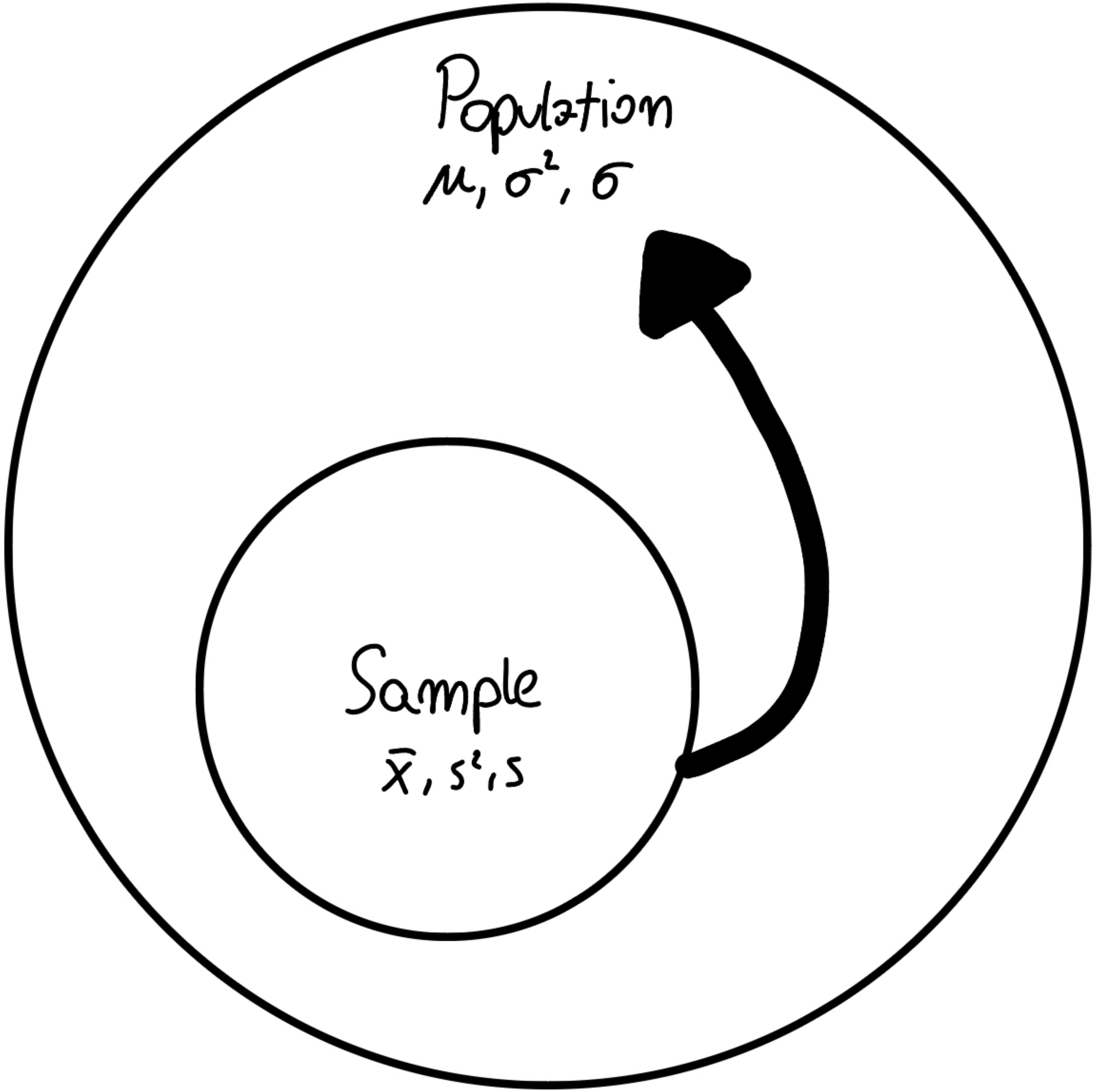
Statistics

Group of Horribly Optimistic STatisticians

Jędrzej Ogrodowski

Statistics is extremely broad subject. In our course we
we will talk only about key aspects with as less theory as possible.
I recommend read about these topics. For understanding of statistics, it is necessary
to know some topics from Probabilistic Methods. (2nd semester on AI and CS)

For self study, recommend StatQuest on Youtube.


Population
μ, σ^2, σ
Sample
$\bar{x}, s^2, s$

We will focus on dependency between parameters and
working with sample data and decide
if our sample result is likely in whole population.

