

PROBABILITY

MSc 2023/2024

J E D S M L
O S O M A 2
C E M S C O
2 0 1 C S 2
0 3 2 1 E 3

WHY DO WE CARE ABOUT PROBABILITY?

Mental exercise:

<https://www.youtube.com/watch?v=IG4VkPoG3ko>

Imagine that you go undergo a diagnostic test with an accuracy of 90%. Your test comes back positive.

What are the chances that you really have the disease?

Keeping in mind that there is no other information other than the test result

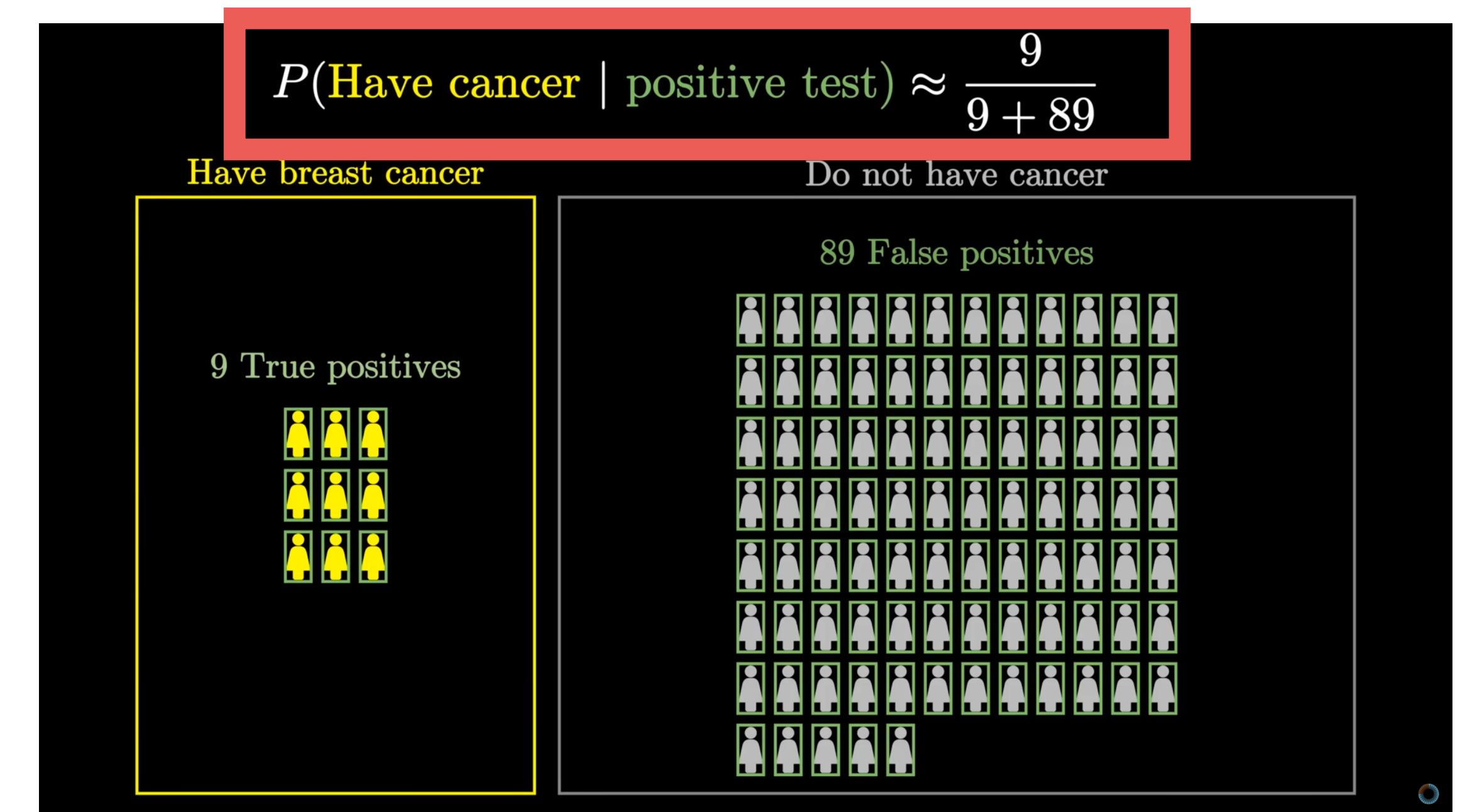
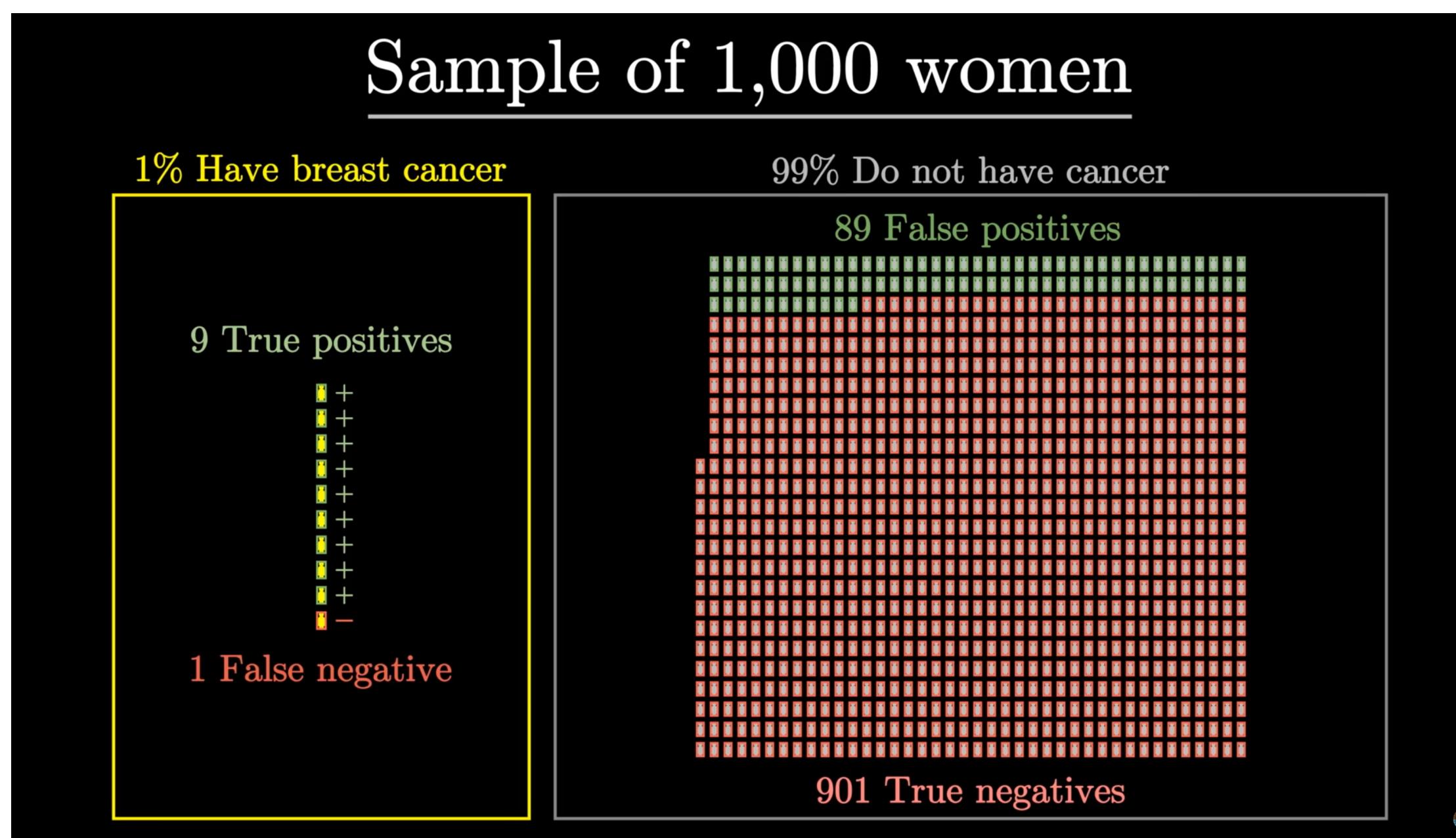
WHY DO WE CARE ABOUT PROBABILITY?

Mental exercise:

<https://www.youtube.com/watch?v=IG4VkPoG3ko>

Imagine that you go undergo a diagnostic test with an accuracy of 90%. Your test comes back positive.

What are the chances that you really have the disease?



Objectives of the day:

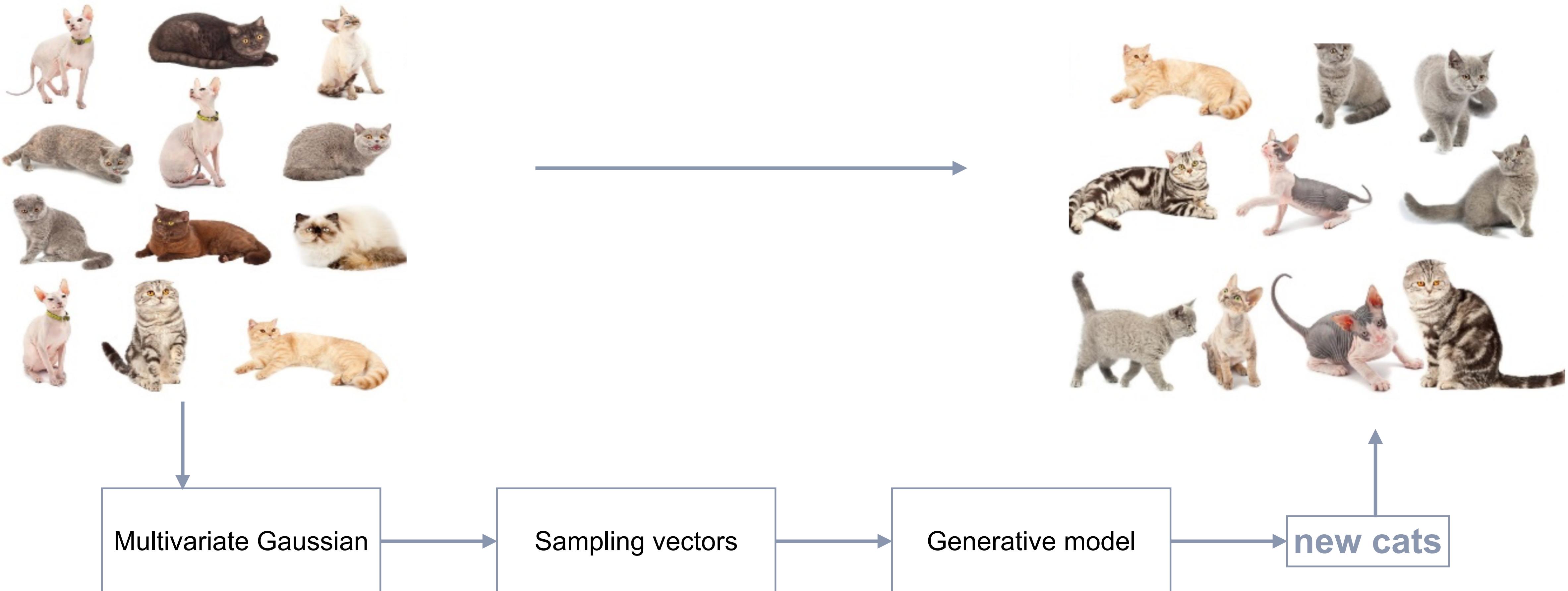
- Introduce uniform, Gaussian, and Bernoulli distributions
- Maximum likelihood estimation in the context of Machine Learning
- Learn how to compare probability density functions

GENERATIVE MODELS

Context:

Given a training set with lots of cats,

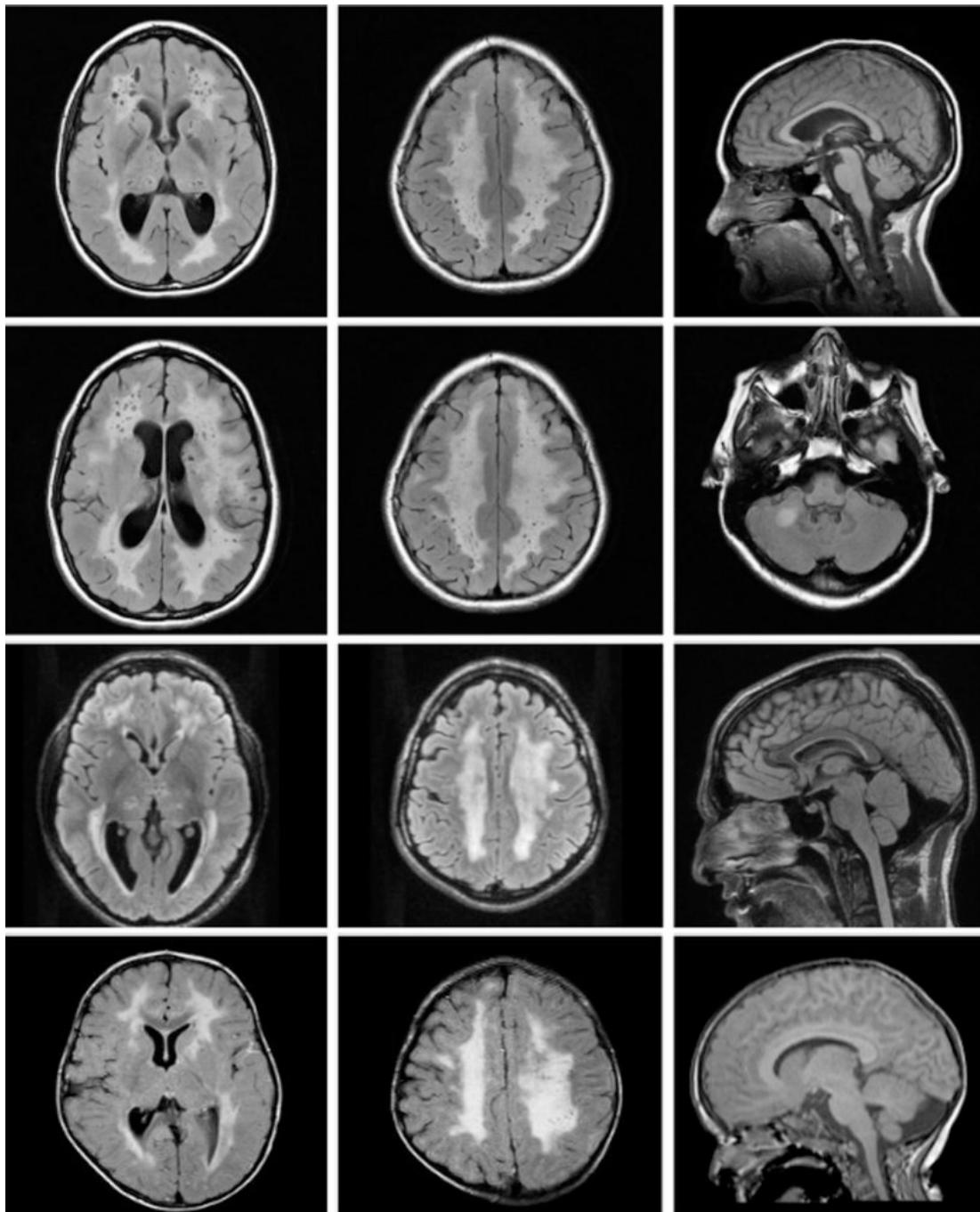
how can I generate new cats?



