xP(x) =$$

$$10.7\frac{1}{20} + 4.8\frac{1}{20} + 15.6\frac{1}{20} + 5.5\frac{1}{20} + (-36.6)\frac{1}{20} + \cdots + (-18)\frac{1}{20} + 26.1\frac{1}{20} = \mathbf{11.04}$$

Mean value for bonds:

$$\mathbf{E}_{P(y)}\{Y\} = \sum_y yP(y) = 10.4\frac{1}{20} + 5.3\frac{1}{20} + \cdots + 8.7\frac{1}{20} = \mathbf{5.98}$$

Variance and standard deviation

Variance:

$$\text{var}\{X\} = \mathbf{E}_{P(x)}\left\{\left(X - \mathbf{E}_{P(x)}(X)\right)^2\right\} = \sum_x \left(x - \mathbf{E}_{P(x)}(X)\right)^2 P(x)$$

Standard deviation:

$$\sigma_X = \sqrt{\text{var}\{X\}}$$

Standard deviation for SP500:

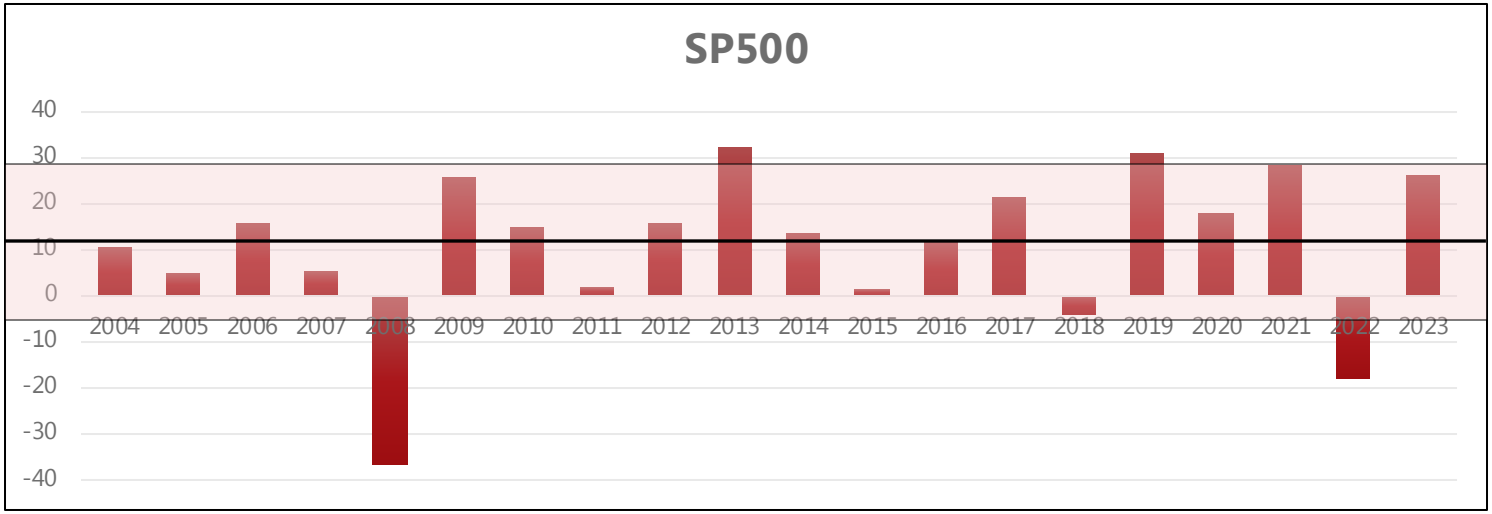
$$\sigma_X = \sqrt{\sum_y \left(x - \mathbf{E}_{P(x)}(X)\right)^2 P(x)} =$$

$$\sqrt{\frac{1}{20}(10.7 - 11.04)^2 + \cdots + \frac{1}{20}(26.1 - 11.04)^2} = \mathbf{16.46}$$