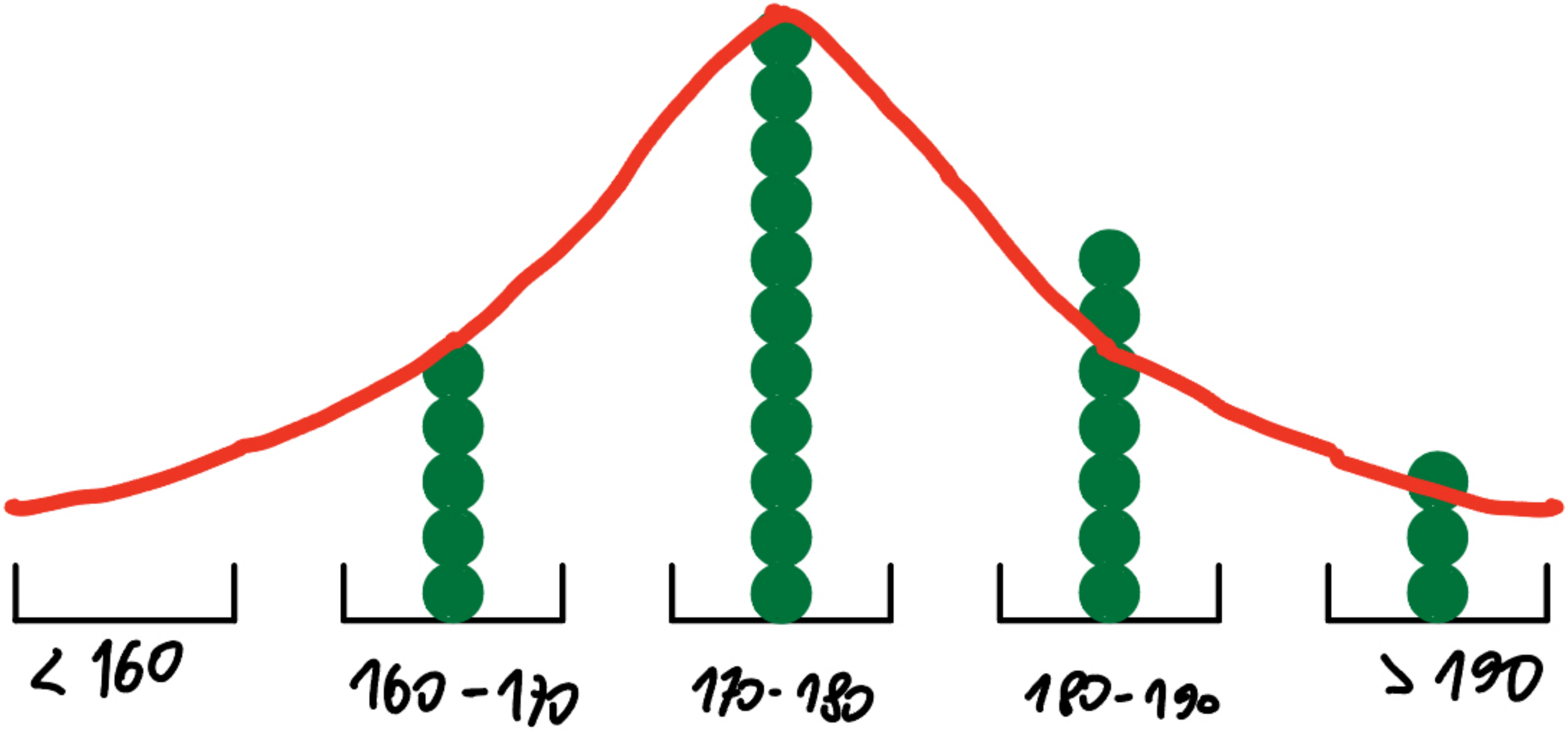
Random variables

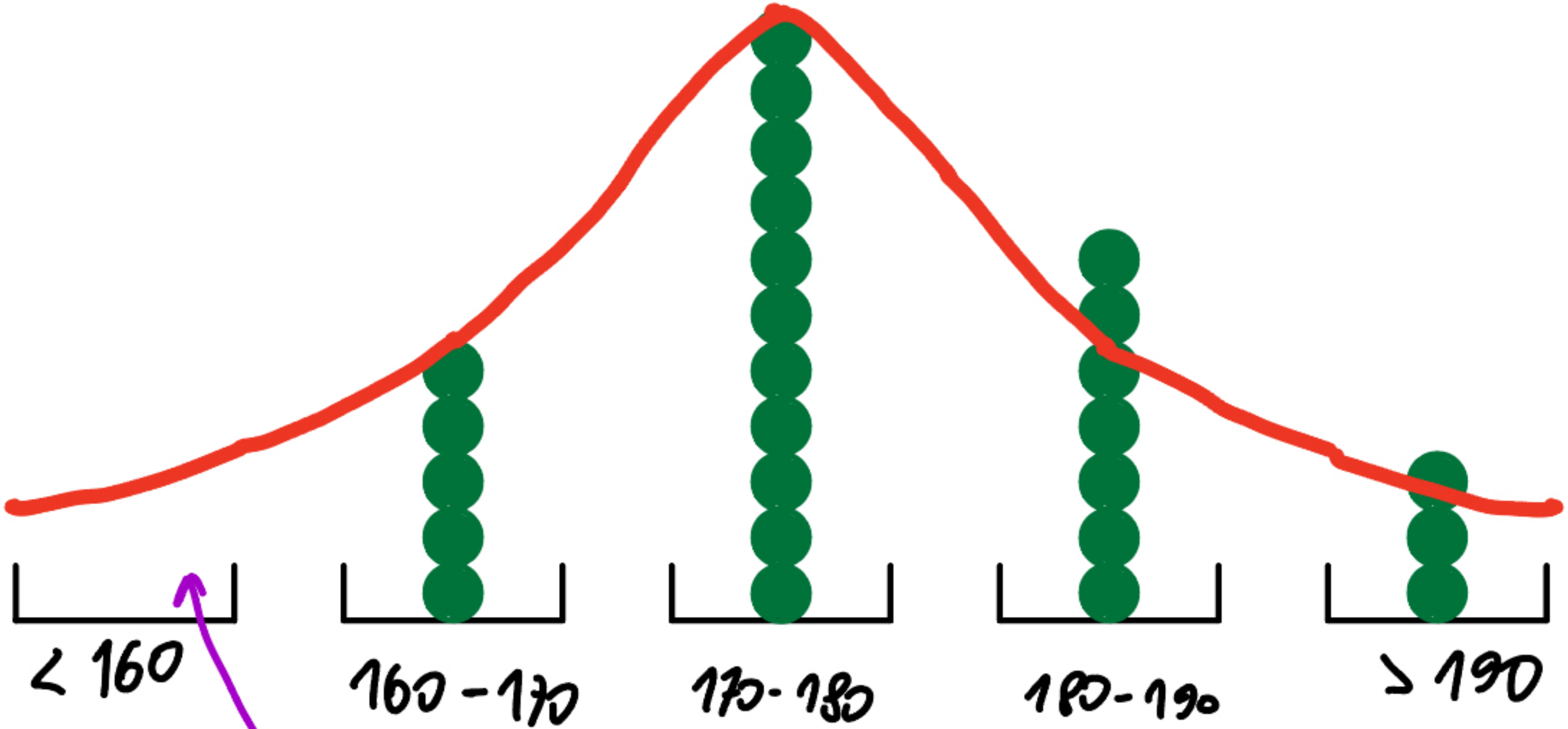
A **random variable** is a numerical value that represents the outcome of a random event or experiment. It’s called "random" because its value depends on the outcome, which isn’t certain until we observe it.

A random variable is closely related to a **distribution**, which describes all the possible values the random variable can take and the probabilities associated with each of these values. The distribution gives us a "picture" of the behavior of the random variable by showing how often we can expect each outcome.

Normal Distribution

It is fundamental concept in statistics. It's a probability distribution that is symmetric about the mean, meaning that data points closer to the mean are more frequent than those farther away.


< 160
160 - 170
170 - 180
180 - 190
> 190



If nothing is there, probability of someone below 160 cm is 0? NO. That's why we apply curve.


-3σ
-2σ
-1σ
μ
+1σ
+2σ
+3σ

Ubiquity in Nature: Many natural phenomena follow a normal distribution.

Central Limit Theorem: This theorem states that the distribution of sample **means** approaches a normal distribution as the sample size increases, regardless of the underlying population distribution.

Foundation for Statistical Tests: Many statistical tests, like t-tests and ANOVA, assume that the data is normally distributed.