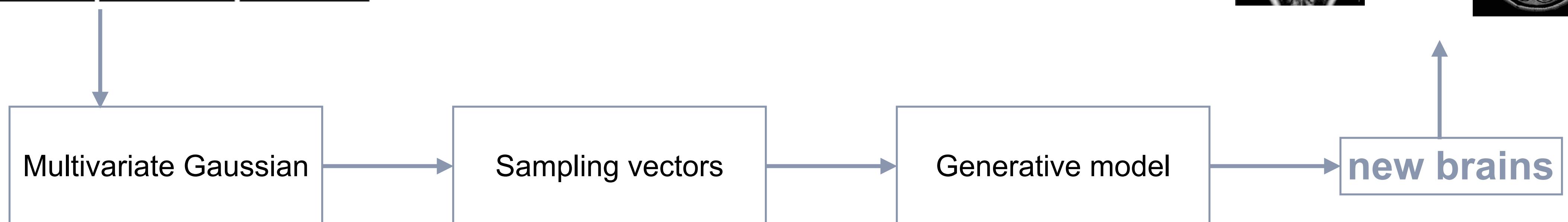
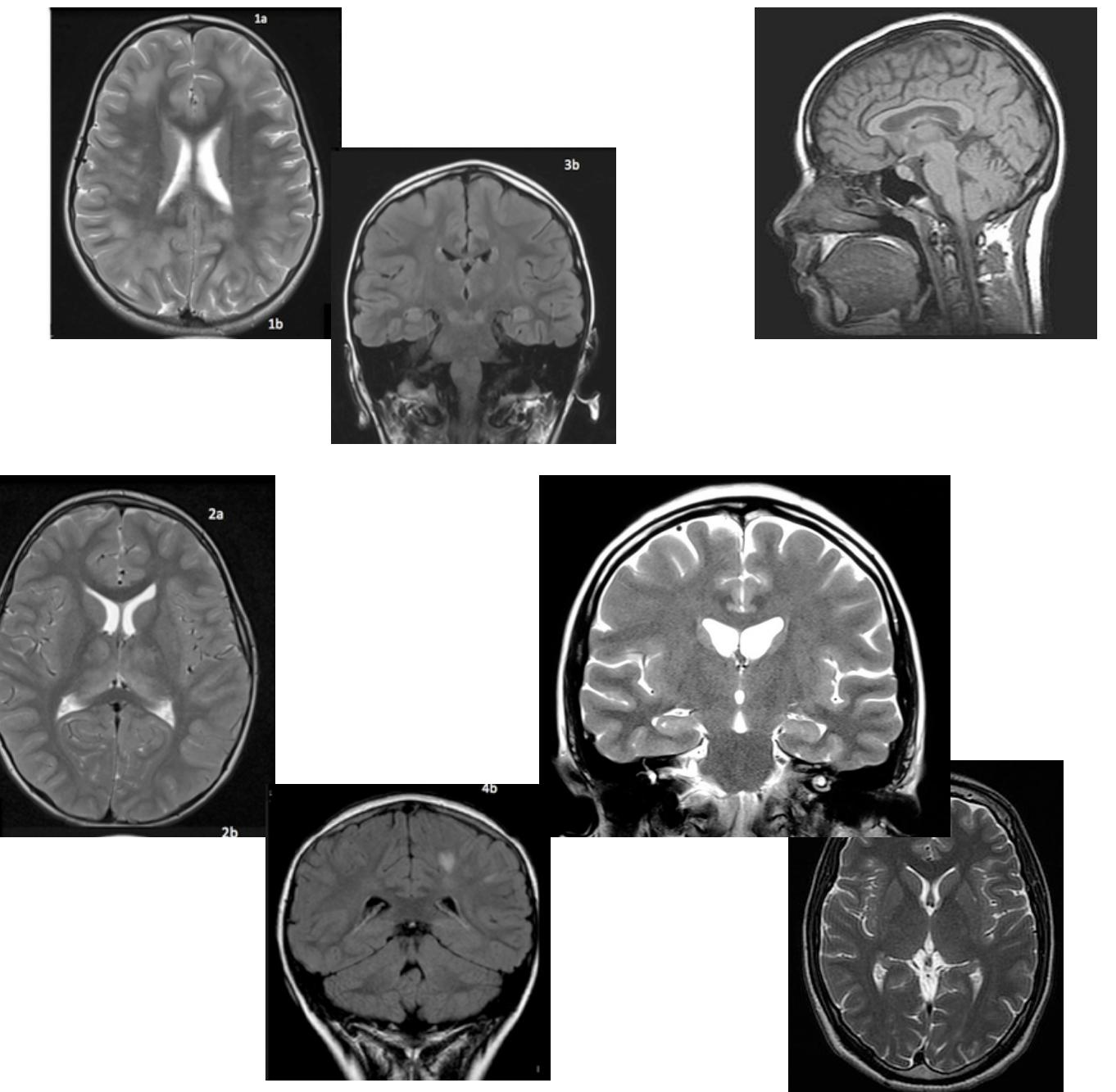
GENERATIVE MODELS

Context:

Given a training set with lots of brains,



how can I **generate new brains?**



GENERATIVE MODELS

In coming sessions we will introduce a few generative model architectures:

- ▶ Variational AutoEncoders (**VAEs**)
- ▶ Generative Adversarial Networks (**GANs**)
- ▶ **Diffusion** models
- ▶ **Transformers** (can be used as generative models)

DEEP GENERATIVE MODELS

Why do we need probabilities?

- Deep Generative Models (DGM) are generative models that use deep neural networks.
- All of these models have something in common: they use **random sampling of probability distributions** that are **representative** of the data they want to generate (audio, images, text, time series, etc).
- **Training** these models allows us to **capture these distributions** and use them to generate new original data samples.

1. Probability distributions
2. Maximum likelihood
3. Comparison of probability density functions

PROBABILITY DISTRIBUTIONS

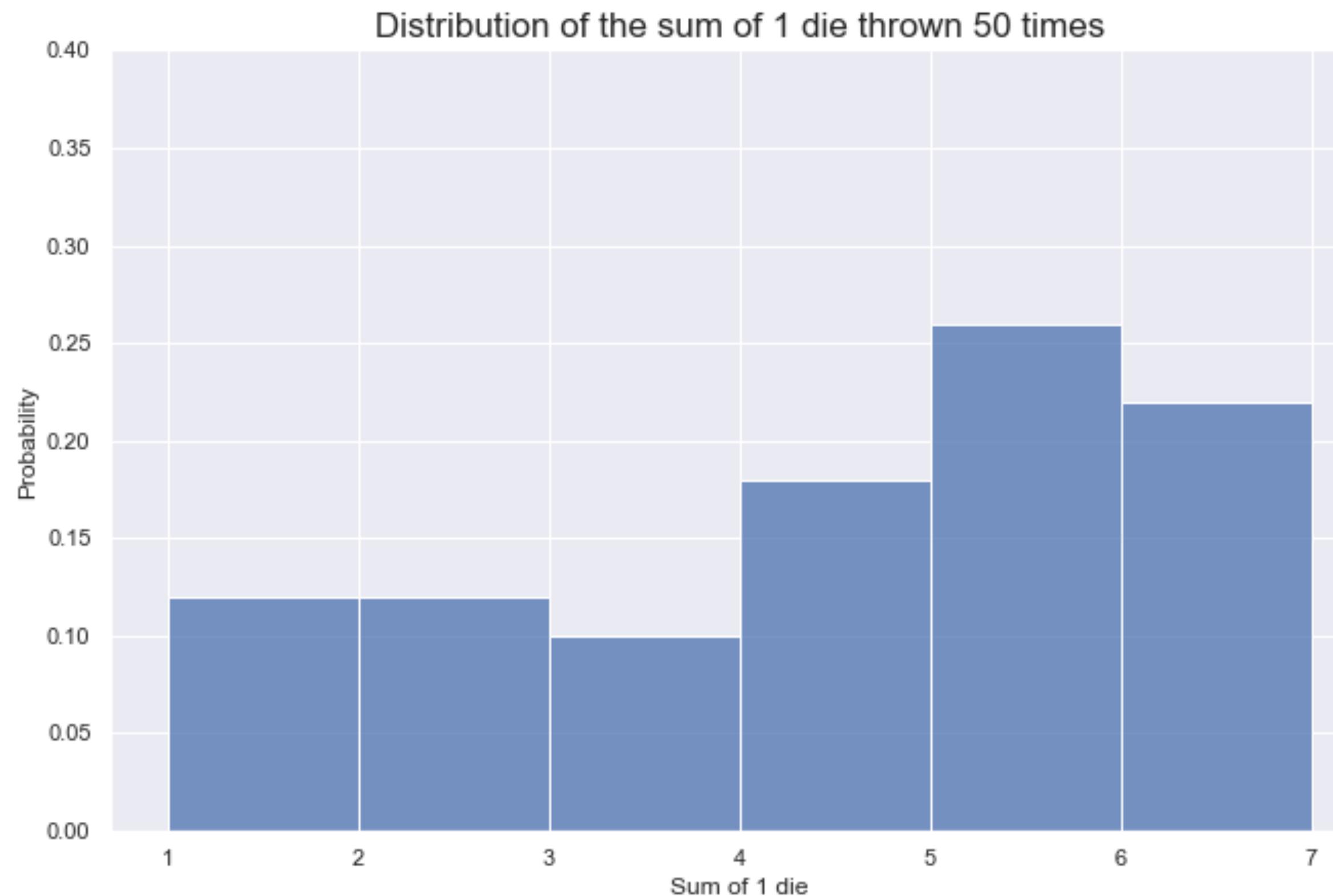
What is a probability distribution?

Definition:

A probability distribution is a mathematical function that describes all the possible outcomes of an experiment as well as the likelihoods of each possible outcome.

PROBABILITY DISTRIBUTIONS

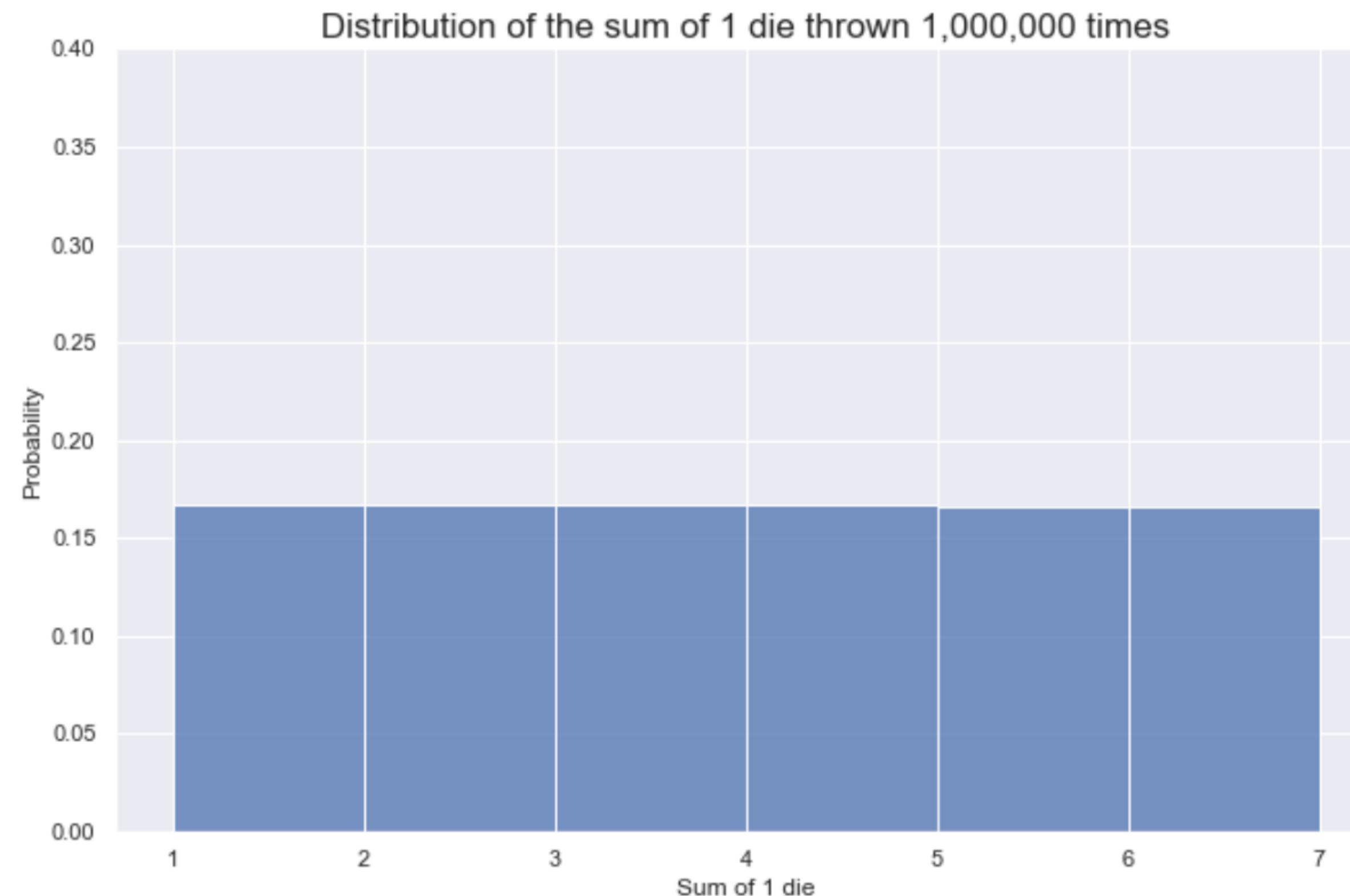
Example: what is the probability distribution of the possible outcomes when throwing 1 die:



<https://www.cantorsparadise.com/what-to-expect-when-throwing-dice-and-adding-them-up-5231f3831d7>

PROBABILITY DISTRIBUTIONS

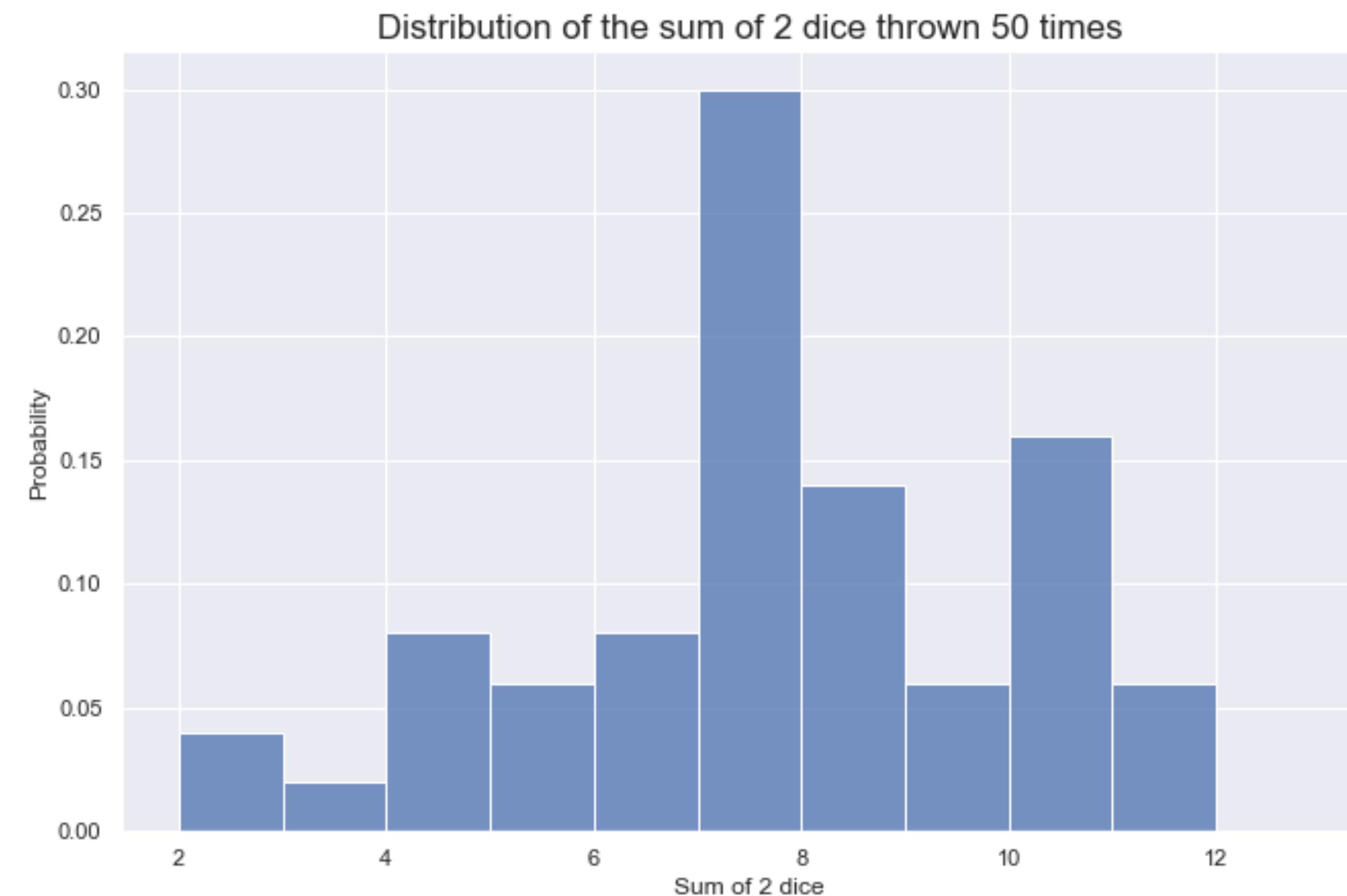
Example: what is the probability distribution of the possible outcomes when throwing 1 die:



<https://www.cantorsparadise.com/what-to-expect-when-throwing-dice-and-adding-them-up-5231f3831d7>

PROBABILITY DISTRIBUTIONS

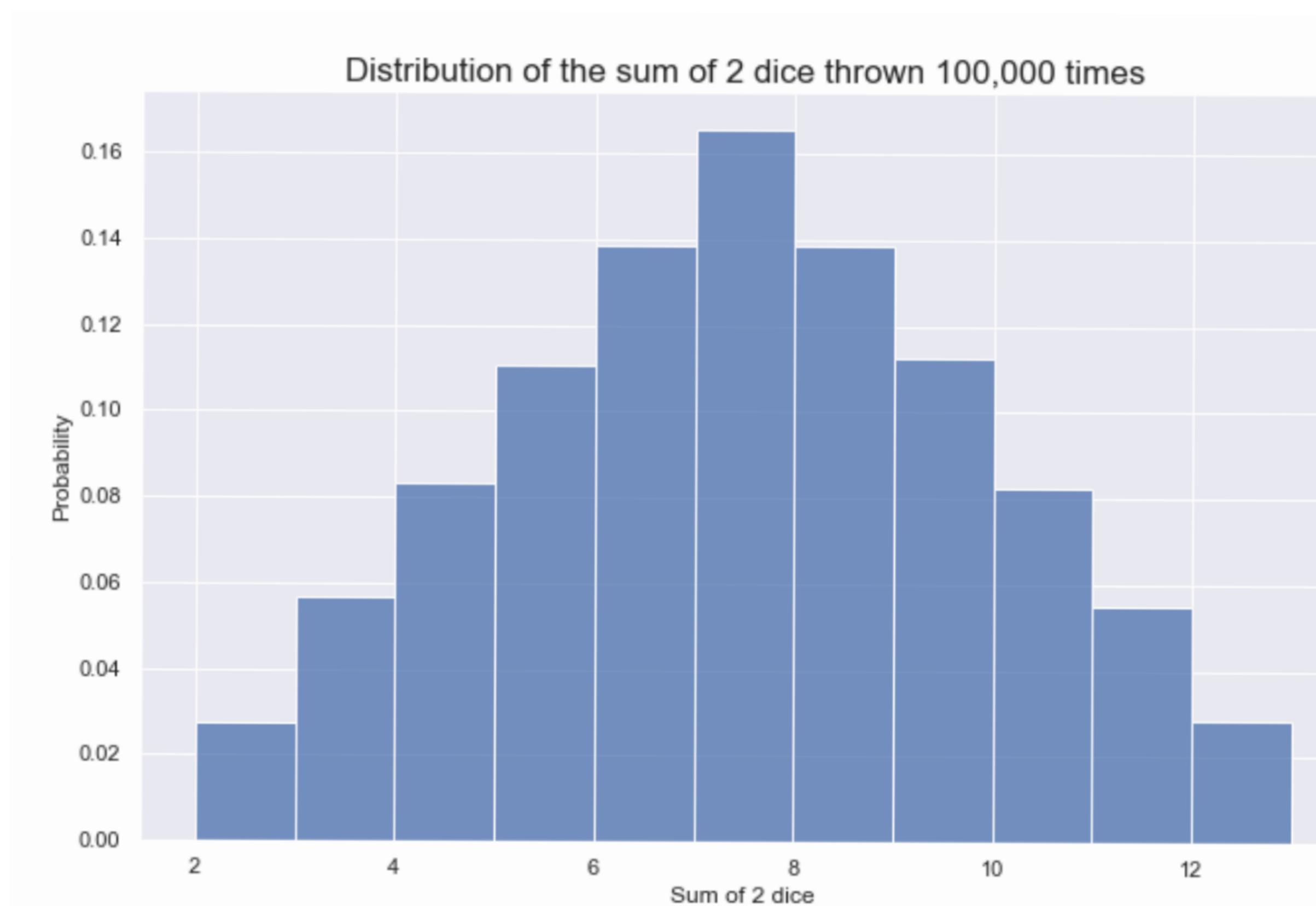
Example: what is the probability distribution of the possible outcomes when throwing 2 dice:



<https://www.cantorsparadise.com/what-to-expect-when-throwing-dice-and-adding-them-up-5231f3831d7>

PROBABILITY DISTRIBUTIONS

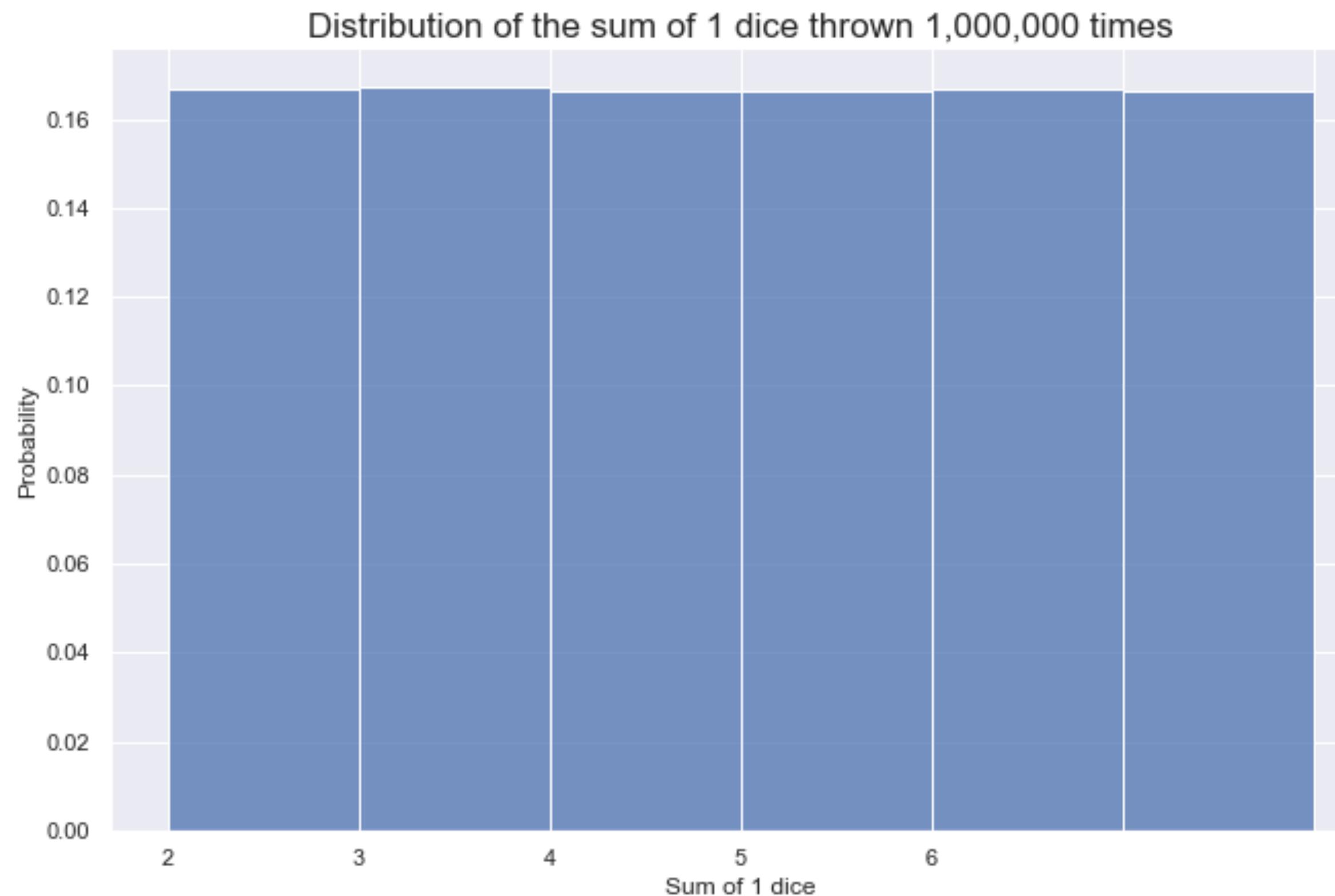
Example: what is the probability distribution of the possible outcomes when throwing 2 dice:



<https://www.cantorsparadise.com/what-to-expect-when-throwing-dice-and-adding-them-up-5231f3831d7>

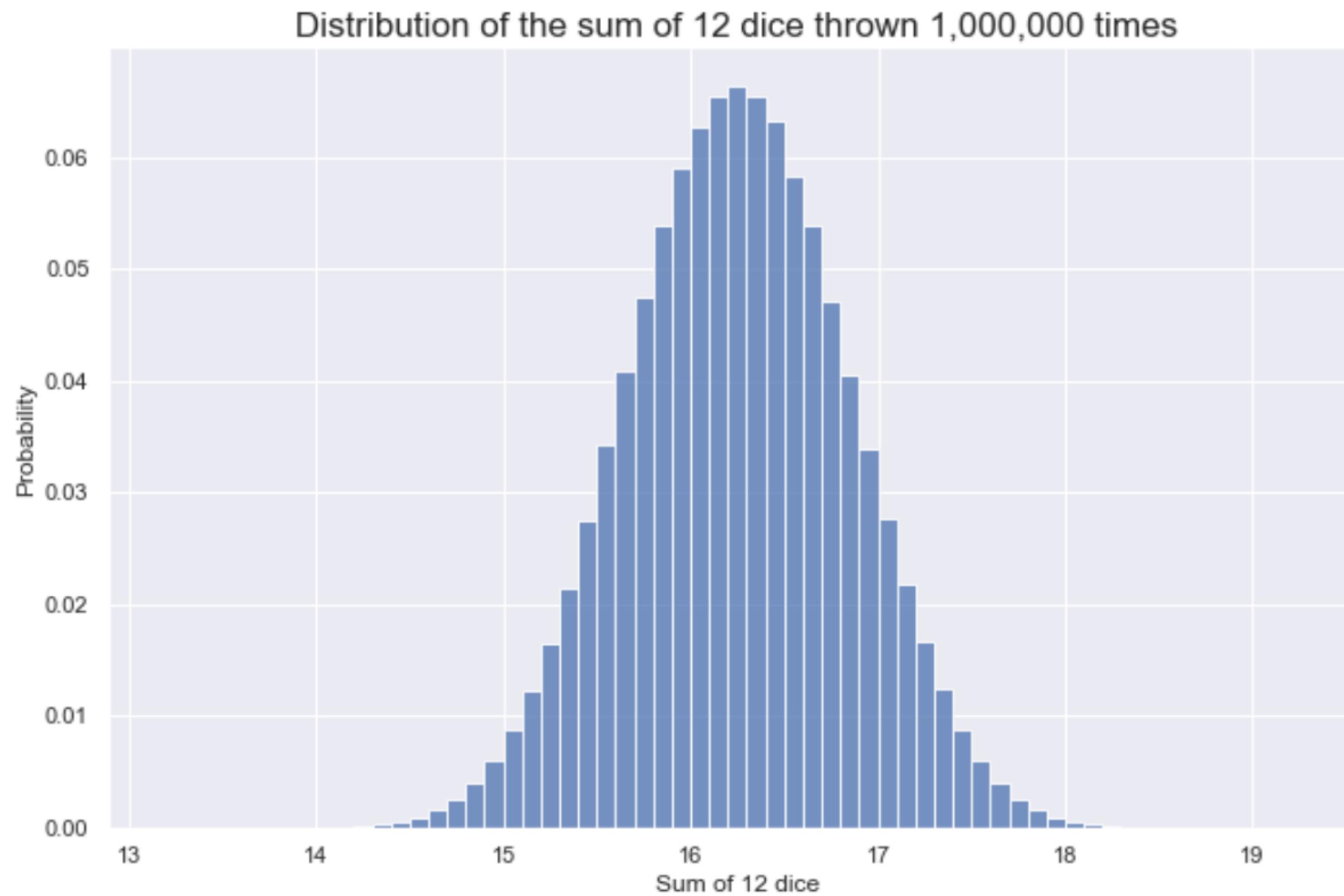
PROBABILITY DISTRIBUTIONS

Example: what is the probability distribution of the possible outcomes when throwing an increasing number of dice:



<https://www.cantorsparadise.com/what-to-expect-when-throwing-dice-and-adding-them-up-5231f3831d7>

Example: what is the probability distribution of the possible outcomes when throwing an increasing number of dice:



<https://www.cantorsparadise.com/what-to-expect-when-throwing-dice-and-adding-them-up-5231f3831d7>

DEFINITIONS

Probability

$$P(a < X < b) = \int_a^b f(x) dx \quad \text{and} \quad \int_{-\infty}^{\infty} f(x) dx = 1$$

Expectation (or mean)

$$E(X) = \mu = \int_{-\infty}^{\infty} x f(x) dx$$

Variance

$$Var(X) = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = E[(X - \mu)^2] = E(X^2) - \mu^2$$

DEFINITIONS

Probability

$$P(a < X < b) = \int_a^b f(x) dx \quad \text{and} \quad \int_{-\infty}^{\infty} f(x) dx = 1$$

Expectation (or mean)

$$E(X) = \mu = \int_{-\infty}^{\infty} x f(x) dx$$

X : random variable

x : regular or constant variable

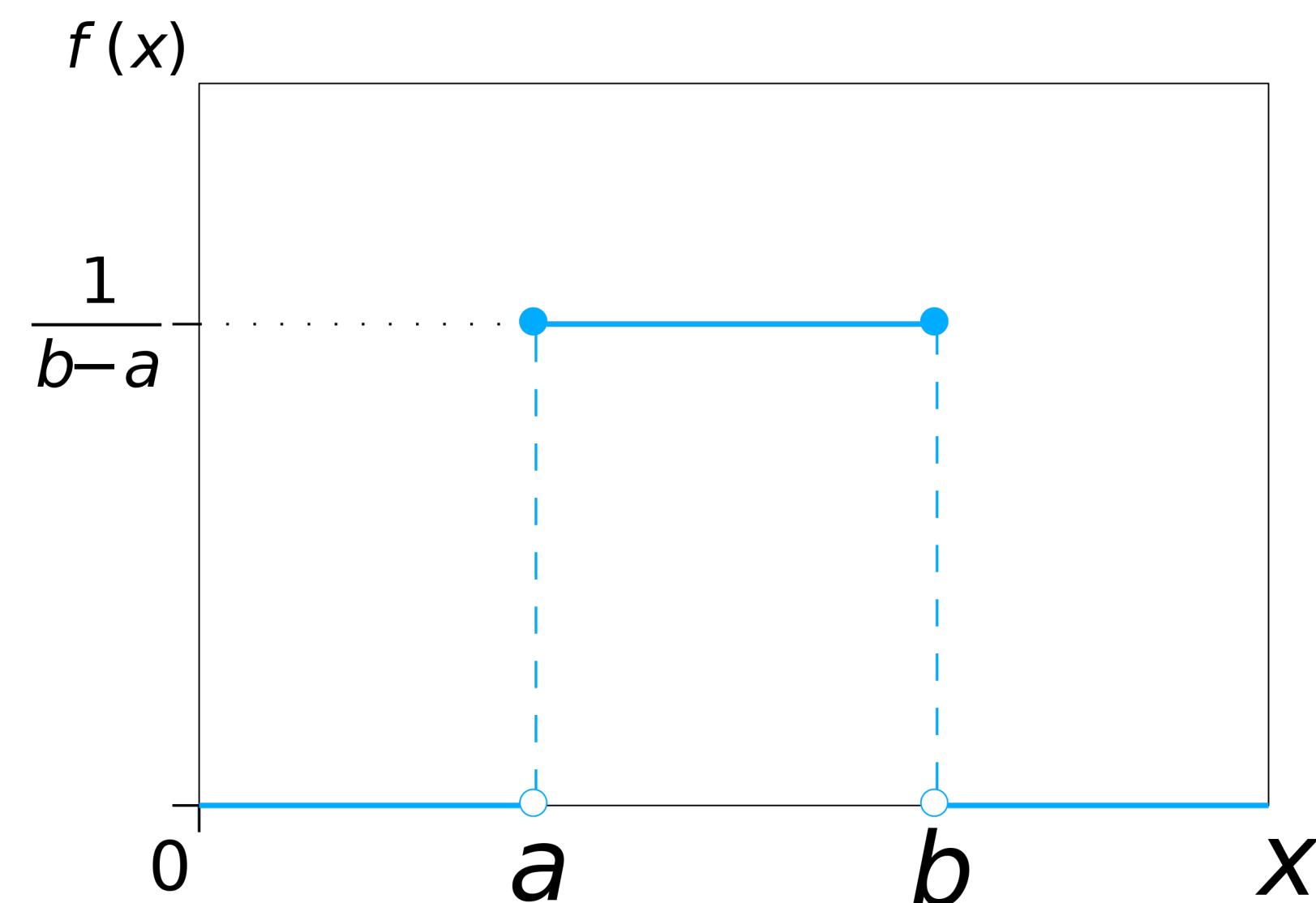
f : probability density function

Variance

$$Var(X) = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = E[(X - \mu)^2] = E(X^2) - \mu^2$$