Mean and standard deviation

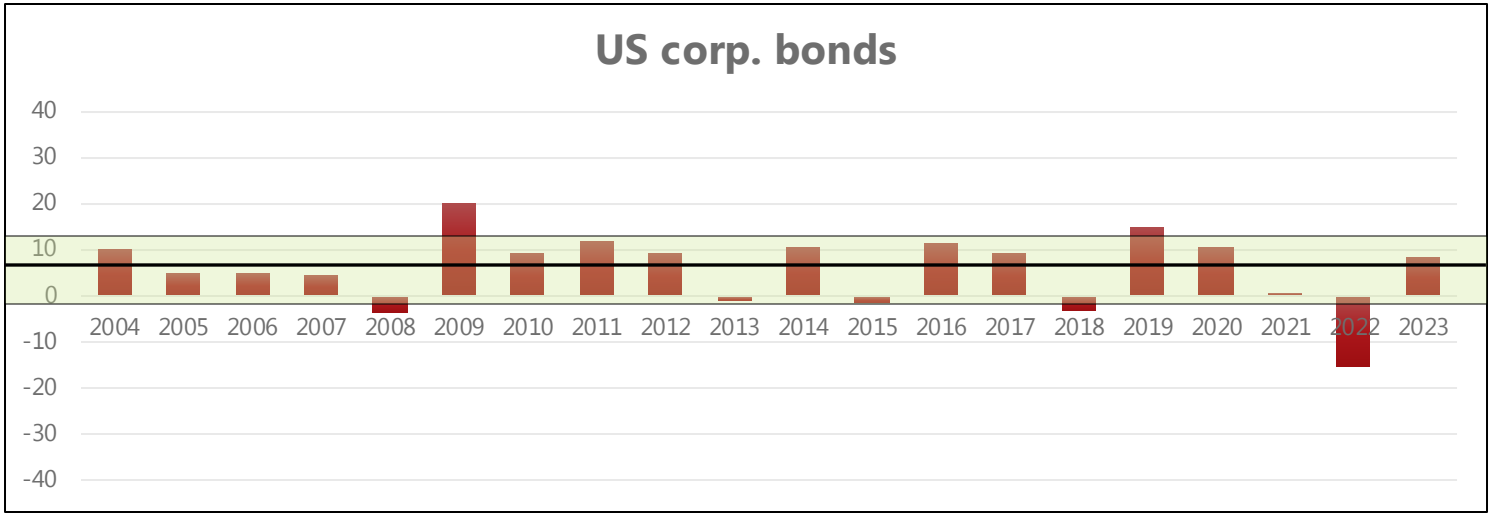
$\mathbf{E}_{SP500} = 11.04$

$\sigma_{SP500} = 16.46$

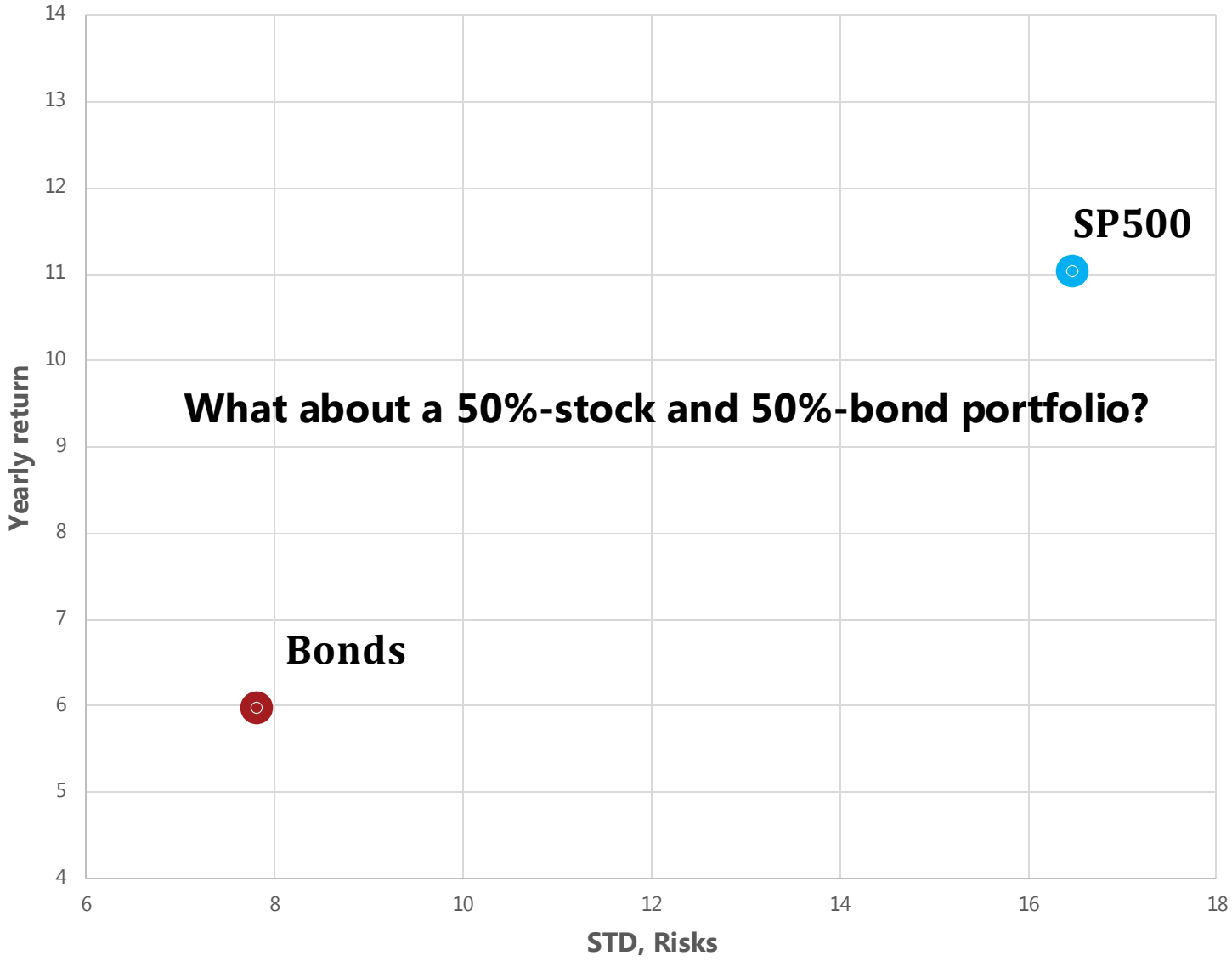


$\mathbf{E}_{bonds} = 5.98$

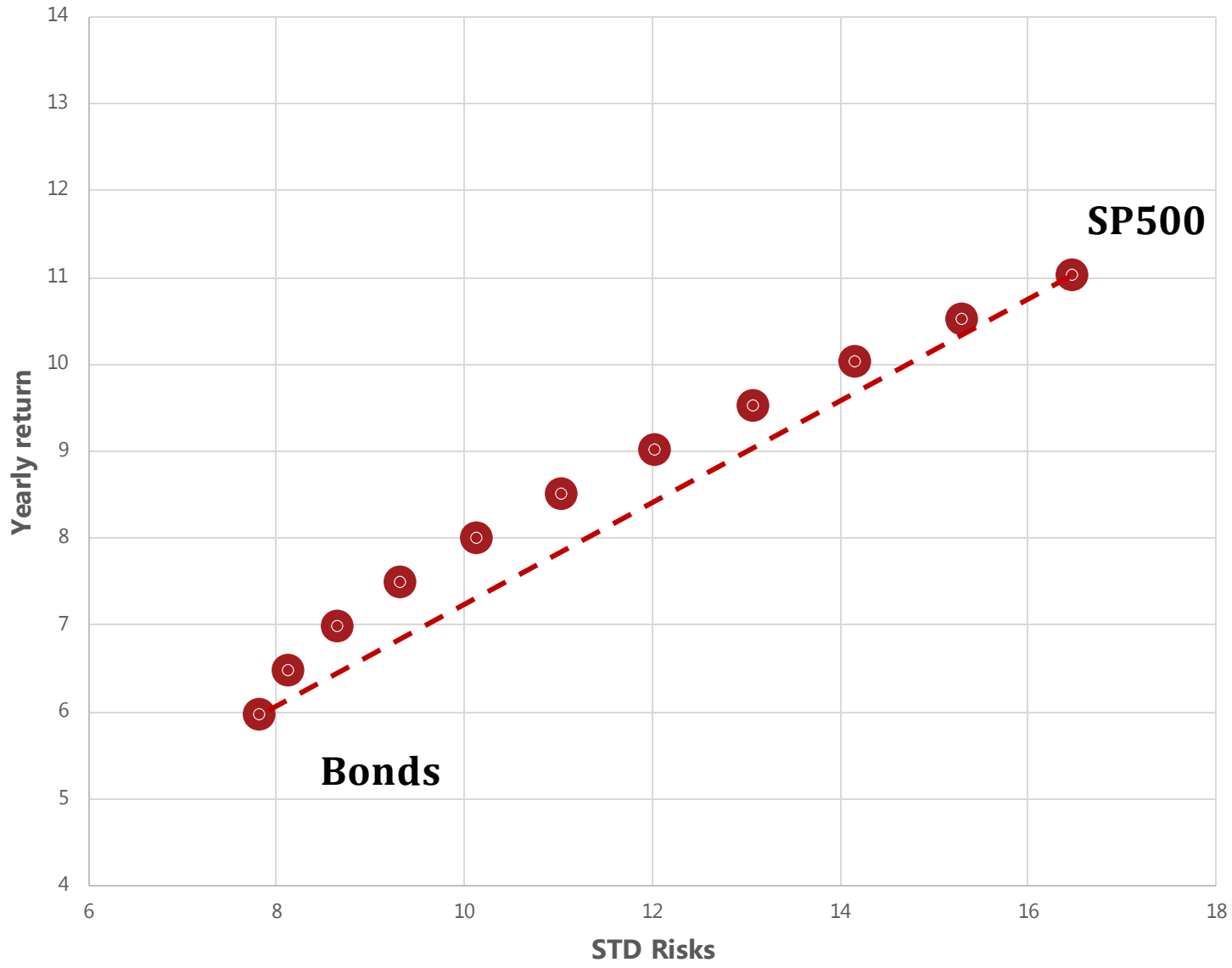
$\sigma_{bonds} = 7.80$