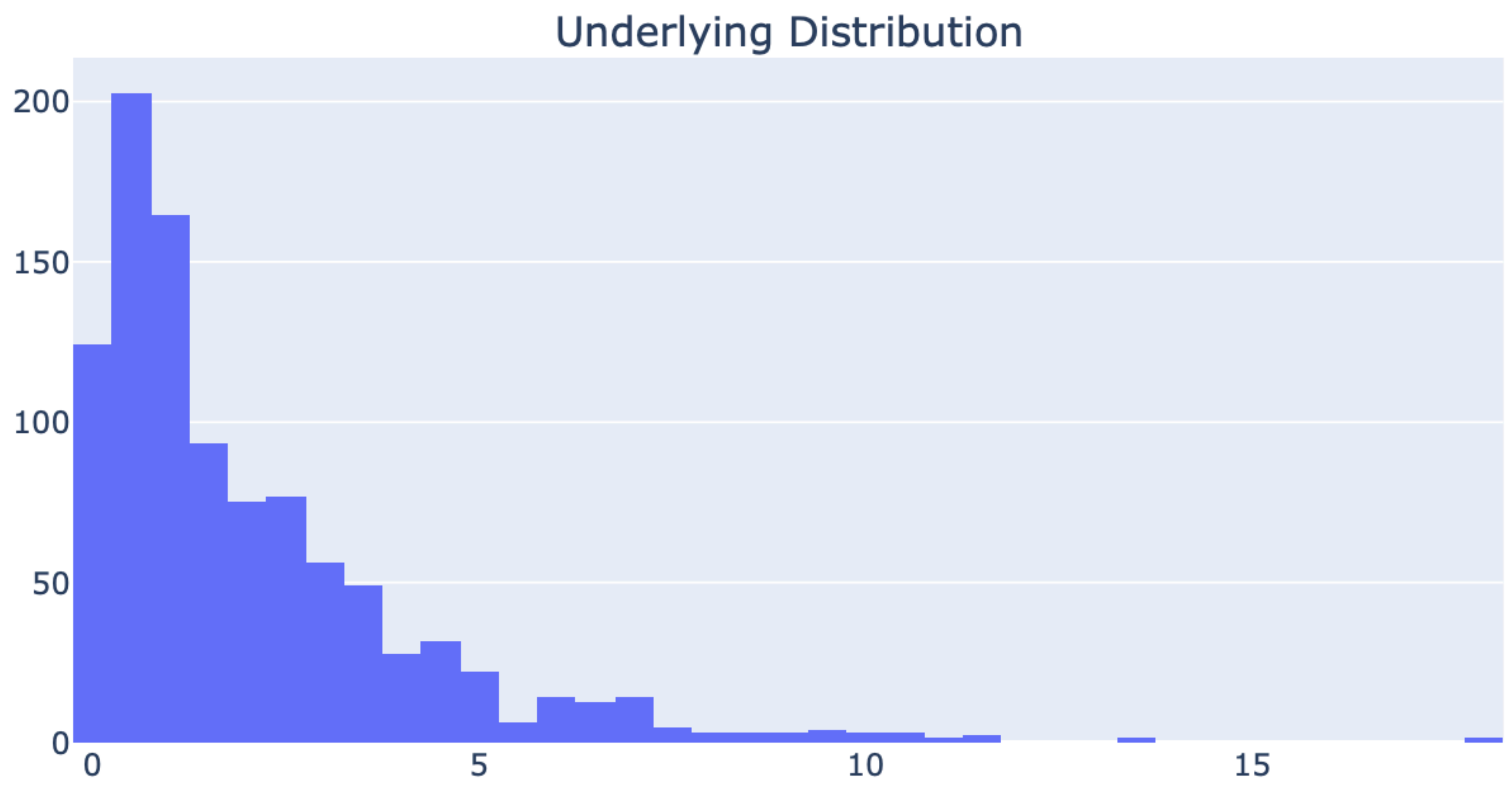
Central Limit Theorem

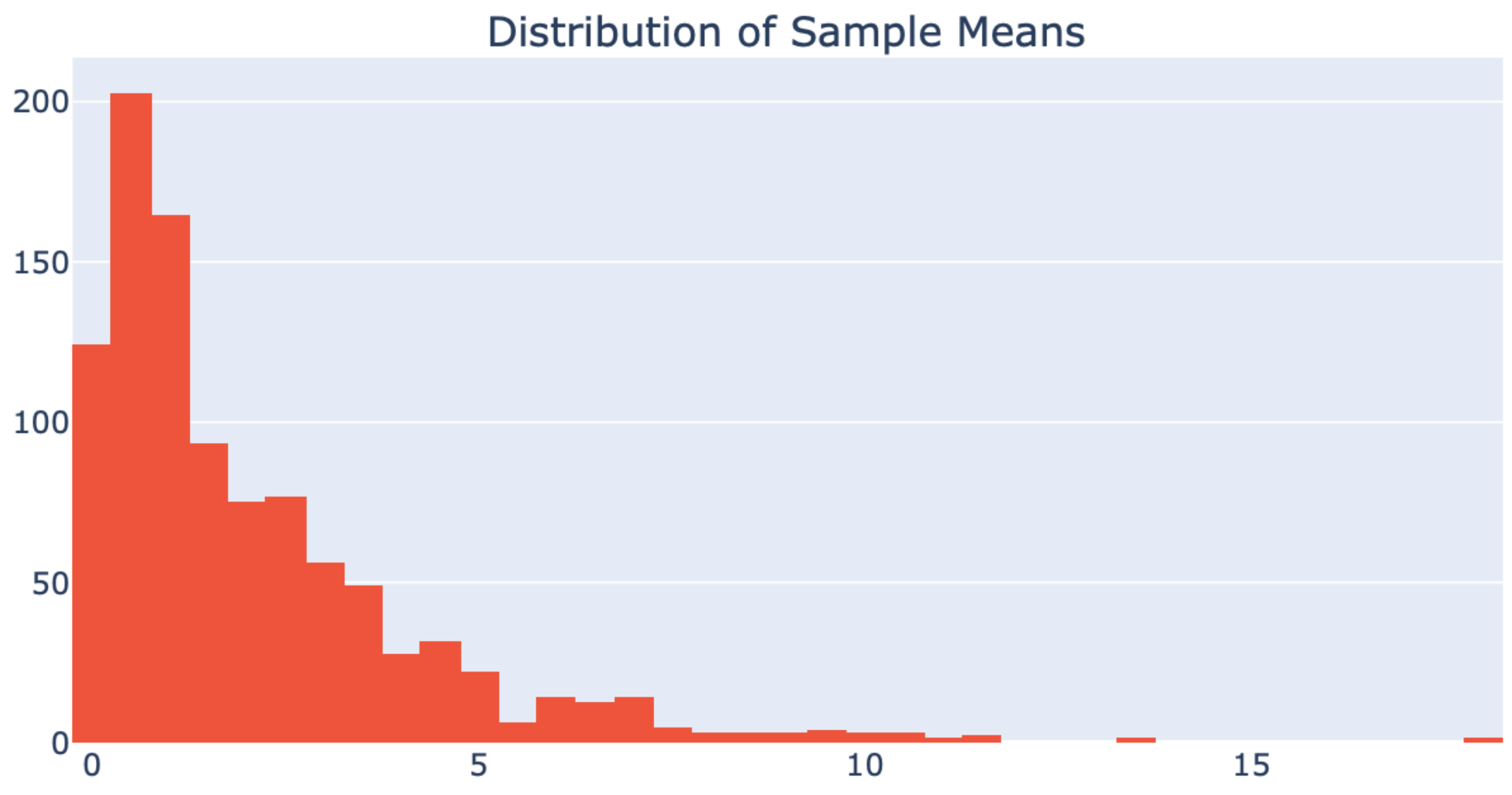
Sample Mean Distribution

Select a distribution and the size of each sample.

