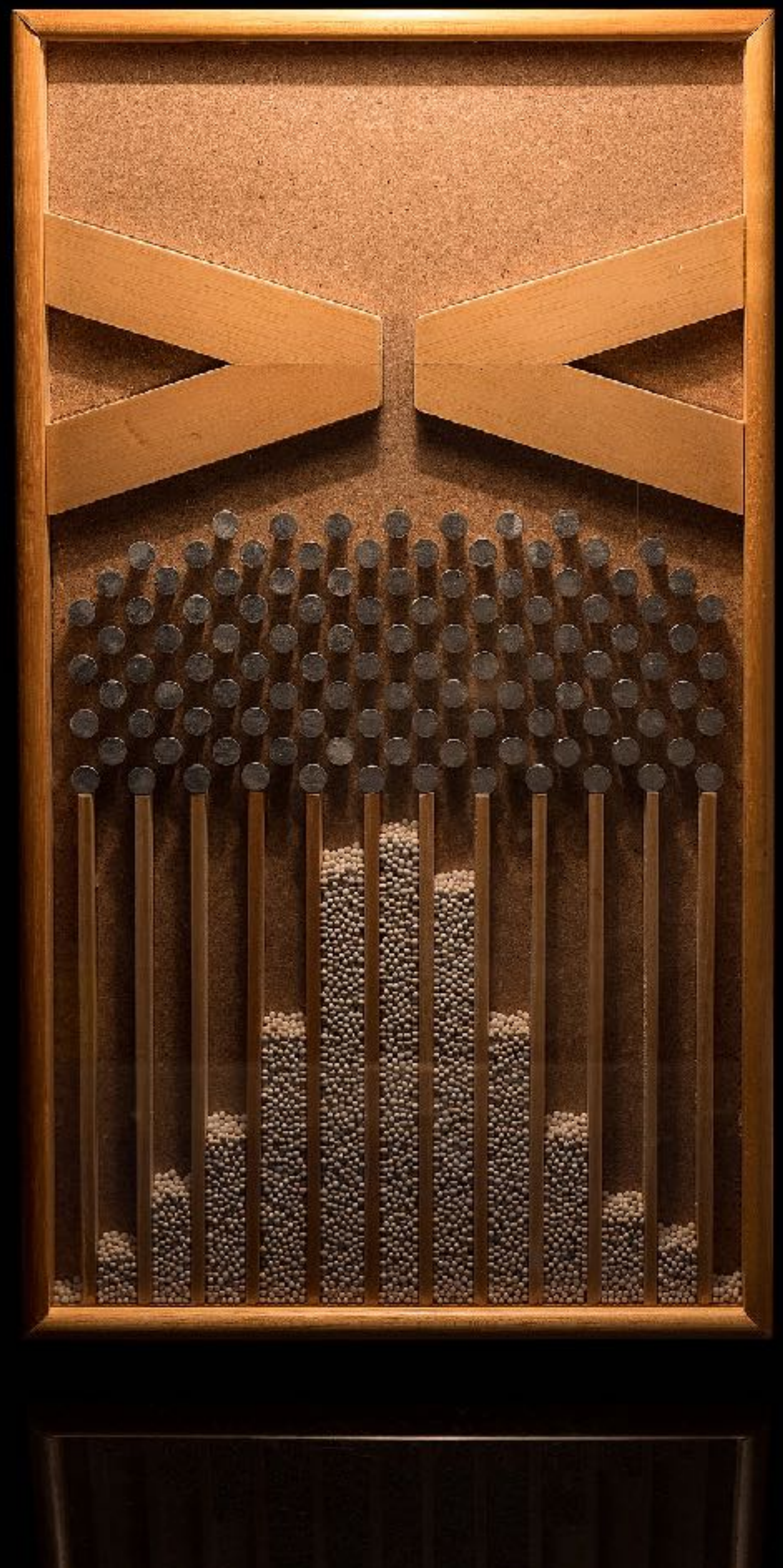
Distributies

Raoul Grouls, 12 maart 2025

Distributions
Definition

A probability distribution is:

- A mathematical description



Distributions

Definition

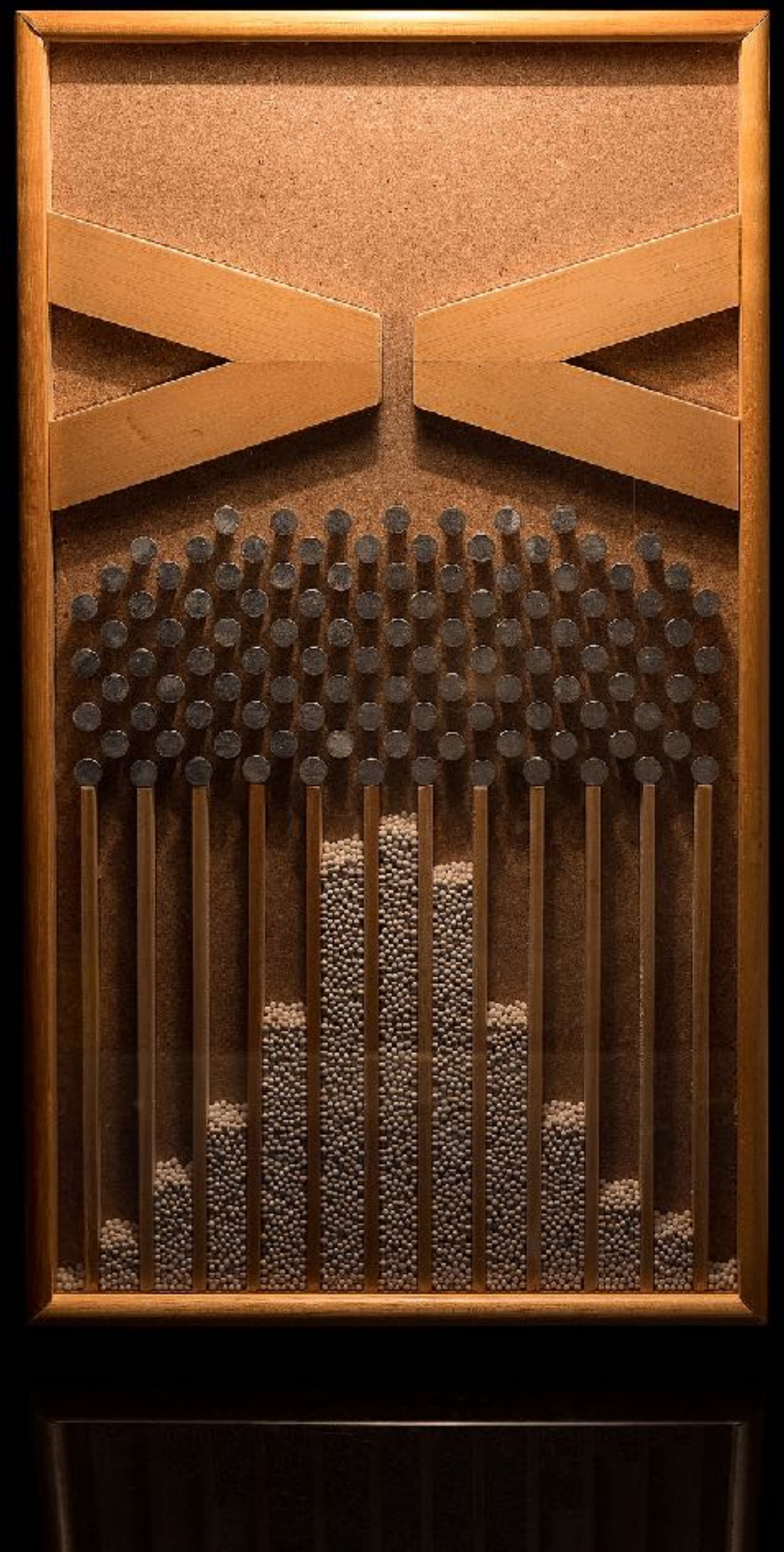
A probability distribution is:

- A mathematical description
- Of a random phenomenon

Distributions
Definition

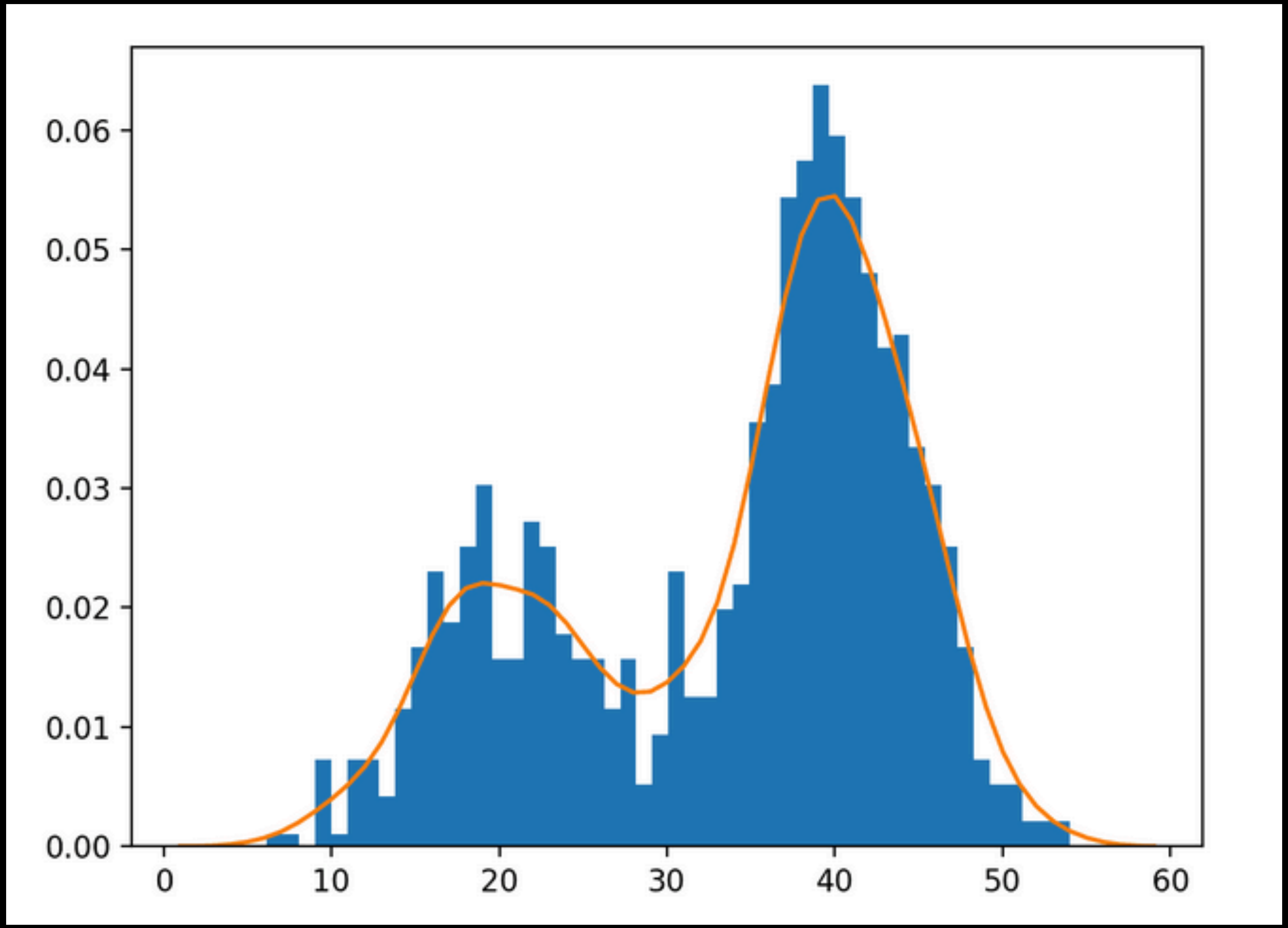
A probability distribution is:

- A mathematical description
- Of a random phenomenon
- In terms of all its possible outcomes and their associated probabilities



Distributions

- In this image, what are:
  - All the possible outcomes?
  - The associated probabilities?



Distributions

Types

The main types of distributions are:

- **Discrete** : when an outcome can only take discrete values (e.g. number of birds)

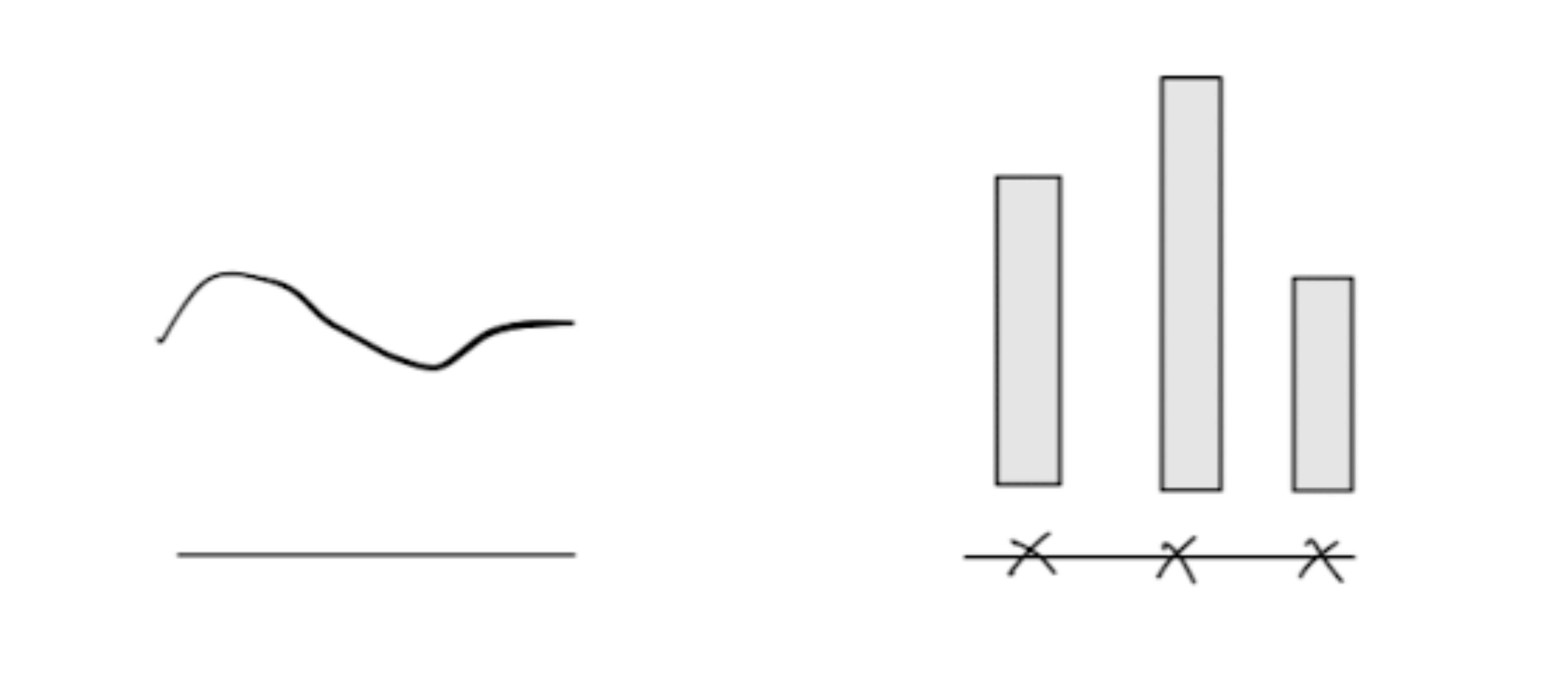
- **Continuous** : when outcomes take continuous values (e.g. blood pressure)

Distributions

Basic visualisation type

Every horizontal line you draw can be interpreted as a continuous distribution.
Every barplot as a discrete distribution.