Combining mean and standard deviation



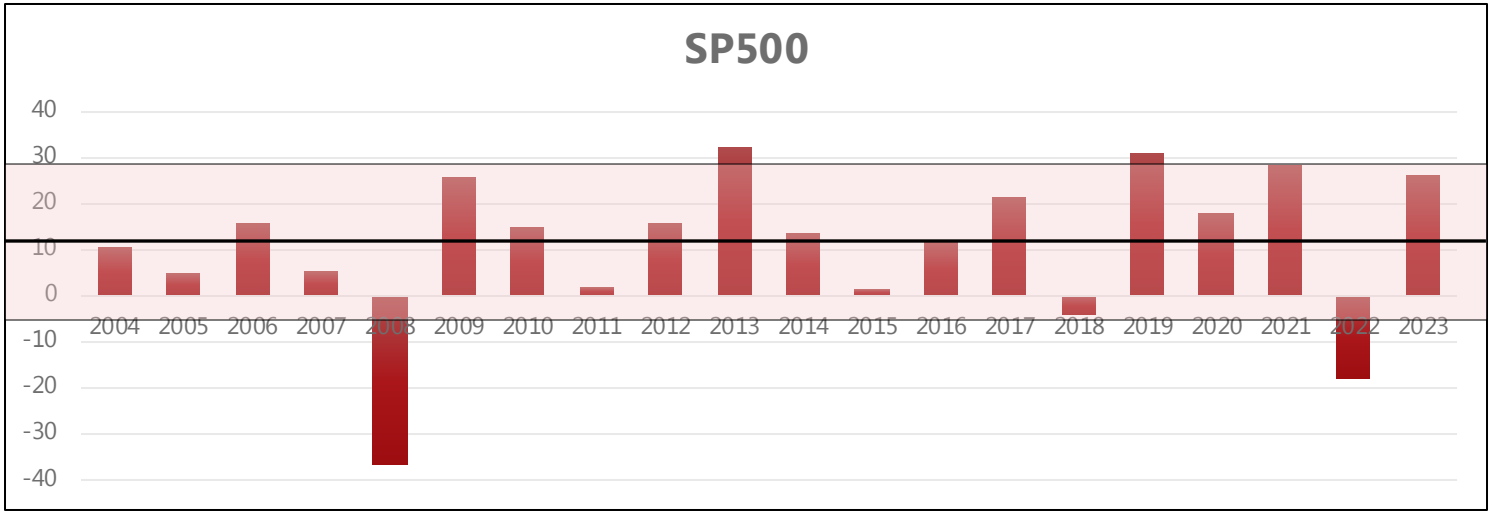
Combining mean and standard deviation



Covariance

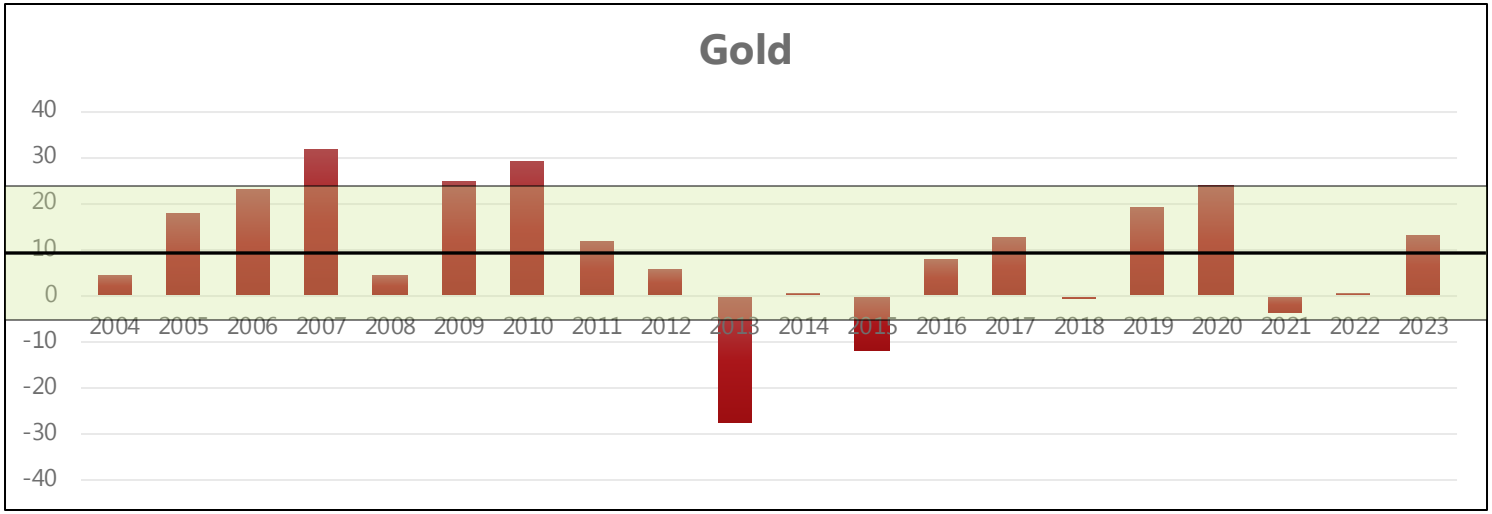
$\mathbf{E}_{SP500} = 11.04$

$\sigma_{SP500} = 16.46$