Exponential

1





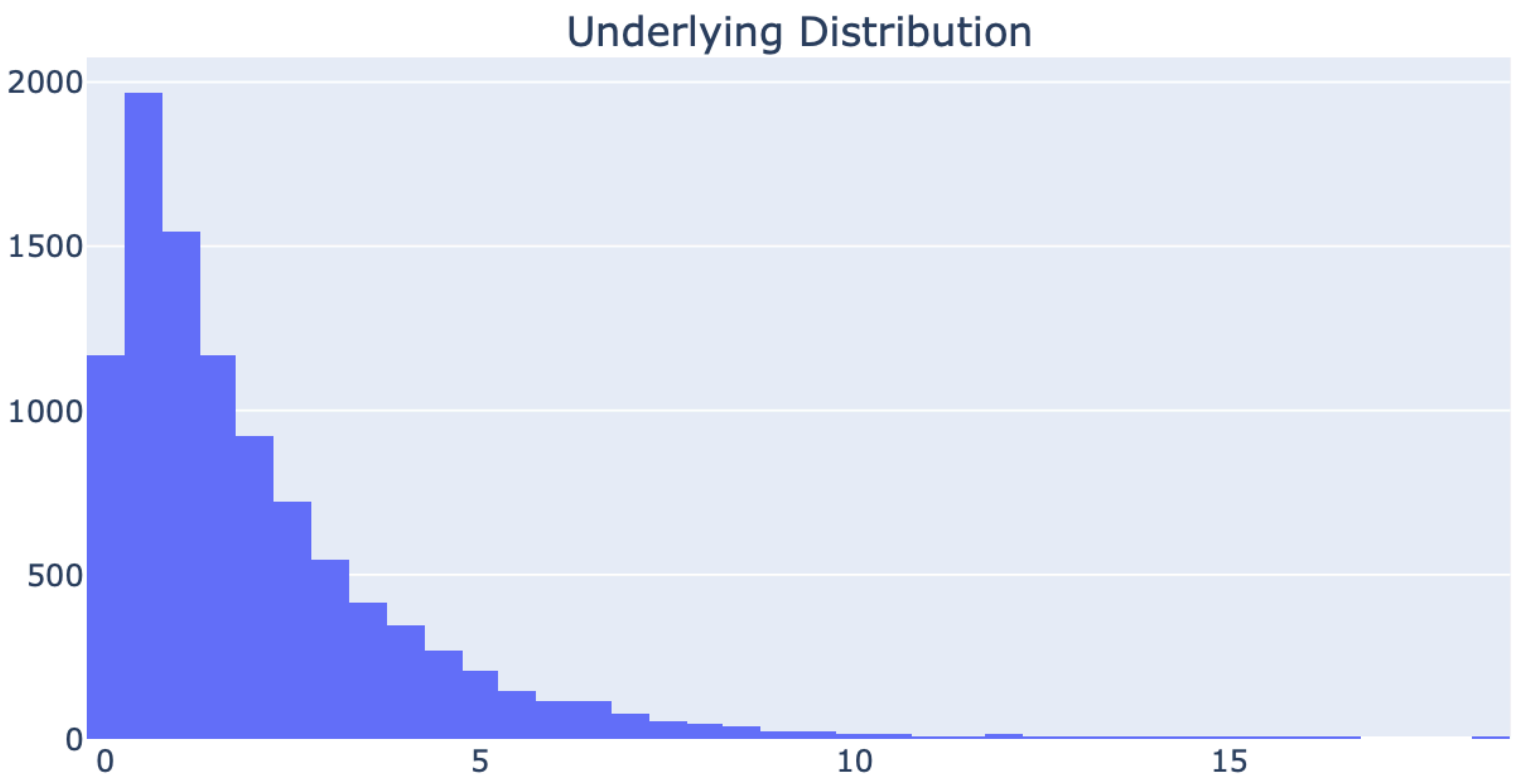
https://cltapp.fly.dev/

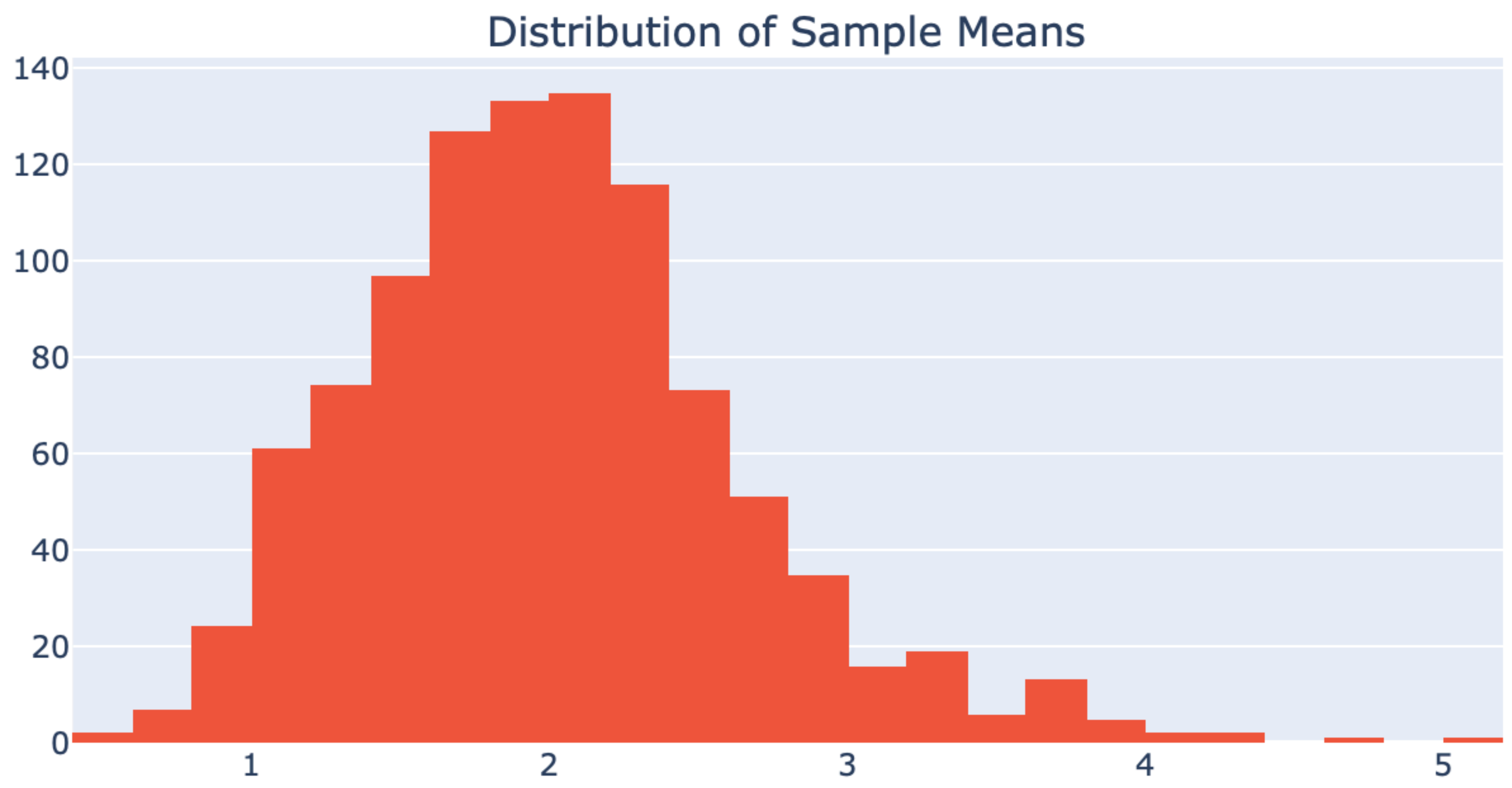
Sample Mean Distribution

Select a distribution and the size of each sample.

