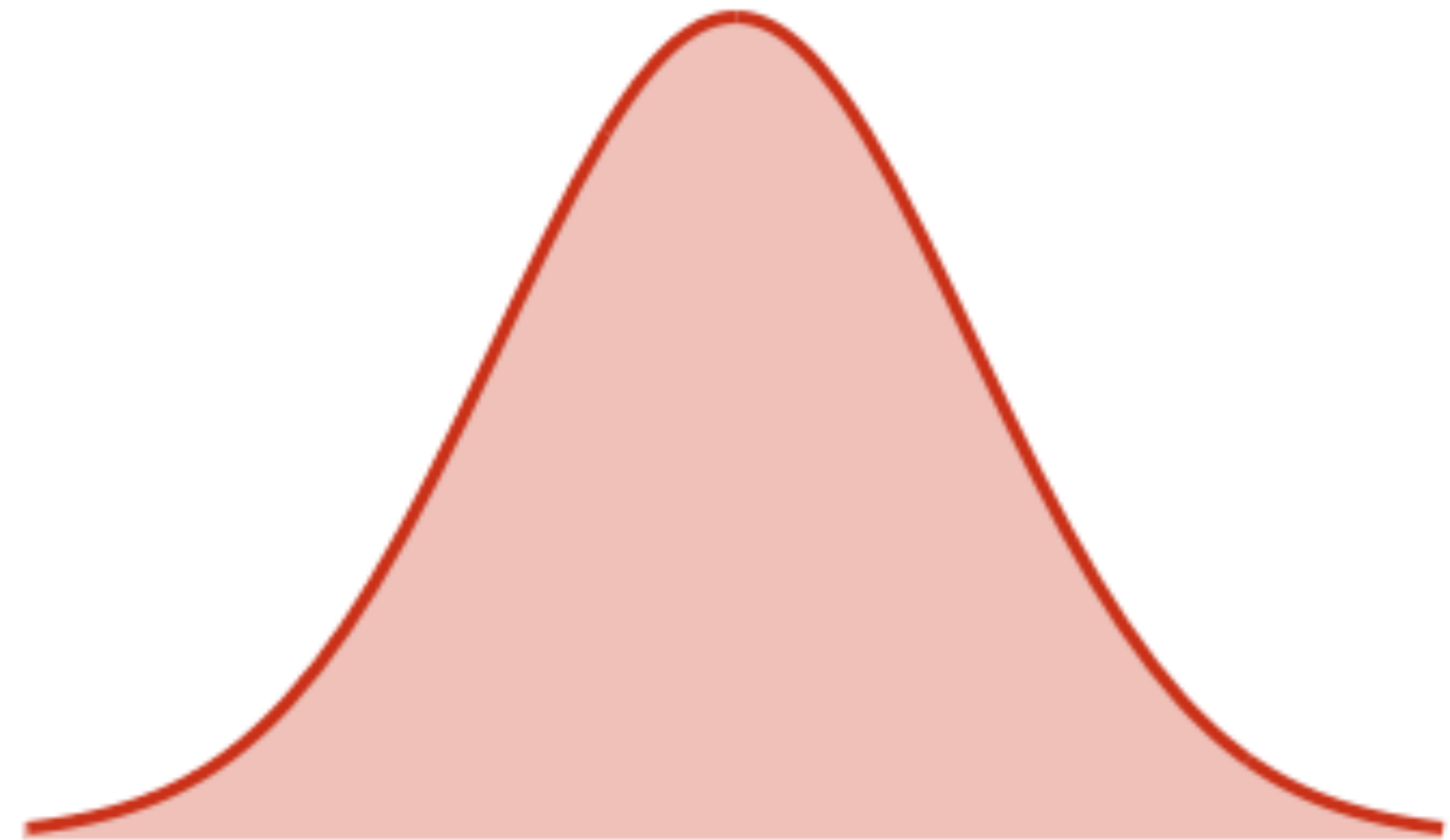


Lecture 13: Normal distributions

Instructor: Michael Szell

Oct 11, 2023



Today you will learn more about (normal) distributions

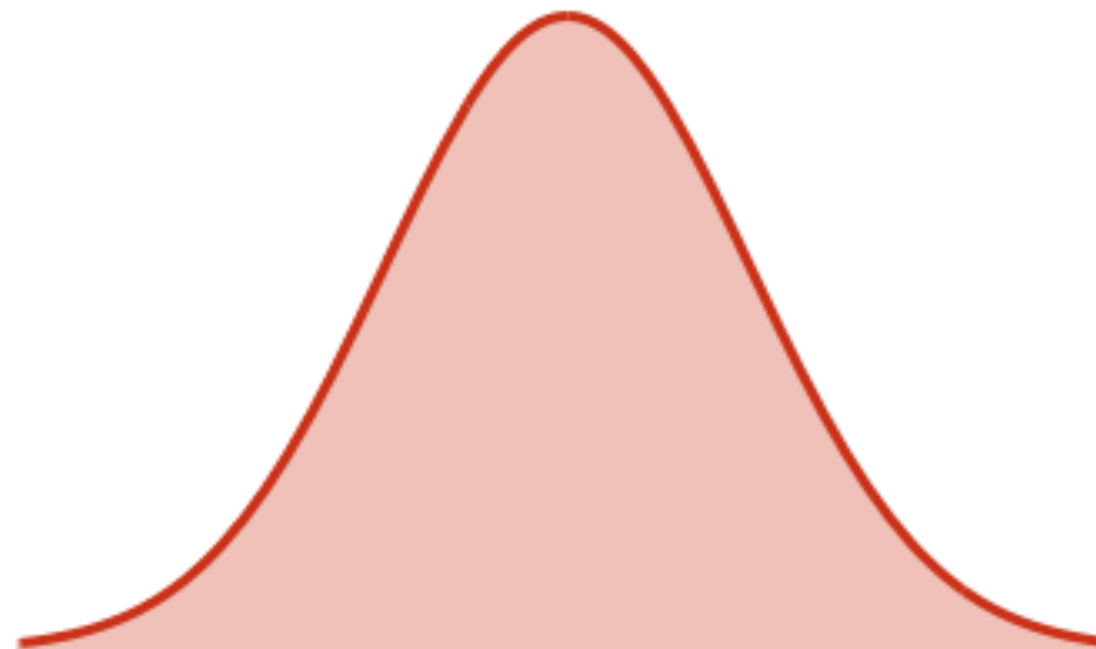
Fundamentals of probability theory



Standardization

$$z = \frac{x - \mu}{\sigma}$$

Normal distributions



The **mean** \bar{x} is the average value

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum x_i$$

It gives an idea of the "center" of the distribution

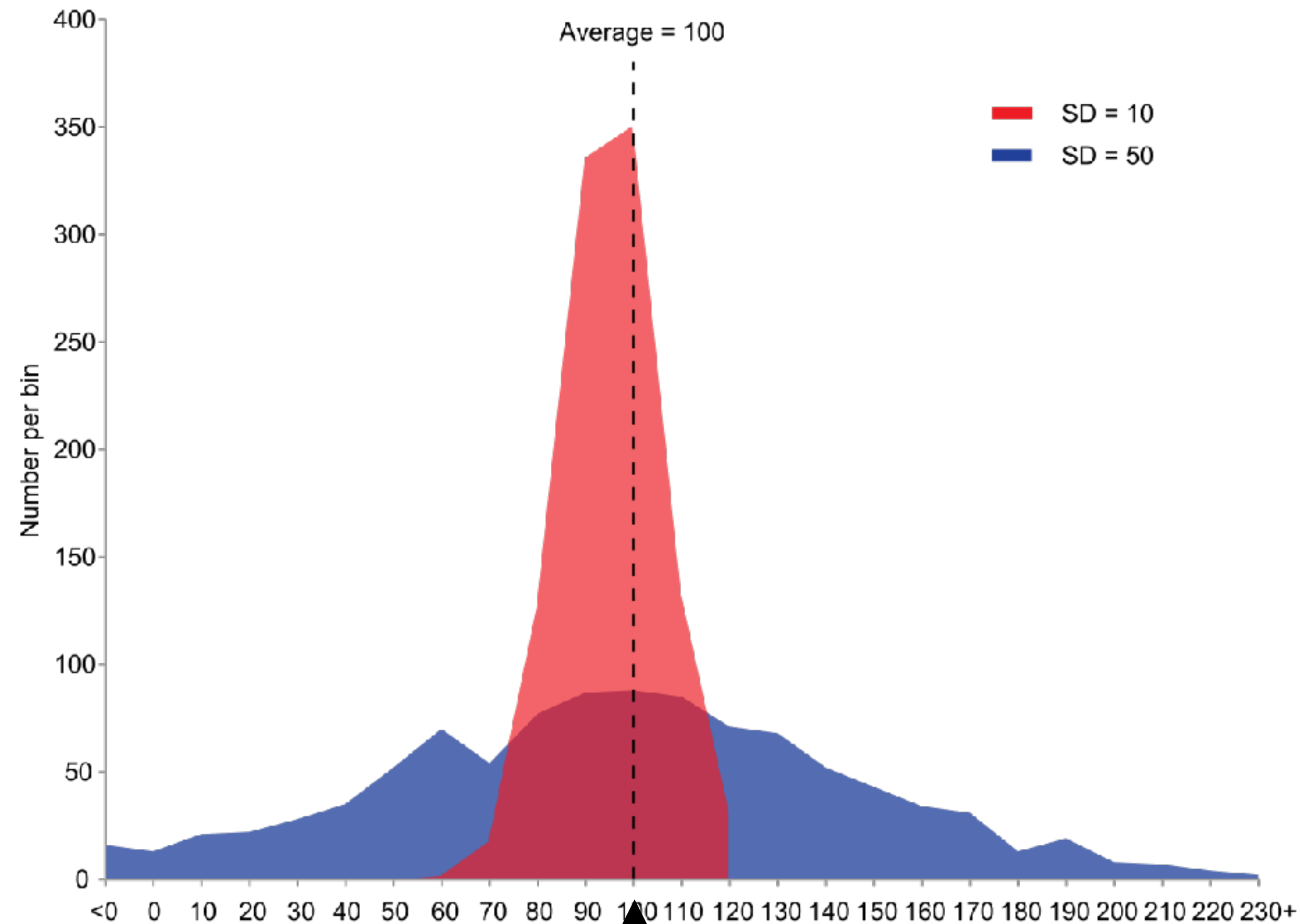
The **standard deviation s** measures spread

variance: $s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$

$$= \frac{1}{n - 1} \sum (x_i - \bar{x})^2$$

$$s = \sqrt{\frac{1}{n - 1} \sum (x_i - \bar{x})^2}$$

Mean and deviation give you a rough idea of the distribution

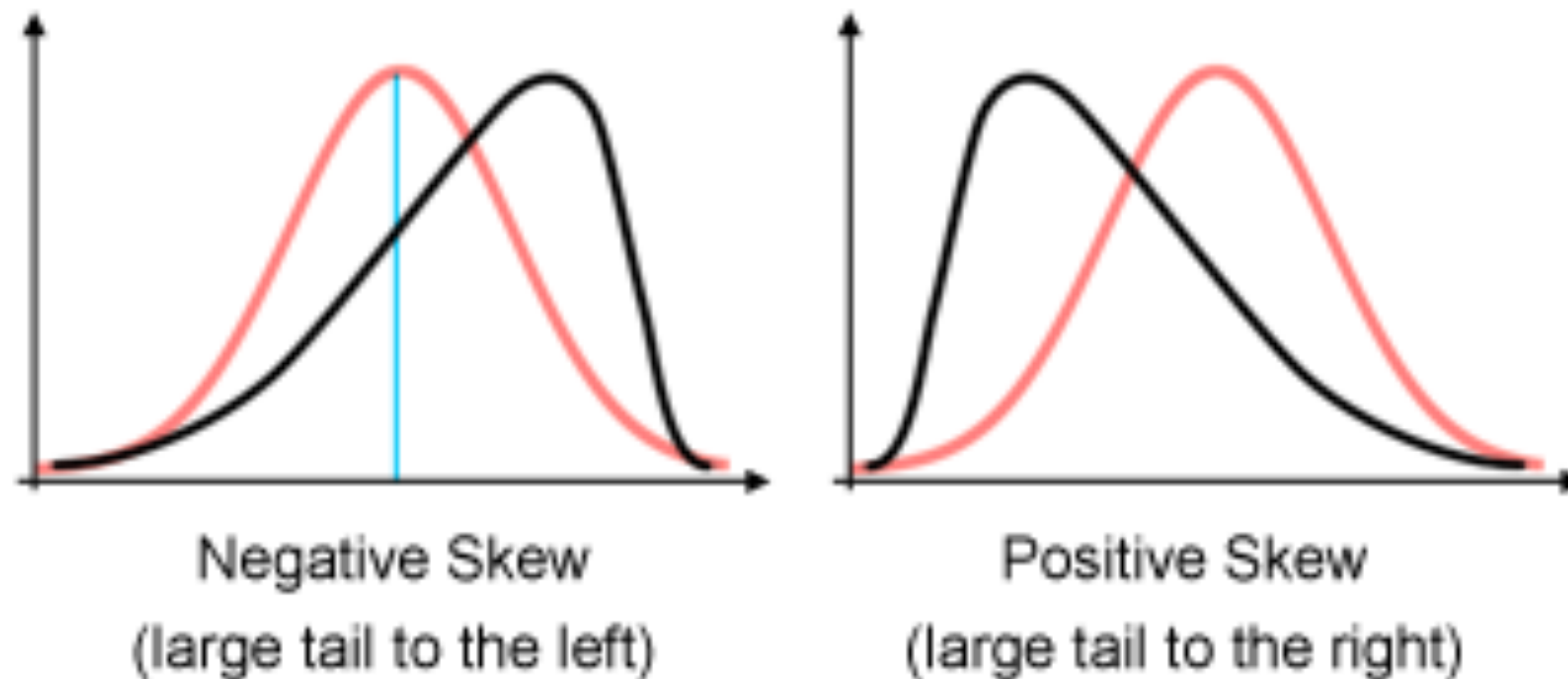


Center

Spread

Beware: Mean and deviation are often not useful measures

- 1) They are not robust to outliers
- 2) They are inadequate for skewed distributions:



Our recipe for exploring data on a single quantitative variable:

- 1) Plot a histogram
- 2) Look for the overall pattern and deviations, outliers
- 3) Calculate a numerical summary to describe center and spread

Our recipe for exploring data on a single quantitative variable:

- 1) Plot a histogram
- 2) Look for the overall pattern and deviations, outliers
- 3) Calculate a numerical summary to describe center and spread
- 4) Sometimes the distribution is so regular that we can describe it by a smooth curve

Probability theory

(Skipping some intro because covered in the probability course)

A **random variable** X assigns to each outcome a numerical value x , formalizing the notion of a measurement

Example: Rolling two 4-sided dice
where X is the maximum roll



A **random variable** X assigns to each outcome a numerical value x , formalizing the notion of a measurement

Example: Rolling two 4-sided dice
where X is the maximum roll



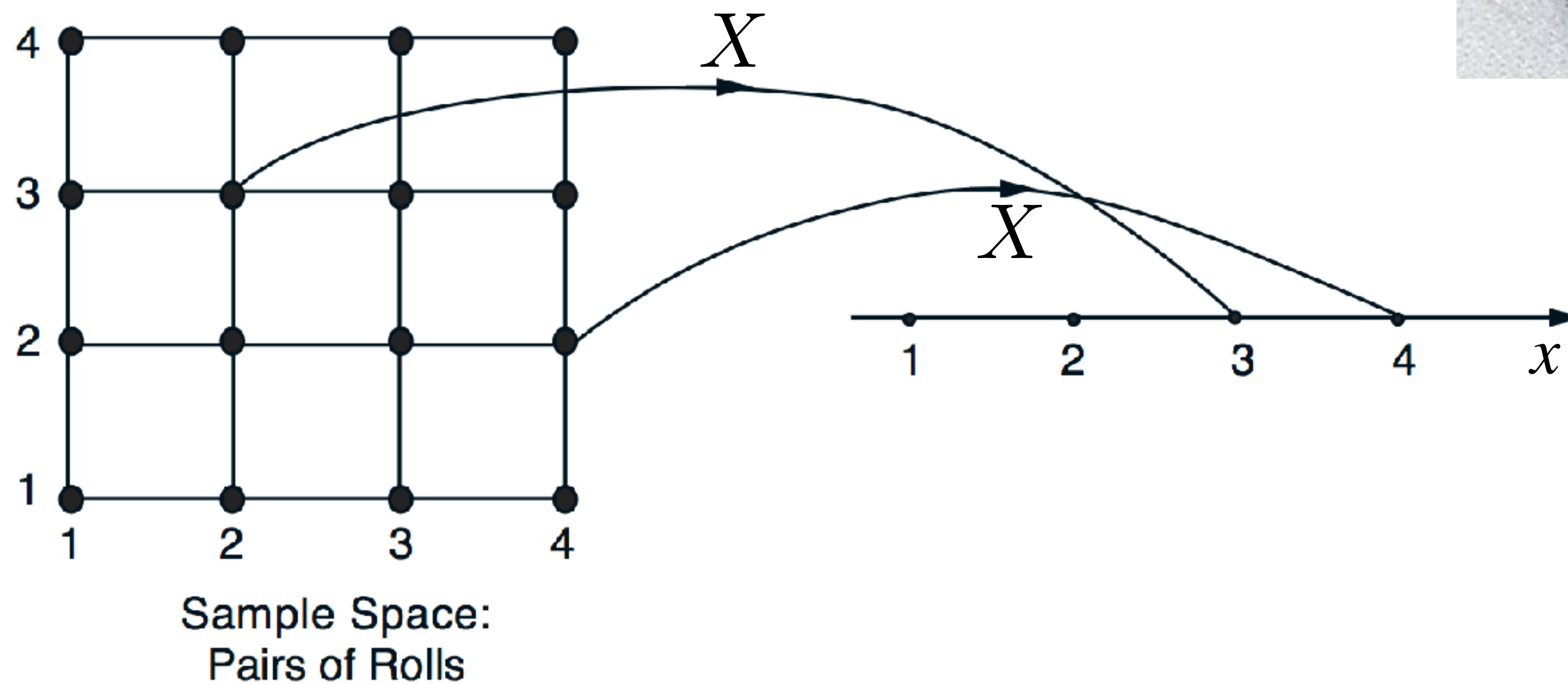
EXERCISE: in groups of 2:

What is the sample space?