UNIFORM DISTRIBUTION

The **uniform distribution** assigns the same probability to all the possible outcomes:



so that the area under the curve is 1

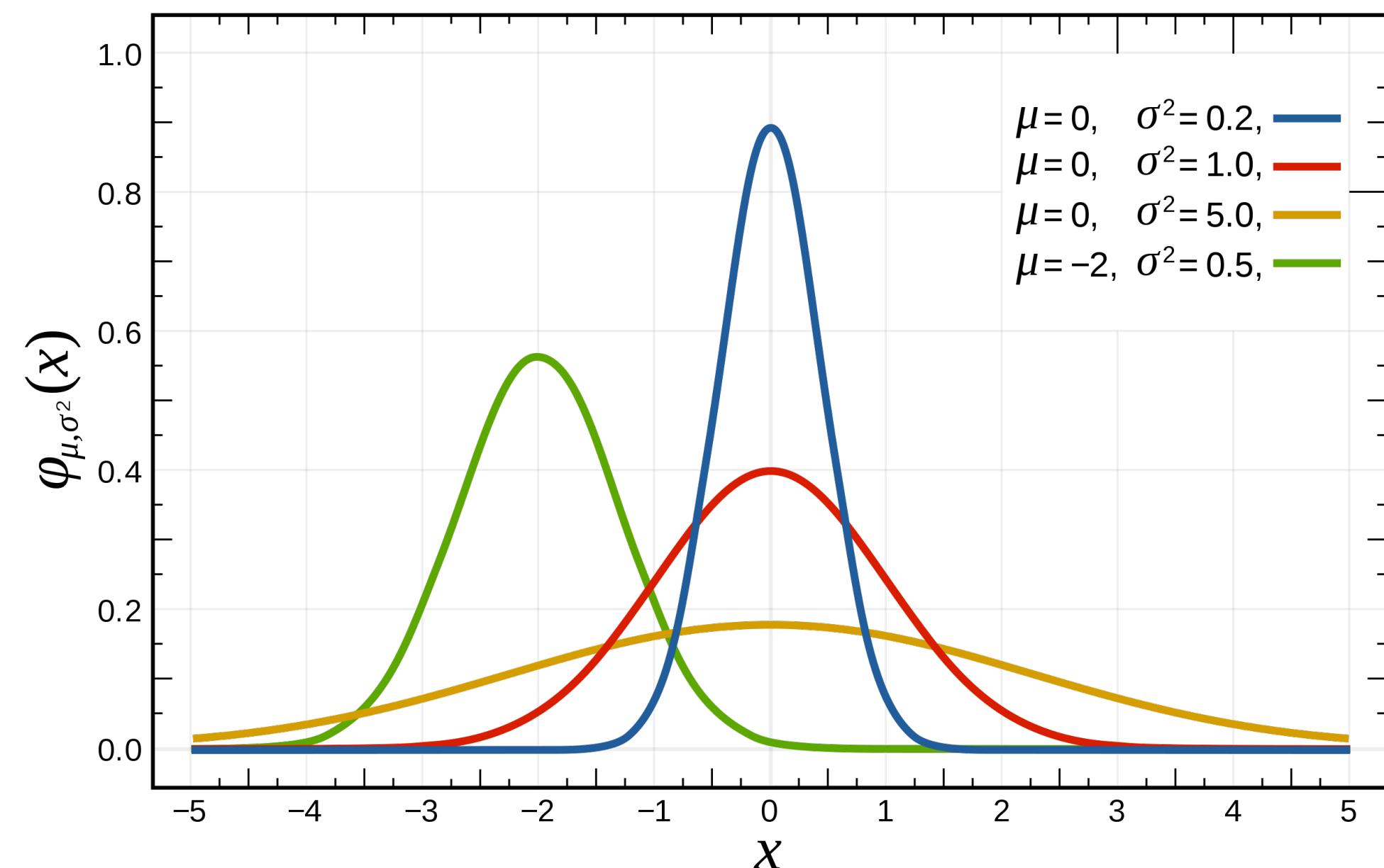
$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{if } x \notin [a, b] \end{cases}$$

$f(x)$ is also written $U(x; a, b)$

one die

GAUSSIAN DISTRIBUTION

The **normal (or Gaussian) distribution** describes a symmetric probability distribution uniquely defined by its **mean μ** and its **standard deviation σ** :



‘infinite’ dice

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

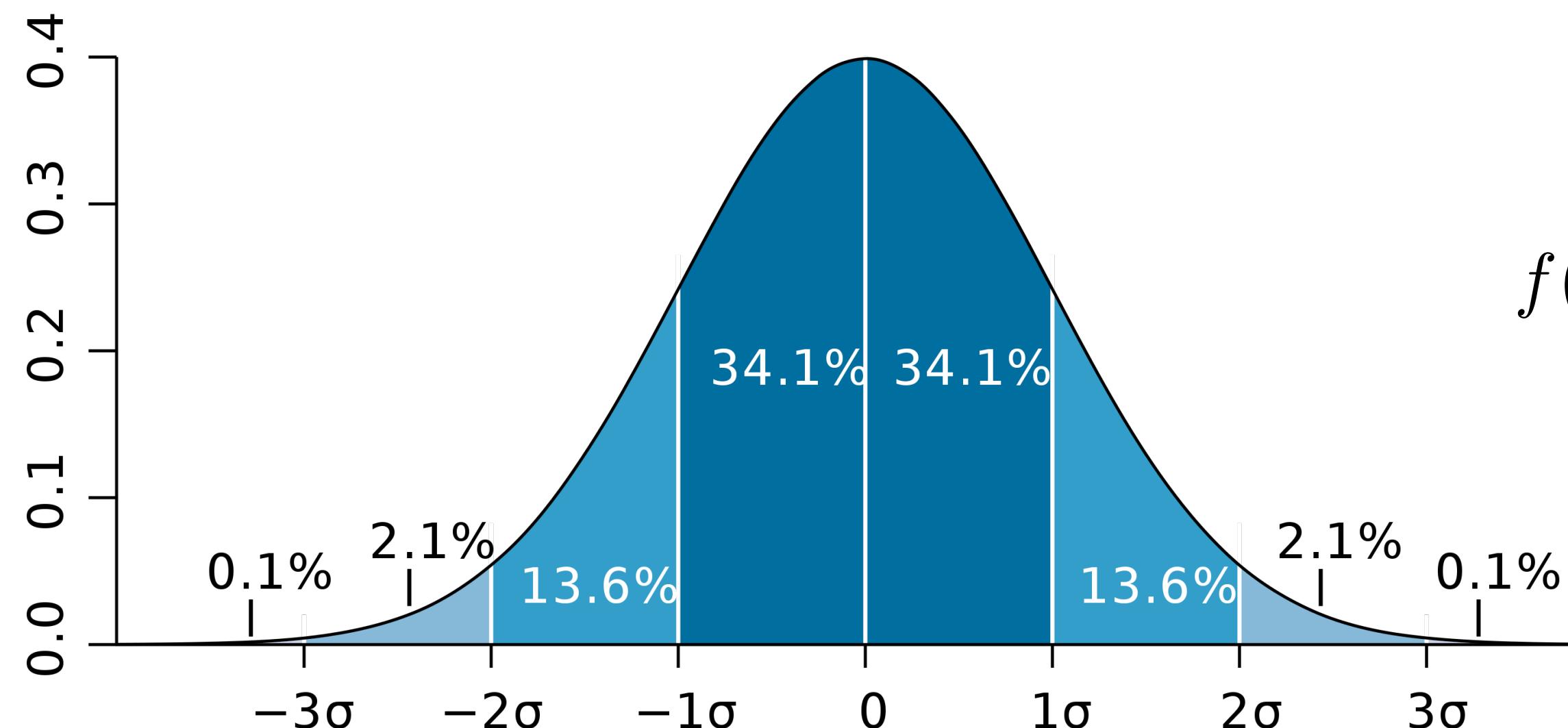
$f(x)$ is also written $N(x; \mu, \sigma^2)$

https://en.wikipedia.org/wiki/Normal_distribution

https://en.wikipedia.org/wiki/Central_limit_theorem

GAUSSIAN DISTRIBUTION

The **normal (or Gaussian) distribution** describes a symmetric probability distribution uniquely defined by its **mean μ** and its **standard deviation σ** :



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

$f(x)$ is also written $N(x; \mu, \sigma^2)$

‘infinite’ dice

https://en.wikipedia.org/wiki/Normal_distribution

https://en.wikipedia.org/wiki/Central_limit_theorem

MEAN AND VARIANCE IN UNIFORM AND NORMAL DISTRIBUTIONS

Uniform distribution:

$$\mu(X) = \frac{1}{2}(a + b)$$

$$Var(X) = \frac{1}{12}(b - a)^2$$

Normal distribution:

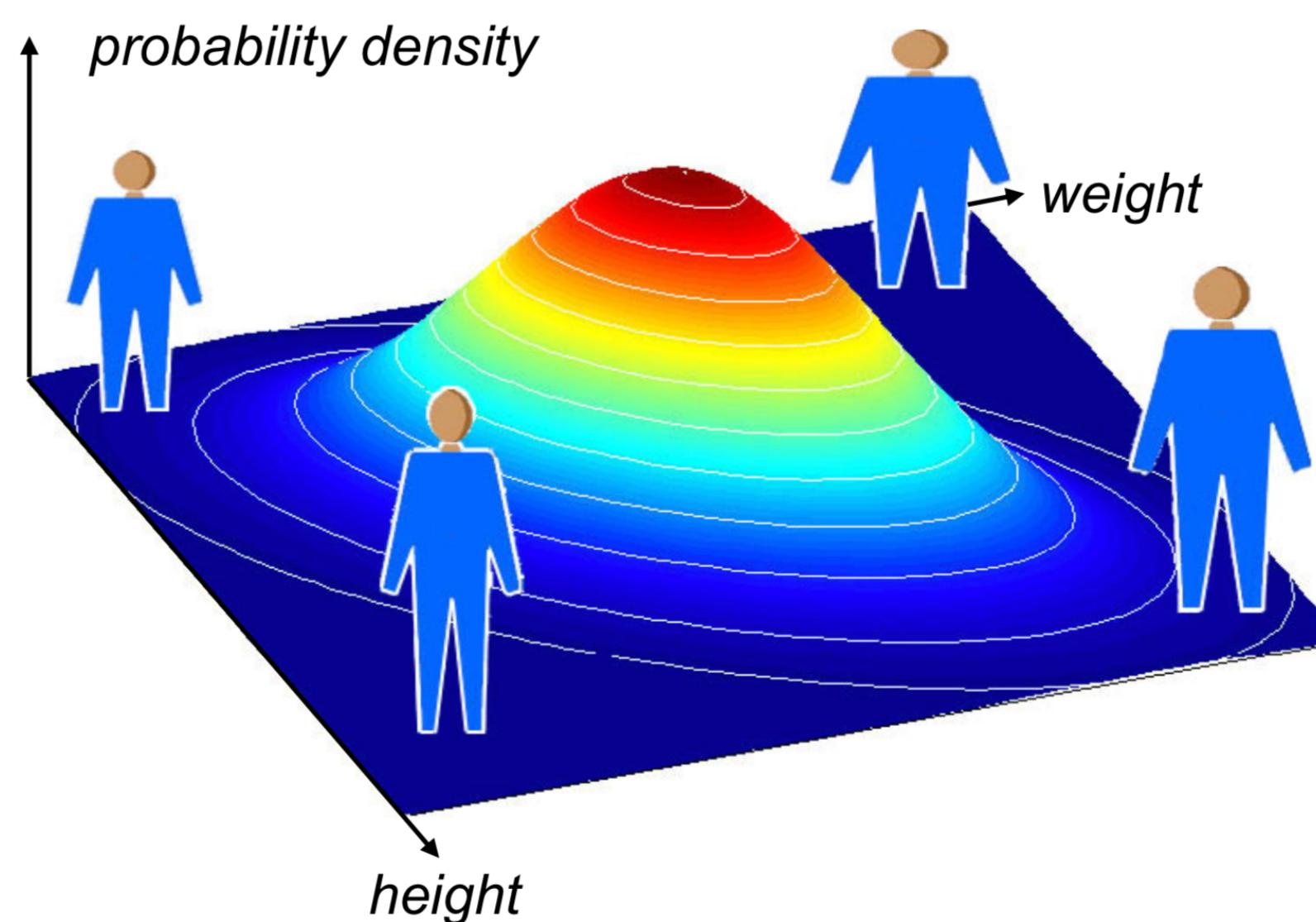
$$\mu(X) = \mu$$

$$Var(X) = \sigma^2$$

MULTIVARIATE NORMAL DISTRIBUTION

Multivariate normal distributions represent probabilities of random variables in several dimensions.