Exponential

10





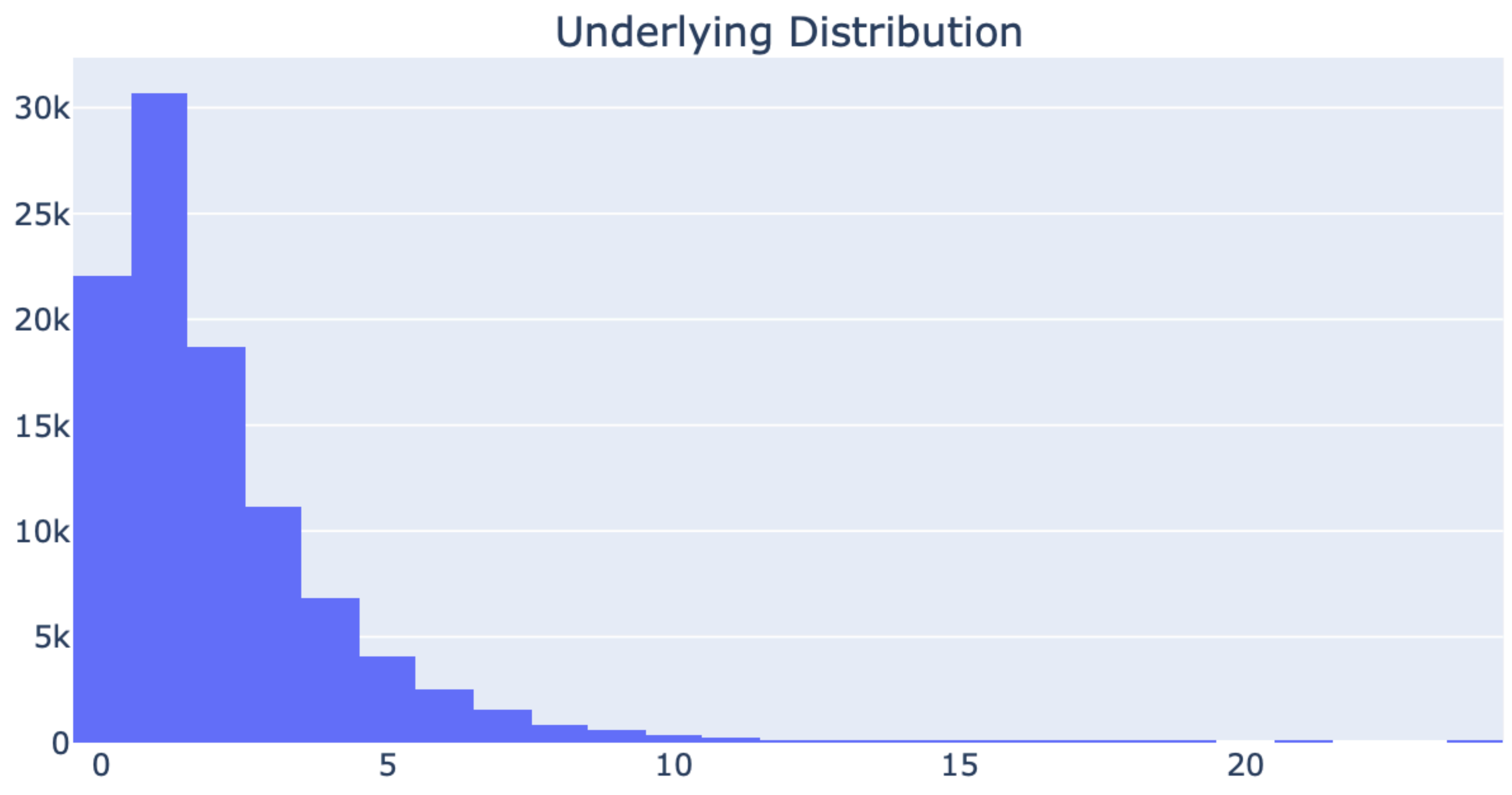
https://cltapp.fly.dev/

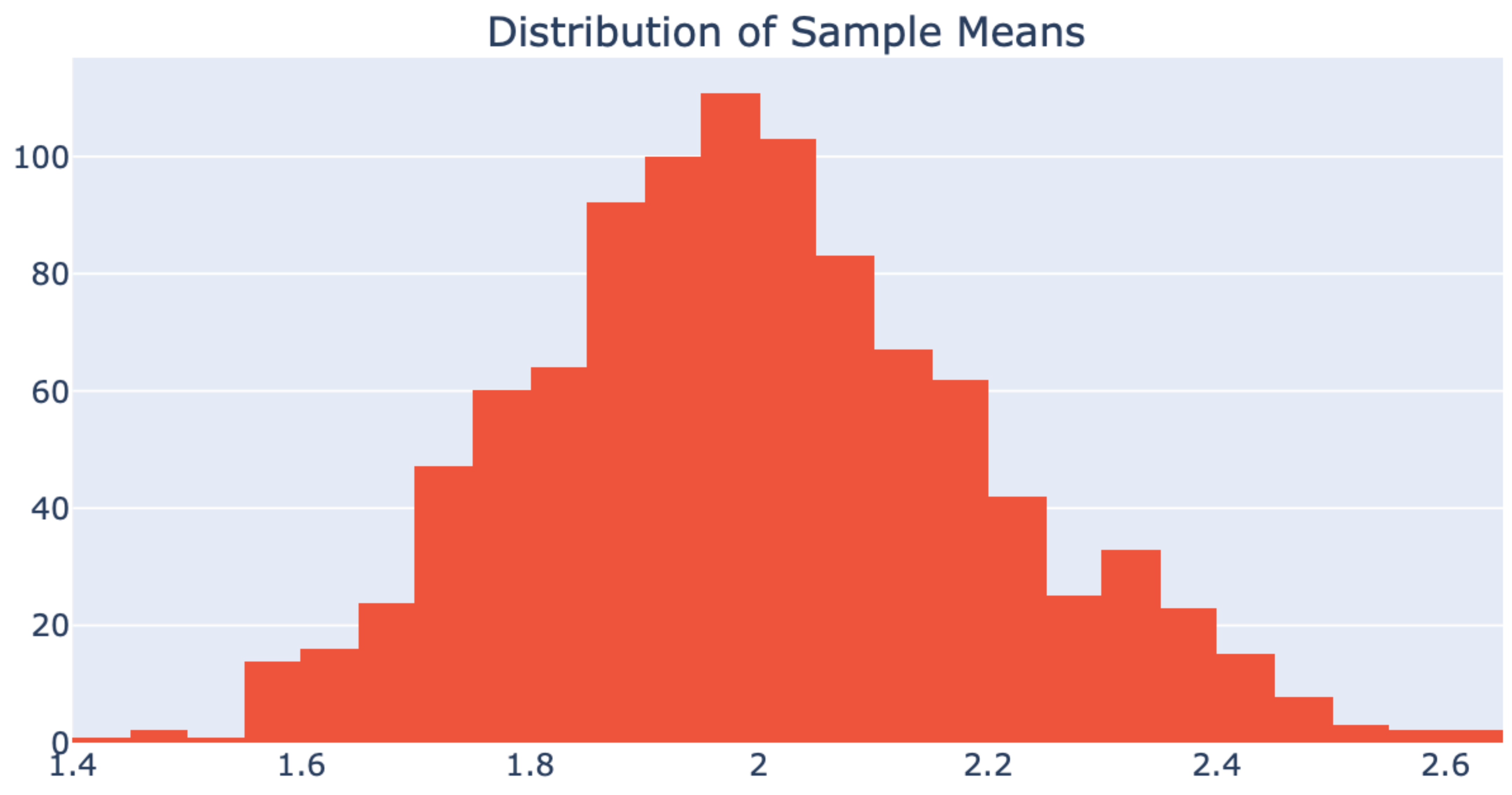
Sample Mean Distribution

Select a distribution and the size of each sample.

Exponential

100





https://cltapp.fly.dev/

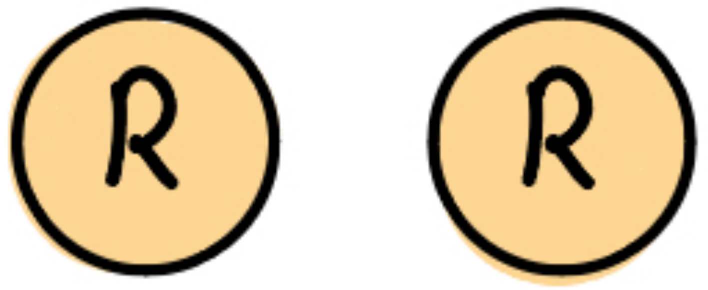
Law of Large Numbers

As the number of trials of a random experiment increases, the average of the results obtained will converge to the expected value.

The larger the sample size, the closer the sample mean will be to the population parameter.

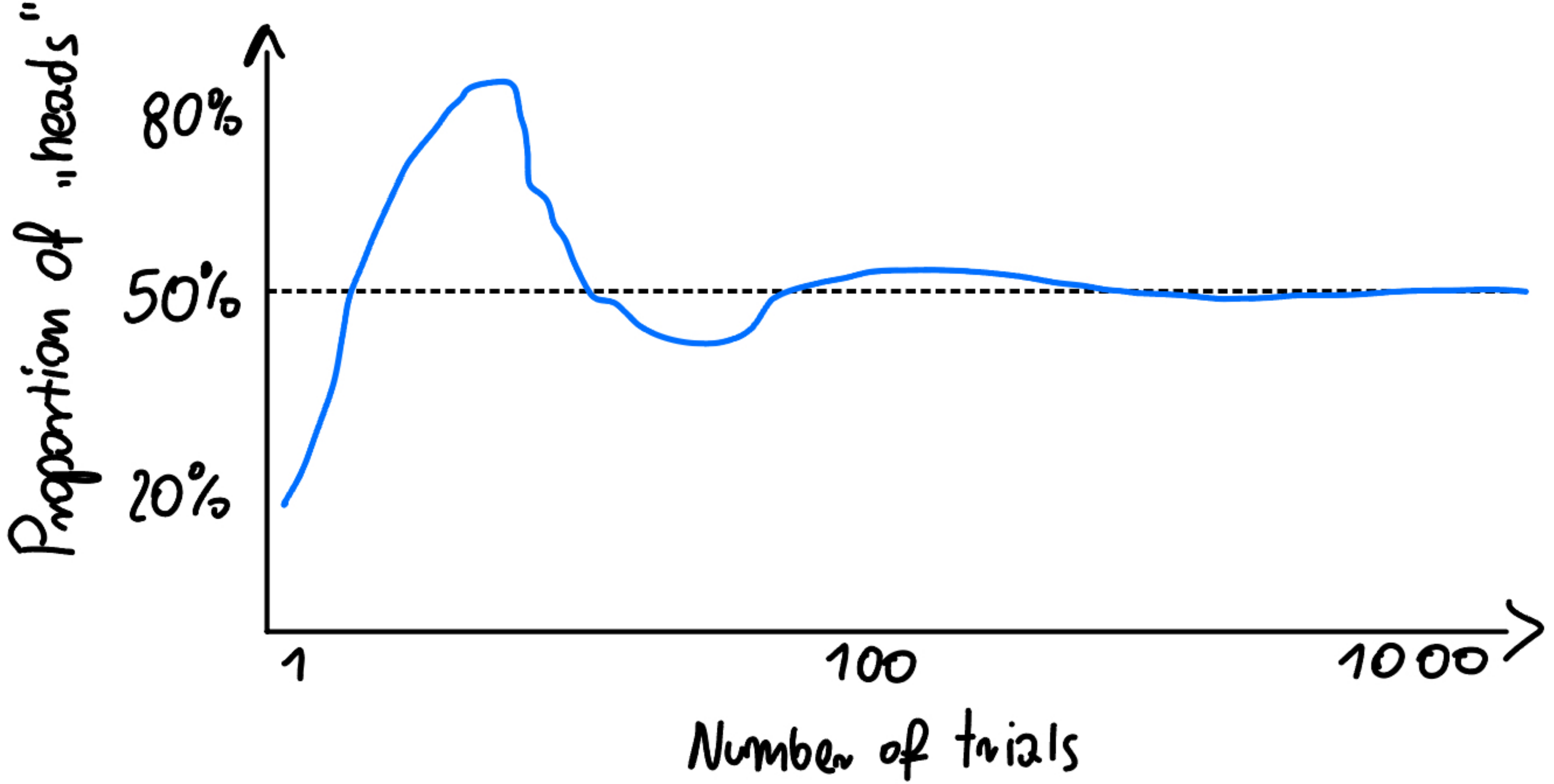
Imagine flipping a coin.

2 throws: R: 100%, O: 0%



3 throws: R: 67%, O: 33%

X throws: R: 50,5%, O: 49,5%



$$\bar{X}_n \longrightarrow \mu \text{ as } n \longrightarrow \infty$$

Sampling

Operating on whole population is either non possible or very exstensive. Thus, we need to extract samples. Our goal is to make unbiased samples which represent population parameters well.