All the distributions we are going to discuss are variations of these two basic types!



Distributions

Basic visualisation type

For *parametric distributions*, we have a formula that describes the line / bars. You just put in the parameters, and the output is the line / bars.

Discrete distributions
PMF

A **probability mass function** (pmf) describes the probability distribution of discrete variables.

Consider a toin coss:

$$f(x) = \begin{cases} 0.5 & x \text{ is head} \\ 0.5 & x \text{ is tails} \end{cases}$$

This is the pmf of the *Bernoulli distribution*

Conditions for a PMF

Plain English

1. An event cannot have a negative probability

Conditions for a PMF

Plain English

1. An event cannot have a negative probability
2. The sum of probabilities of all events must be 1

Conditions for a PMF

Plain English

1. An event cannot have a negative probability
2. The sum of probabilities of all events must be 1
3. The probability of a subset X of outcomes T is the same as adding the probabilities of the individual elements.

Conditions for a PMF

Mathematical

The probability is a function f over the sample space $\mathcal{S}$ of a discrete random variable X, which gives the probability that X is equal to a certain value.

$$f(x) = P(X = x)$$

Each pmf satisfies these conditions:

1. $f(x) \geq 0, \forall x \in X$
2. $\Sigma_{x \in \mathcal{S}} f(x) = 1$
3. For a collection $\mathcal{A}$, $P(\mathcal{A} \in \mathcal{S}) = \Sigma_{x_i \in \mathcal{A}} f(x_i)$

Continuous Distributions
PDF

For continuous distributions, we use a probability density function (pdf)

Continuous Distributions

PDF

For continuous distributions, we use a probability density function (pdf)

1. $f(x) > 0, \forall x \in X$

2. The integral of the probabilities of all possible events must be 1 (area under the curve)

3. The probability X of values in the interval $[a, b]$ is the integral from a to b

Continuous Distributions

PDF

This might look like unnecessary mathematical details. But it is actually important to understand the difference.

Example: can you answer the question “What is the probability your body temperature is 37.0 C?”

Continuous Distributions

PDF

The answer might be unexpected: 0!

Let’s say your answer is 25%. But what if your temperature is 37.1? does that count? Or 37.01?

Continuous Distributions

PDF

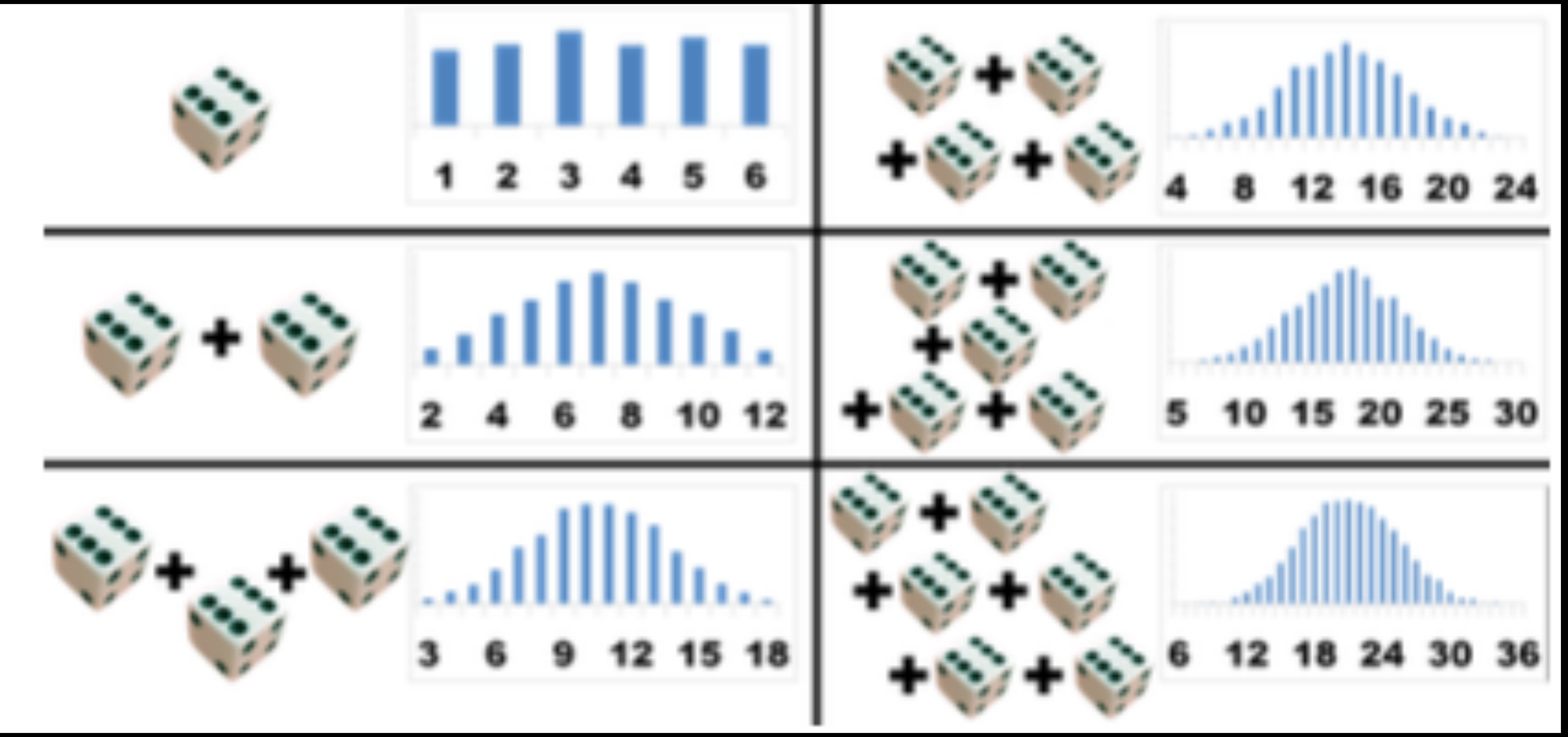
- Because the distribution is *continuous* you can only say something about the *range*
- "What is the probability your temperature is between 36.5 and 37.2 C?"

Quiz time

Normal distributions
Central limit theorem

The central limit theorem states that:

- the distribution of a normalized version of the sample mean <u>converges to a standard normal distribution</u>.
- This holds even if the original variables themselves are not normally distributed.