What is x for each possible event?

A **random variable** X assigns to each outcome a numerical value x , formalizing the notion of a measurement

Example: Rolling two 4-sided dice
where X is the maximum roll



When X is a **discrete** random variable and the probability $p(x)$ is known for all possible x , then $p(x)$ is called the **probability mass function (PMF)**

Example: Rolling two 4-sided dice
where X is the maximum roll



$p(x)$ is short for: $P(X = x)$

When X is a **discrete** random variable and the probability $p(x)$ is known for all possible x , then $p(x)$ is called the **probability mass function (PMF)**

Example: Rolling two 4-sided dice
where X is the maximum roll

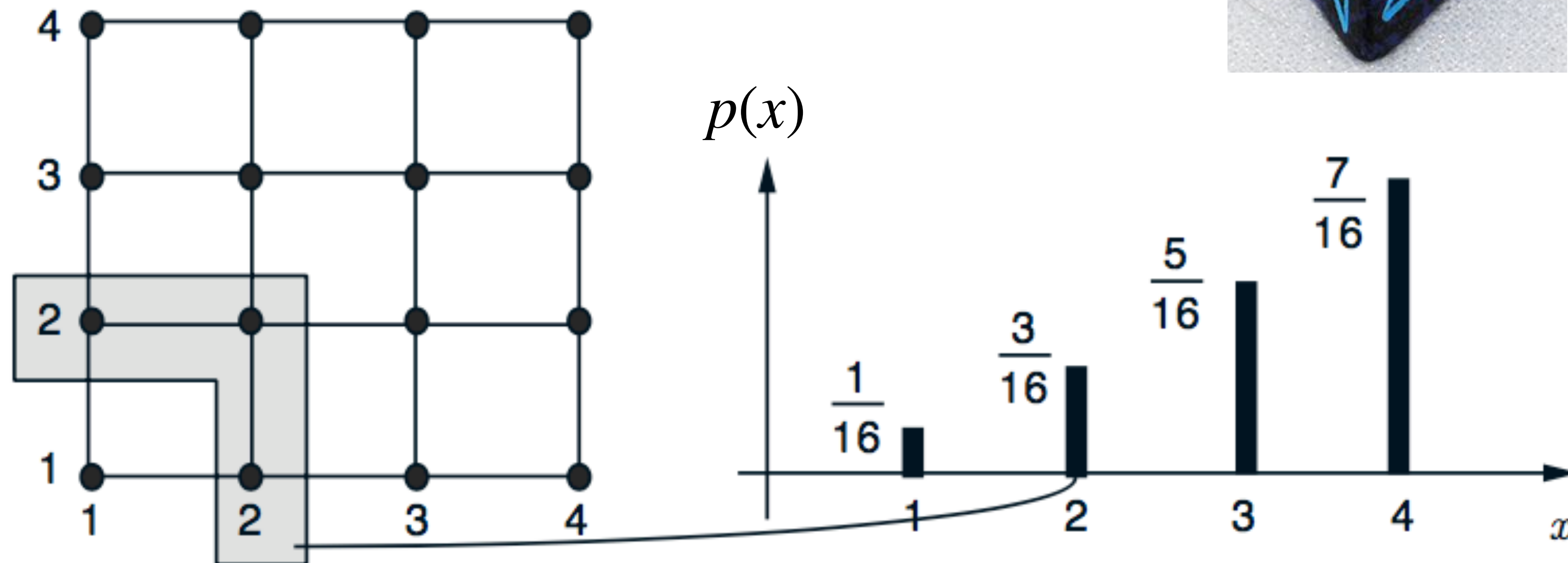


EXERCISE:

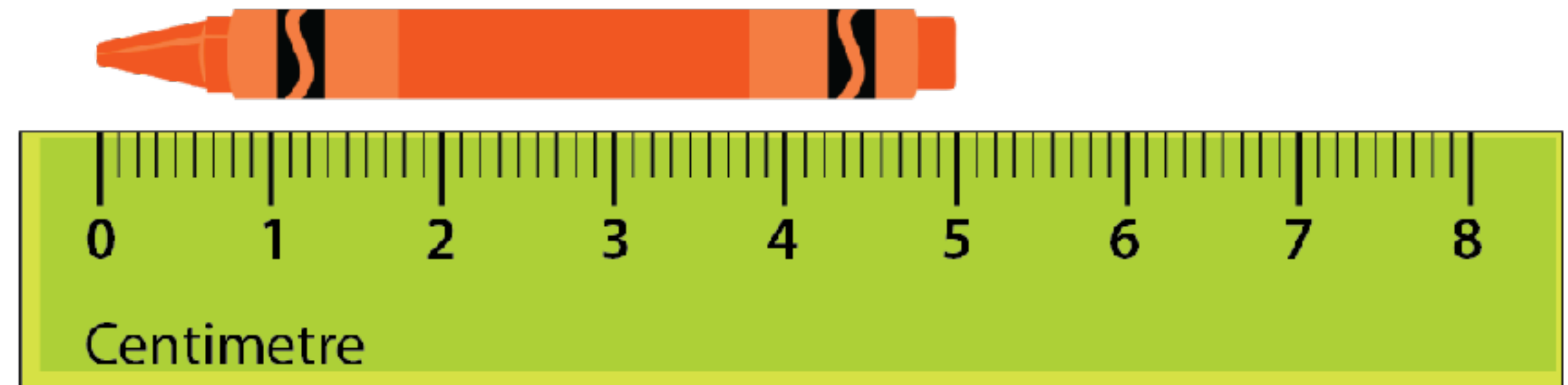
What is $p(x)$ here?

When X is a **discrete** random variable and the probability $p(x)$ is known for all possible x , then $p(x)$ is called the **probability mass function (PMF)**

Example: Rolling two 4-sided dice
where X is the maximum roll

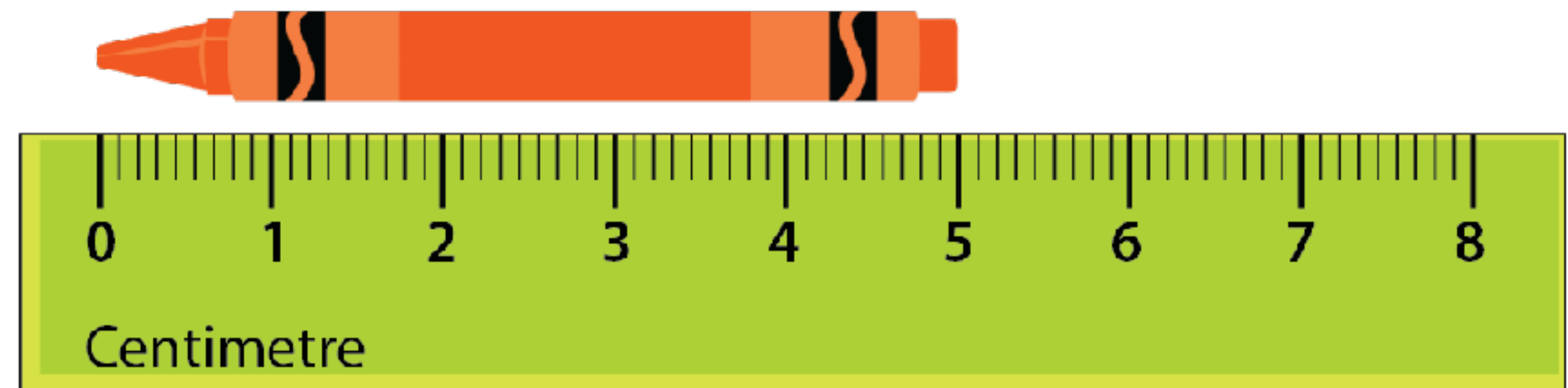


When X can take any value along an interval, it is **continuous**



Here the probability of measuring a specific value is effectively 0.

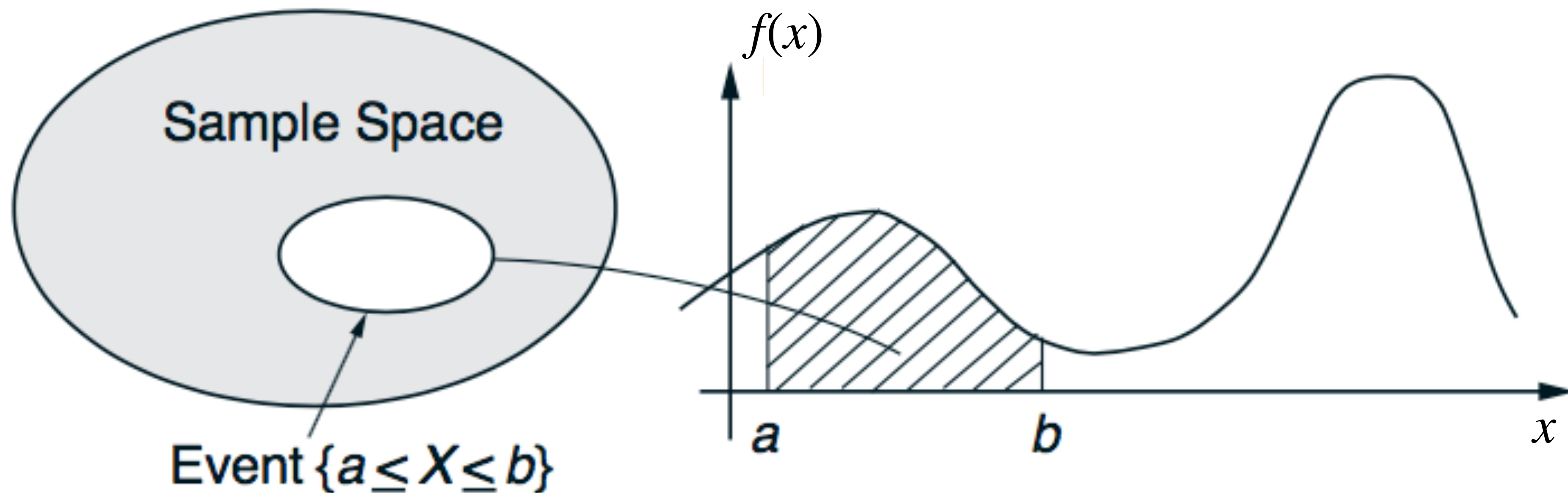
When X can take any value along an interval, it is **continuous**



Here the probability of measuring a specific value is effectively 0.
So we use intervals. We ask: What is $P(a \leq X \leq b)$?

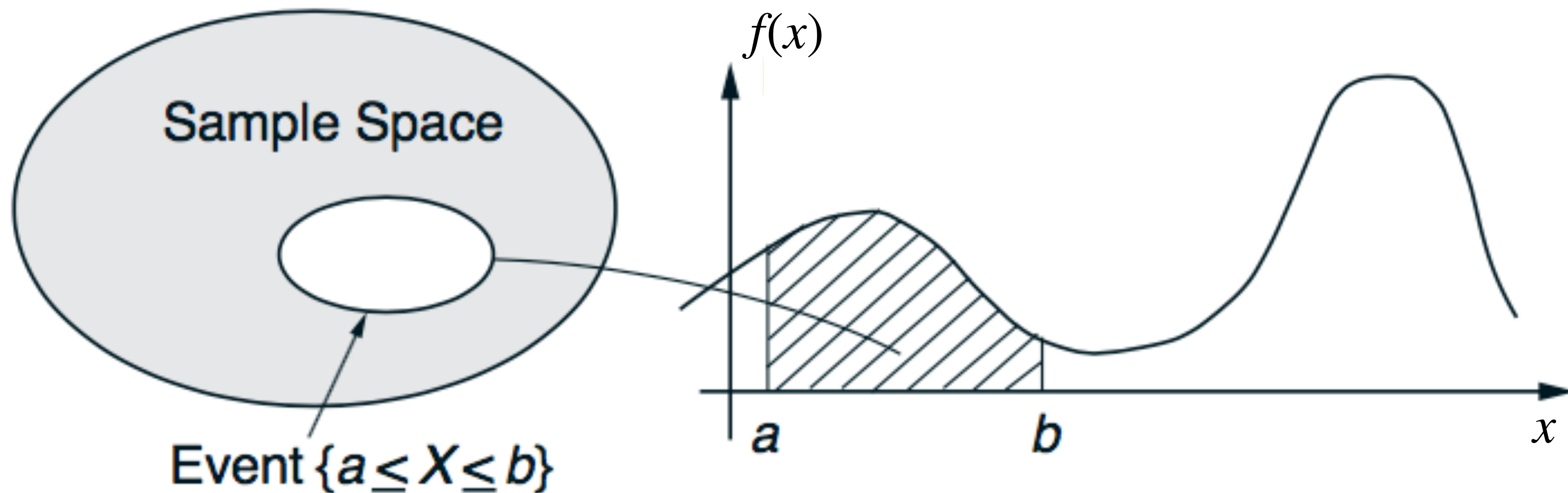
When X is a **continuous** random variable, we get the probability of X falling into an interval by integrating the **probability density function (PDF) $f(x)$**