HOW?

- Random Sampling
- Systematic Sampling
- Stratified Sampling
- Cluster Sampling

- Convenience Sampling
- Snowball Sampling


NM
AR
MS
LA
El Paso
Dallas
Fort Worth
Austin
Houston
San Antonio
Corpus Christi
Votes
Percentage
6,375,318
56.2%
4,806,441
42.4%

The greater the accuracy required for the results, the larger the sample size should be.

The higher the confidence level, the larger the sample.

If the population is very diverse, a larger sample is needed to account for all this diversity.

Time, budget, and data availability often limit sample size.

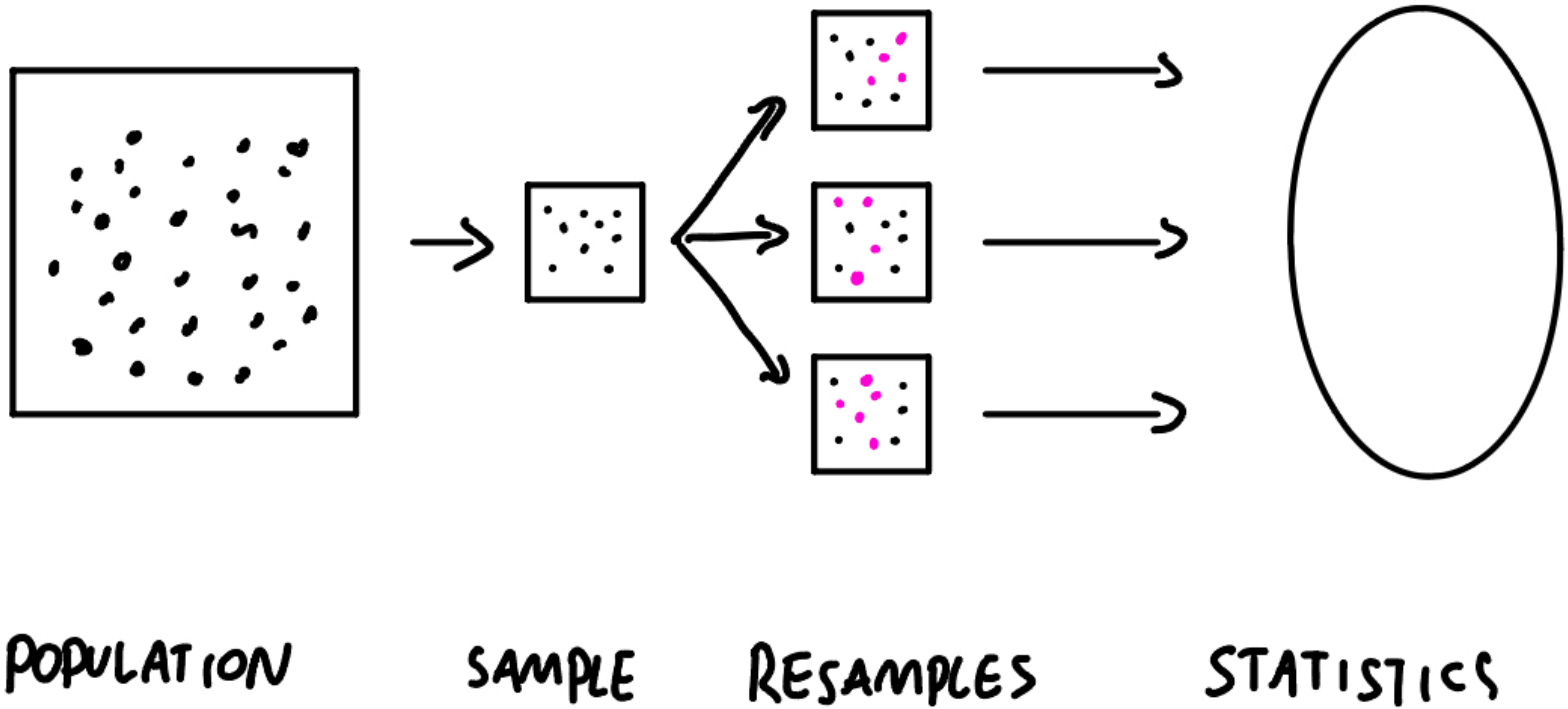
Random and systematic sampling implementations on GitHub.

Resampling

Resampling methods are useful when training models on limited datasets. These techniques involve repeatedly drawing samples from the original data to create multiple training sets, helping to improve model performance and reduce overfitting.

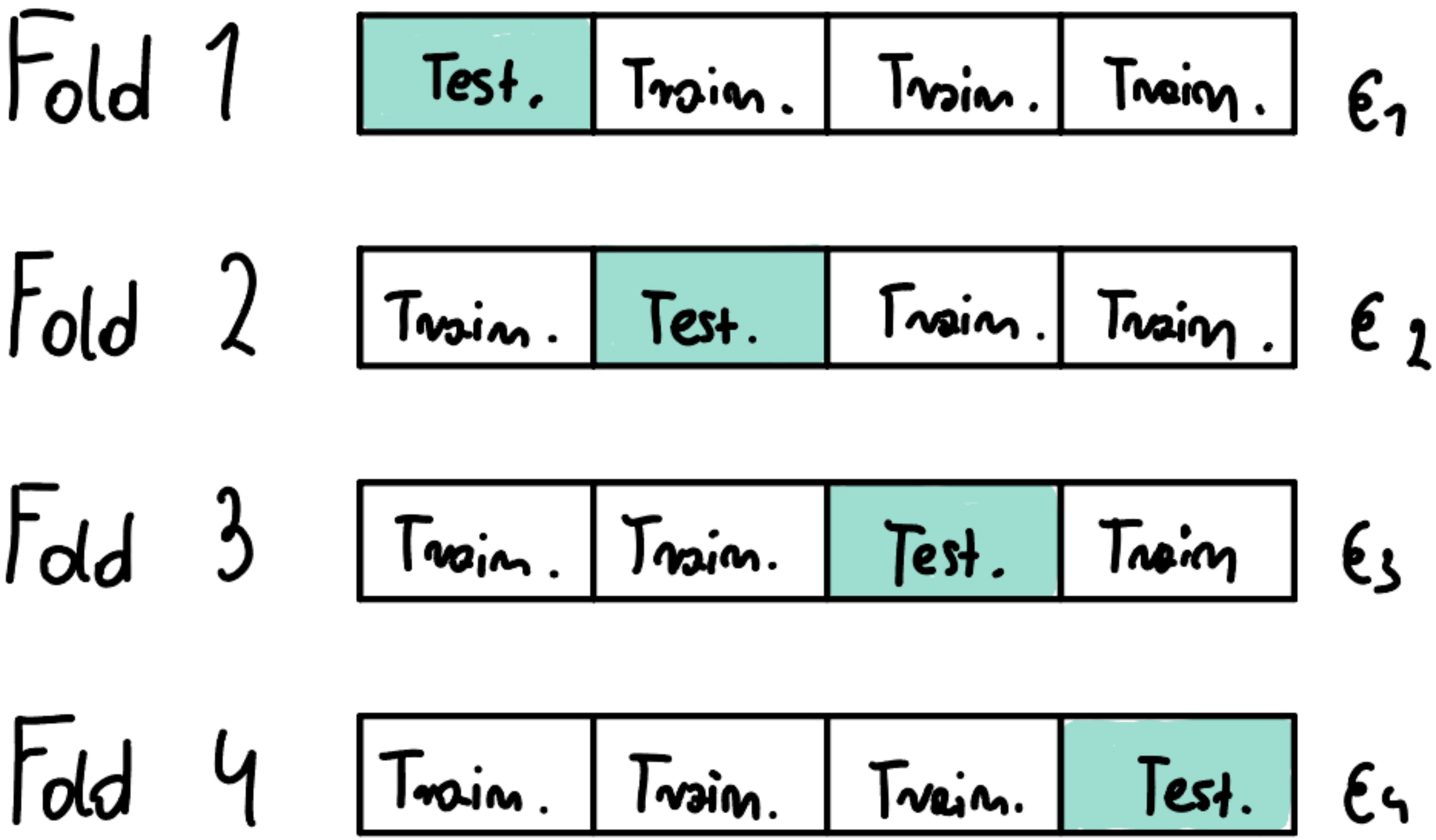
Bootstraping

Bootstrapping is a resampling technique where multiple samples are drawn with replacement from the original dataset. Each resample is the same size as the original dataset. The statistic of interest is calculated for each resample, and the distribution of these calculated statistics is used to estimate the sampling distribution of the original statistic.