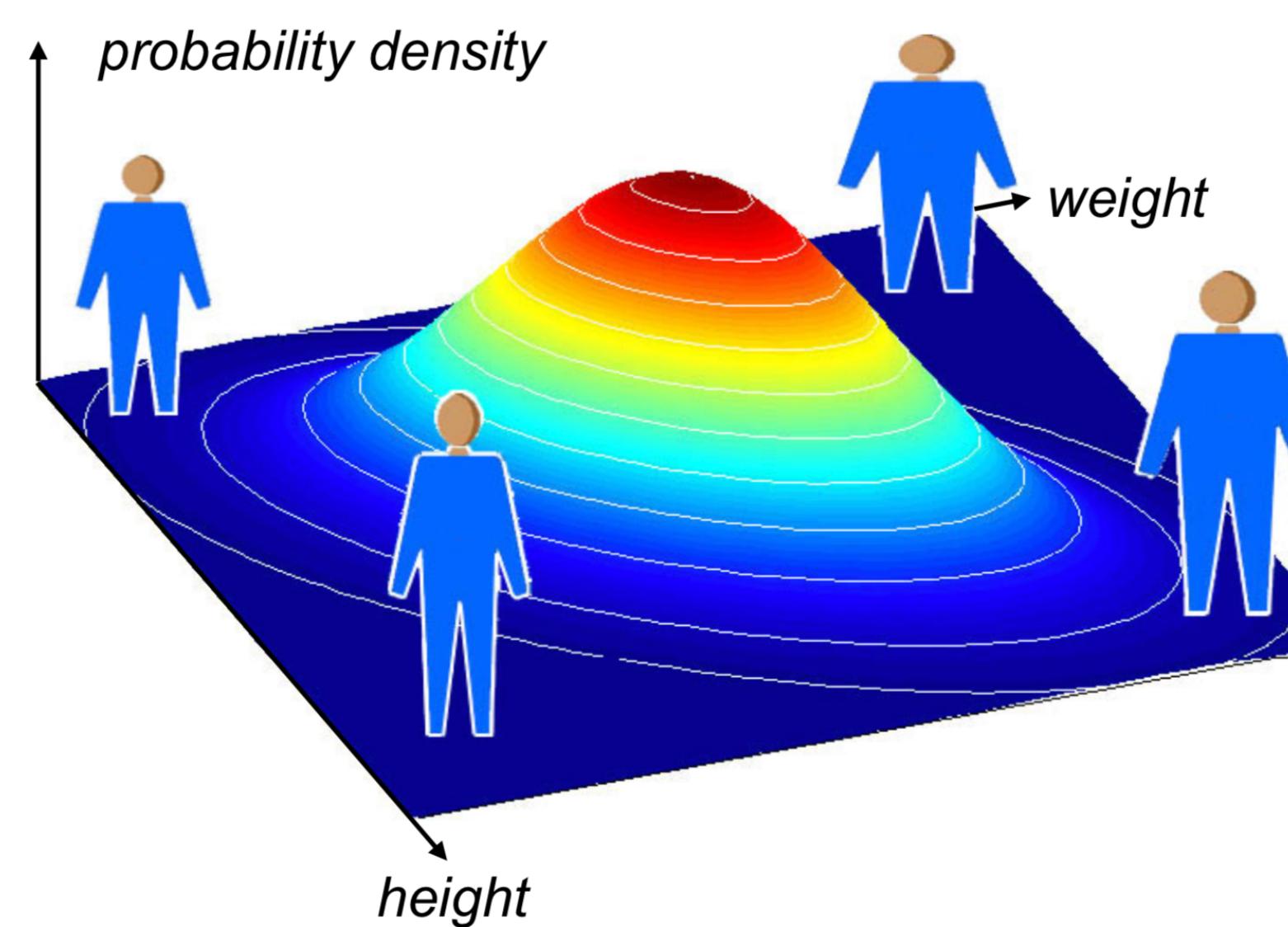
Example: population distribution of height and weight



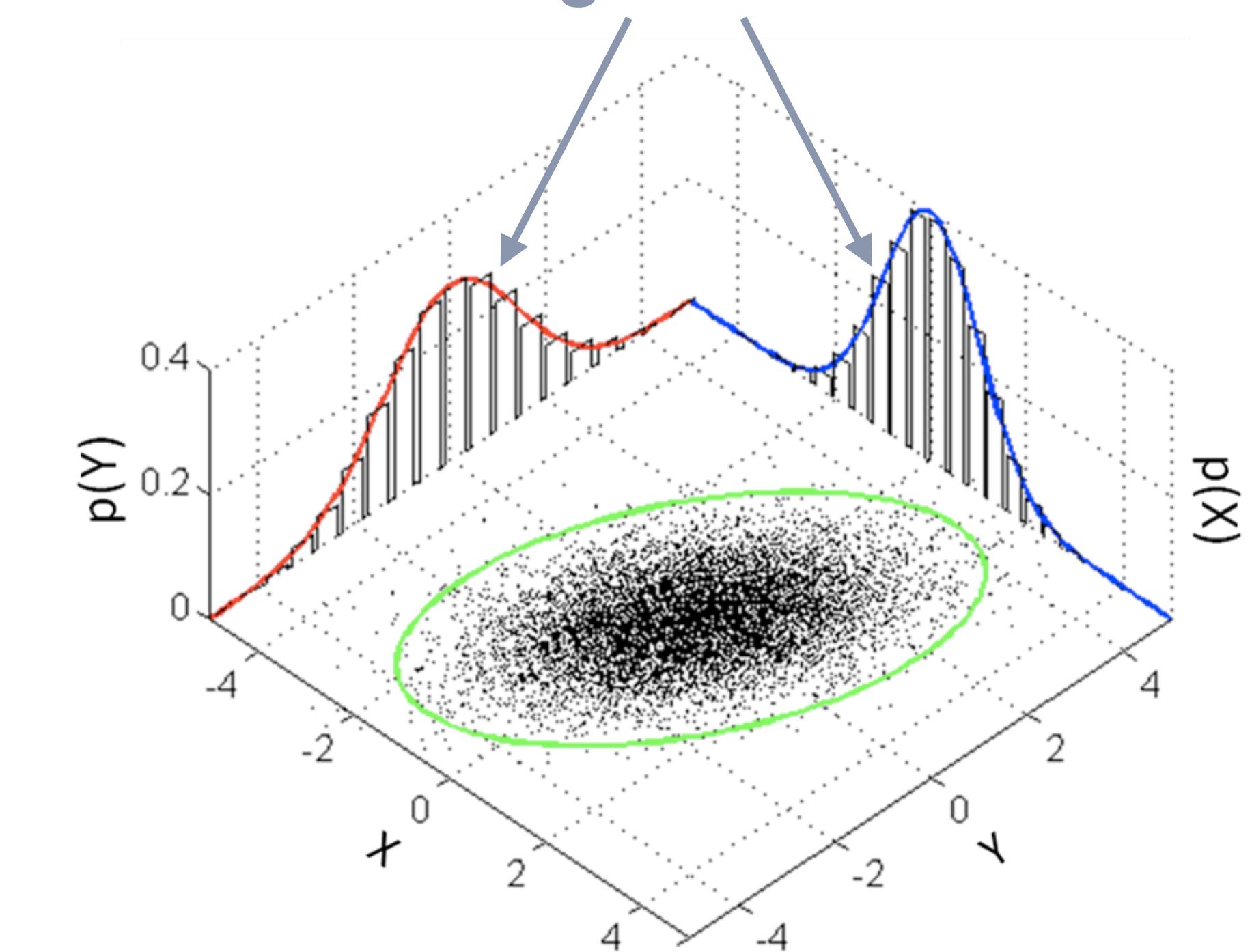
MULTIVARIATE NORMAL DISTRIBUTION

Multivariate normal distributions represent probabilities of random variables in several dimensions.

Example: population distribution of height and weight



Marginalisation



marginalisation is the process of
'projecting out' one variable

MULTIVARIATE NORMAL DISTRIBUTION

How do we define a multivariate Gaussian distribution?

We need to define the **covariance matrix** first

MORE DEFINITIONS

Covariance & Correlation

Assume we have two random variables:

X_1 has mean μ_1 and standard deviation σ_1

X_2 has mean μ_2 and standard deviation σ_2

The **covariance** is defined as:

$$Cov(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)]$$

The **correlation coefficient** is defined as:

$$\rho = \frac{Cov(X_1, X_2)}{\sigma_1 \sigma_2}$$

properties of ρ : $-1 \leq \rho \leq 1$

MORE DEFINITIONS

Covariance & Correlation

Assume we have two random variables:

X_1 has mean μ_1 and standard deviation σ_1

X_2 has mean μ_2 and standard deviation σ_2

The **covariance** is defined as:

$$\text{Cov}(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)]$$

The **correlation coefficient** is defined as:

$$\rho = \frac{\text{Cov}(X_1, X_2)}{\sigma_1, \sigma_2}$$

properties of ρ : $-1 \leq \rho \leq 1$

$\text{Cov}(X_1, X_2) = 0$ or $\rho = 0 \longrightarrow X_1$ and X_2 are uncorrelated

$\rho = -1$ (or $+1$) \longrightarrow means perfect linear relation between them

MORE DEFINITIONS

Covariance matrix (aka variance-covariance matrix, variance matrix, auto-covariance matrix, or dispersion matrix):

$$\Sigma = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_1, X_2) & Var(X_2) \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

MULTIVARIATE NORMAL DISTRIBUTION

And now we have all the ingredients to define a **multivariate normal (or Gaussian) distribution**:

$$f(x_1, x_2) = \frac{1}{\sqrt{2\pi^k |\Sigma|}} \exp \left[-\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right]$$

$$|\Sigma| \equiv \det(\Sigma)$$

When Σ is positive definite, then the multivariate normal distribution is non-degenerate and has density (it can be interpreted as a probability density function)

<https://distill.pub/2019/visual-exploration-gaussian-processes/>

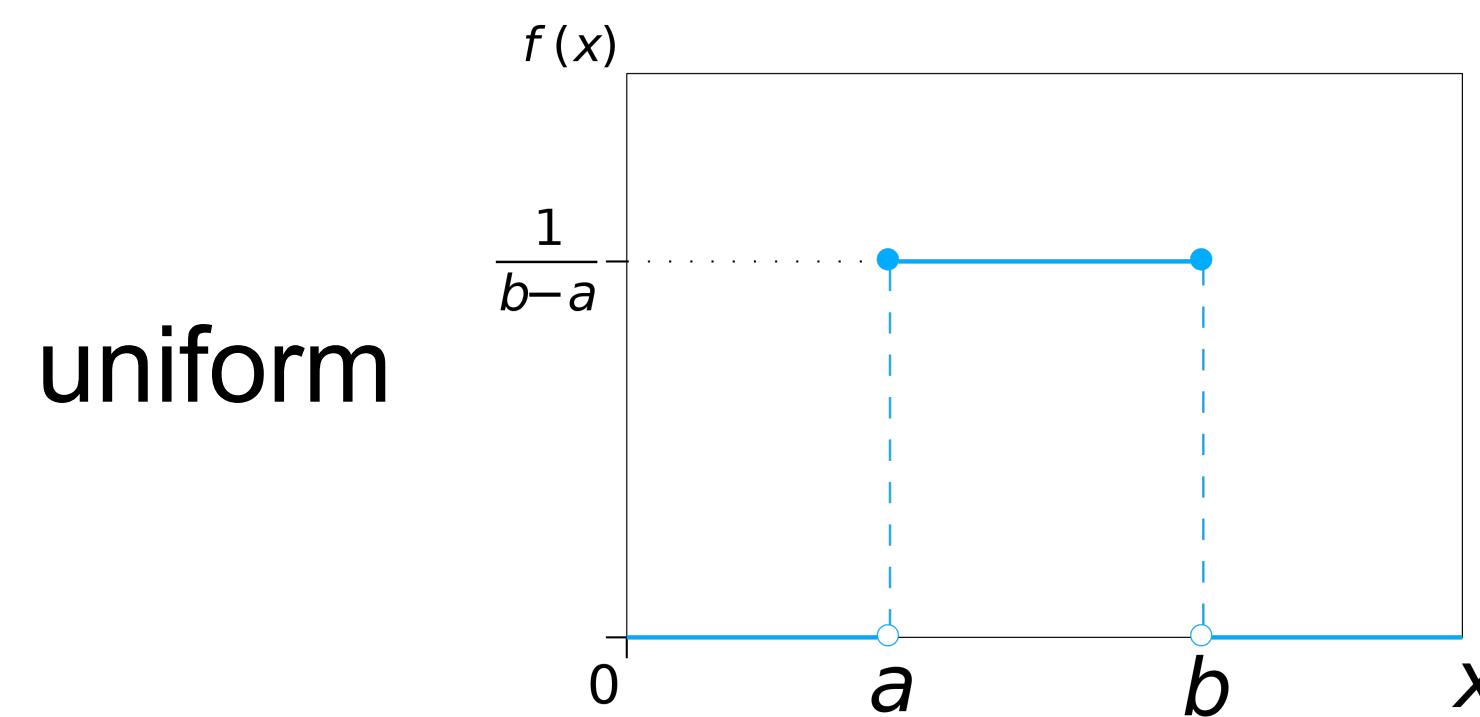
QUICK RECAP

Let's revisit a few concepts before we move on:

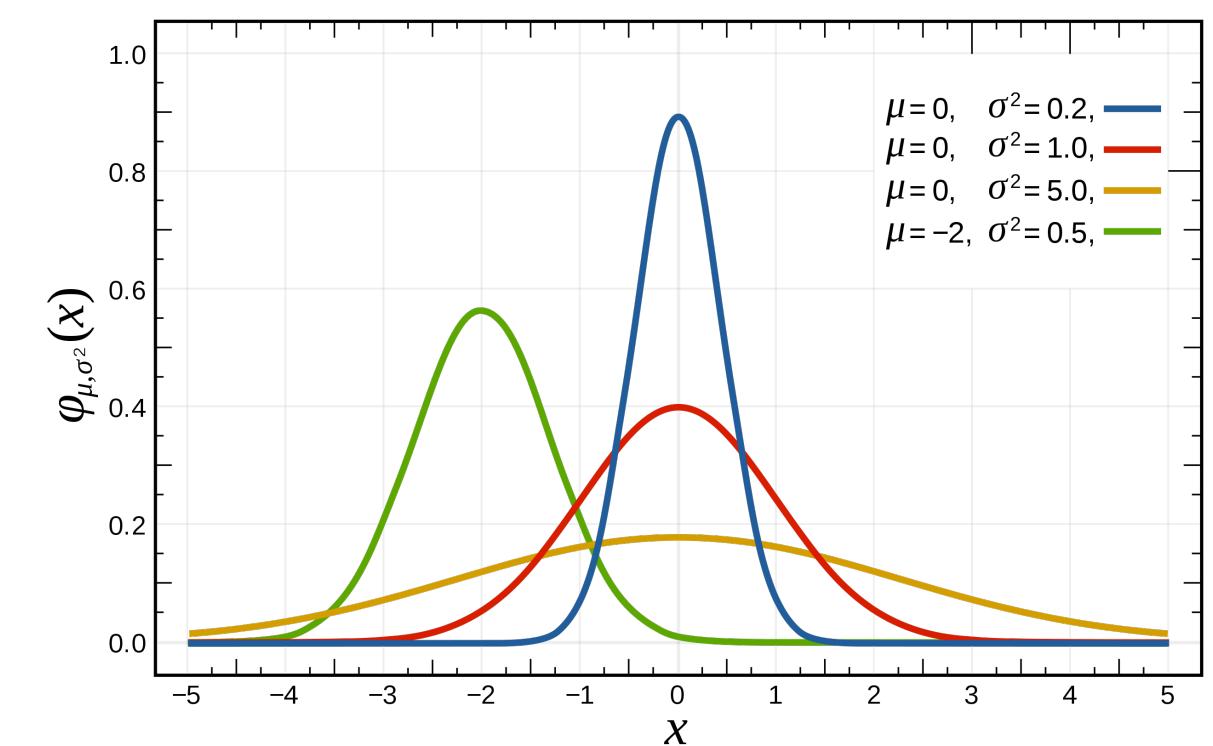
- random variables: a random variable X takes a value x

X : random variable
 x : regular or constant variable

- univariate distribution: probability distribution describing a variable with one dimension



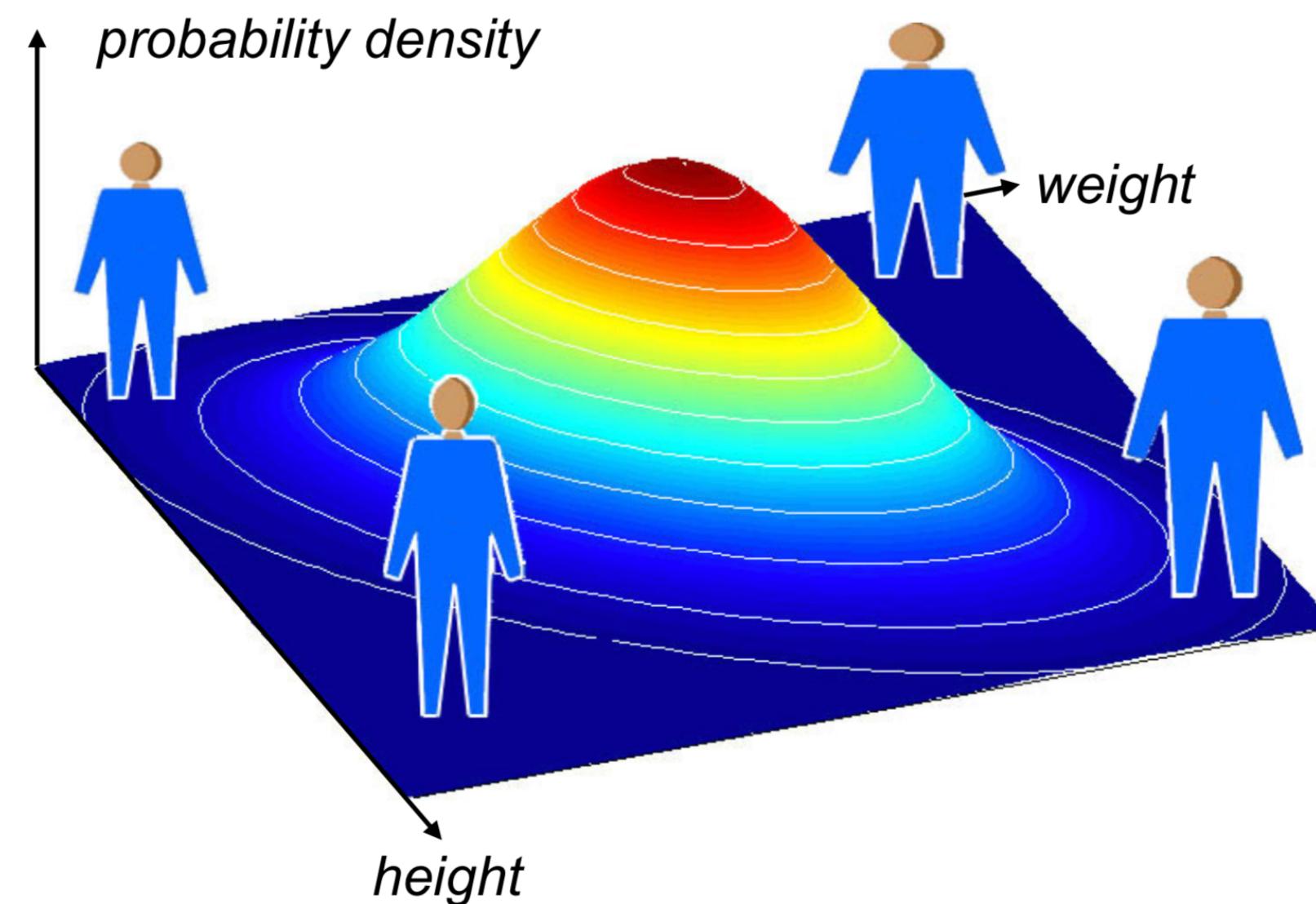
normal
or
Gaussian



QUICK RECAP

Let's revisit a few concepts before we move on:

- **Multivariate distribution:** probability distribution of variables with multiple dimensions.