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

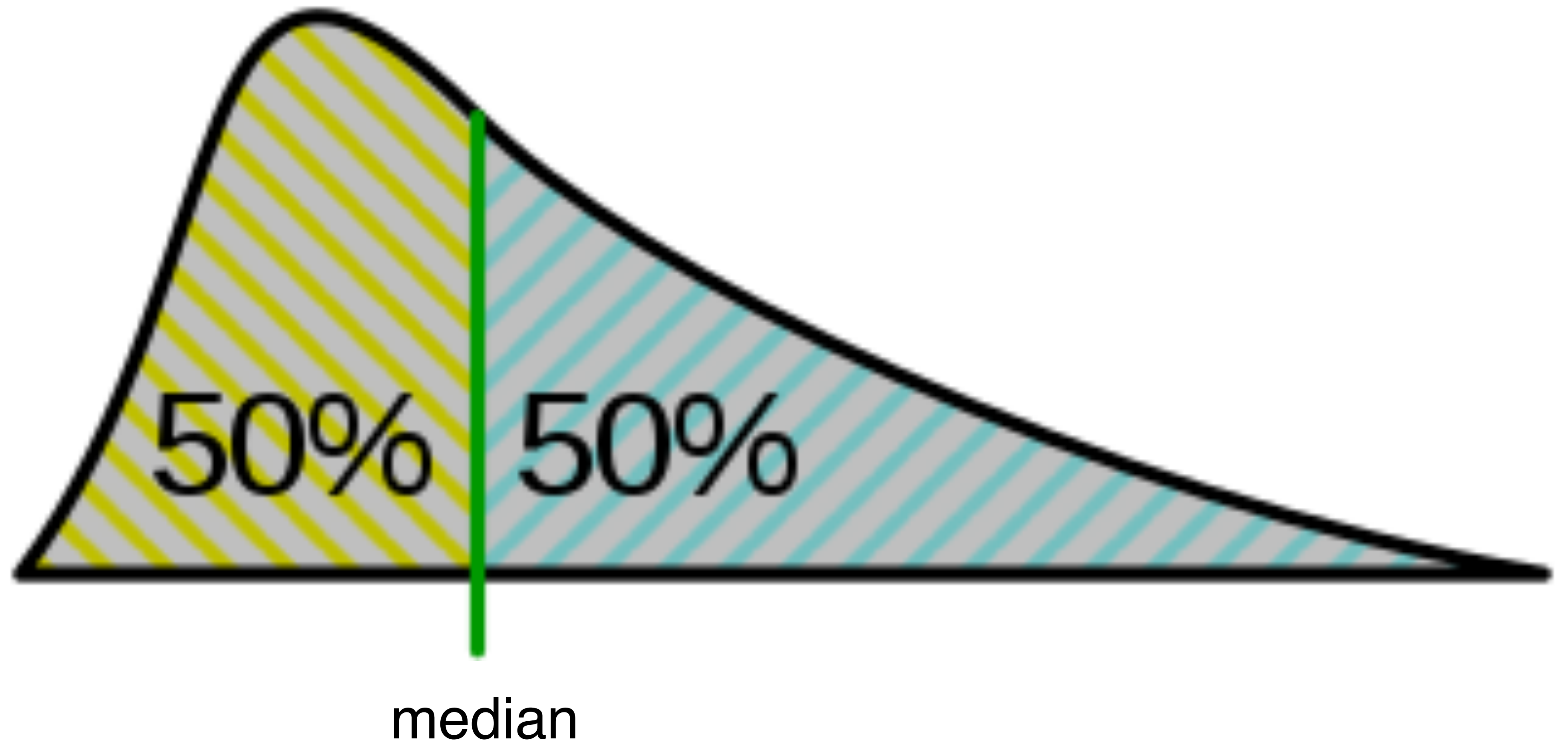


A density curve must be everywhere non-negative
and the entire area below it must be 1

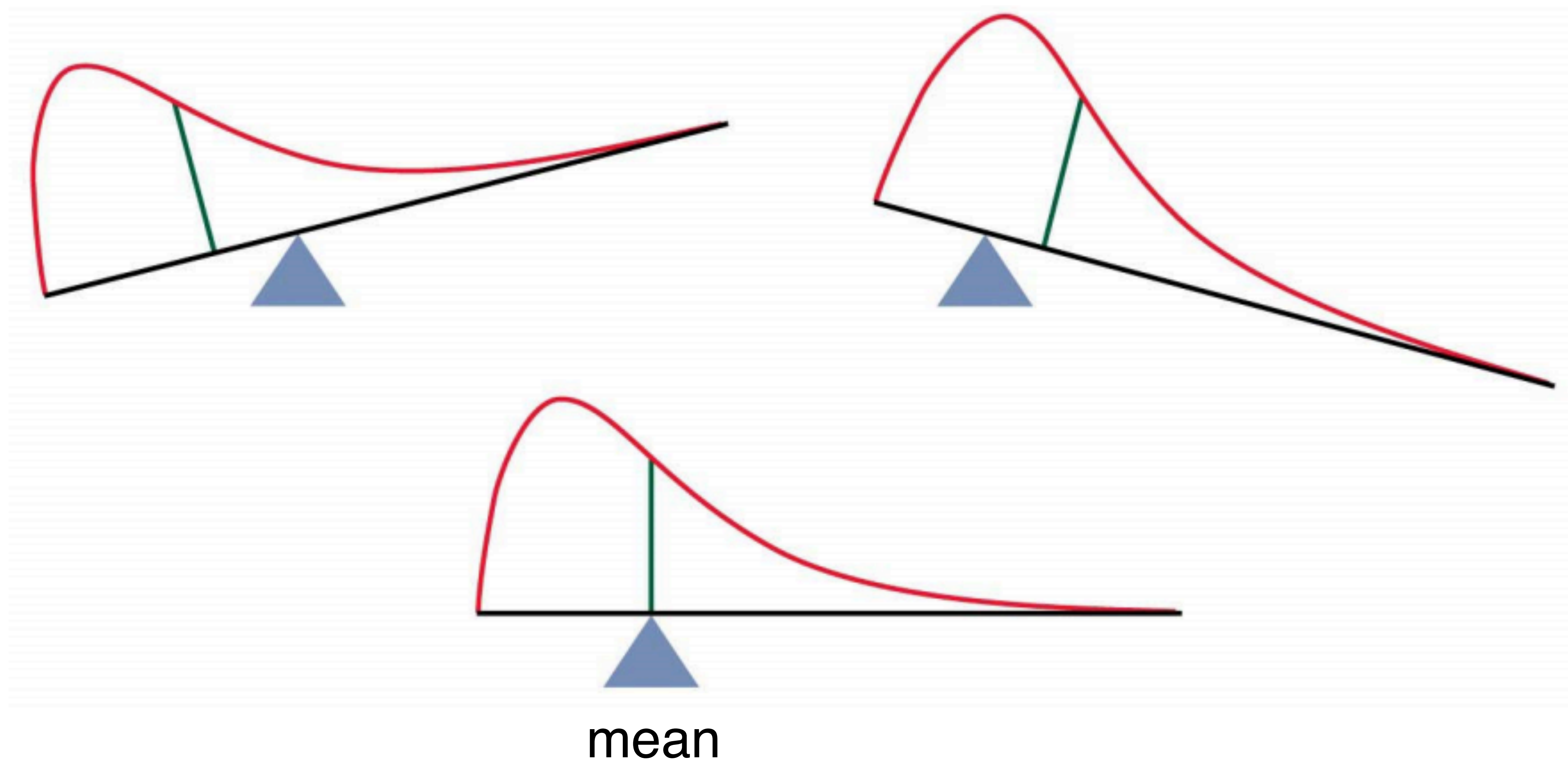
$$\int_{-\infty}^{\infty} f(x) dx = 1$$



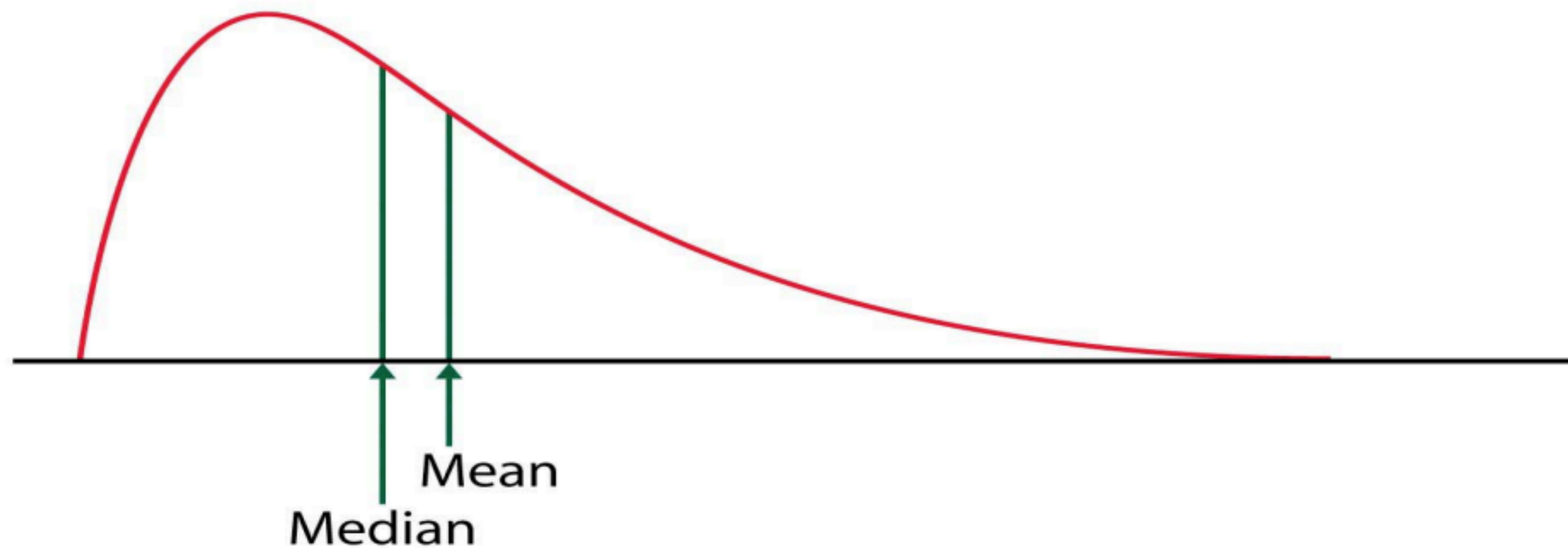
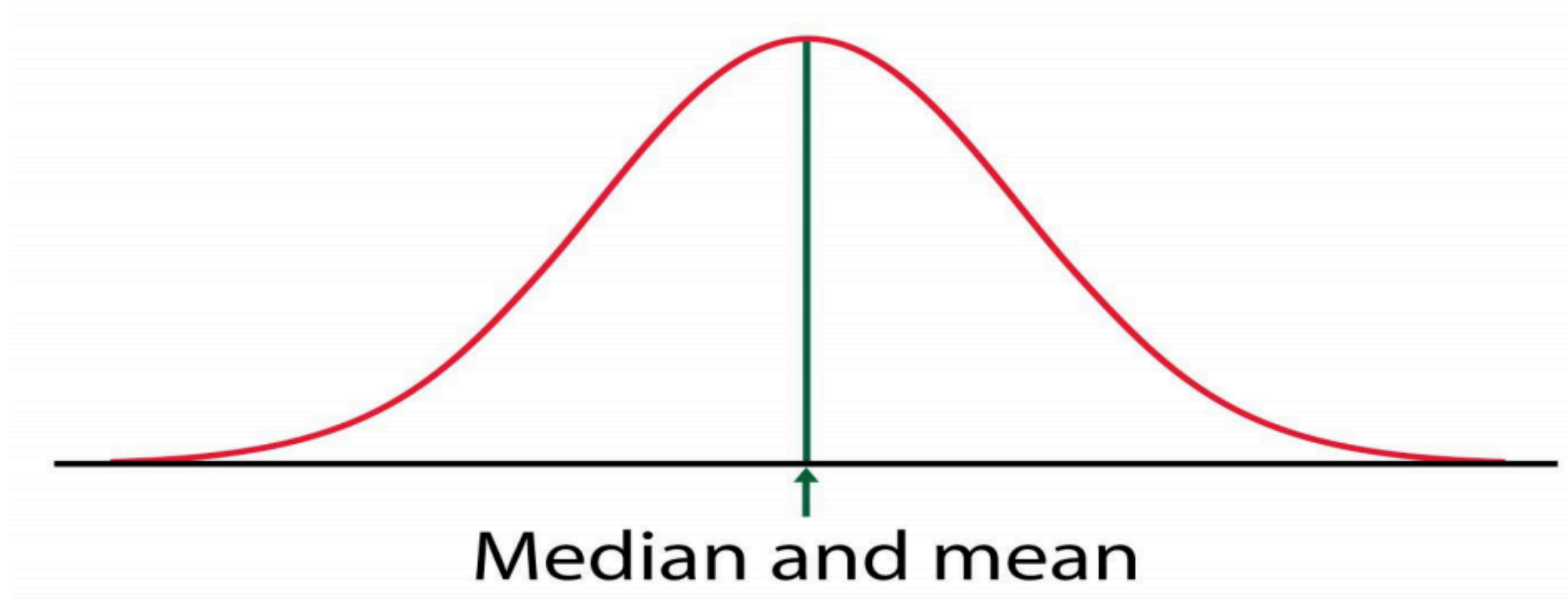
The median of a density curve cuts the area in half