POPULATION
SAMPLE
RESAMPLES
STATISTICS

Cross Validation

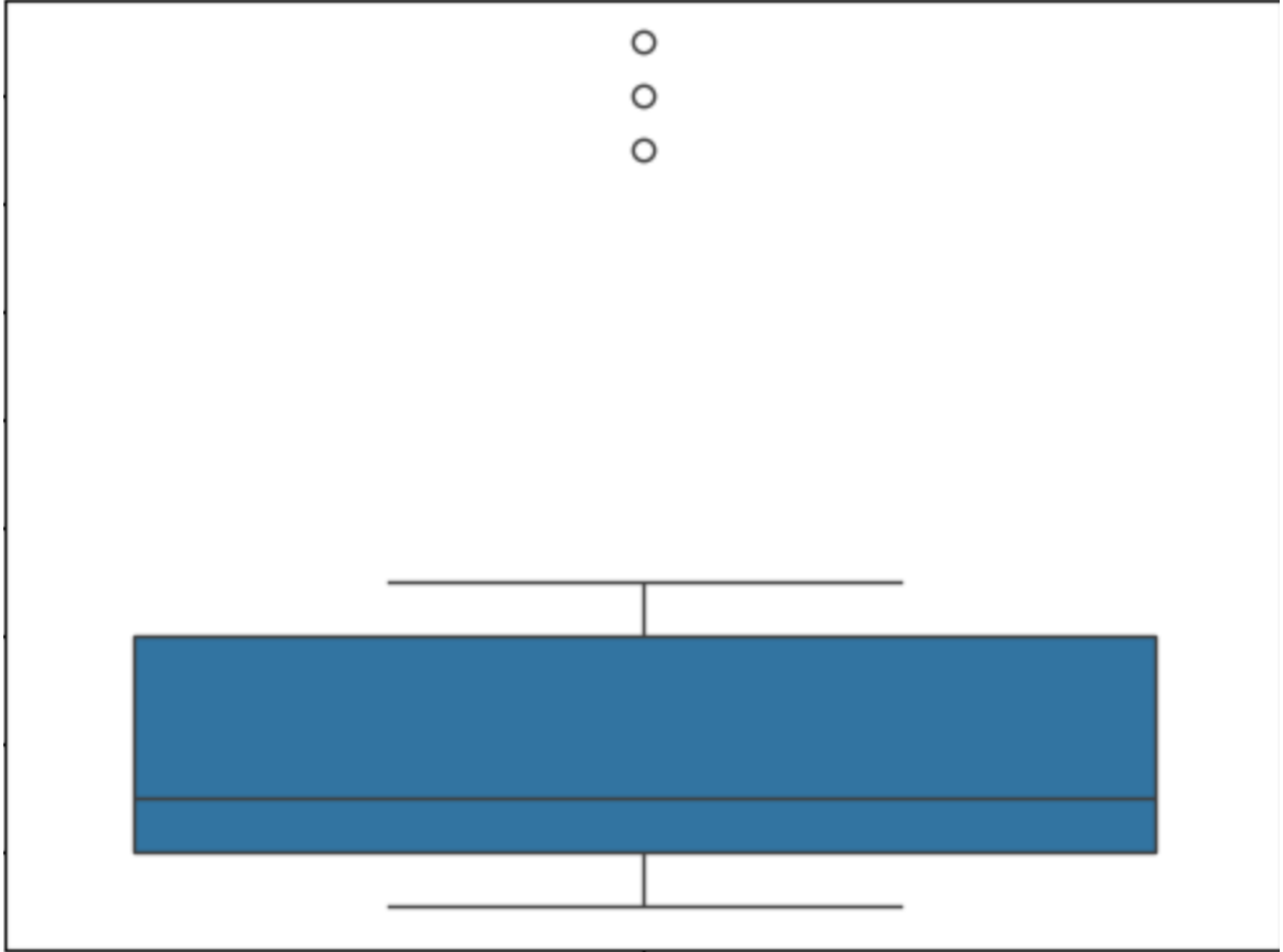
Cross-validation is a method for evaluating machine learning models. By dividing the dataset into multiple folds, we can train and test the model on different subsets of data. This helps mitigate overfitting and provides a more reliable estimate of the model's generalization performance.


Fold 1
Test.
Train.
Train.
Train.
ϵ_1
Fold 2
Train.
Test.
Train.
Train.
ϵ_2
Fold 3
Train.
Train.
Test.
Train
ϵ_3
Fold 4
Train.
Train.
Train.
Test.
ϵ_4

Outliers

to drop or not to drop

An outlier is data that deviates from the rest significantly. Many statistics (e.g. mean) are very sensitive for outstanding data. Thus, as Data Scientists, we have to take care of it to prevent inaccuracy. But remember, removing outliers is legitimate only for specific reasons.



If the outlier is

… a natural part of the population you are studying, you should not remove it.

… a measurement error or data entry error, correct the error if possible. If you can’t fix it, remove that observation because you know it’s incorrect.

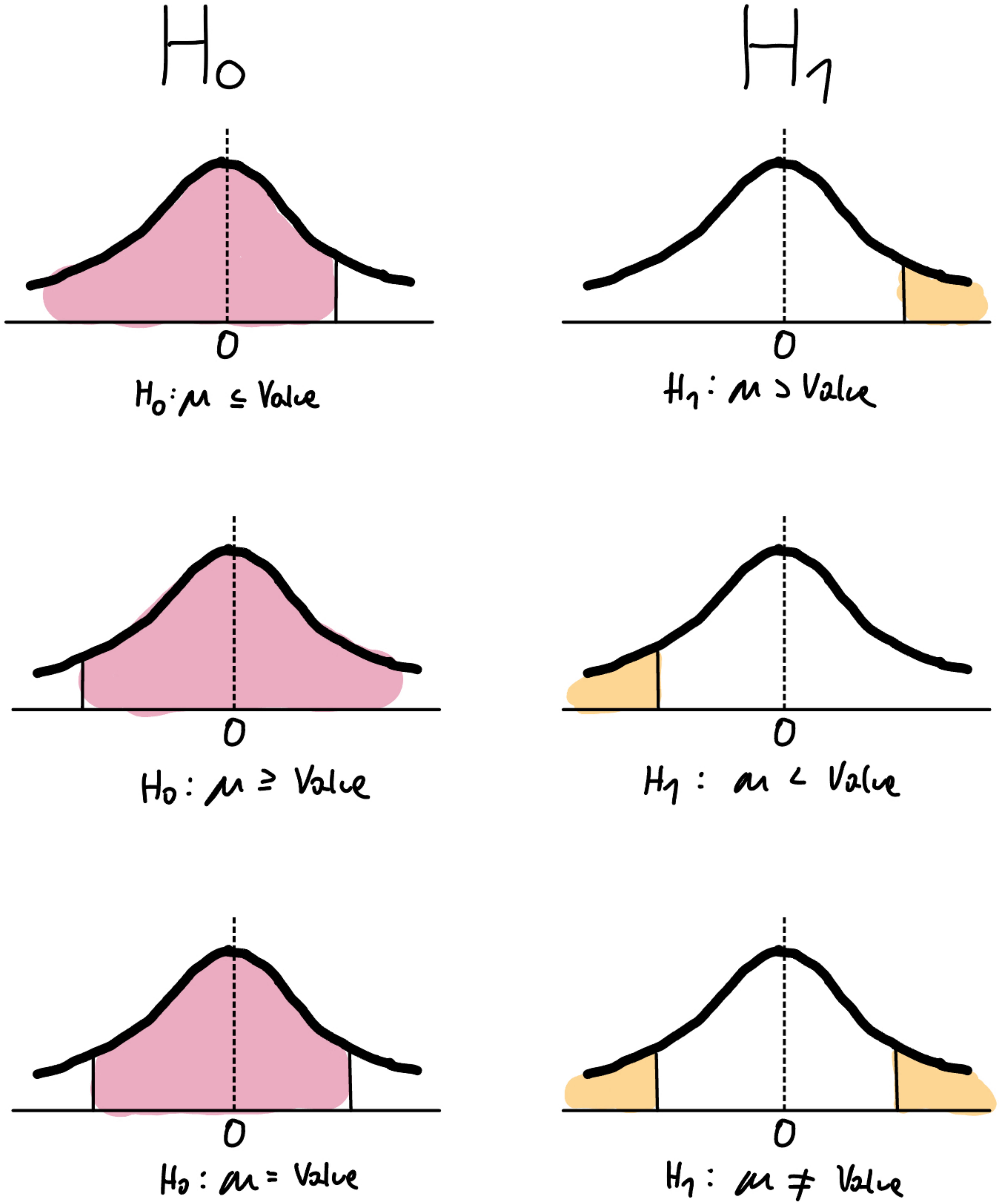
… not a part of the population you are studying (i.e., unusual properties or conditions), you can legitimately remove the outlier.