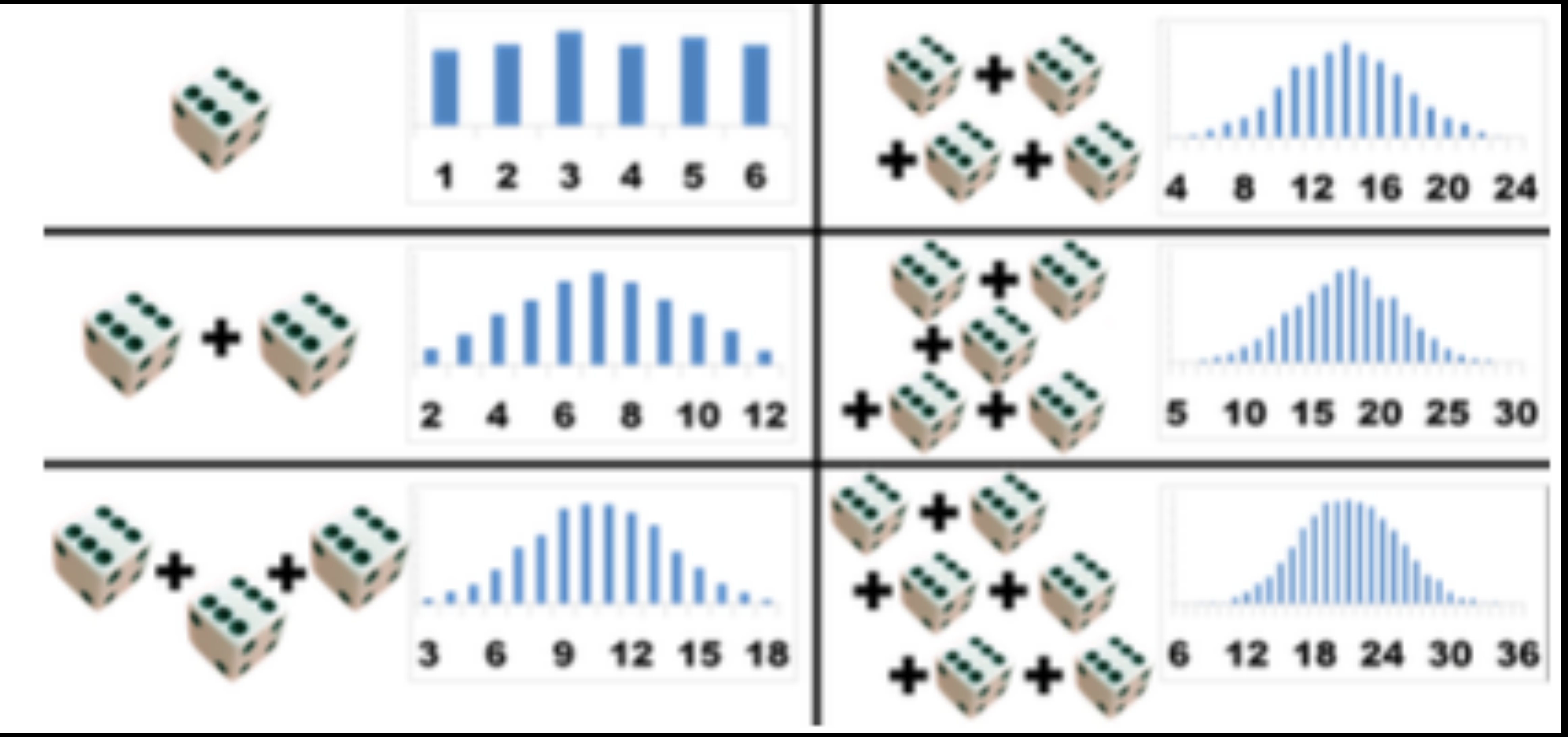


Normal distributions

Central limit theorem

The Normal distribution is one of the distributions that is used most often.

A major reason for this is, that if you keep sampling and **adding** from a population you *always* end up with a normal distribution.

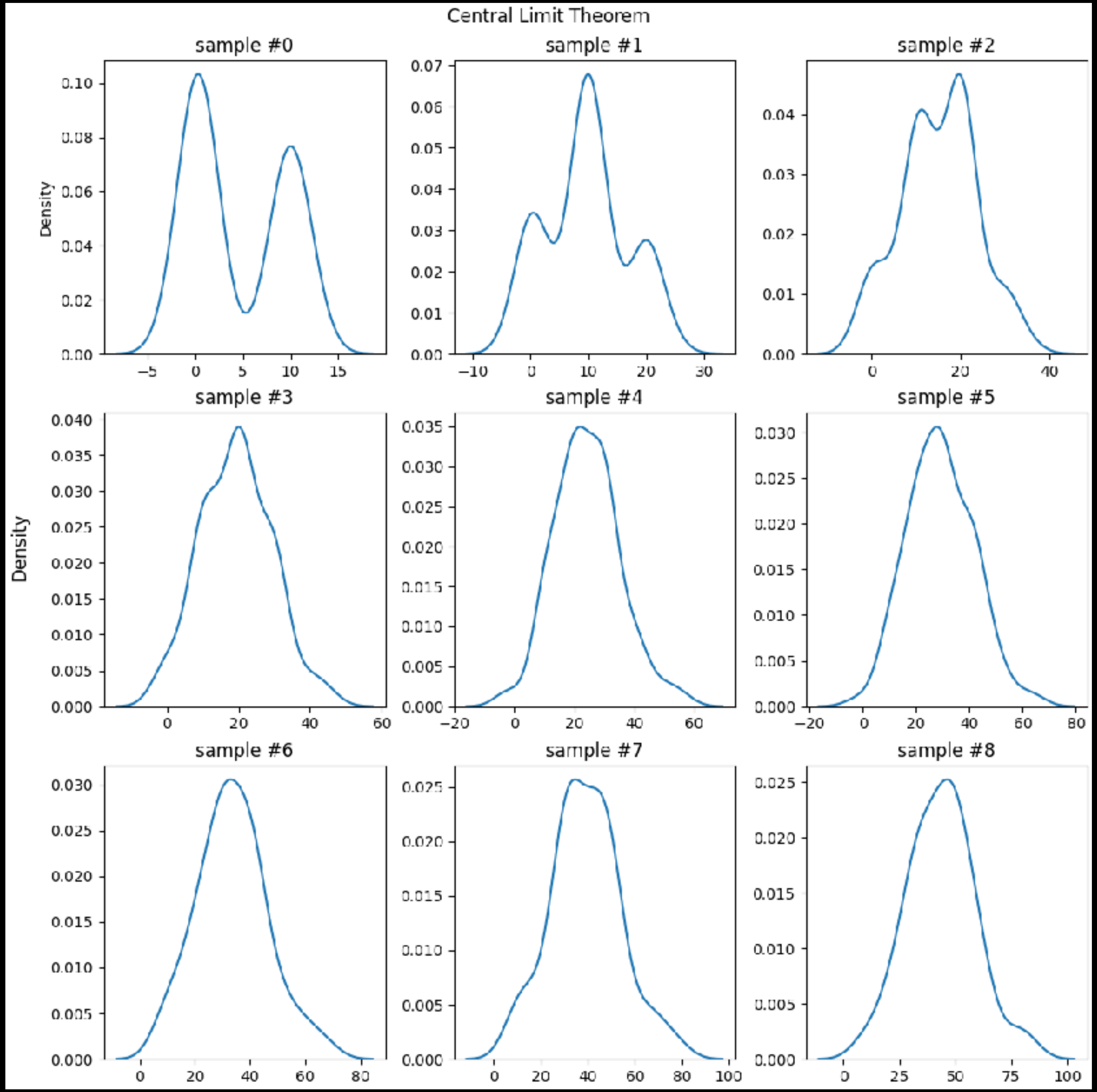


Normal distributions
Central limit theorem

Take a persons height.