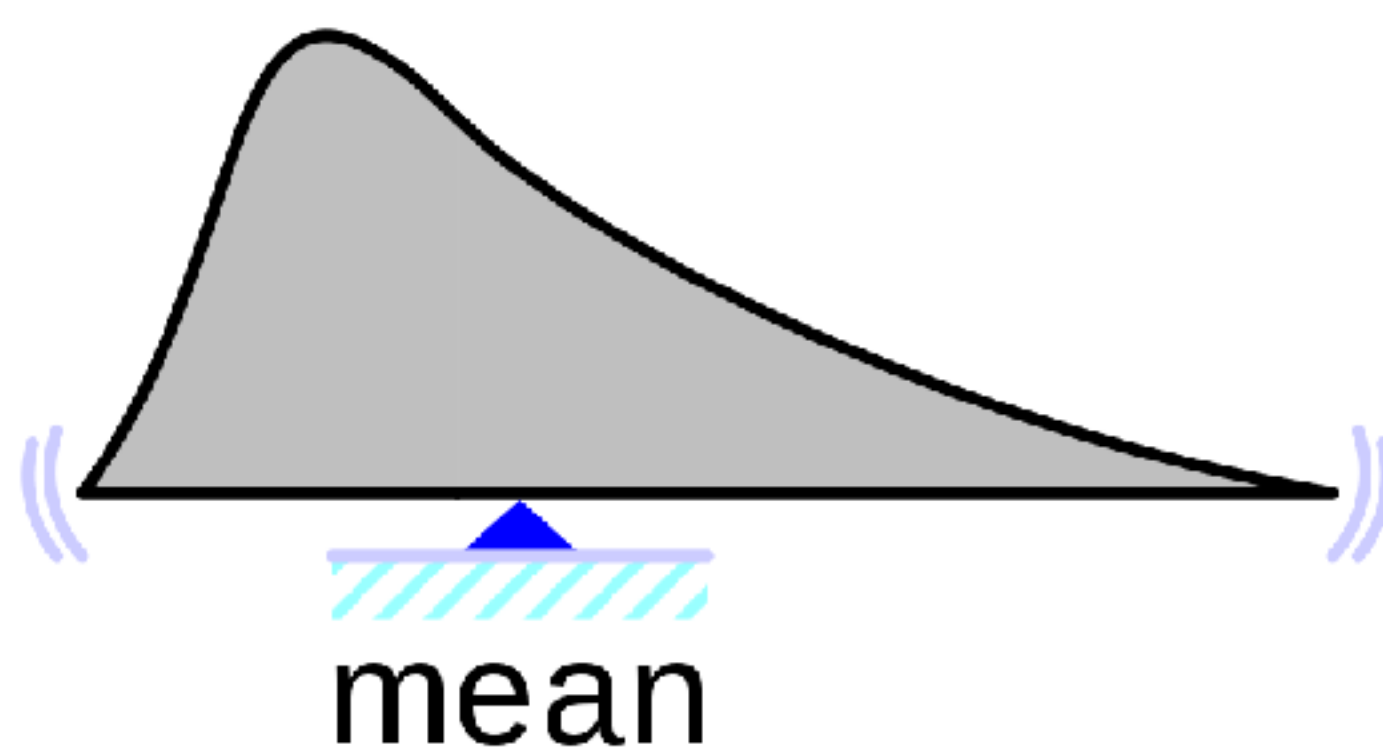
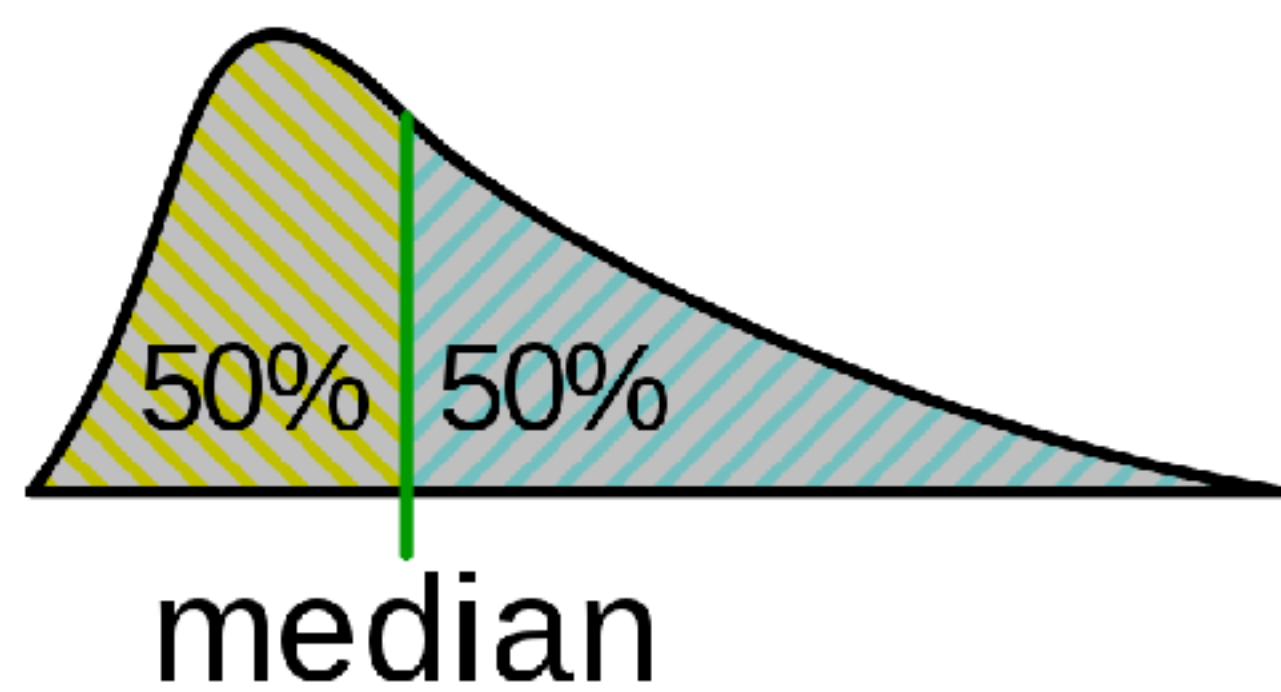
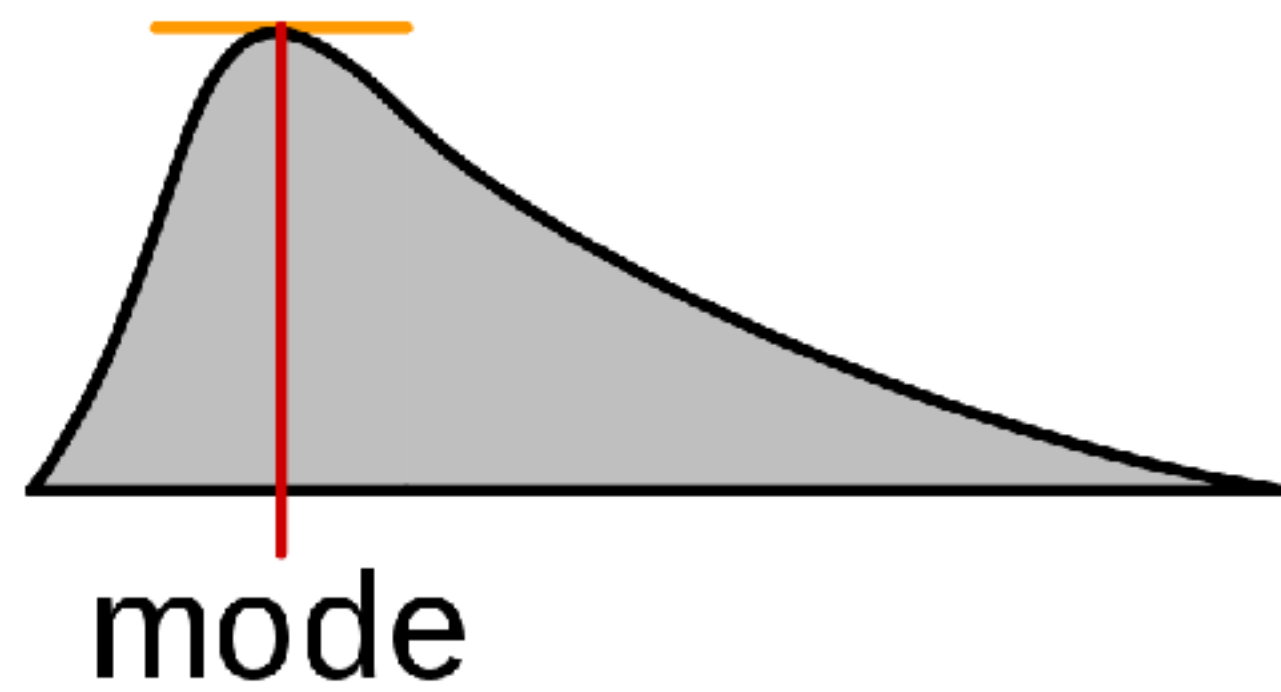
The mean of a density curve is the balance point



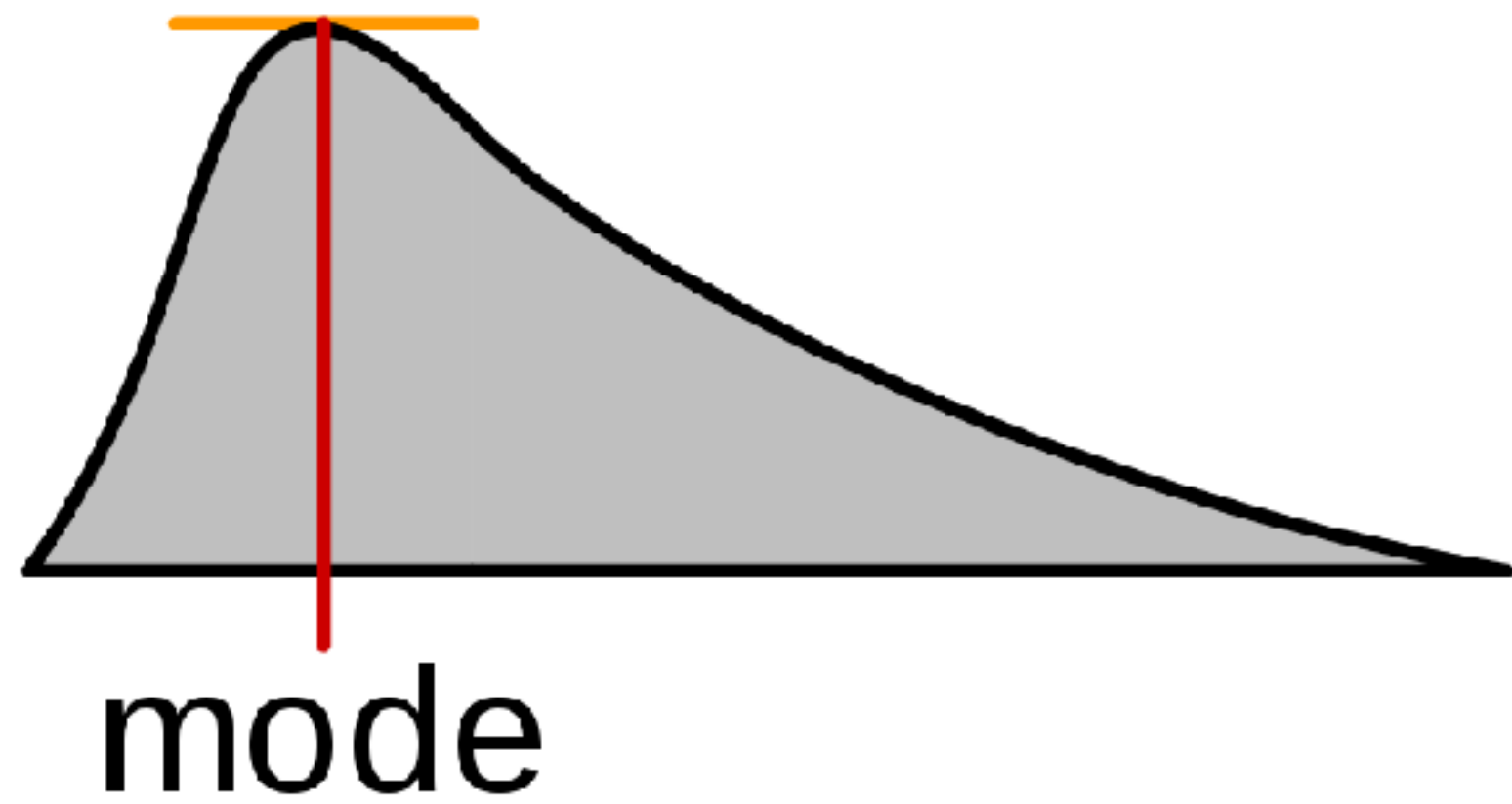
Often mean and median are not the same



The **mode** is the value that appears most often



A **unimodal** distribution has one "hump"
A **multimodal** distribution has multiple "humps"



unimodal



bimodal



trimodal

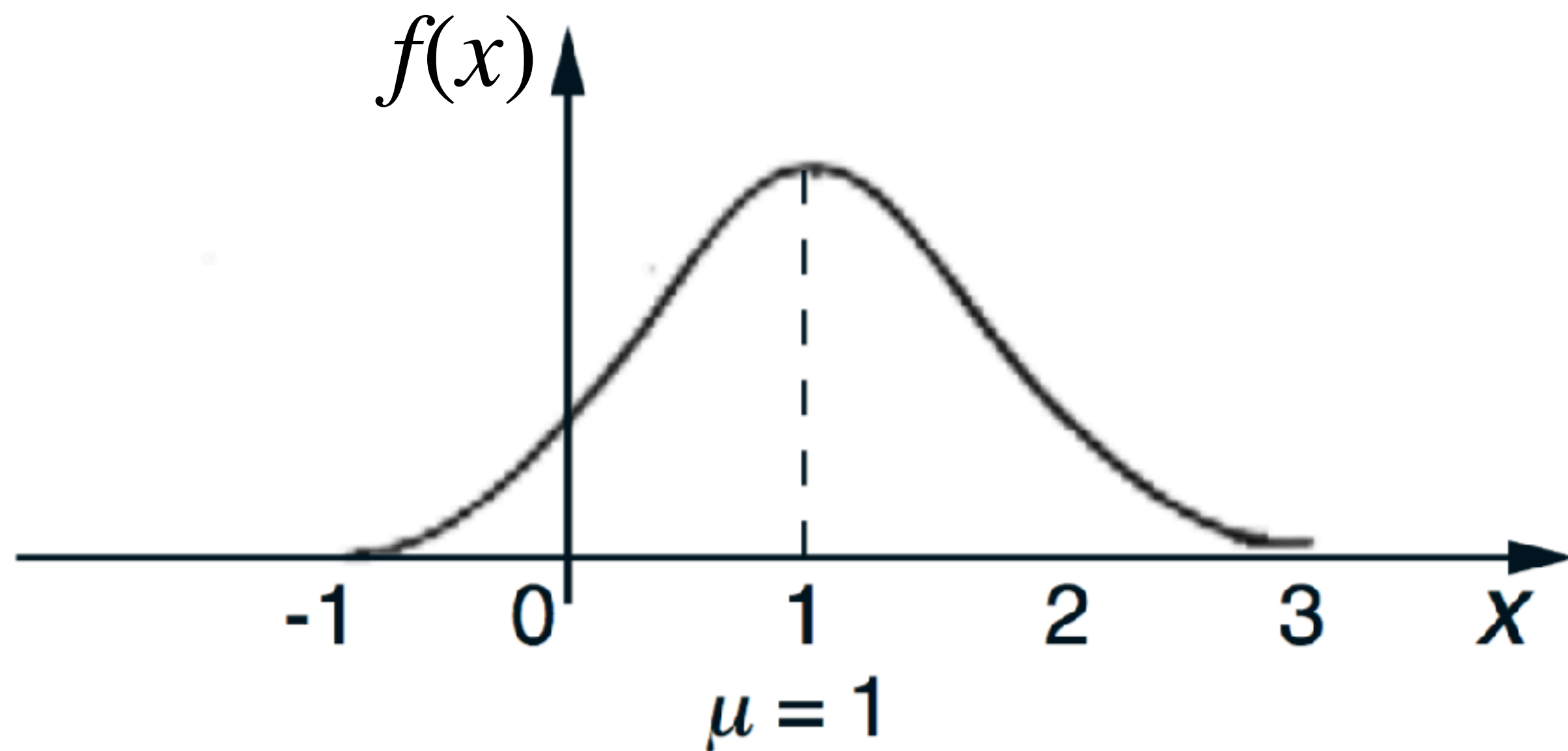
Jupyter

A **normal distribution** with parameters μ and σ^2 is a continuous random variable X with the PDF:

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

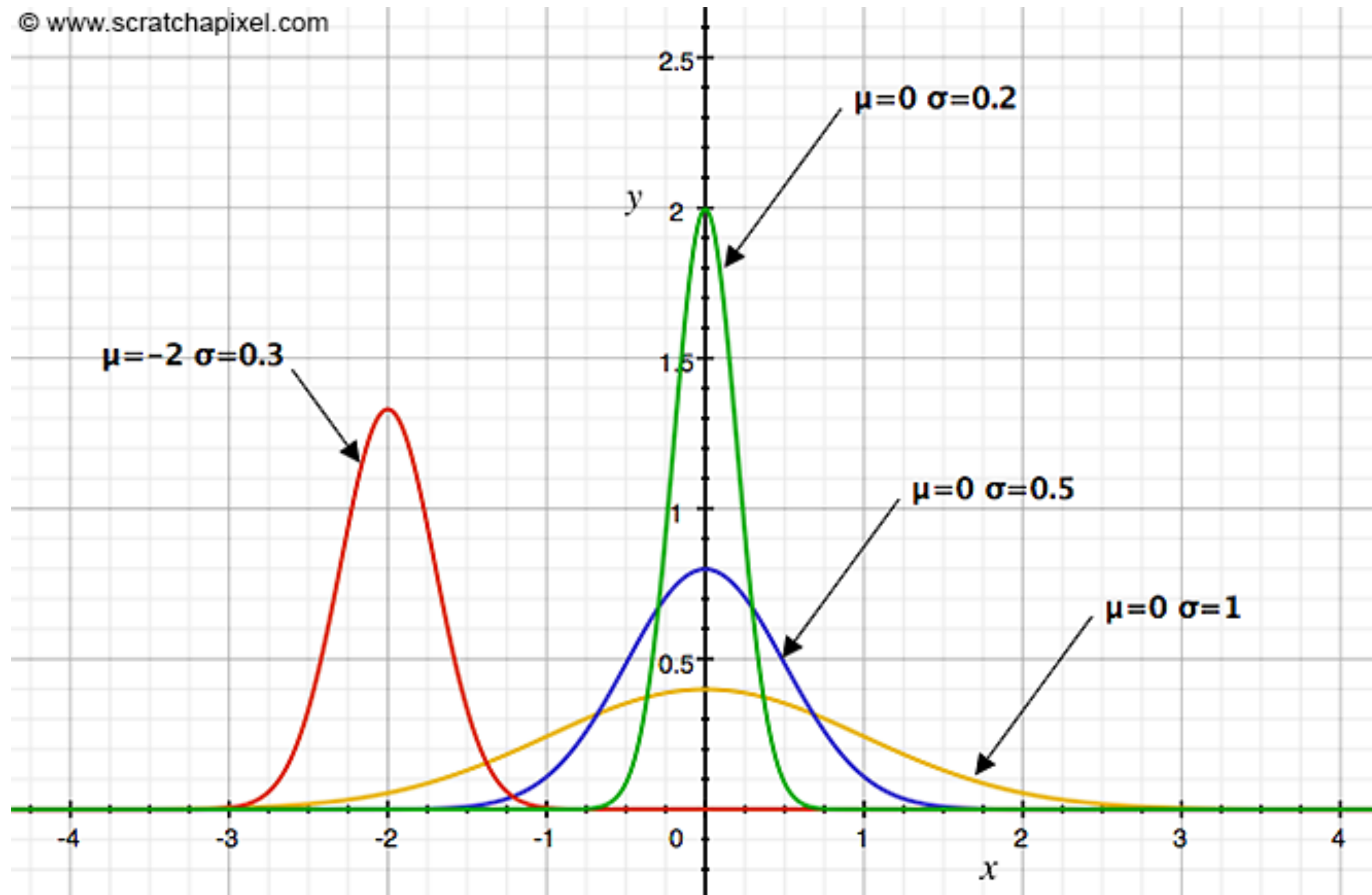
A **normal distribution** with parameters μ and σ^2 is a continuous random variable X with the PDF:

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



$$X \sim \mathcal{N}(\mu, \sigma^2)$$

The two parameters completely determine the curve



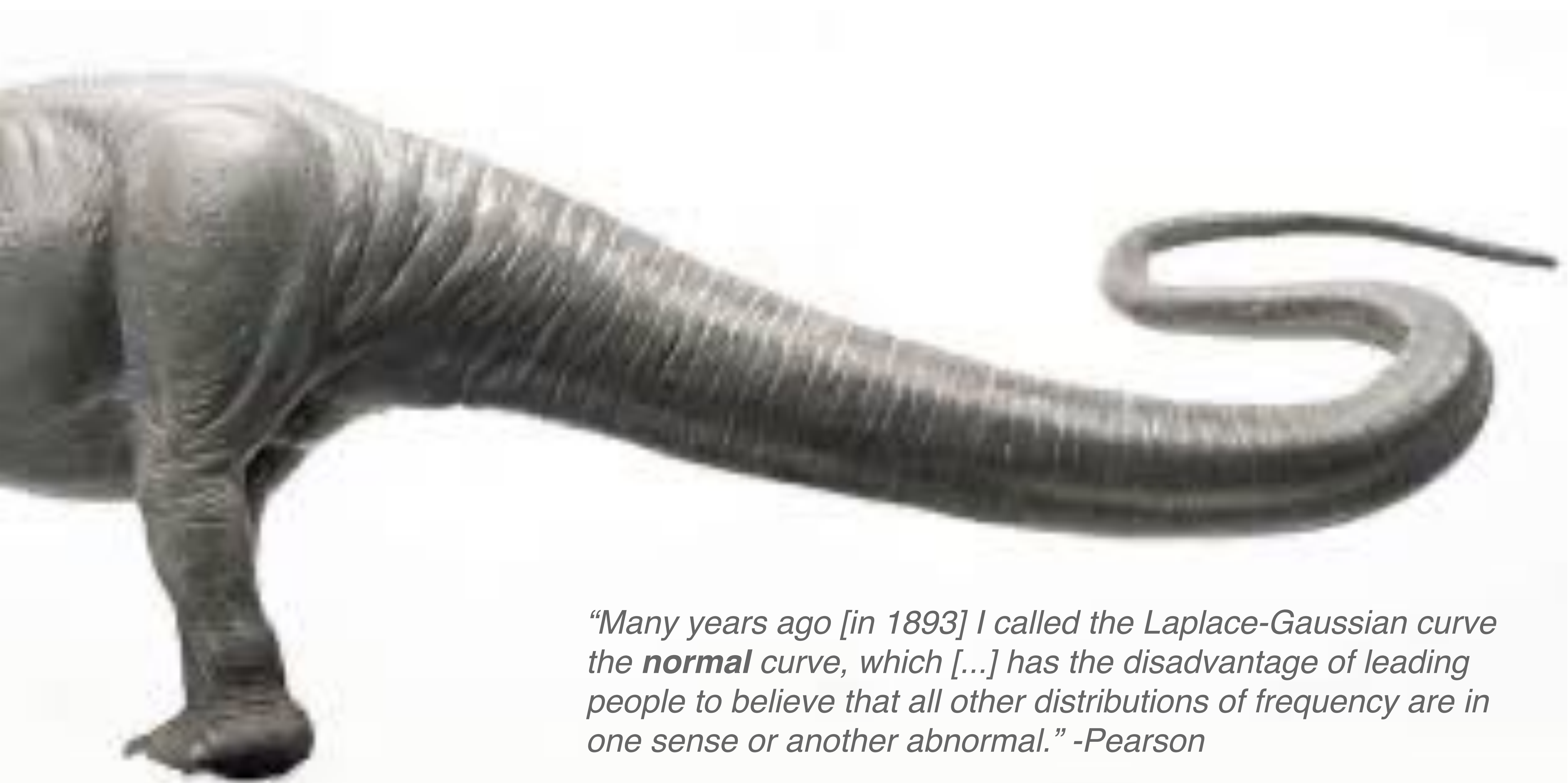
A **parameter** is numerical characteristic of a statistical model

Normal distributions are important in statistics

- 1) They describe well some real data sets
- 2) They approximate well many chance outcomes
- 3) They are useful for statistical inference of many symmetric distributions

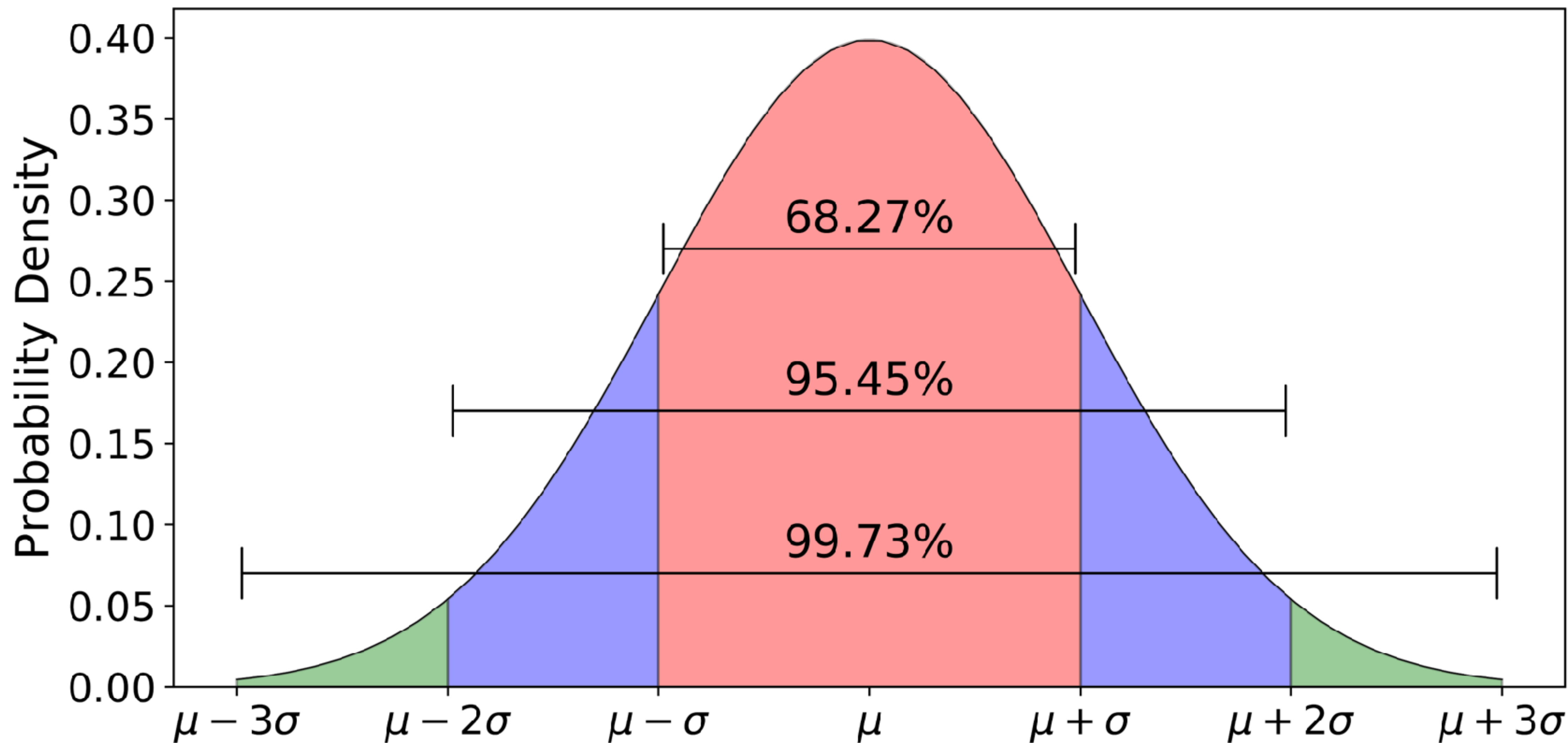


Real data are often VERY MUCH NOT normal



*“Many years ago [in 1893] I called the Laplace-Gaussian curve the **normal** curve, which [...] has the disadvantage of leading people to believe that all other distributions of frequency are in one sense or another abnormal.” -Pearson*

68-95-99.7 Rule



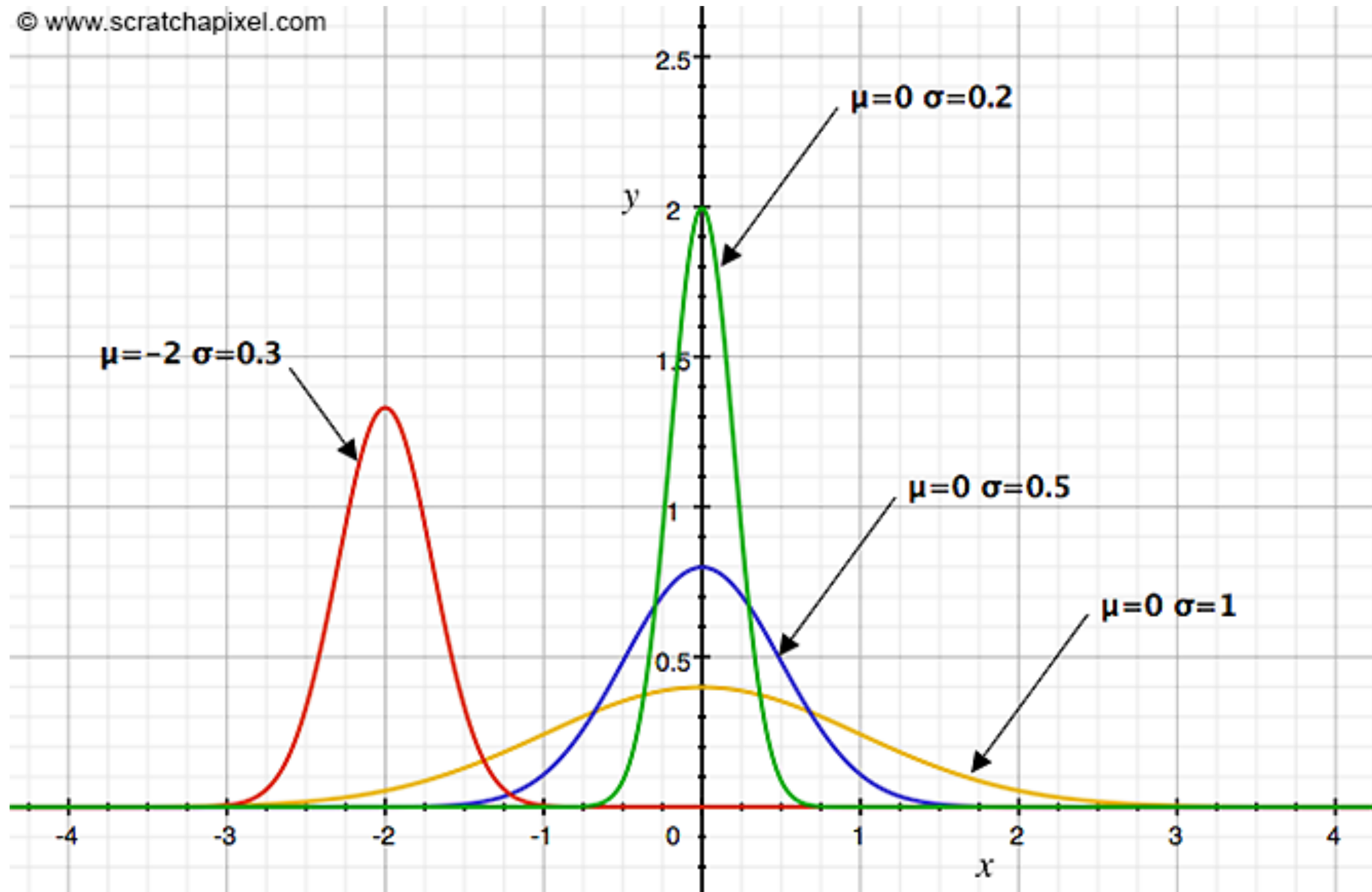
Linear transformation of a normal distribution is a normal distribution

$$x_{\text{new}} = a + bx$$

It only:

- adds a to the center
- multiplies center and spread by b .

All normal distributions are the same if we measure in units of size σ around the mean μ as center



The **standard normal distribution** $\mathcal{N}(0,1)$ has $\mu = 0$ and $\sigma = 1$

We can turn any normally distributed variable $X \sim \mathcal{N}(\mu, \sigma)$ into $\mathcal{N}(0,1)$ by standardizing it:

$$Z = \frac{X - \mu}{\sigma}$$

The **standard normal distribution** $\mathcal{N}(0,1)$ has $\mu = 0$ and $\sigma = 1$



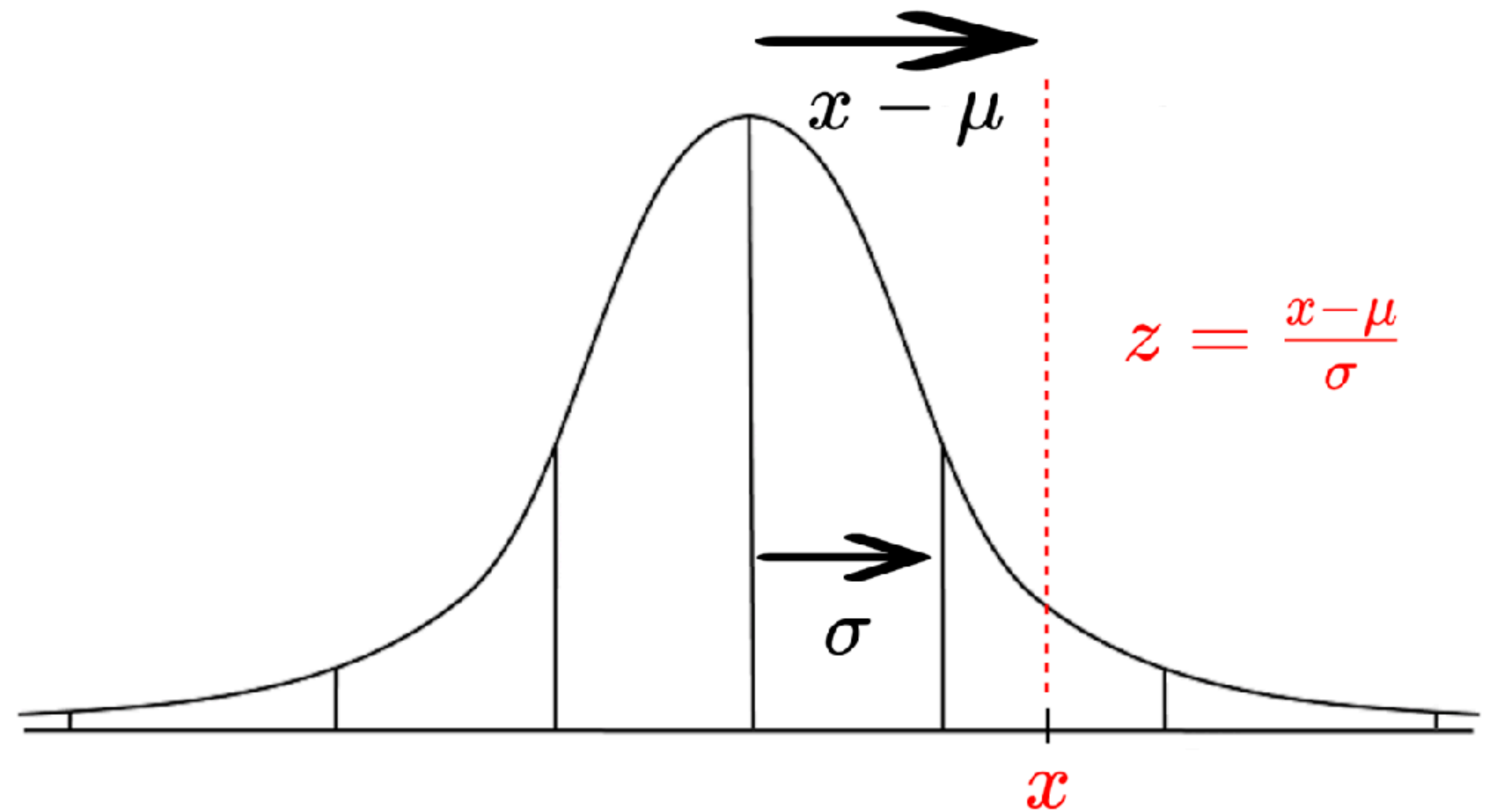
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