Industry knowledge can be very useful

Hypothesis testing

Statistical significance is a concept used in data analysis and statistics to determine whether observed results are likely due to chance or reflect a real relationship between variables. Simply put, statistical significance helps us assess how confident we can be that an effect observed in a sample also exists in the broader population.

- Null hypothesis (H_0) – assumes that there is no difference or relationship between the variables
- Alternative hypothesis (H_1) – assumes that there is a real difference or relationship between the variables.
- Significance level (α) – this is the threshold set by the researcher before analysis, usually at 0.05 (5%). It means that if the probability of the result happening by chance is less than 5%, we reject the null hypothesis and consider the result statistically significant.
- p-value – this is a measure that tells us how likely it is that the observed result could have occurred by chance.


H0
H1
0
H0 : μ ≤ Value
0
H1 : μ > Value
0
H0 : μ ≥ Value
0
H1 : μ < Value
0
H0 : μ = Value
0
H1 : μ ≠ Value

Example (The data are examples and not true)

- Null hypothesis (H_0) - no difference between employment on Universities
- Alternative hypothesis (H_1) – there is difference between employment on Universities
- Significance level (α) – 0.05

PUT		U2	
empdyed	nonemploяed	empdyed	nonemploяed
1578	200	600	1200
89%		30 %	

$p = 0{,}03$

$p < \alpha$, we reject H_0

PUT		U2	
empdyed	nonemploяed	empdyed	nonemploяed
1200	70	1600	150
94%		91%	

$p = 0{,}9$

$p > \alpha$, H_0 is good

Standard verification procedure:

- Formulate the null hypothesis
  and the alternative hypothesis
- Select the appropriate test statistic
- Set the significance level
- Determine the critical region
- Calculate the test
- Make decision

Correlation

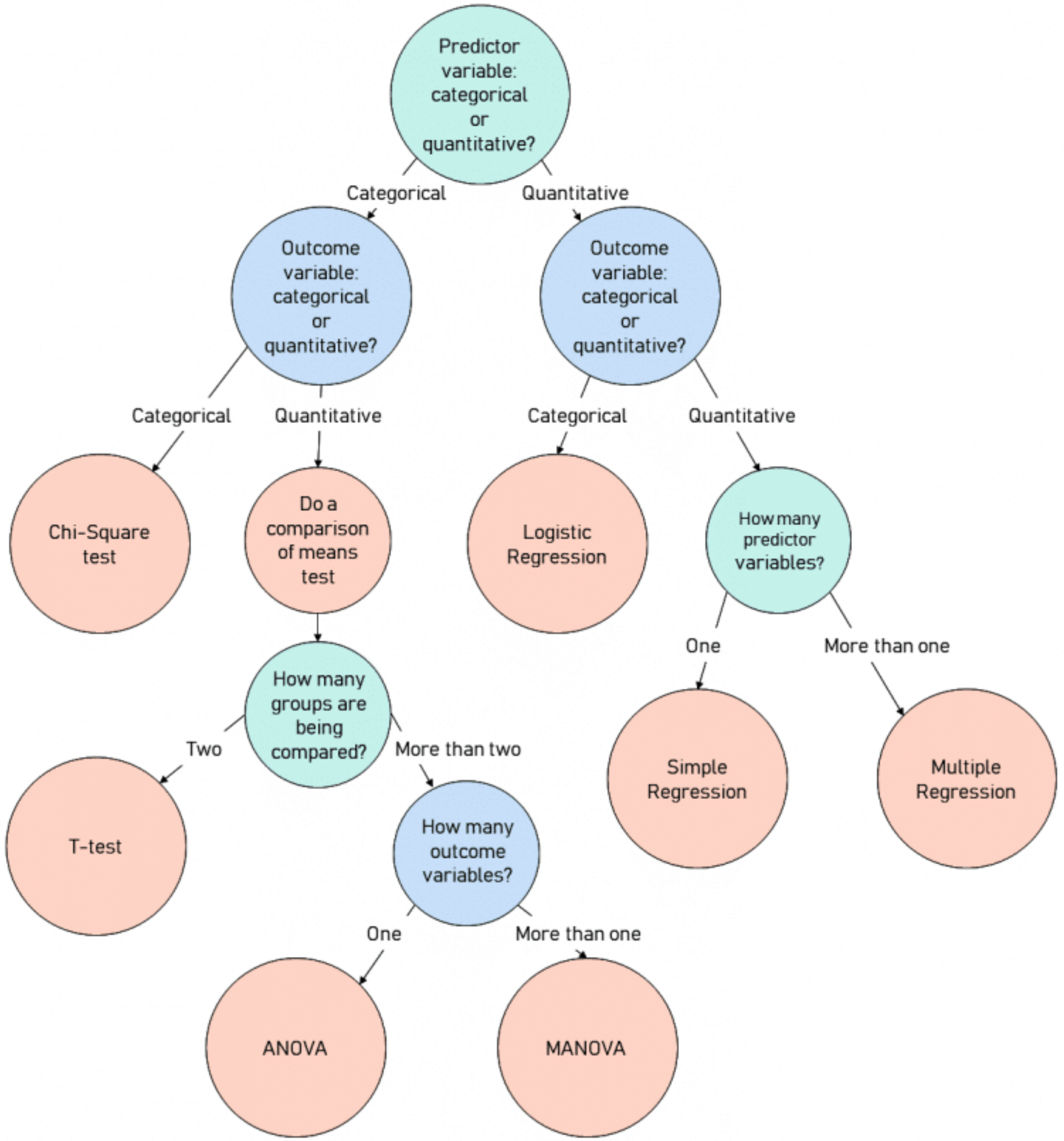
Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate)

Standard Deviation

Measure that quantifies the amount of variation or dispersion in a set of data values. It indicates how much the individual data points deviate from the mean (average) value of the dataset

Parametric tests

Choosing a statistical test



https://www.scribbr.com/statistics/statistical-tests/

Normality test - Shapiro-Wilk

Normality tests are statistical procedures used to assess whether a data set is well-modeled by a normal distribution and to evaluate how likely it is that the underlying random variable follows a normal distribution. These tests are important because many statistical methods, assume that the data are normally distributed.

T test

Used when we want to compare a sample mean with a population mean. A one-sample t-test examines whether the mean of a sample is statistically different from a known or hypothesized population mean

ANOVA, MANOVA

ANOVA is a powerful statistical method used to assess whether there are significant differences between the means of various groups. It is particularly useful when analyzing data across multiple populations influenced by one or more factors simultaneously. By examining variance within and between groups, ANOVA isolates sources of variability and identifies whether specific factors contribute to observed differences among group means.

TYPE	NUM. OF DEPENDENT VARIABLES	NUM. OF INDEPENDENT VARIABLES	INDEPENDENT/ CORRELATED SAMPLES
ONE - WAY ANOVA	ONE	ONE	INDEPENDENT
TWO WAY ANOVA	ONE	MORE THAN ONE	INDEPENDENT
MANOVA	MORE THAN ONE	ONE	CORRELATED
REPEATED MEASURE ANOVA	ONE	MORE THAN ONE	INDEPENDENT

Nonparametric tests

Chi-Square

The Chi-square test is a non-parametric test used to determine if there is a significant difference between observed and expected frequencies in categorical data. It is commonly applied in hypothesis testing to see if the distribution of data aligns with expectations under the null hypothesis. By comparing observed values against expected values, the test quantifies the discrepancies and helps evaluate whether they are due to random variation or if they suggest a meaningful association or pattern.

And a lot more...

Live coding!