- This is determined by a combination of 180 genes.
- One gene will contribute to a longer neck, the other to longer legs
- If we assume the genes contribute independently, height equals the sum of 180 genes.

Thus, height will be normally distributed. So will the weight of wolves or the length of a penguins wing.



Log Normal distribution

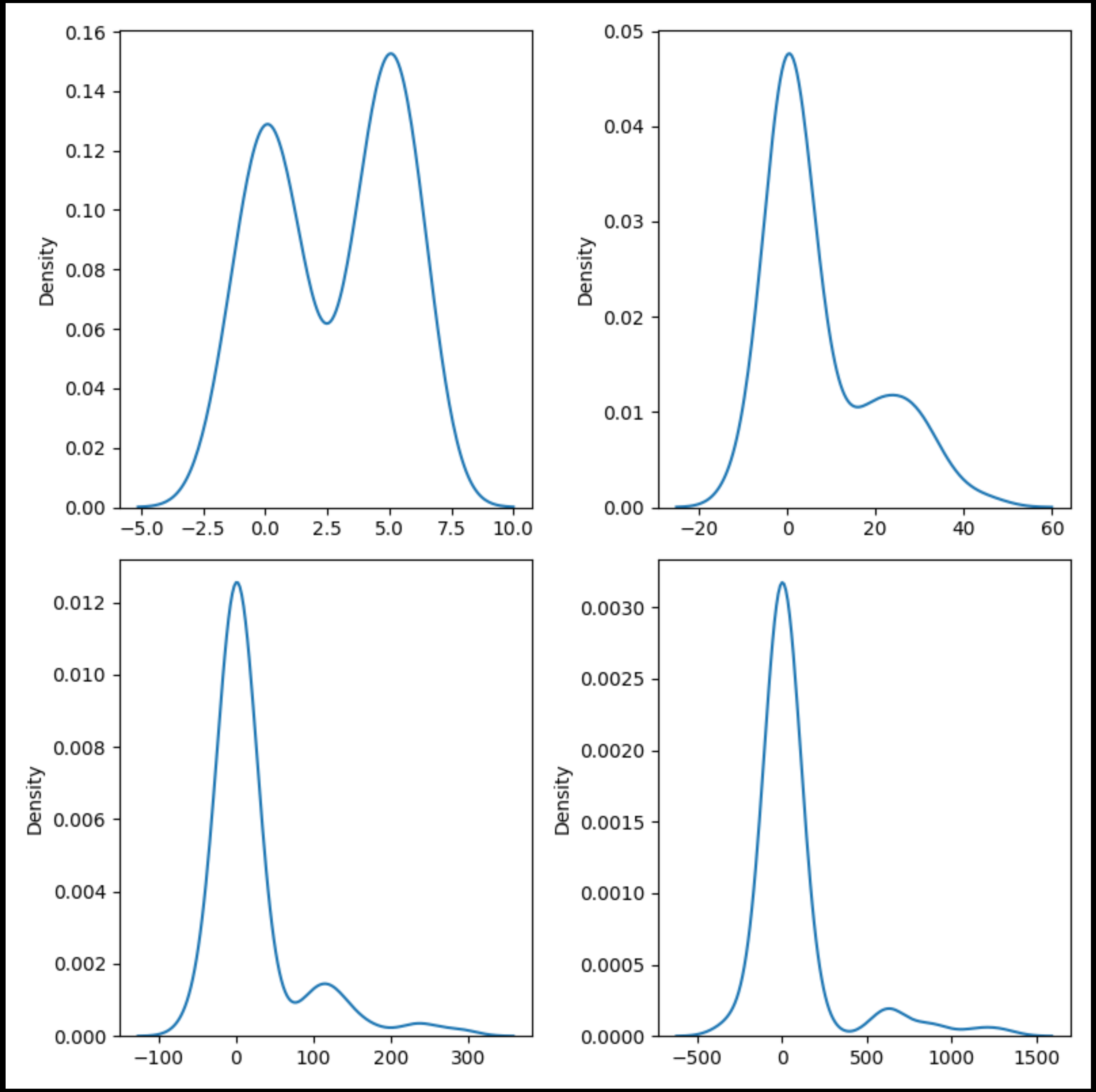
However, **multiplying** values will give you a long tail!

This is the case when variables interact in some way, and are not independent.