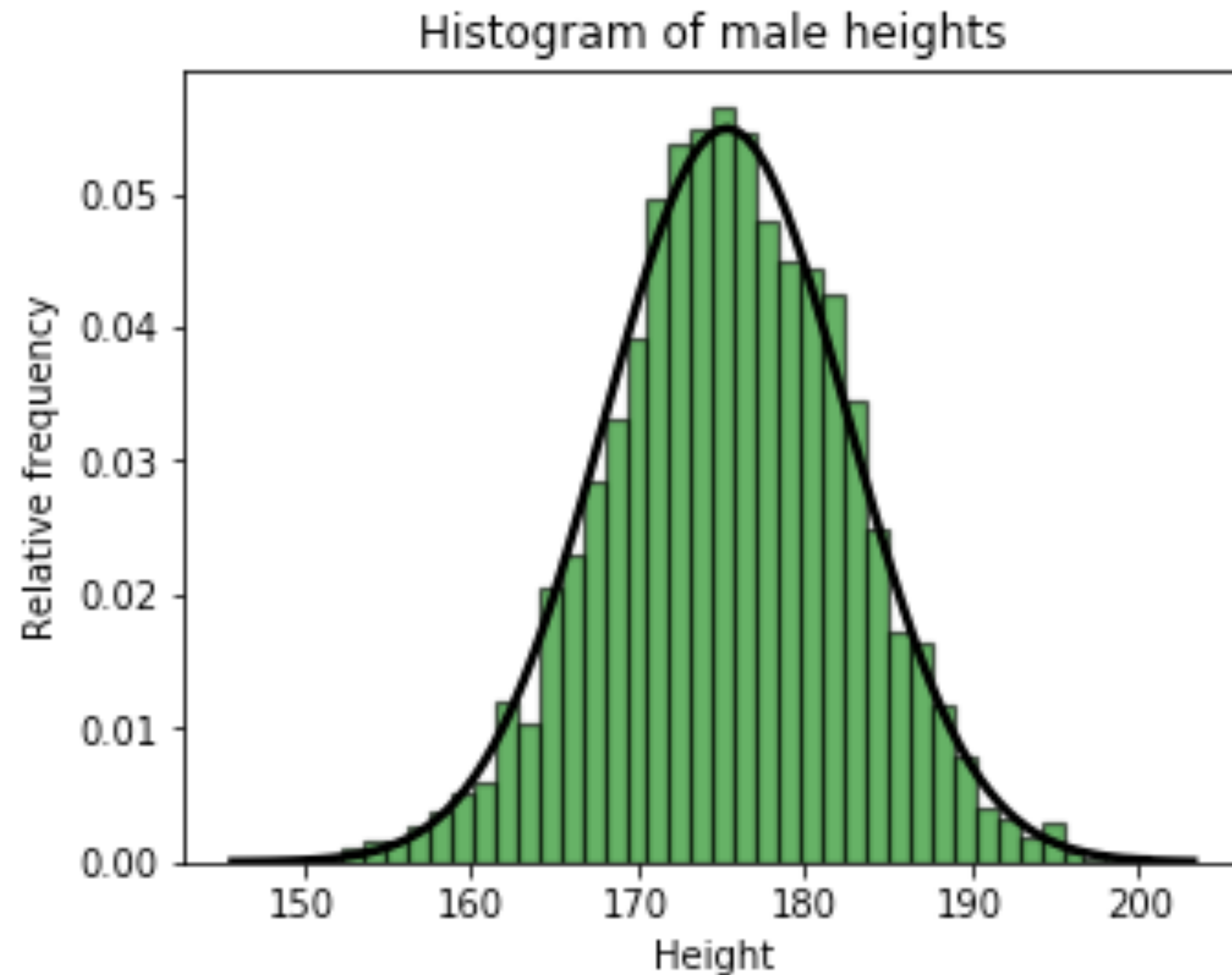
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

The **z-score (standard score)** is the number of standard deviations that an observed value x is away from the mean of a reference distribution

$$z = \frac{x - \mu}{\sigma}$$

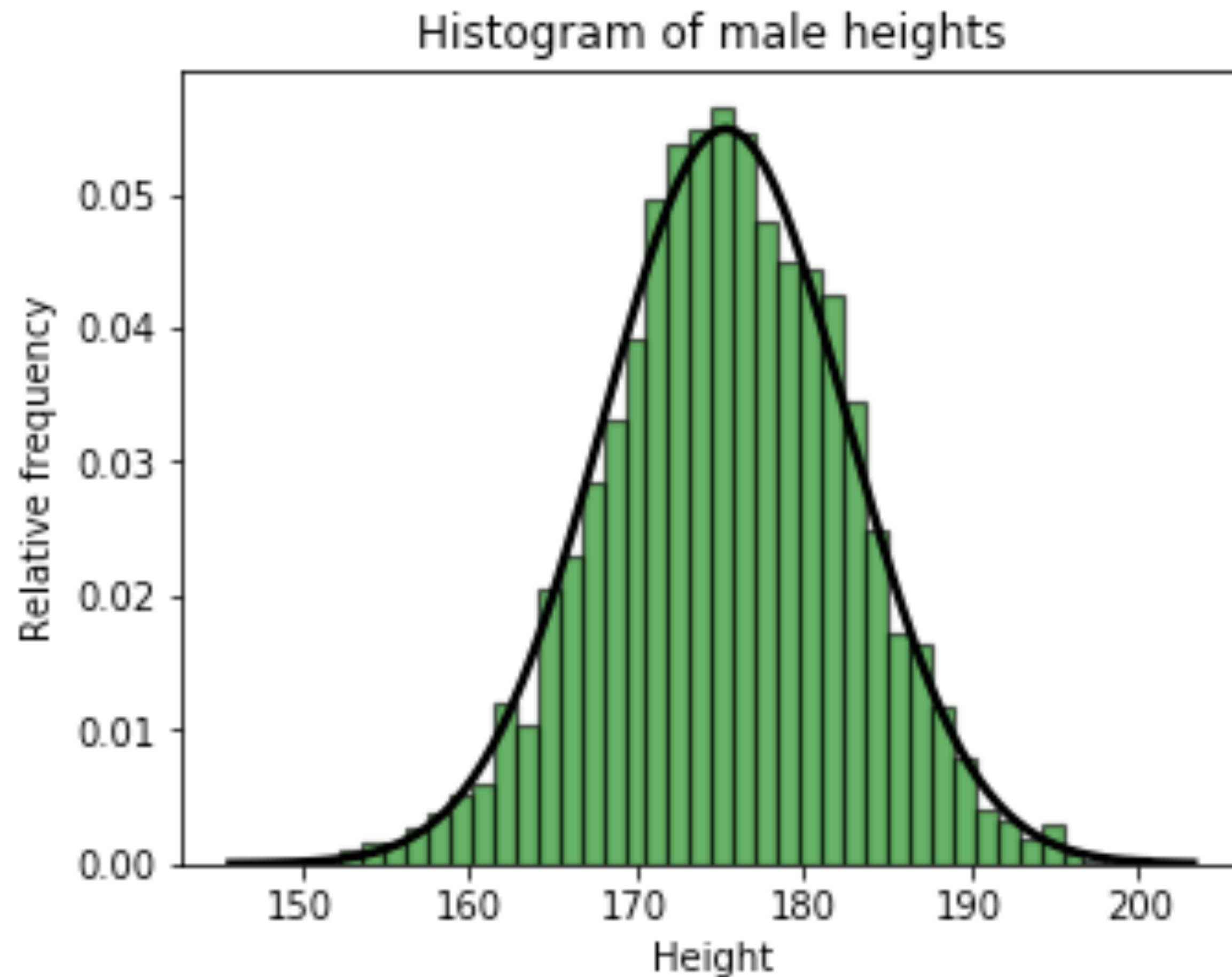


Now we know what normal distributions are, but how to reliably spot one?

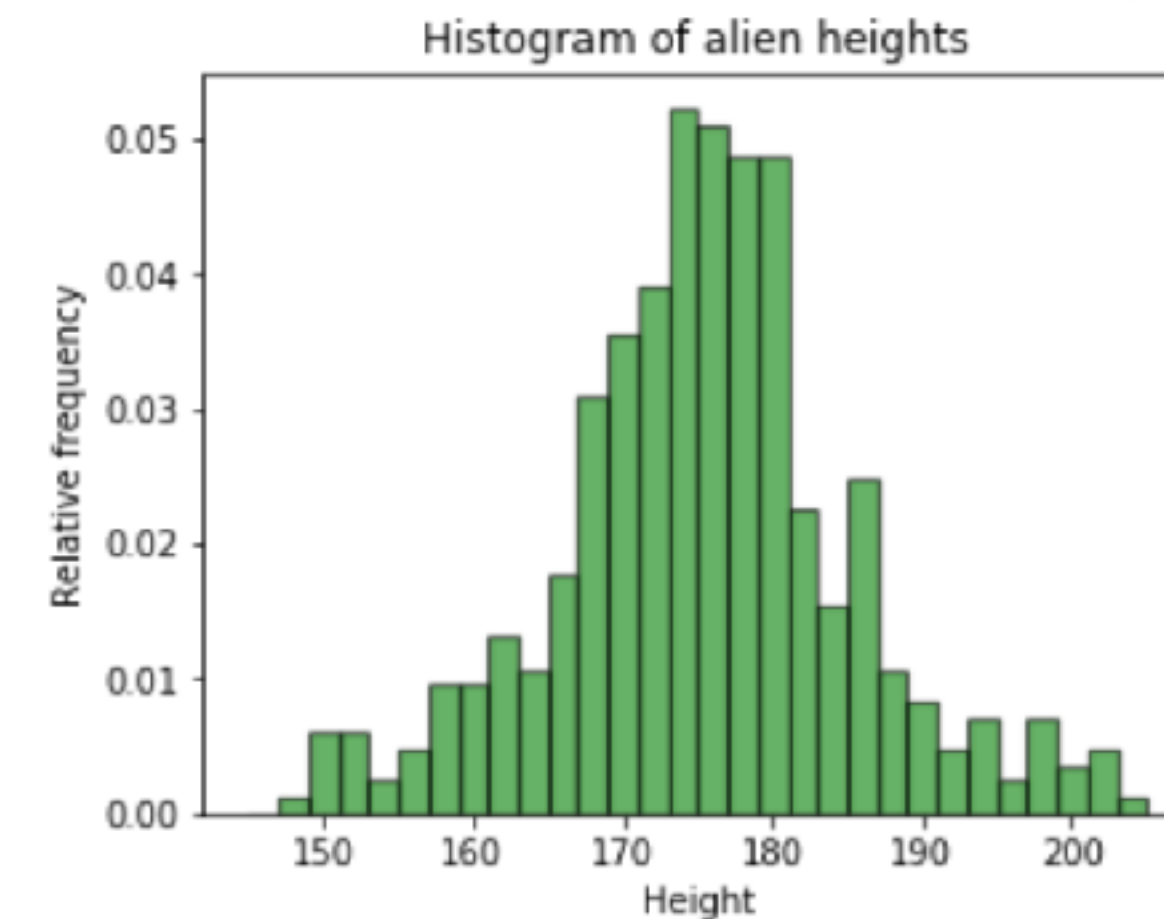


How can we be sure that the normal distribution is a good model for the data?

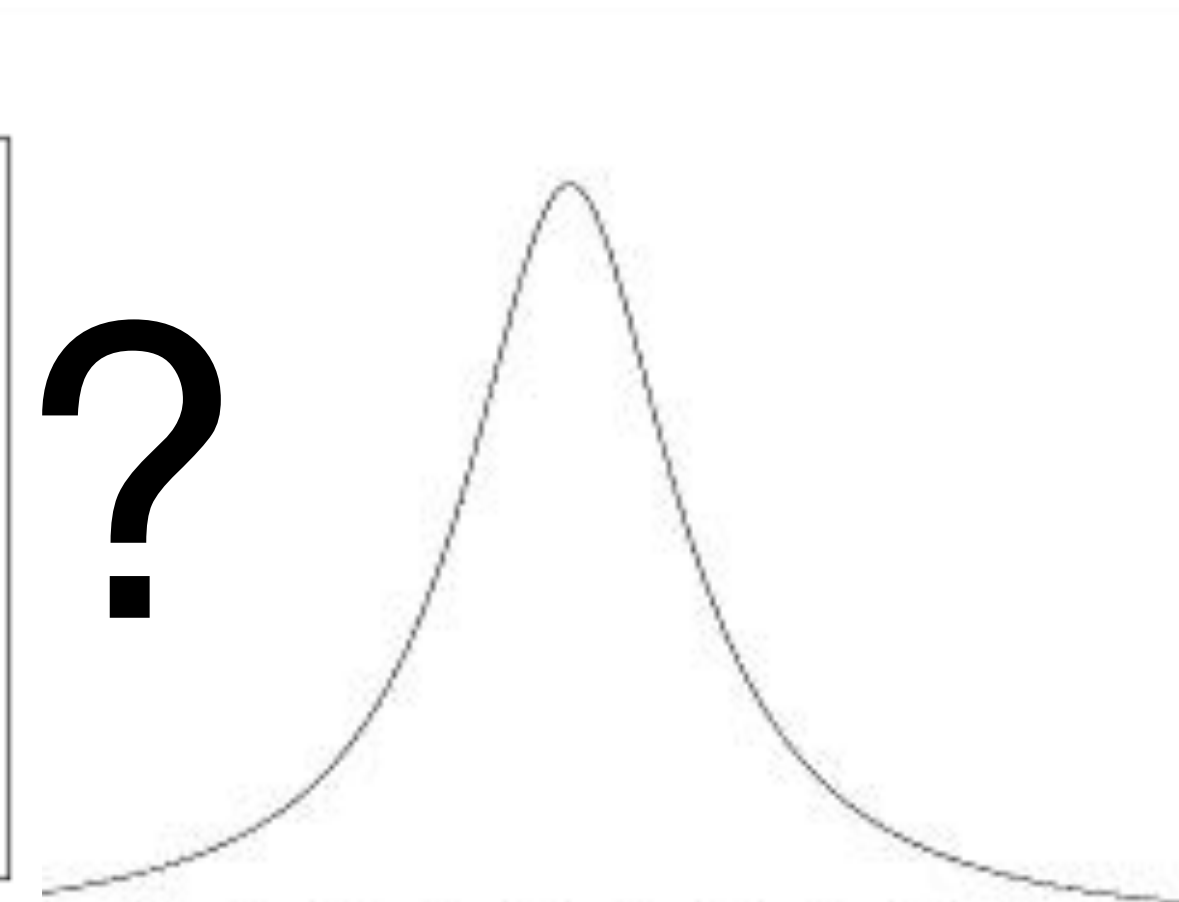
Now we know what normal distributions are, but how to reliably spot one?