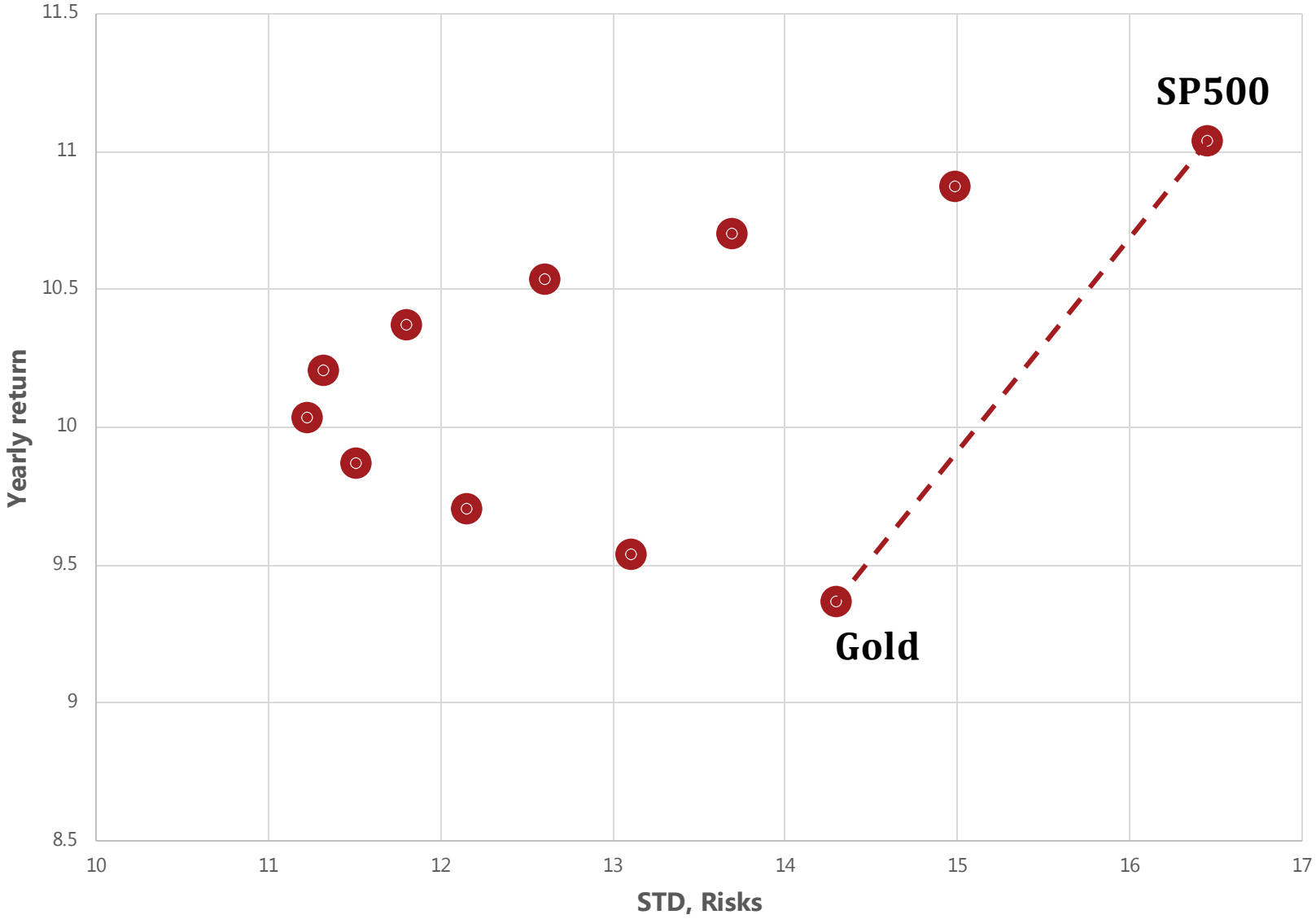


$\mathbf{E}_{gold} = 9.37$

$\sigma_{gold} = 14.30$



Covariance


11.5
11
10.5
10
9.5
9
8.5
Yearly return
10
11
12
13
14
15
16
17
STD, Risks
SP500
Gold

Covariance

$$\mathrm{cov}\{X, Y\} = \frac{\sum_{i=1}^{N}\big(x_i - \mathbf{E}(X)\big)\big(y_i - \mathbf{E}(Y)\big)}{N}$$

Mean returns =
[11.04, 7.80, 9.37]

	SP500	Bonds	Gold
2004	10.7	10.4	4.7
2005	4.8	5.3	17.8
2006	15.6	5.2	23.2
2007	5.5	4.8	31.9
2008	-36.6	-3.5	4.3
	.....		
2022	-18.0	-15.1	0.6
2023	26.1	8.7	13.3