For example, you have m variables (m individuals) that contain two values each (weight and height) and I want to find what is the multivariate, or bivariate in this case, normal distribution that best fits my data (my m samples or m individuals)

QUICK RECAP

Let's revisit a few concepts before we move on:

- **Multivariate distribution:** data, mean, and variance matrices

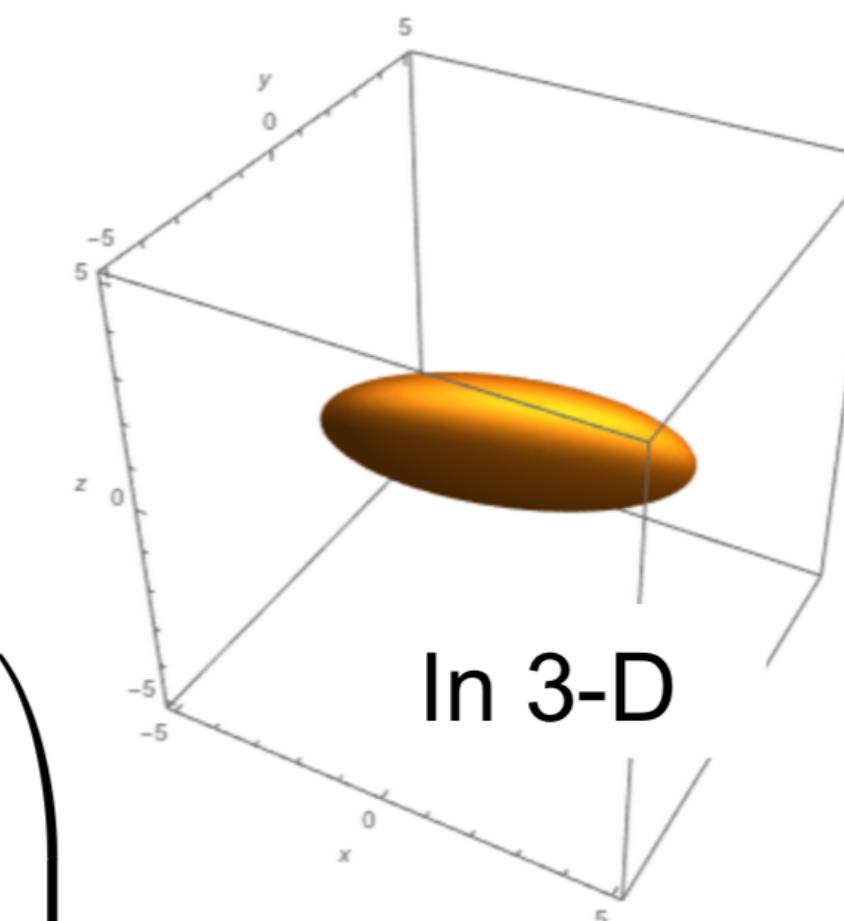
x is the $n \times 1$ vector

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

μ is the $n \times 1$ expectation vector

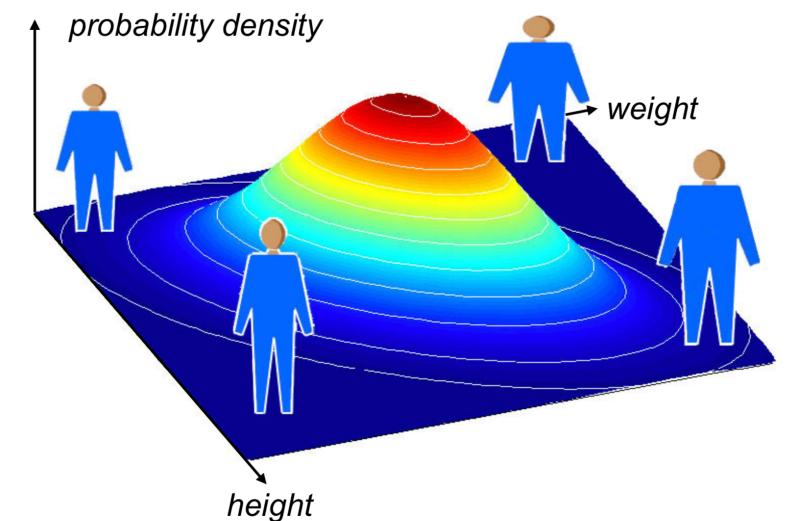
$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix}$$

Σ is the $n \times n$ variance-covariance matrix $\Sigma = \left(\left(Cov(X_i, X_j) \right)_{i,j=1,\dots,n} \right)$



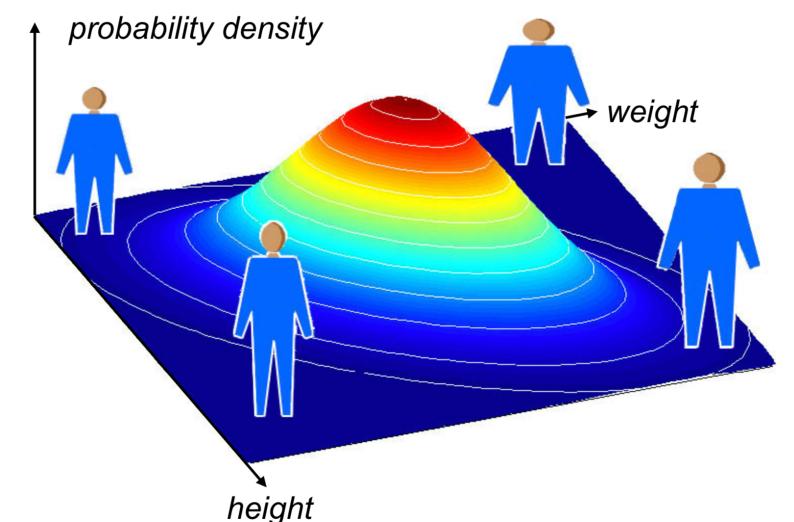
QUICK RECAP

Exercise. If I have the height and weight data from 5 individuals:



1. How many random variables I need to describe this distribution?
2. What is the size of the covariance matrix Σ ?
3. If two people have the exact same weight and height, then the variables are no longer random. True or false?
4. What would be the volume under the Gaussian bell if $\rho = 1$ (correlation coefficient)

Exercise. If I have the height and weight data from 5 individuals:



1. How many random variables I need to describe this distribution?
2, and they take values (x_1^i, x_2^i) for each of the 5 samples ($i=1,2, \dots, 5$)
2. What is the size of the covariance matrix Σ ?
2x2
3. If two people have the exact same weight and height, then the variables are no longer random. True or false?
False. The fact that the variables are randomly distributed is a prior assumption.
4. What would be the volume under the Gaussian bell if $\rho = 1$ (correlation coefficient)
0, although the area under the bell will be > 0.

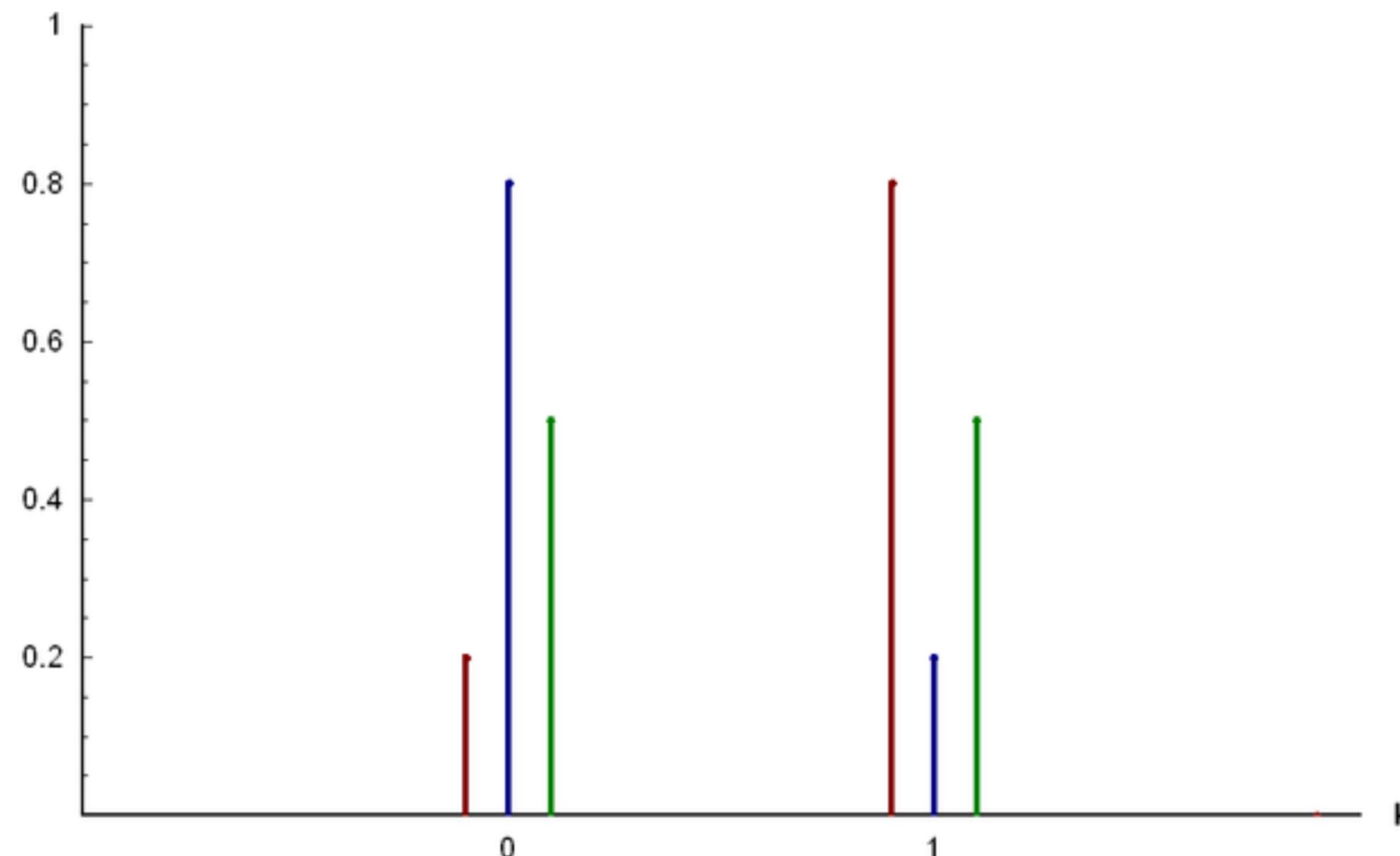
BERNOULLI DISTRIBUTION

A Bernoulli random variable only takes two possible values: 0 or 1

Probability mass function



For discrete random variables, we use the term pmf
(probability mass function)



Three examples of Bernoulli distribution:

■ $P(x = 0) = 0.2$ and $P(x = 1) = 0.8$

■ $P(x = 0) = 0.8$ and $P(x = 1) = 0.2$

■ $P(x = 0) = 0.5$ and $P(x = 1) = 0.5$

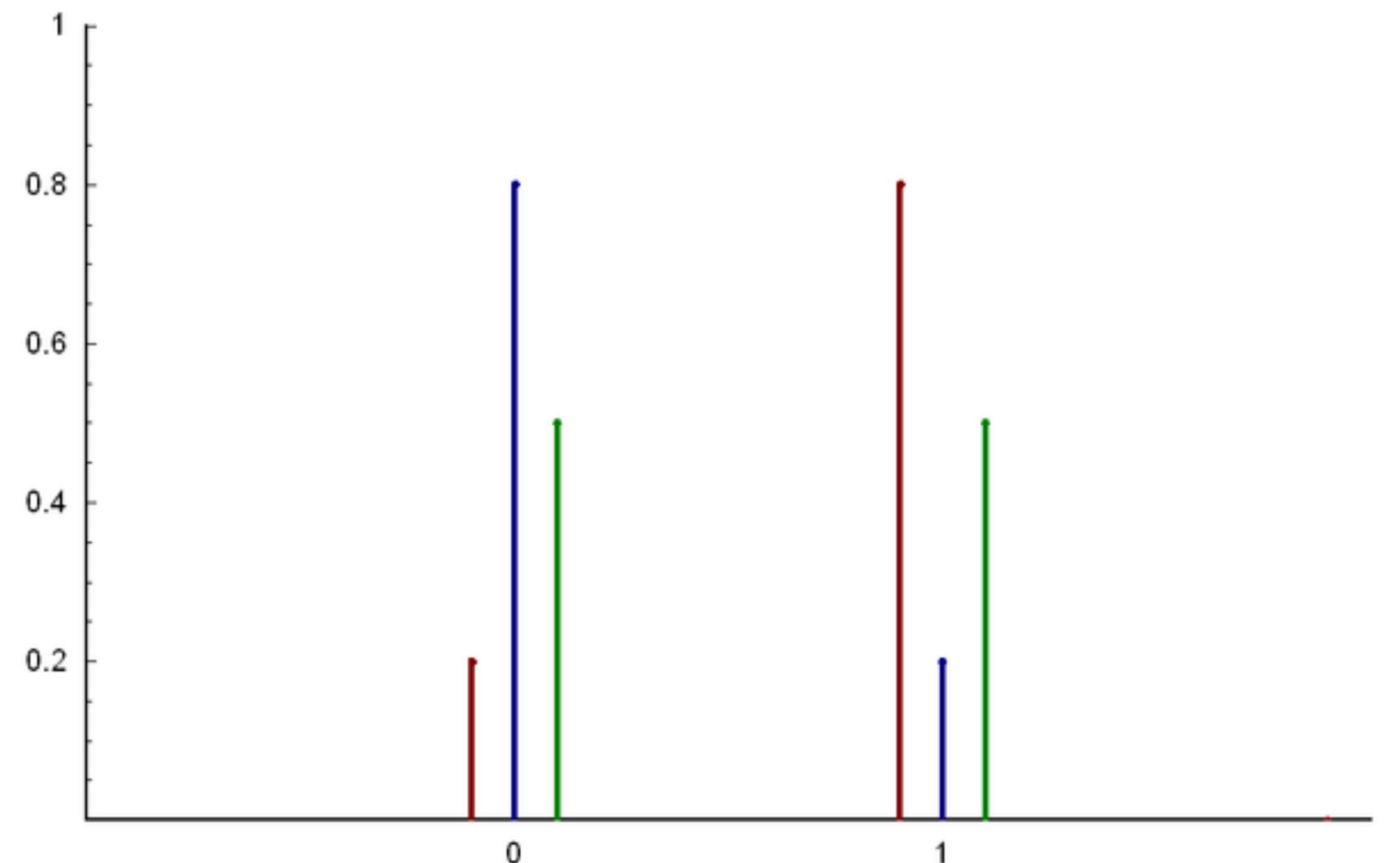
BERNOULLI DISTRIBUTION

A Bernoulli random variable only takes two possible values: 0 or 1

Probability mass function



For discrete random variables, we use the term pmf (probability mass function)



A Bernoulli random variable takes the value 1 with probability p and 0 with probability $(1-p)$.