Unimodal
Symmetric

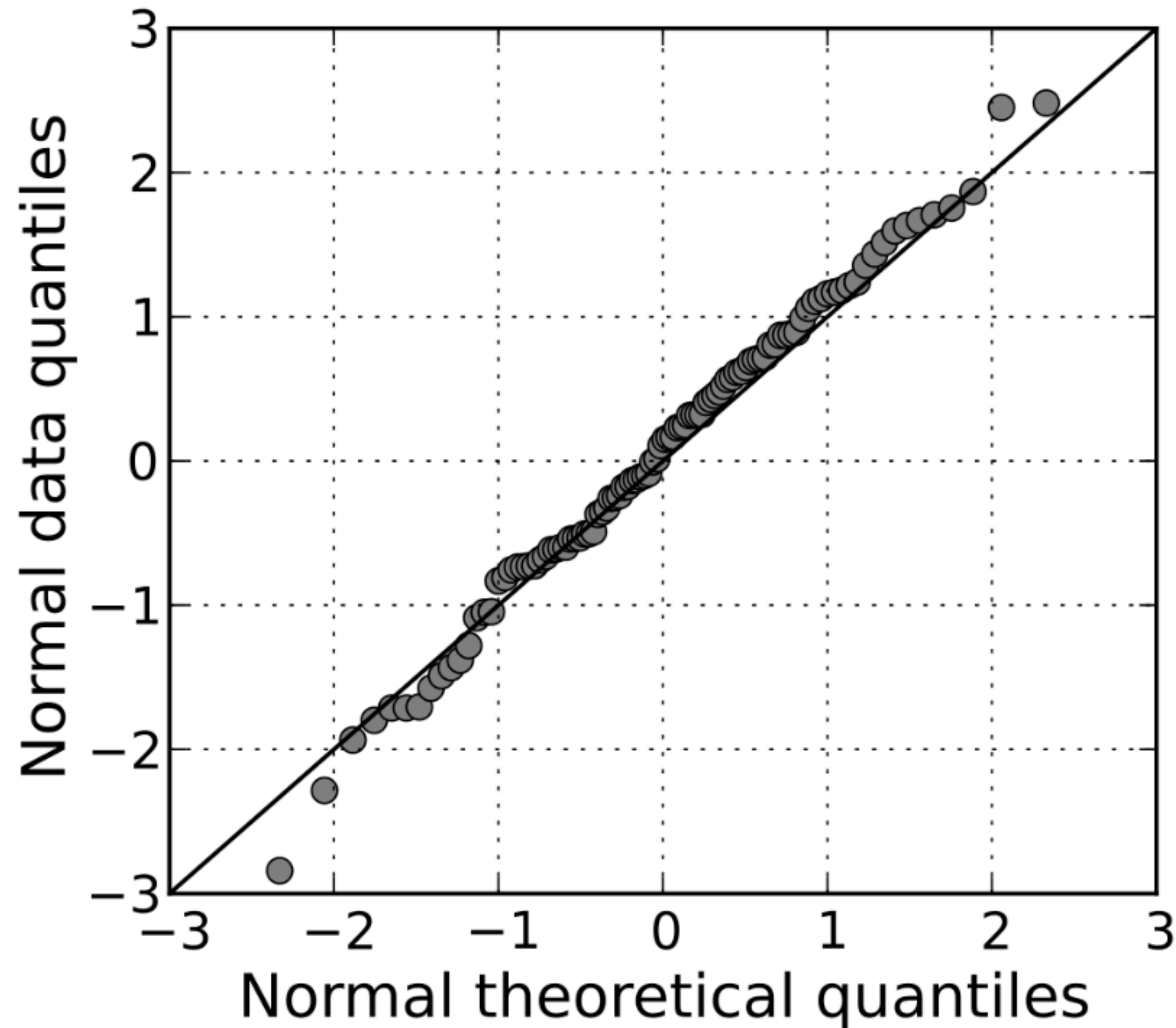


?



How can we be sure that the normal distribution is a good model for the data?

The Q-Q plot (quantile-quantile) assesses normality visually

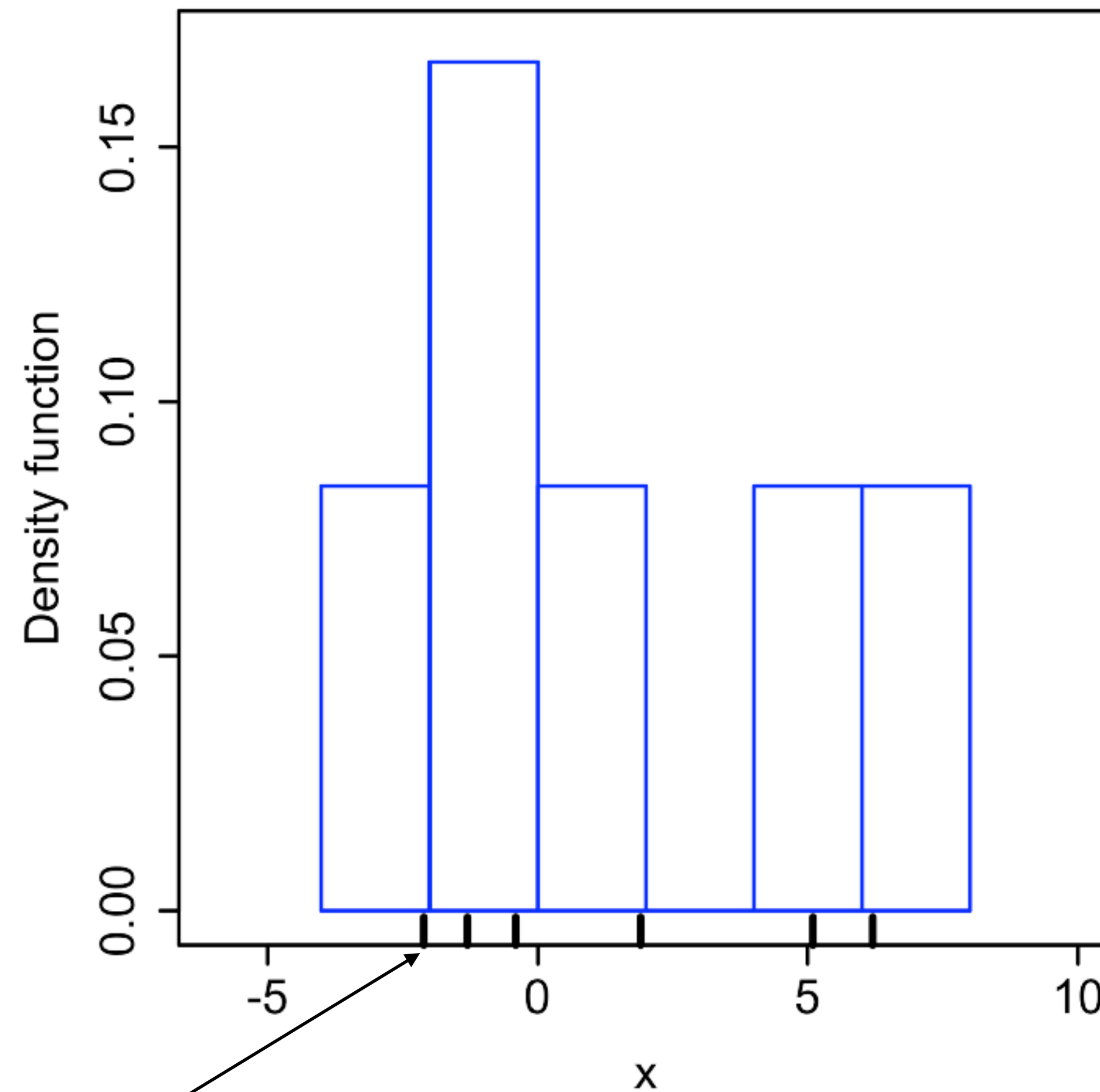


- 1) Order data points and calculate their quantiles
- 2) Calculate z-scores of theoretical normal distributions at same quantiles
- 3) Compare the two. If on the diagonal, we have a normal distribution

Jupyter

Kernel density estimation (KDE)

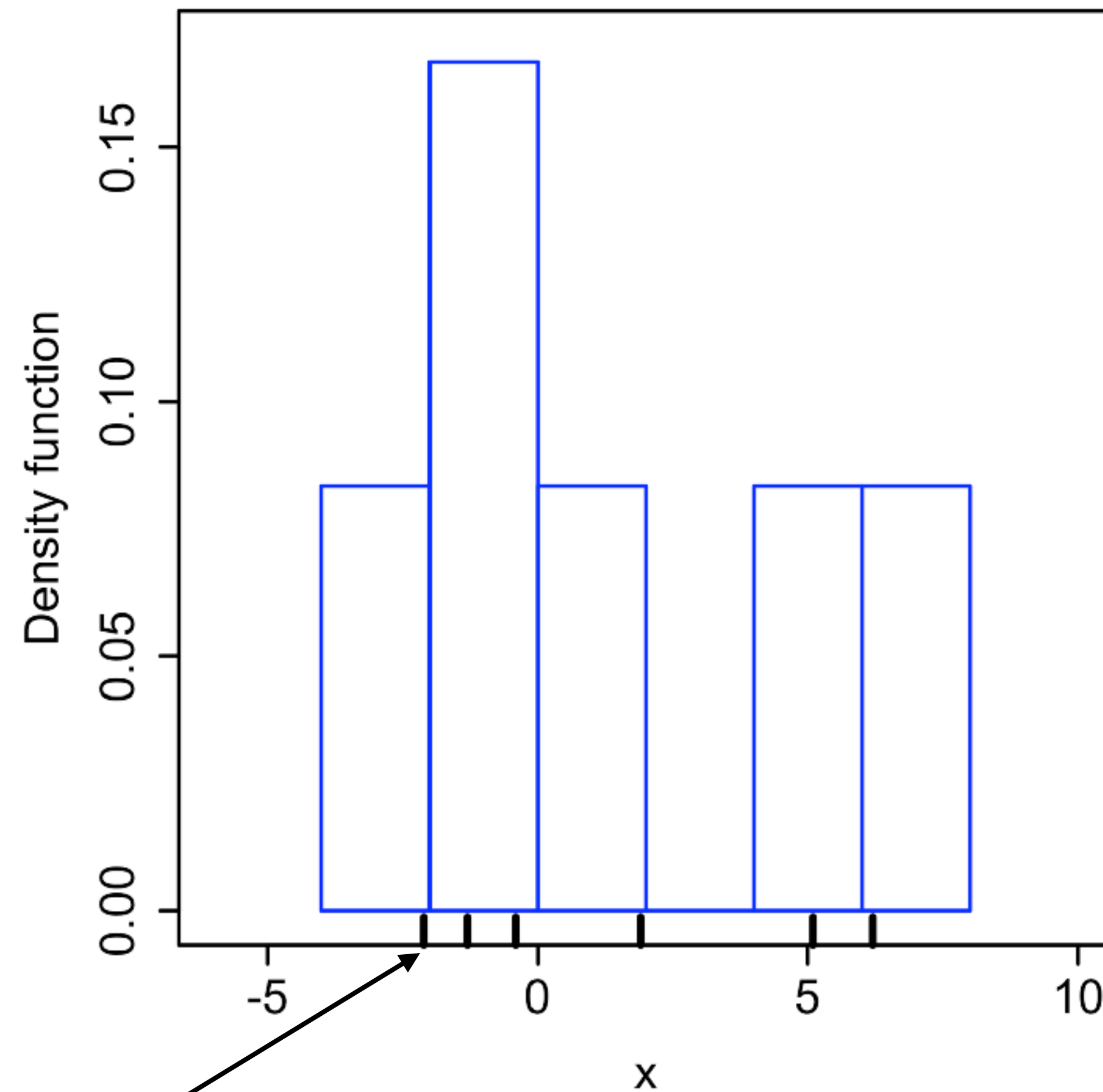
Histogram



Rug plot

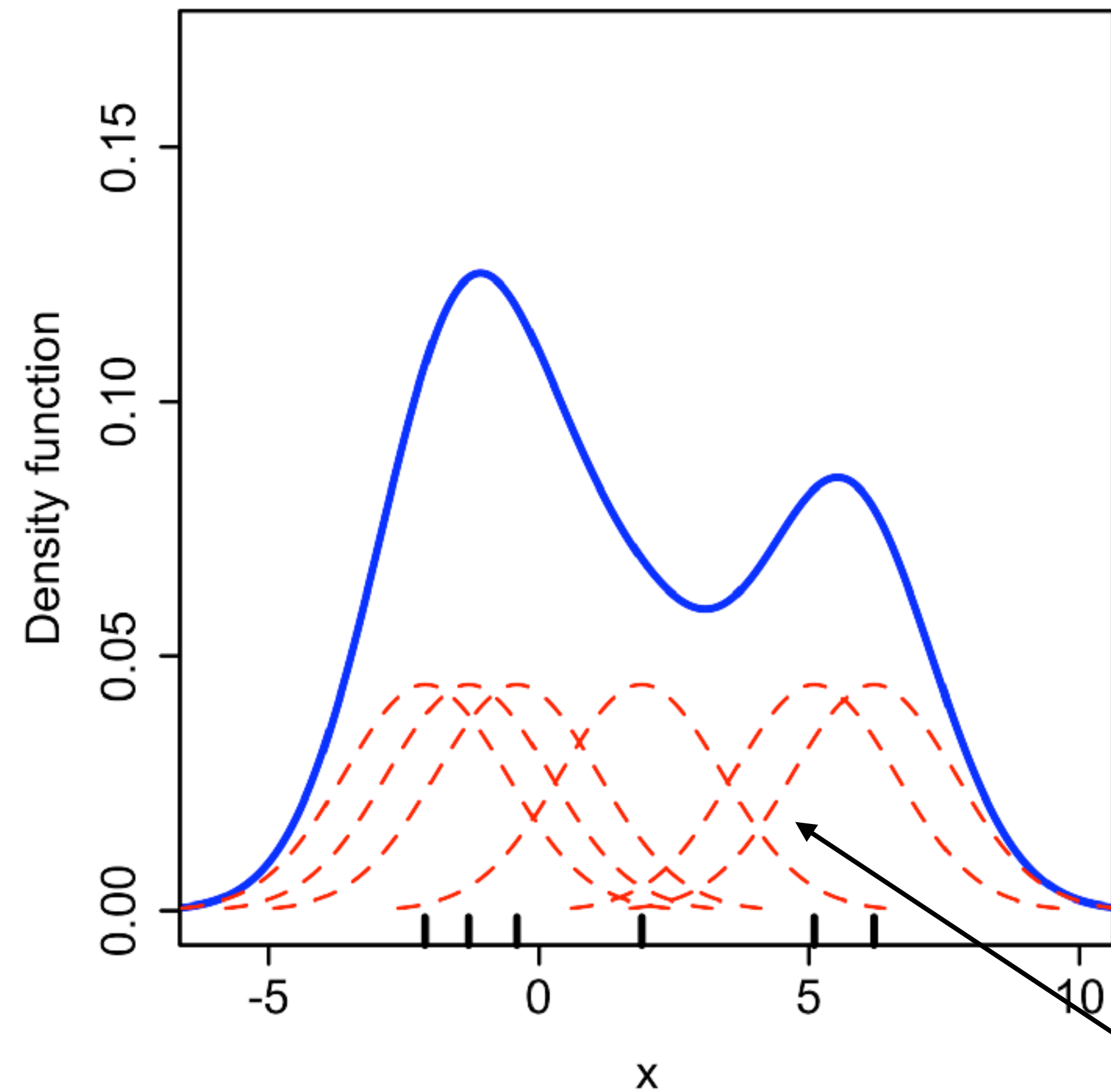
Kernel density estimation creates a smooth pdf to data points