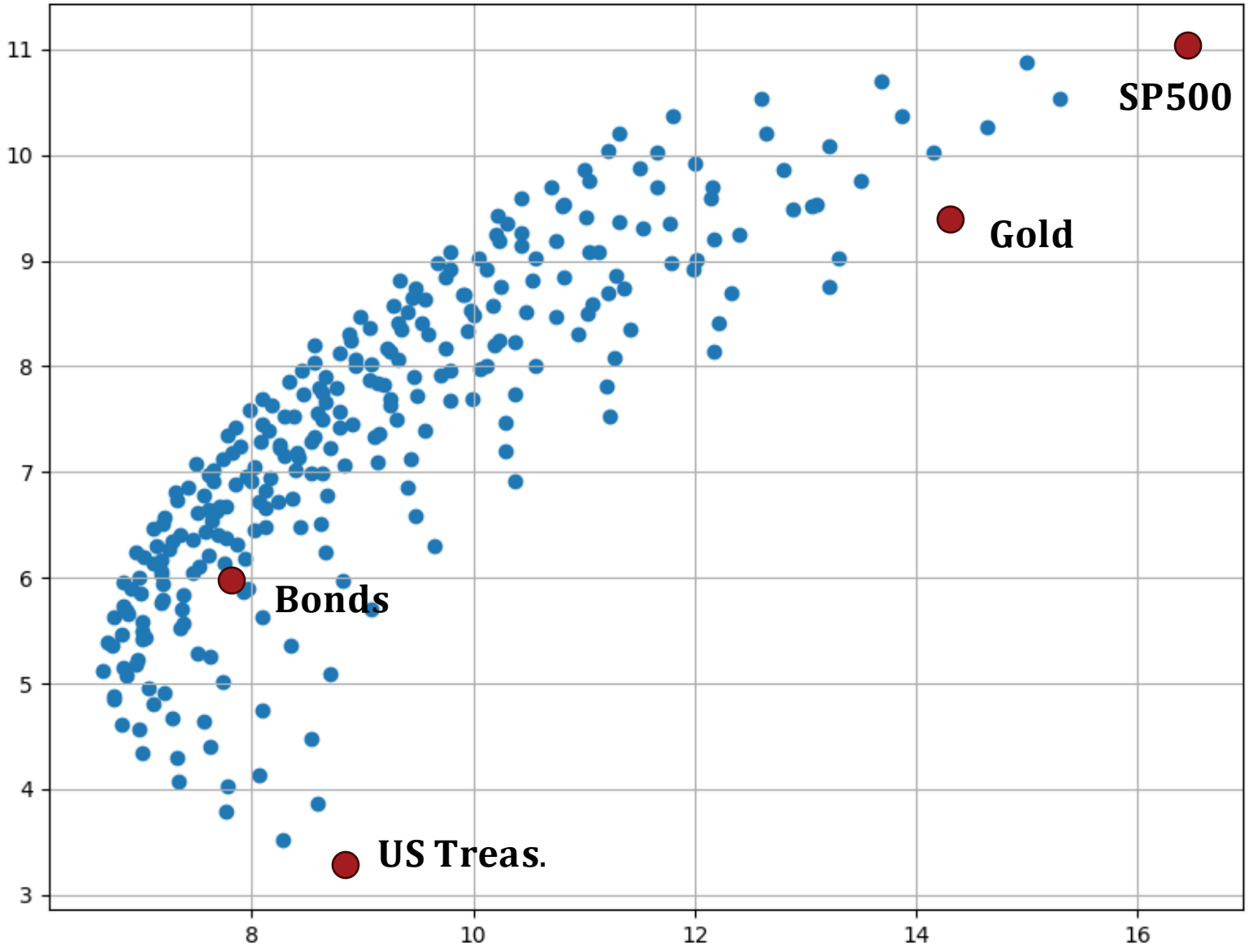
Matrix of correlations

Correlation $p\{X,Y\}$ between vectors X and Y is covariance $\mathrm{cov}\{X,Y\}$ normalized to standard deviations, σ_X and σ_Y :