Probability that a Bernoulli variable takes the value x is:

$$b(x) = p^x (1 - p)^{1-x}$$

Three examples of Bernoulli distribution:

■ $P(x = 0) = 0.2$ and $P(x = 1) = 0.8$

with $x \in \{0, 1\}$

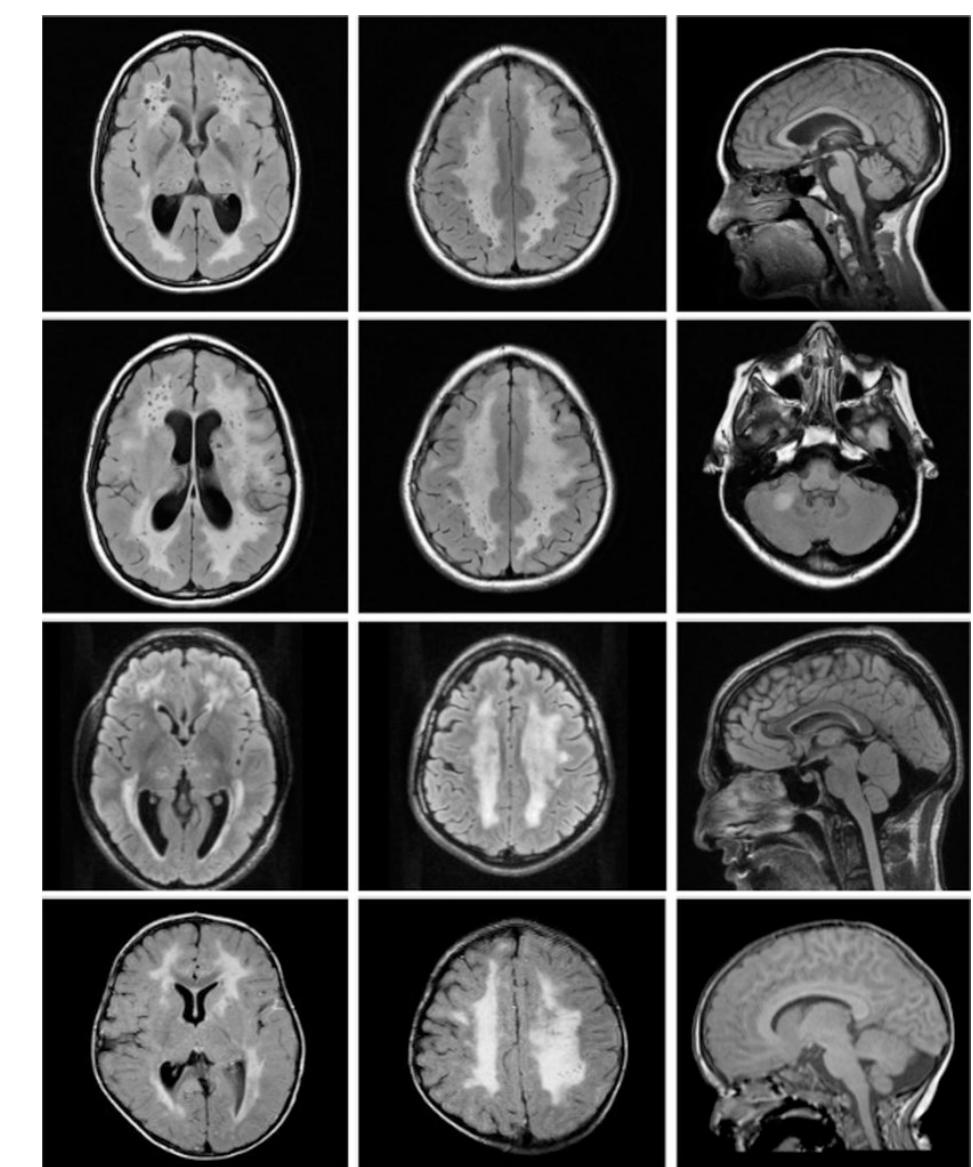
■ $P(x = 0) = 0.8$ and $P(x = 1) = 0.2$

■ $P(x = 0) = 0.5$ and $P(x = 1) = 0.5$

INDEPENDENT AND IDENTICALLY DISTRIBUTED (IID)

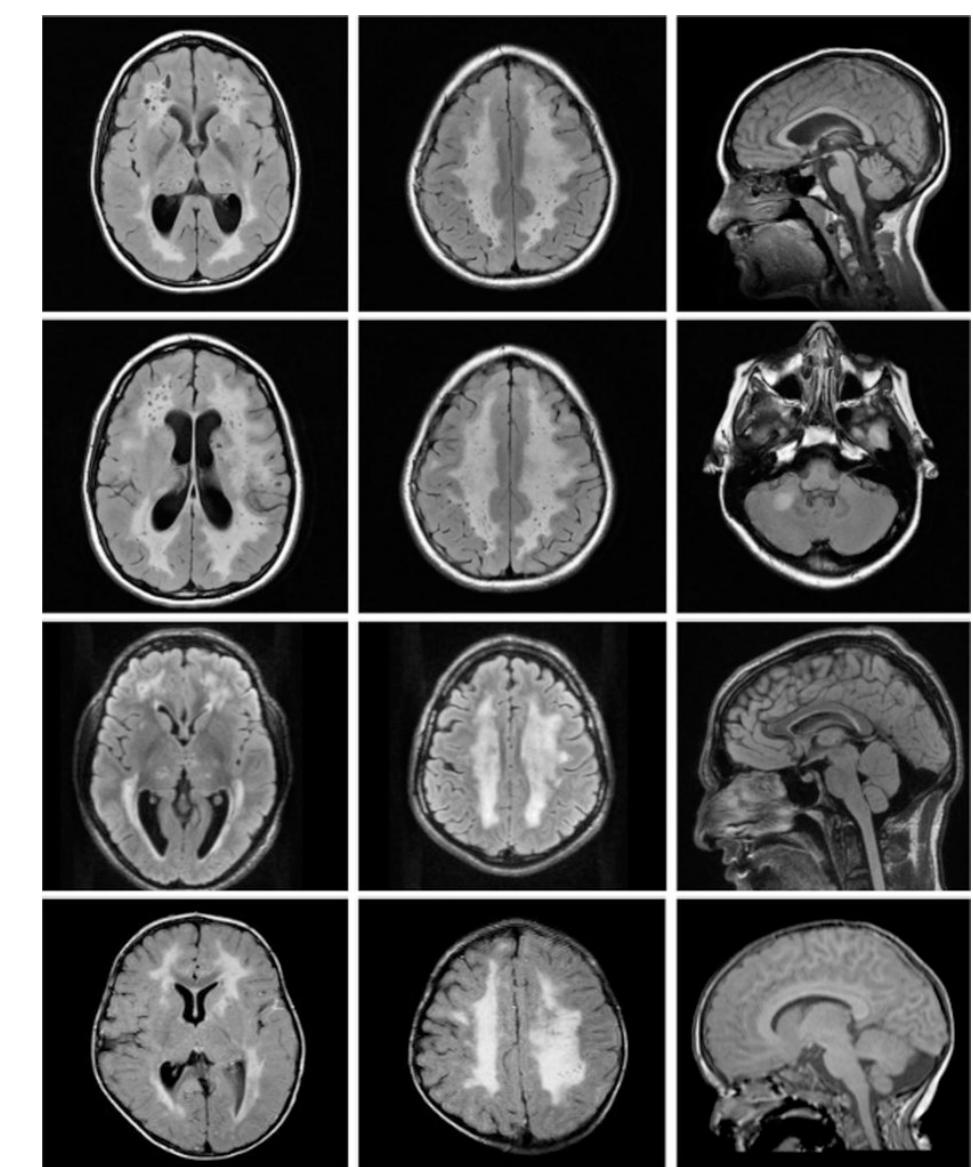
In probability theory, a sequence or collection of random variables is ***independent and identically distributed (i.i.d., iid, or IID)*** if each random variable has the same probability distribution as the others and they are all mutually independent.

We almost always assume that samples from a training or test dataset are ***IID***.



INDEPENDENT AND IDENTICALLY DISTRIBUTED (IID)

In these IID images below, each **pixel** can be treated as a **dimension of a multivariate distribution**.



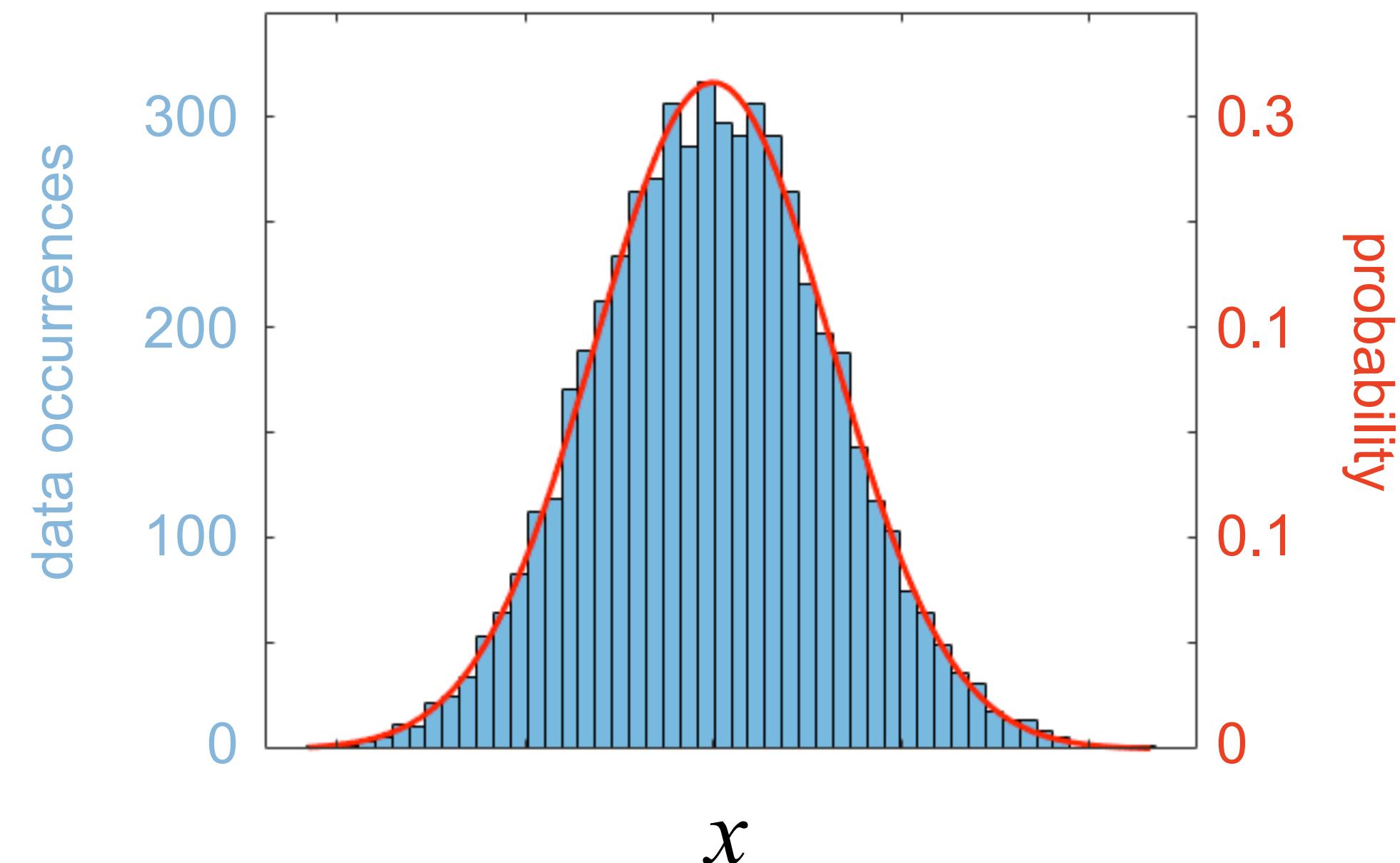
PROBABILITY

1. Probability distributions
2. Maximum likelihood
3. Comparison of probability density functions

PROBABILITY DENSITY ESTIMATION

LED SML
OSOMA2
CEMSCO
201CS2
0321E3

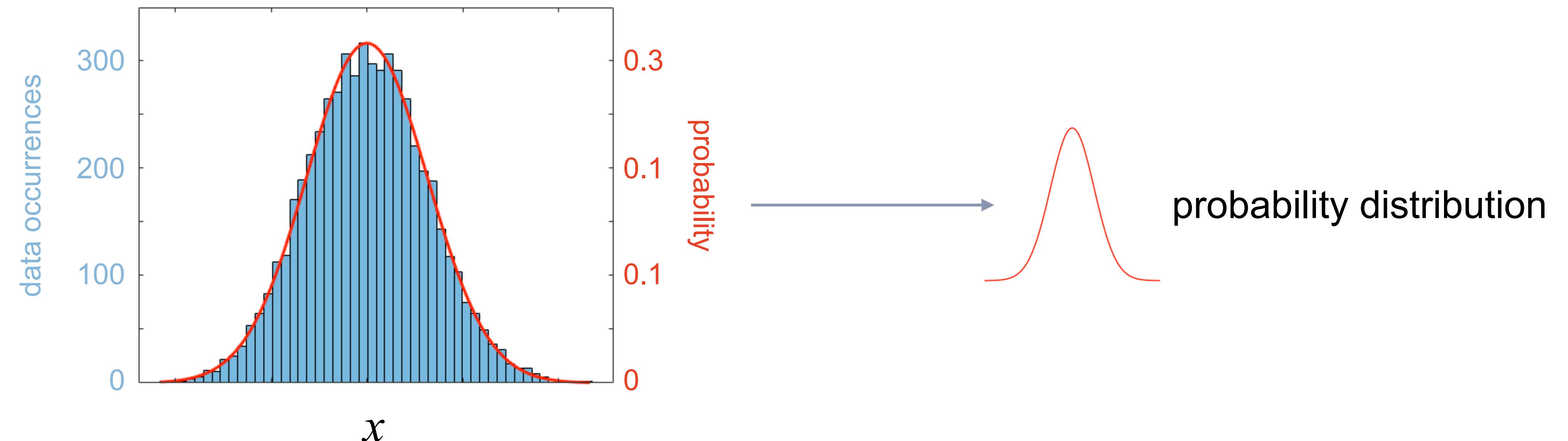
Probability density estimation aims at estimating the distribution that best fits our data:



PROBABILITY DENSITY ESTIMATION

1 EDSML
2 SOMA2
3 CEMSCO
201 CS2
0321E3

Probability density estimation aims at estimating the distribution that best fits our data:



MAXIMUM LIKELIHOOD ESTIMATION

The two main methods to solve the probability density estimation are:

- **MLE**, or maximum likelihood estimation: **frequentist school** $\rightarrow \theta_{MLE} = \arg \max_{\theta} P(D|\theta)$
- **MAP**, or maximum a posteriori: **Bayesian school** $\rightarrow \theta_{MAP} = \arg \max_{\theta} P(\theta|D)$

Bayes' theorem:
$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)}$$

We will focus on **MLE**, but if you have strong feelings about this (ie, you are Bayesian) I recommend that you read **E.T. Jaynes'** book: *The logic of science*

<https://bayes.wustl.edu/etj/prob/book.pdf>

MAXIMUM LIKELIHOOD ESTIMATION

The two main methods to solve the probability density estimation are:

- **MLE**, or maximum likelihood estimation: **frequentist school** $\rightarrow \theta_{MLE} = \arg \max_{\theta} P(D|\theta)$
- **MAP**, or maximum a posteriori: **Bayesian school** $\rightarrow \theta_{MAP} = \arg \max_{\theta} P(\theta|D)$

$$\text{posterior } P(\theta|D) = \frac{\text{prior } P(\theta) \text{likelihood } P(D|\theta)}{\text{marginal likelihood } P(D)}$$

We will focus on **MLE**, but if you have strong feelings about this (ie, you are Bayesian) I recommend that you read **E.T. Jaynes'** book: *The logic of science*

<https://bayes.wustl.edu/etj/prob/book.pdf>

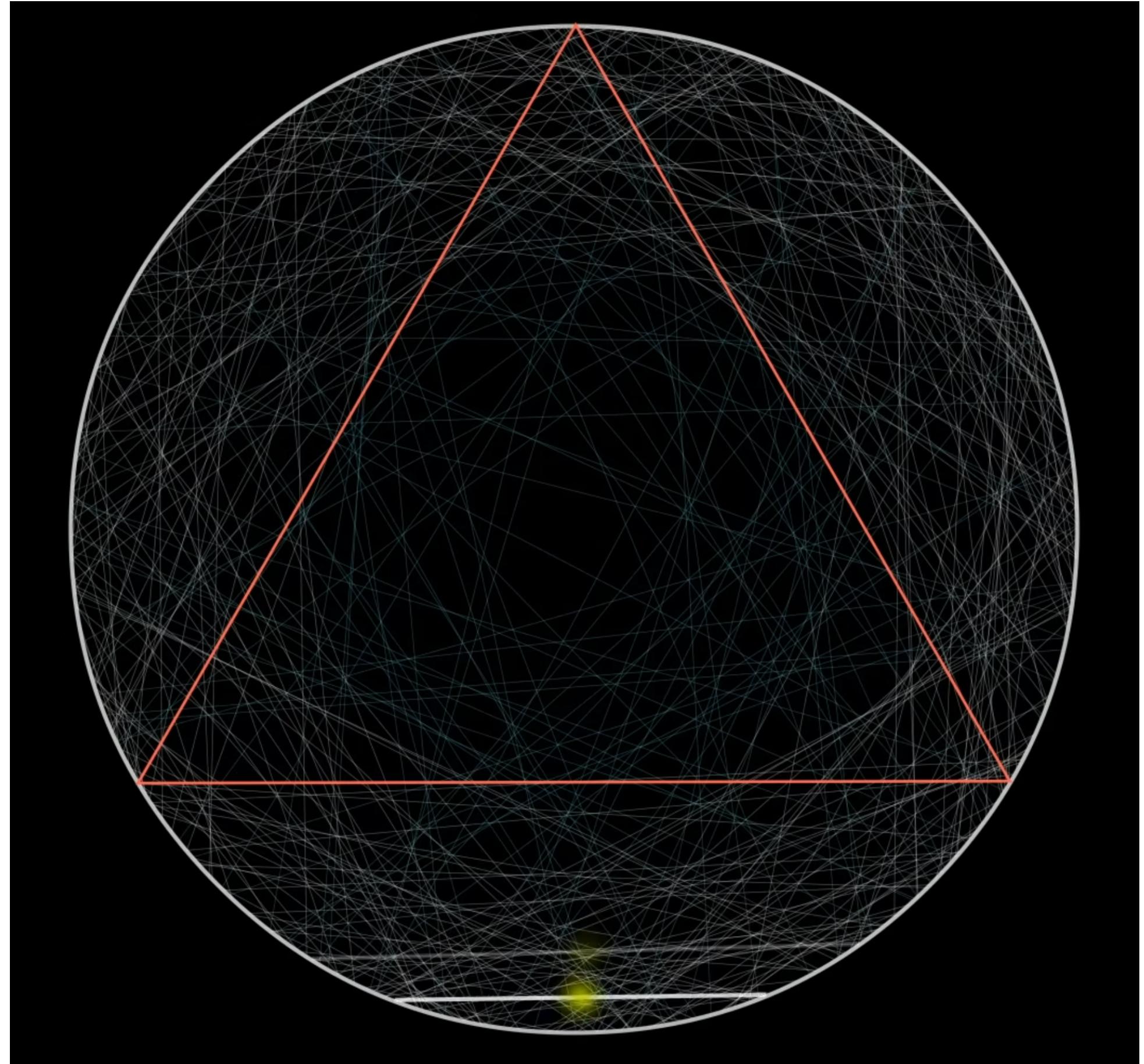
Why does it matter?

Because it can lead to different answers.

For example:

Bayesian vs **Frequentist** results on what is the average length of cord in a circle **differ**.

This is known as **Bertrand's paradox**



Bertrand's paradox:

<https://www.youtube.com/watch?v=mZBwsm6B280>

<https://www.youtube.com/watch?v=pJyKM-7lgAU>

MAXIMUM LIKELIHOOD ESTIMATION

In MLE, we want to maximise the probability of my data X given a distribution θ , or in other words, to **maximise the likelihood**:

—likelihood—

$$p_{\theta}(x) \equiv P(X = x | \theta) \equiv P(x | \theta) \equiv L(\theta | x)$$

MAXIMUM LIKELIHOOD ESTIMATION

Example:

Assume we have m *independent (IID)* data points $x^{(1)}, x^{(2)}, \dots, x^{(m)}$

Each data point is $x^{(i)}$ is composed of n features.

We want to fit a multivariate (for instance multivariate Gaussian) pdf $p_{\theta}(x)$ of dimension n to these m samples.

Maximum likelihood consists of calculating the parameters θ such that the m samples maximize their likelihood.

Example:

Assume we have 4 *independent (IID)* data points $x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}$ with:

$$x^{(1)} = -1, x^{(2)} = 0, x^{(3)} = 1, x^{(4)} = 2$$

Here each data point $x^{(i)}$ is composed of just $n = 1$ feature. We want to fit a Normal distribution $N(x; \mu, \sigma^2)$ to these four data points.

The likelihood of $x^{(1)}$ is: the value of $N(x; \mu, \sigma^2)$ for $x = x^{(1)} = -1$:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x^{(1)}-\mu}{\sigma}\right)^2} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{-1-\mu}{\sigma}\right)^2}$$

Example:

And the likelihood of the IID sequence $(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)})$ is:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{-1-\mu}{\sigma}\right)^2} \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{0-\mu}{\sigma}\right)^2} \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{1-\mu}{\sigma}\right)^2} \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{2-\mu}{\sigma}\right)^2}$$

The **likelihood** is the **multiplication** of the conditional probability for each sample given θ (ie, μ and σ)

A good tool to visually understand likelihood:

<https://seeing-theory.brown.edu/bayesian-inference/index.html>

FROM LIKELIHOOD TO MAXIMUM LIKELIHOOD

For one single image:

$$x^{(1)} = (x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \dots, x_n^{(1)})$$

Likelihood is:

$$p_{\theta}(x^{(1)}) = p_{\theta}(x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \dots, x_n^{(1)})$$

FROM LIKELIHOOD TO MAXIMUM LIKELIHOOD

For one single image:

$$x^{(1)} = (x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \dots, x_n^{(1)})$$

Likelihood is:

$$p_{\theta}(x^{(1)}) = p_{\theta}(x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \dots, x_n^{(1)})$$

Maximum likelihood estimate:

$$\theta_{ML} = \operatorname{argmax}_{\theta} p_{\theta}(x^{(1)}) \longleftrightarrow \boxed{\theta_{ML} = \operatorname{argmax}_{\theta} \log(p_{\theta}(x^{(1)}))}$$

FROM LIKELIHOOD TO MAXIMUM LIKELIHOOD

For m images, the maximum likelihood estimate is:

$$\theta_{ML} = \operatorname{argmax}_{\theta} \left(p_{\theta}(x^{(1)})p_{\theta}(x^{(2)})p_{\theta}(x^{(3)})...p_{\theta}(x^{(m)}) \right)$$

$$\theta_{ML} = \operatorname{argmax}_{\theta} \left(\log \left(p_{\theta}(x^{(1)})p_{\theta}(x^{(2)})p_{\theta}(x^{(3)})...p_{\theta}(x^{(m)}) \right) \right)$$

$$\theta_{ML} = \operatorname{argmax}_{\theta} \left(\sum_{i=1}^m \log p_{\theta}(x^{(i)}) \right)$$

LOG-LIKELIHOOD

If we now assume a normal distribution, the log-likelihood becomes:

$$\begin{aligned} \sum_{i=1}^m \log p_\theta(x^{(i)}) &= \sum_{i=1}^m \log \left(\frac{1}{\sigma \sqrt{2\pi}} \exp \left(-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right) \right) \\ &= -\frac{1}{2} \sum_{i=1}^m \left(\frac{x - \mu}{\sigma} \right)^2 - m \log \sigma - \frac{m}{2} \log 2\pi \end{aligned}$$

LOG-LIKELIHOOD

If we now assume a normal distribution, the log-likelihood becomes:

$$\begin{aligned} \sum_{i=1}^m \log p_\theta(x^{(i)}) &= \sum_{i=1}^m \log \left(\frac{1}{\sigma \sqrt{2\pi}} \exp \left(-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right) \right) \\ &= -\frac{1}{2} \sum_{i=1}^m \left(\frac{x - \mu}{\sigma} \right)^2 - m \log \sigma - \frac{m}{2} \log 2\pi \end{aligned}$$



this should look familiar

MLE AND L2 LOSS RELATION

Now we can link MLE with the L2 norm in our network training. But first, we will need to make an important assumption:

$$y^{(i)} = \hat{y}^{(i)} + \epsilon = f(x^{(i)}, \theta) + \epsilon$$

where my assumption is that the noise ϵ follows a normal distribution:

$$y^{(i)} - f(x^{(i)}, \theta) = \epsilon \sim \mathcal{N}(\mu, \sigma^2)$$

MLE AND L2 LOSS RELATION

with this assumption, we can now see that:

$$\begin{aligned}\theta_{ML} &= \operatorname{argmax}_{\theta} p_{\theta}(x^{(i)}) \\ &= \operatorname{argmax}_{\theta} \log p_{\theta}(x^{(i)}) \\ &= \operatorname{argmin}_{\theta} -\log p_{\theta}(x^{(i)})\end{aligned}$$



remember assumption
of noise distribution:

$$p_{\theta}(x^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(y^{(i)} - f(x^{(i)}, \theta))^2}{2\sigma^2}\right)$$

MLE AND L2 LOSS RELATION

with this assumption, we can now see that:

$$\theta_{ML} = \operatorname{argmax}_{\theta} p_{\theta}(x^{(i)})$$

$$= \operatorname{argmax}_{\theta} \log p_{\theta}(x^{(i)})$$

$$= \operatorname{argmin}_{\theta} -\log p_{\theta}(x^{(i)})$$

$$= \operatorname{argmin}_{\theta} \frac{1}{2\sigma^2} \sum_{i=1}^m \left(y^{(i)} - f(x^{(i)}, \theta) \right)^2 + \text{constant}$$

$$= \operatorname{argmin}_{\theta} \frac{1}{2} \sum_{i=1}^m \left(y^{(i)} - f(x^{(i)}, \theta) \right)^2$$

remember assumption
of noise distribution:

$$p_{\theta}(x^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left(\frac{(y^{(i)} - f(x^{(i)}, \theta))^2}{2\sigma^2} \right)$$

L2-norm loss (MSE loss)

MLE AND CROSS-ENTROPY RELATION

The likelihood of a Bernoulli distribution was:

$$b(x) = p^x (1 - p)^{1-x} \quad \text{with } x \in \{0, 1\}$$

and its log-likelihood:

$$\log b(x) = x \log p + (1 - x) \log (1 - p)$$

if we have m samples:

$$\sum_{i=1}^m (x^{(i)} \log p + (1 - x^{(i)}) \log (1 - p))$$

minus **binary cross-entropy**
is the same as the negative
log-likelihood Loss.

[**NLLLoss** in PyTorch]

PROBABILITY

1. Probability distributions
2. Maximum likelihood
3. Comparison of probability density functions

Comparing probability distributions with the Kullback-Leibler (KL) Divergence

For two pdfs $p(x)$ and $q(x)$:

$$D_{KL}(p\|q) = \int_{-\infty}^{+\infty} p(x) \log p(x) dx - \int_{-\infty}^{+\infty} p(x) \log q(x) dx$$

Two Fundamental Properties

- $D_{KL}(p\|q)$ is positive, and 0 if $p(x)$ and $q(x)$ identical
- Asymmetry: $D_{KL}(p\|q) \neq D_{KL}(q\|p)$

Exercise

If we have two normal distributions:

$$f = \mathcal{N}(x; \mu_1, \sigma_1^2) \quad \text{and} \quad g = \mathcal{N}(x; \mu_2, \sigma_2^2)$$

What is the KL divergence between them?

Exercise

If we have two normal distributions:

$$f = \mathcal{N}(x; \mu_1, \sigma_1^2) \quad \text{and} \quad g = \mathcal{N}(x; \mu_2, \sigma_2^2)$$

What is the KL divergence between them?

$$D_{KL}(f\|g) = -\frac{1}{2} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} + \log \frac{\sigma_2}{\sigma_1}$$

KL DIVERGENCE AND MLE

If the Training Set consists of m *IID* data points $x^{(1)}, x^{(2)}, \dots, x^{(m)}$, the KL Divergence $D_{KL}(p_d \| p_\theta)$ between the experimental distribution p_d of the Training Set and any theoretical distribution p_θ is:

$$D_{KL}(p_d \| p_\theta) = \int_{-\infty}^{+\infty} p_d(x) \log p_d(x) dx - \int_{-\infty}^{+\infty} p_d(x) \log p_\theta(x) dx$$

Minimizing $D_{KL}(p_d \| p_\theta)$ in the parameters θ is equivalent to maximizing the second term:

$$\int_{-\infty}^{+\infty} p_d(x) \log p_\theta(x) dx$$

why is it equivalent?

KL DIVERGENCE AND MLE

If the Training Set consists of m *IID* data points $x^{(1)}, x^{(2)}, \dots, x^{(m)}$, the KL Divergence $D_{KL}(p_d \| p_\theta)$ between the experimental distribution p_d of the Training Set and any theoretical distribution p_θ is:

$$D_{KL}(p_d \| p_\theta) = \int_{-\infty}^{+\infty} p_d(x) \log p_d(x) dx - \int_{-\infty}^{+\infty} p_d(x) \log p_\theta(x) dx$$

Minimizing $D_{KL}(p_d \| p_\theta)$ in the parameters θ is equivalent to maximizing the second term:

$$\int_{-\infty}^{+\infty} p_d(x) \log p_\theta(x) dx$$

But this is the expression of the expectation of $\log p_\theta(x)$ calculated over the Training Set :

$$\theta = \operatorname{argmax} \left(\frac{1}{m} \sum_{i=1}^m \log p_\theta(x^{(i)}) \right)$$

Minimizing the KL Divergence is equivalent to maximizing the Likelihood

SUMMARY

- The basic pdfs used in Deep Learning are **Bernouilli for discrete variables** and (Multivariate) **Gaussian or Uniform for continuous variables**.

SUMMARY

- The basic pdfs used in Deep Learning are **Bernouilli** for discrete variables and (Multivariate) **Gaussian** or Uniform for continuous variables.
- Maximum Likelihood is a key approach in Deep Learning and applying it to **Bernouilli** and (Multivariate) **Gaussian** random variables leads respectively to the minimization of the **cross-entropy** for classification or the **L2 norm** for regression.

SUMMARY

- The basic pdfs used in Deep Learning are **Bernouilli** for discrete variables and (Multivariate) **Gaussian** or Uniform for continuous variables.
- Maximum Likelihood is a key approach in Deep Learning and applying it to **Bernouilli** and (Multivariate) **Gaussian** random variables leads respectively to the minimization of the **cross-entropy** for classification or the **L2 norm** for regression.
- The **Kullback-Leibner (KL) Divergence** is used to calculate the **dissimilarity between two pdfs**. Minimizing it allows one pdf to be adjusted to another. As we will see, there are other divergences that can be used to train networks. For example the **Jensen-Shannon Divergence JSD**.