Histogram



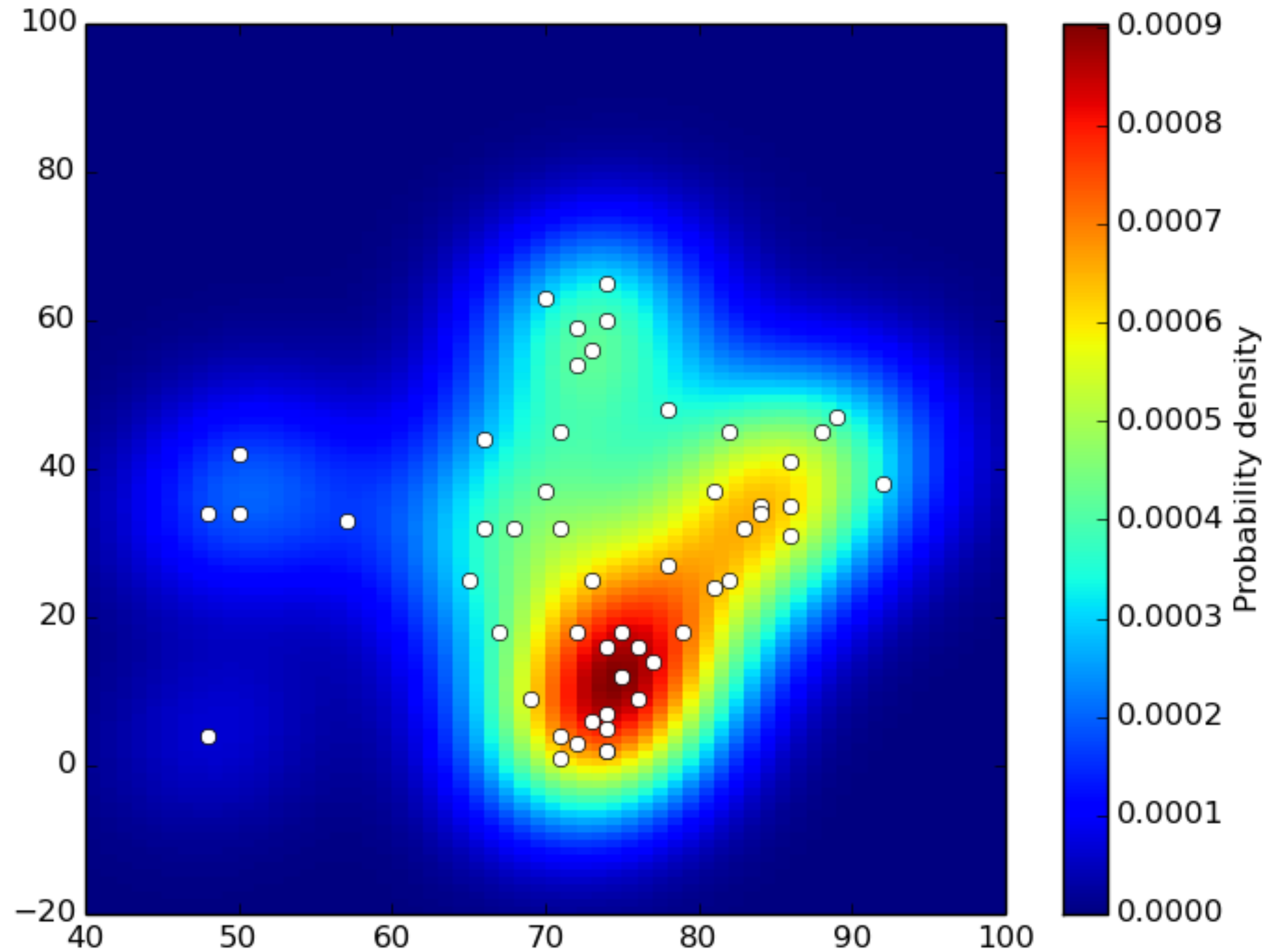
Rug plot

KDE



Kernel

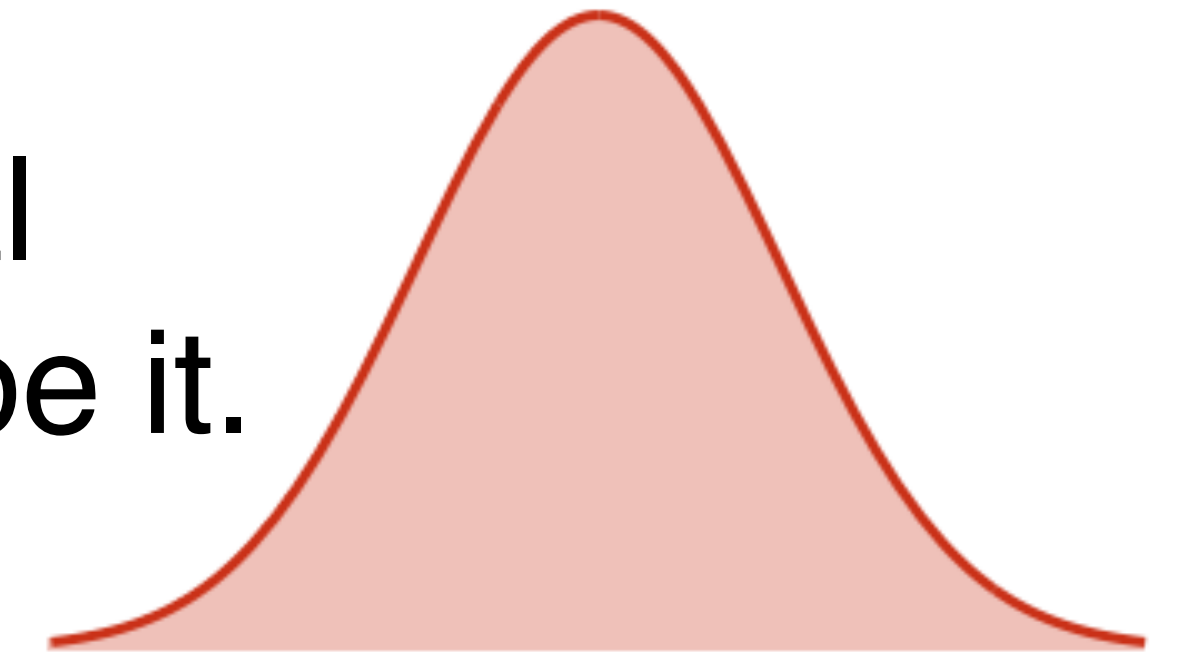
KDE creates heatmaps for geospatial data exploration



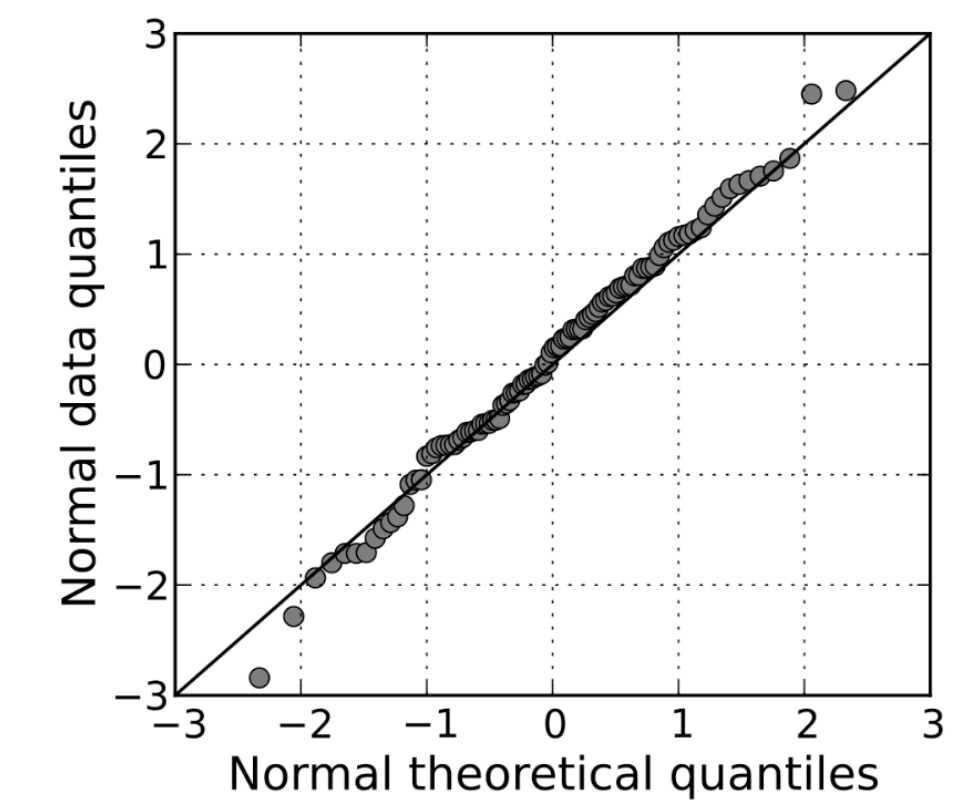
Jupyter

Take-home message

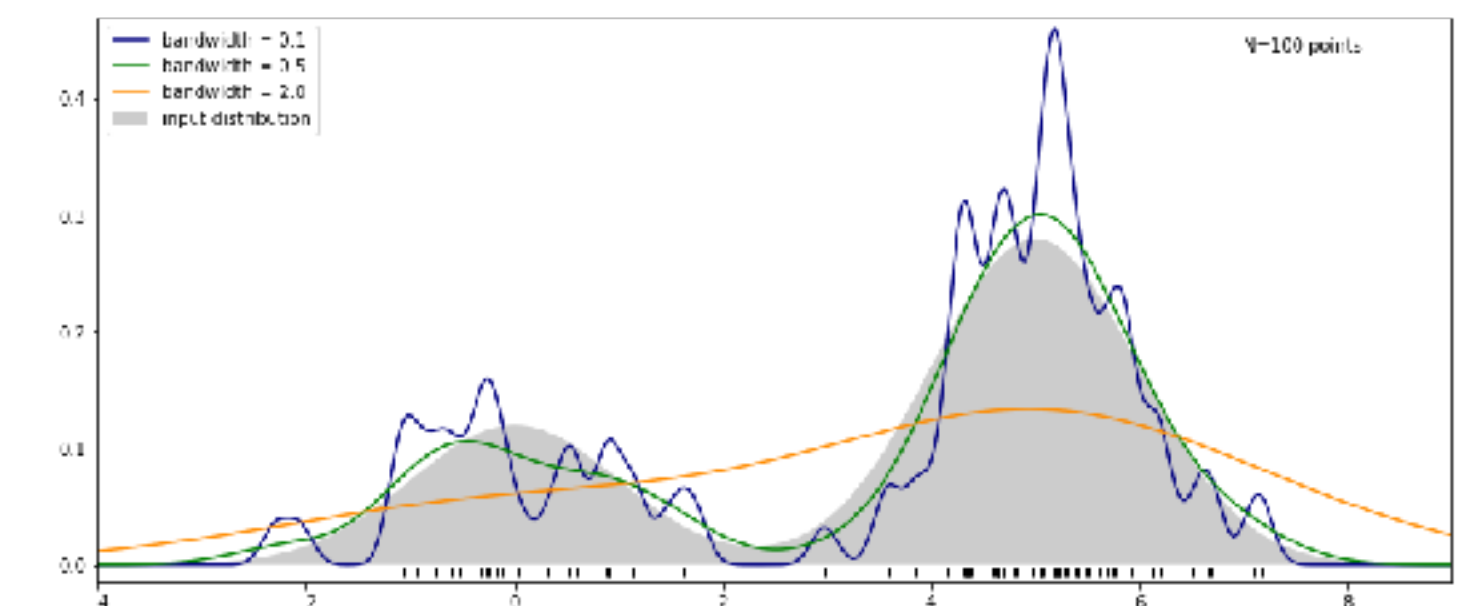
Sometimes data follows a smooth curve. If it's the normal distribution, you just need two parameters μ , σ to describe it.



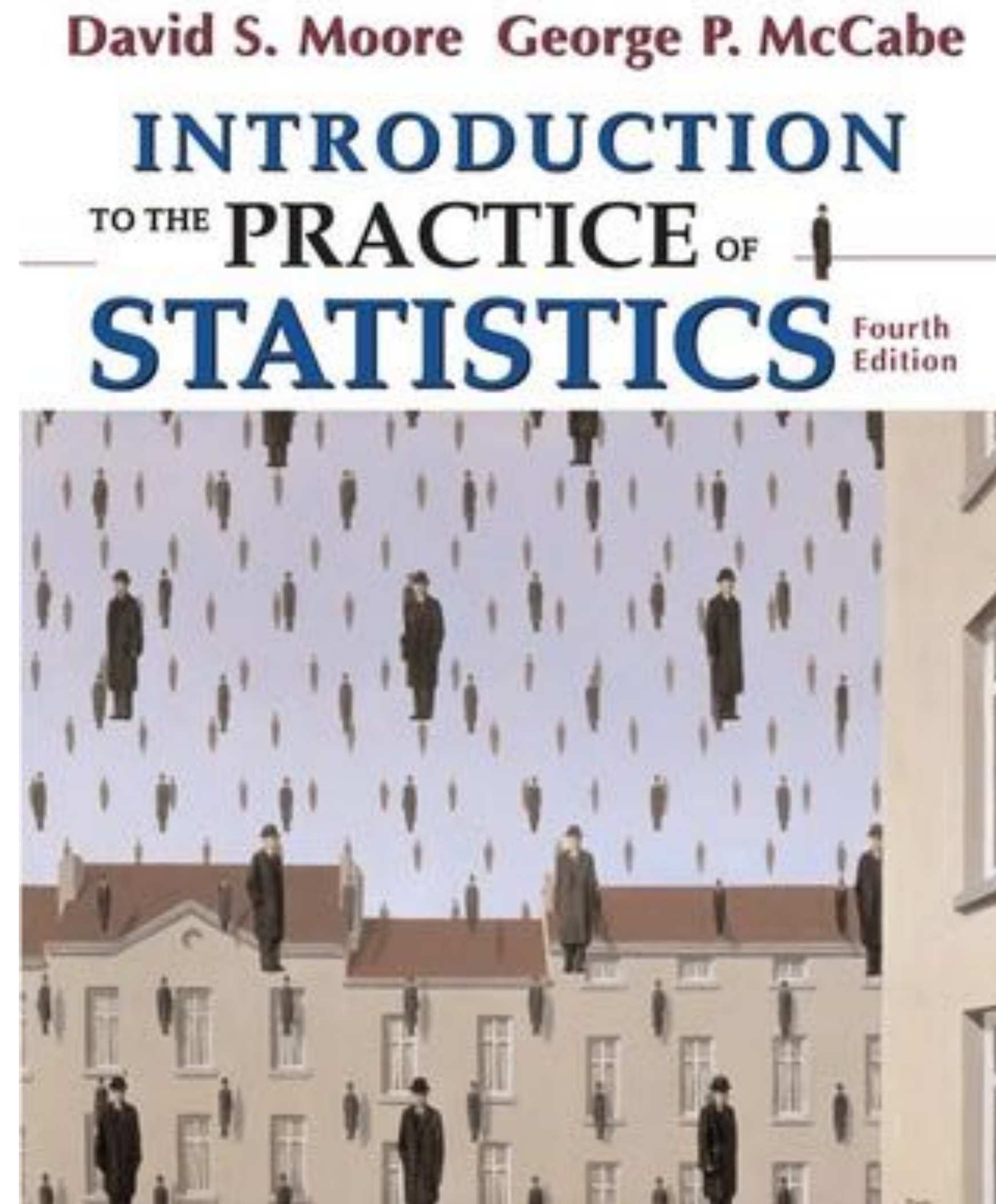
You can always fit any particular distribution, but it might not fit well. Q-Q is a visual tool to check.



The KDE finds a smooth curve fitting arbitrary data.



Sources and further materials for today's class



<https://stackabuse.com/kernel-density-estimation-in-python-using-scikit-learn/>

Chapter 1.3