$$p\{X,Y\} = \frac{\mathrm{cov}\{X,Y\}}{\sigma_X \sigma_Y}$$

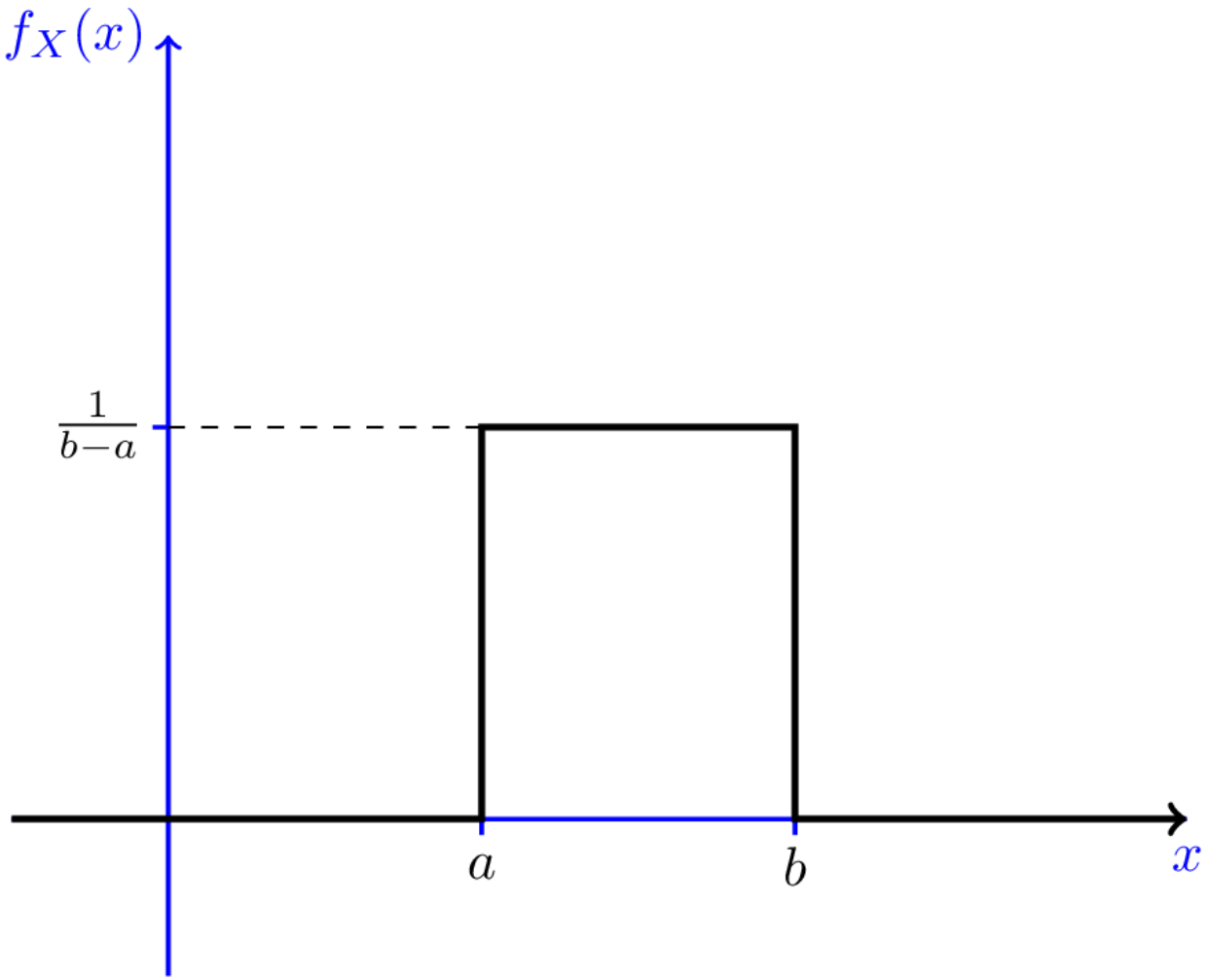
	SP500	Bonds	Golds	US Treas.
SP500	1	0.60	0.08	-0.26
Bonds	0.60	1	0.53	0.33
Gold	0.08	0.53	1	0.35
US Treas.	-0.26	0.33	0.35	1

Matrix of correlations



Simple distributions: Uniform distribution

Let's say we choose a random real number between a and b

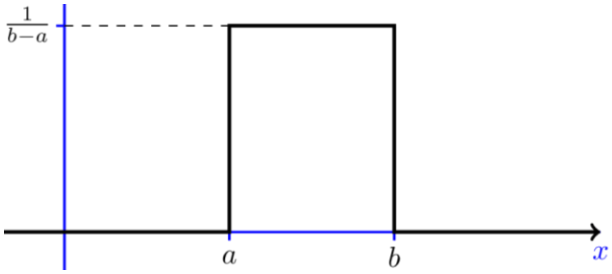


Simple distributions: Uniform distribution

Let's say we choose a random real number between 0 and 1:

- What is the probability of getting a specific number like 0.234234324324?

- The probability of getting a specific real number is zero.
- But we can compute the probability of getting a number in an interval from [0.2, 0.3]