$$4 + 4 + 4 + 4 = 16$$

but

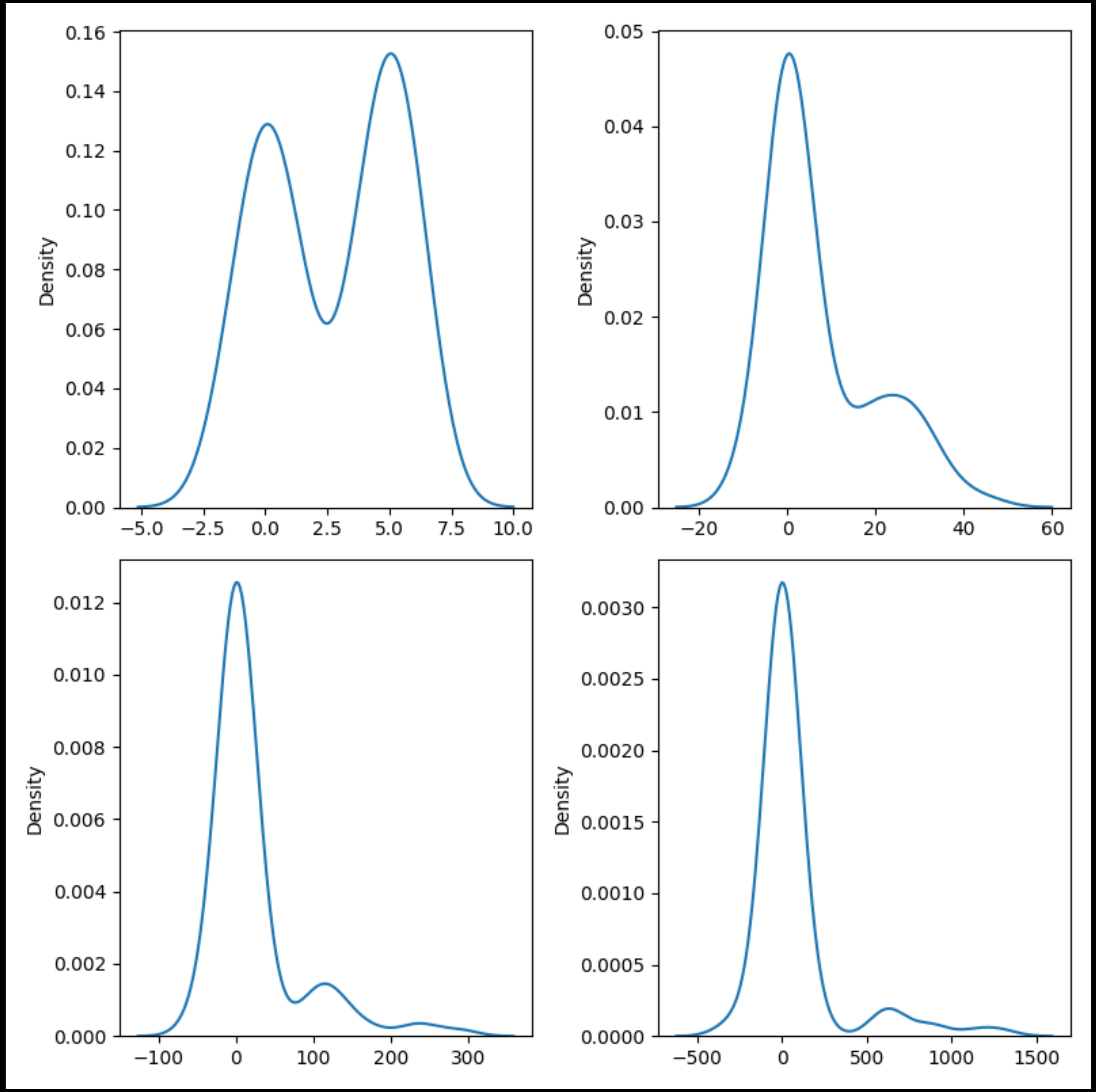
$$4 \times 4 \times 4 \times 4 = 256$$



Log Normal distribution

Multiplying should be expected if variables interact with each other.

Examples are stock prices, failures of machines, ping times on a network, income distribution.



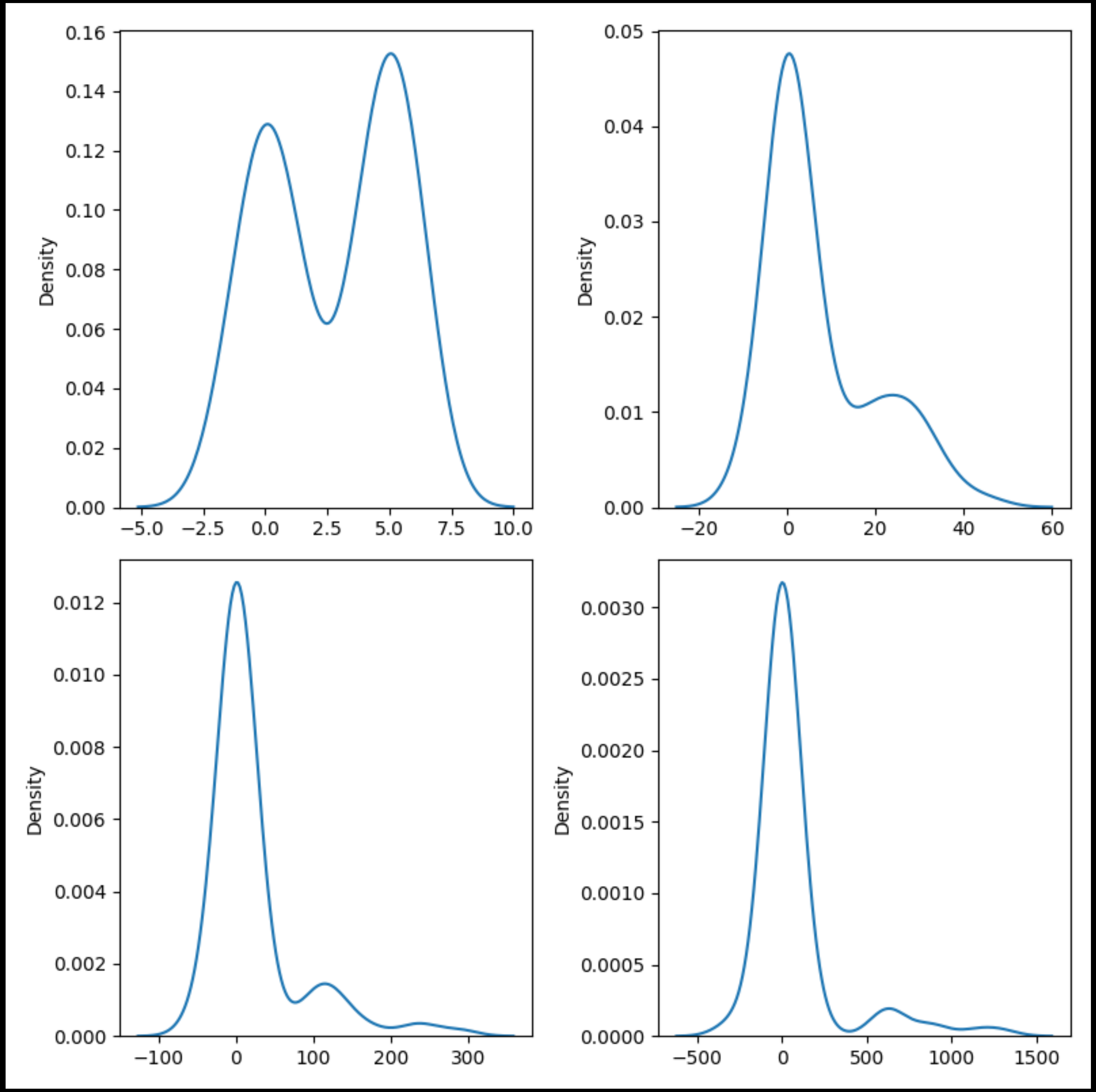
Log Normal distribution

multiplying values will give you a fat-tail distribution! This will typically be a log-normal distribution:

If X is log-normal distibuted, then

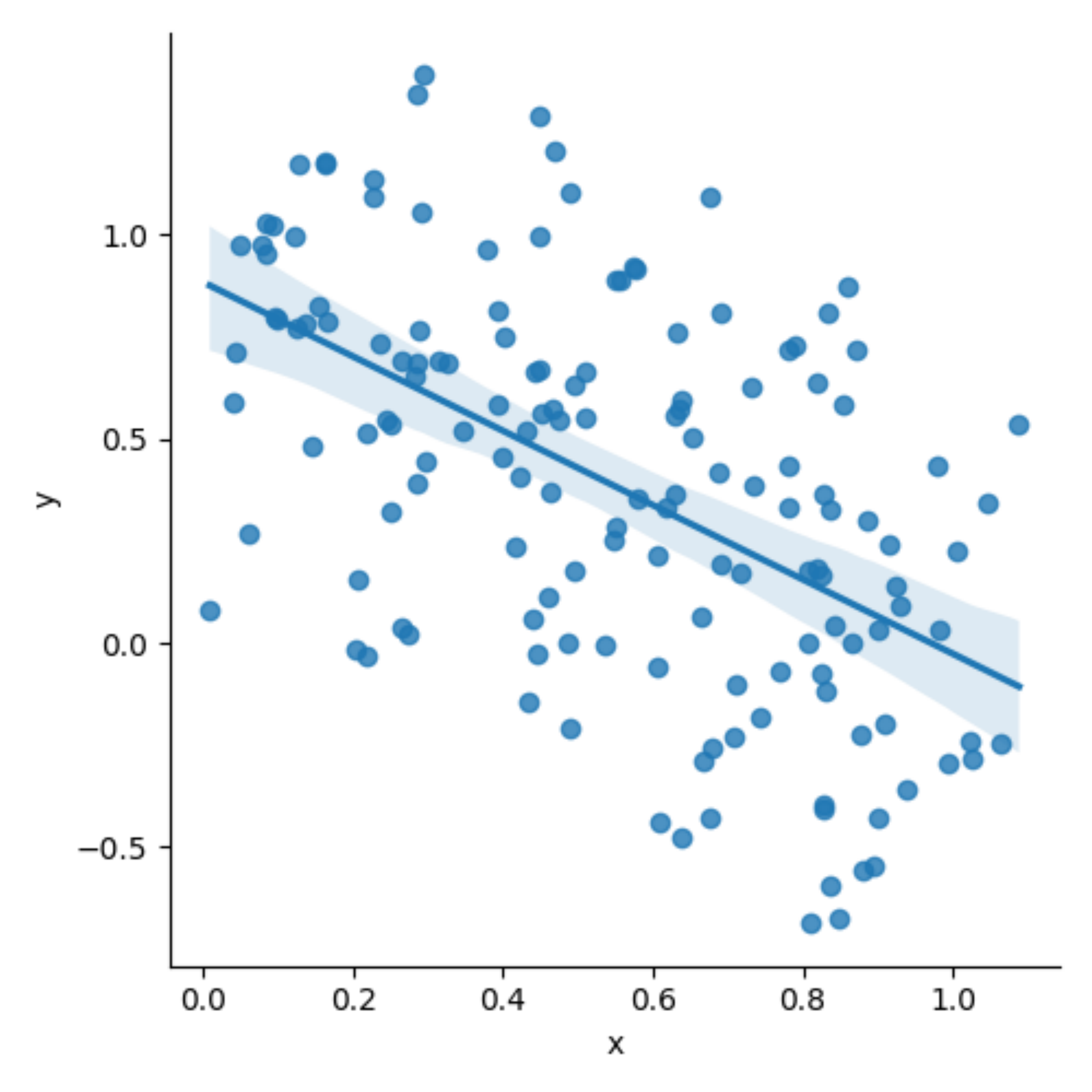
$$y = log(X)$$

will be a normal distribution.



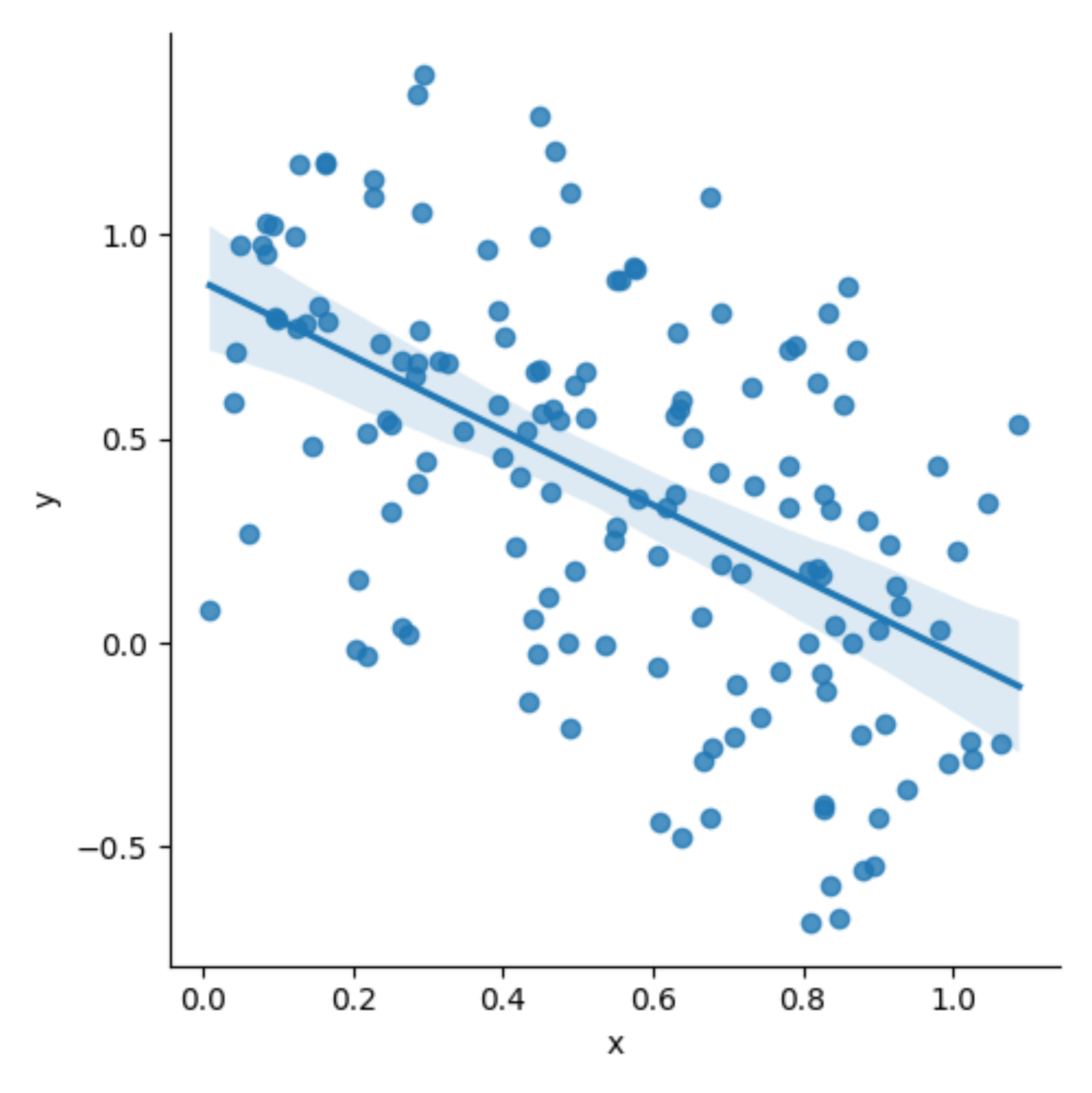
The Simpsons paradox

- The shaded area is the 99% confidence interval of the linear regression.
- What is your conclusion about the data?



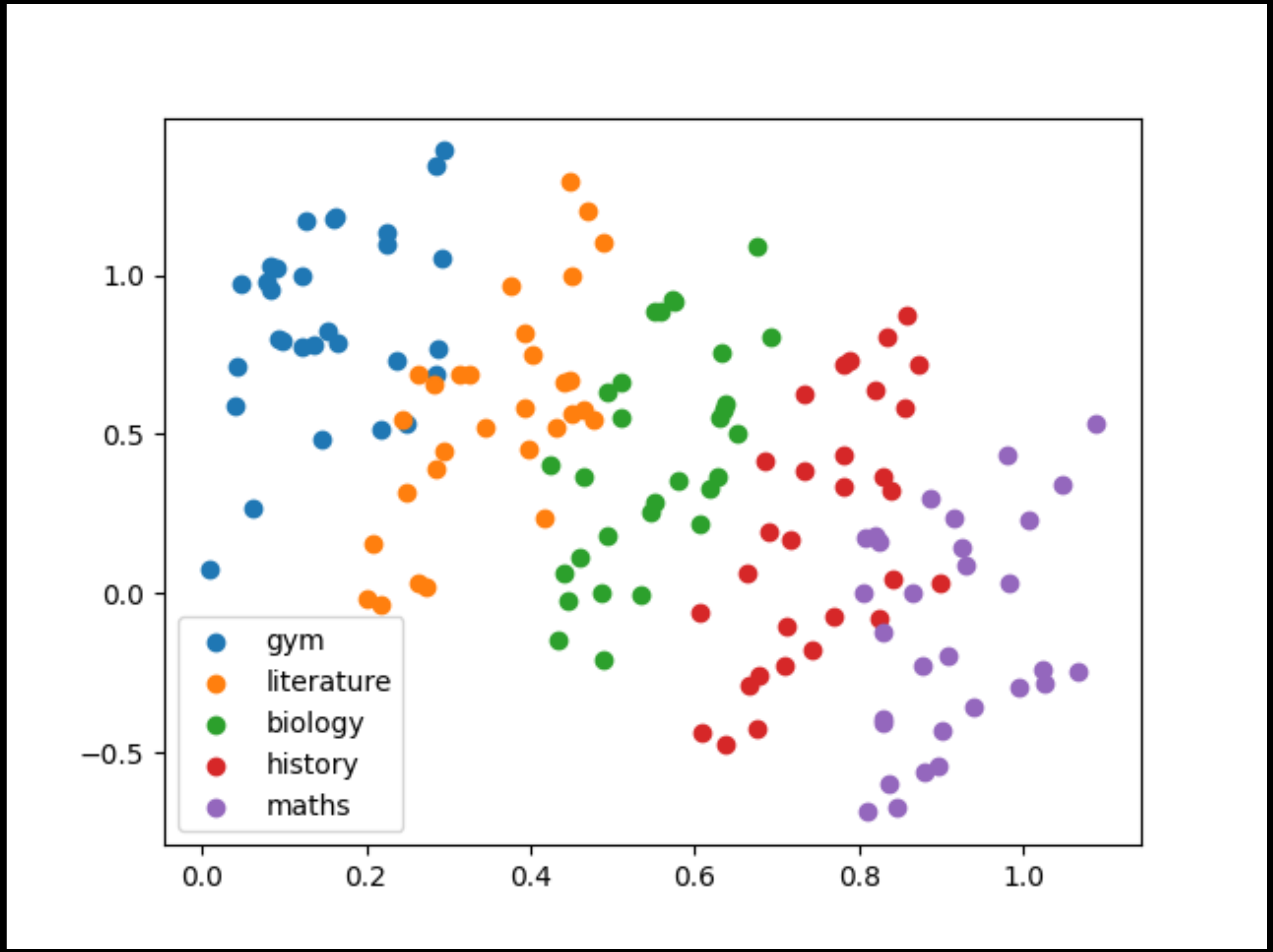
The Simpsons paradox

- The shaded area is the 99% confidence interval of the linear regression.
- Does your conclusion change if I tell you that the x axis is the amount of hours invested in study, and the y axis is the average grade of a student? Why?



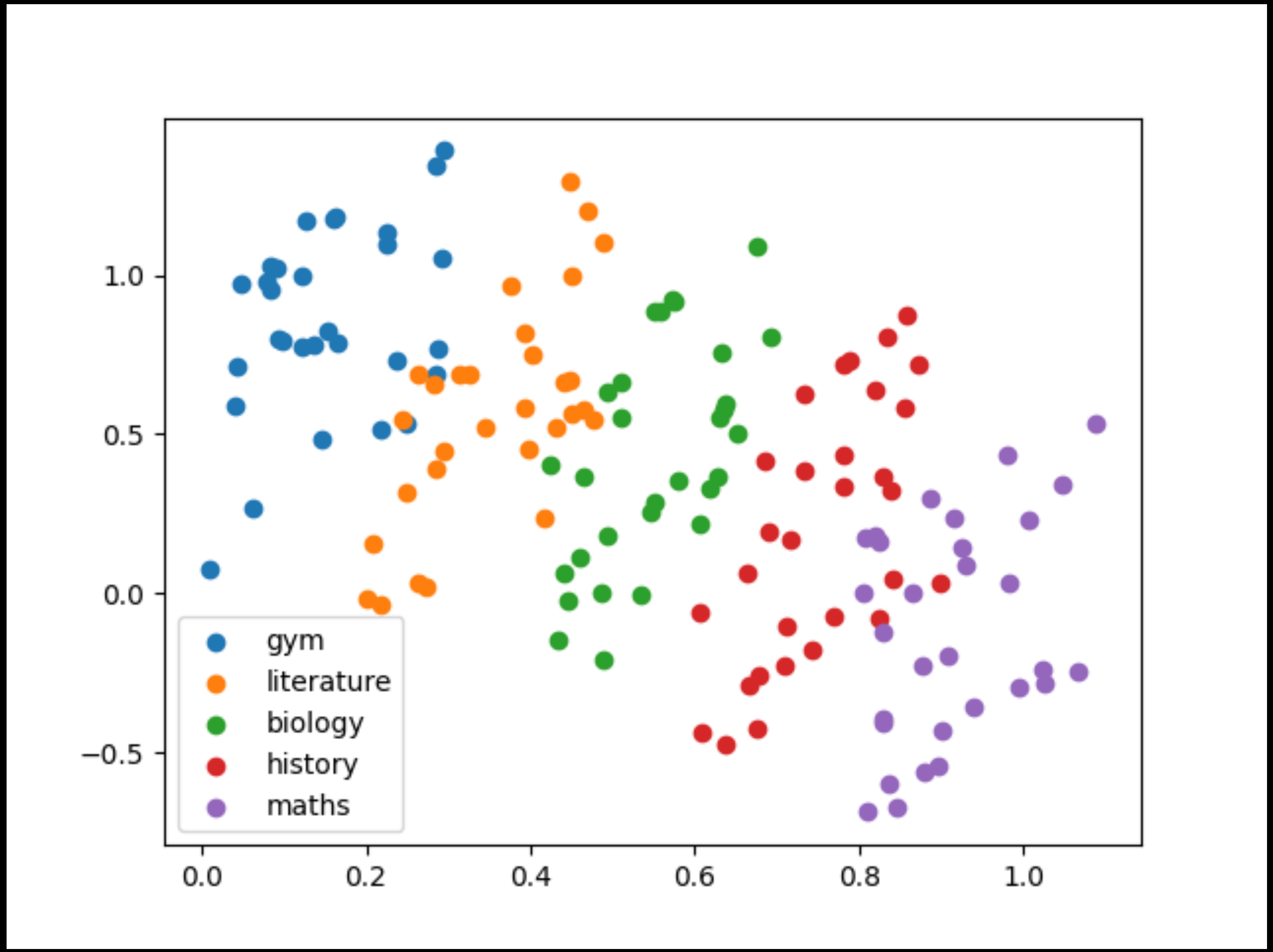
The Simpsons paradox

- The shaded area is the 99% confidence interval of the linear regression.
- the x axis is the amount of hours invested in study, and the y axis is the average grade of a student
- Does changing the colors change your conclusion?
- If so, should you have changed your inital conclusion, even without this extra information?



The Simpsons paradox

- Simpson's paradox is a phenomenon in which a trend appears in several groups of data but **disappears** or **reverses** with different groups.
- This result is often encountered in social-science and medical-science statistics
- It is particularly problematic when frequency data are undeservedly given **causal interpretations**



The Simpsons paradox

UC Berkely Gender Bias (admission fall 1973)

	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	12,763	41%	8,442	44%	4,321	35%

The Simpsons paradox

UC Berkely Gender Bias (admission fall 1973)

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	**825**	62%	108	82%
B	585	63%	**560**	63%	25	68%
C	918	35%	325	37%	**593**	34%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	**393**	24%
F	714	6%	373	6%	341	7%
Total	**4526**	**39%**	**2691**	**45%**	**1835**	**30%**