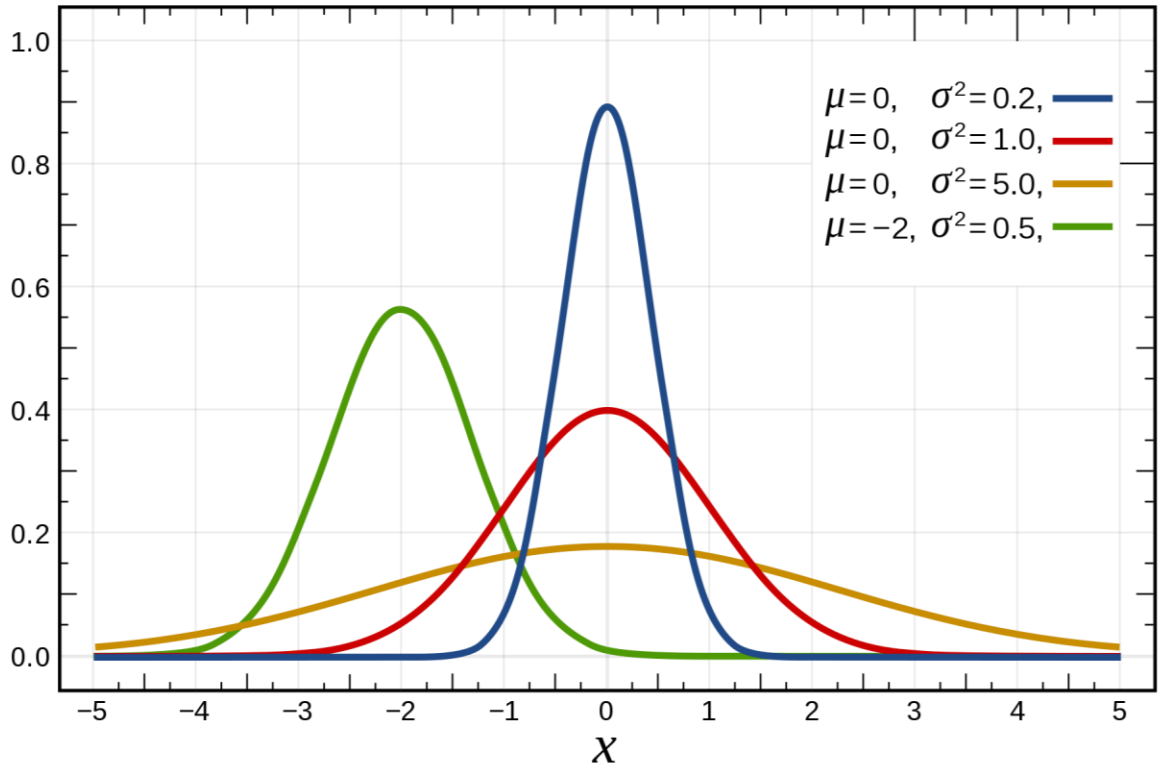


$$P(0.2 \leq x \leq 0.3) = \int_{0.2}^{0.3} \frac{1}{1-0}\, dx = \frac{1}{1}(0.3 - 0.2) = 0.1$$

Simple distributions: Normal distribution

Normal distribution:

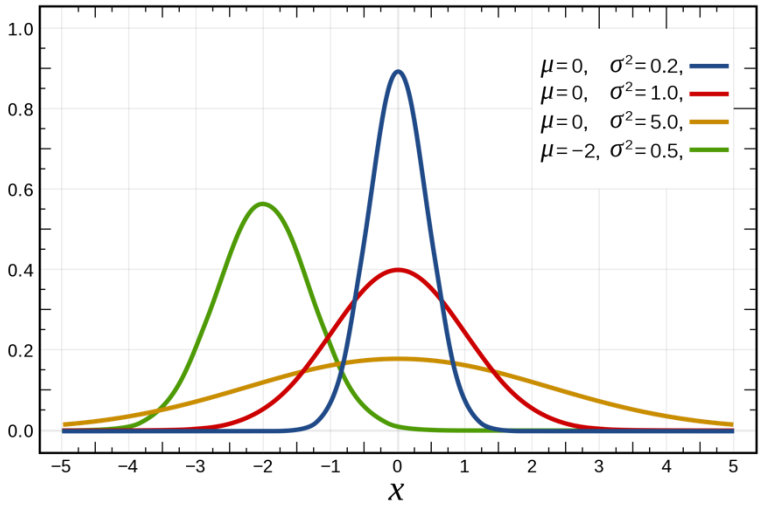
$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



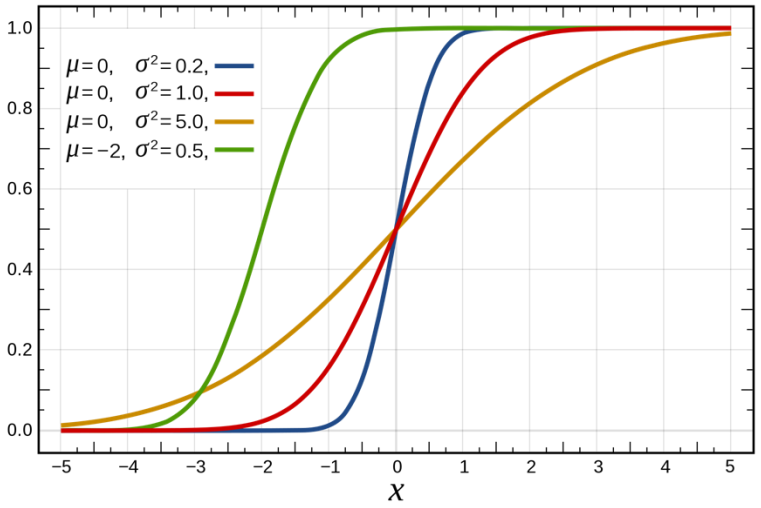
Cumulative distribution function

Plot probability of getting x ≤ t:

$$P(x \leq t) = \int_{-\infty}^{t} f(x)\,dx$$



Probability density function



Cumulative distribution function