

Session 15

Introduction to Mathematics

2. Statistics

Prepared by/ Elzahraa Alaa Tag Eldein

July 3, 2023

Task 1

How to use over-sampling and under-sampling to solve imbalanced datasets?

In your quest to analyze datasets or building machine Learning models, you probably uncounted a situation where you have imbalanced dataset, which means that class label is very imbalanced whereby one class may have abnormally high number of data samples where as another class label will have significantly lower number of data samples.

For example if you have a dataset where you are trying to predict whether a student will pass or not pass an exam on the basis of several parameters.

If there are 2000 students will pass and 200 will not. This mean that students will pass are 10 times higher magnitude than others will not pass. So, this is called imbalanced dataset.

In python, there is a library which include various methods can deal with this problem. Its name is imbalanced-learn.

URL: <https://imbalanced-learn.org/stable/>


Task 2

When to use Mean and when to use Median?

Which is better?

MEAN

$$(5 + 4 + 6 + 10 + 8 + 8 + 7 + 5000 + 1) / 9 = 561$$



| |
|------|
| 5 |
| 4 |
| 6 |
| 10 |
| 8 |
| 8 |
| 7 |
| 5000 |
| 1 |

Mean: 561

Median: 7

Should I
eliminate it?

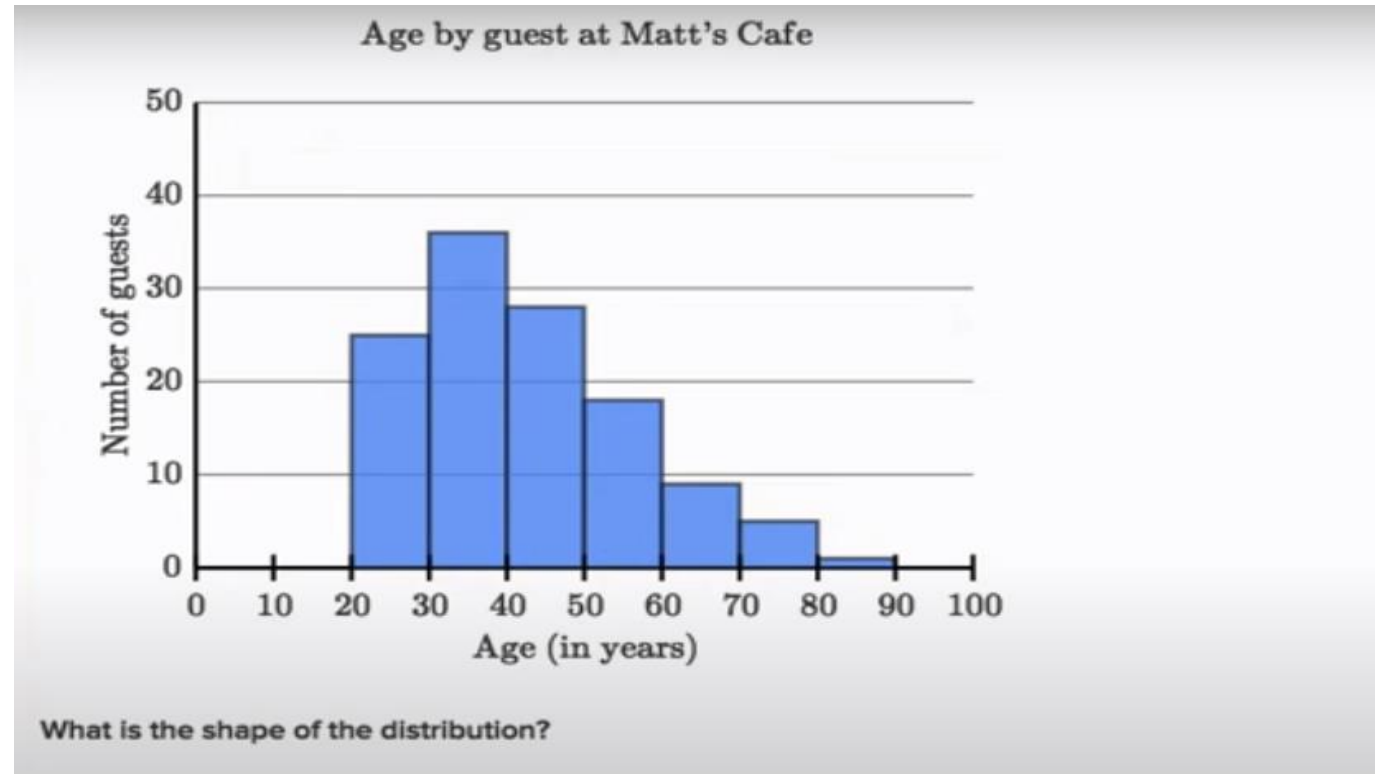


Conclusion: Use them together. Investigate if there is a big difference between them, then decide which of them represent the central tendency for your dataset

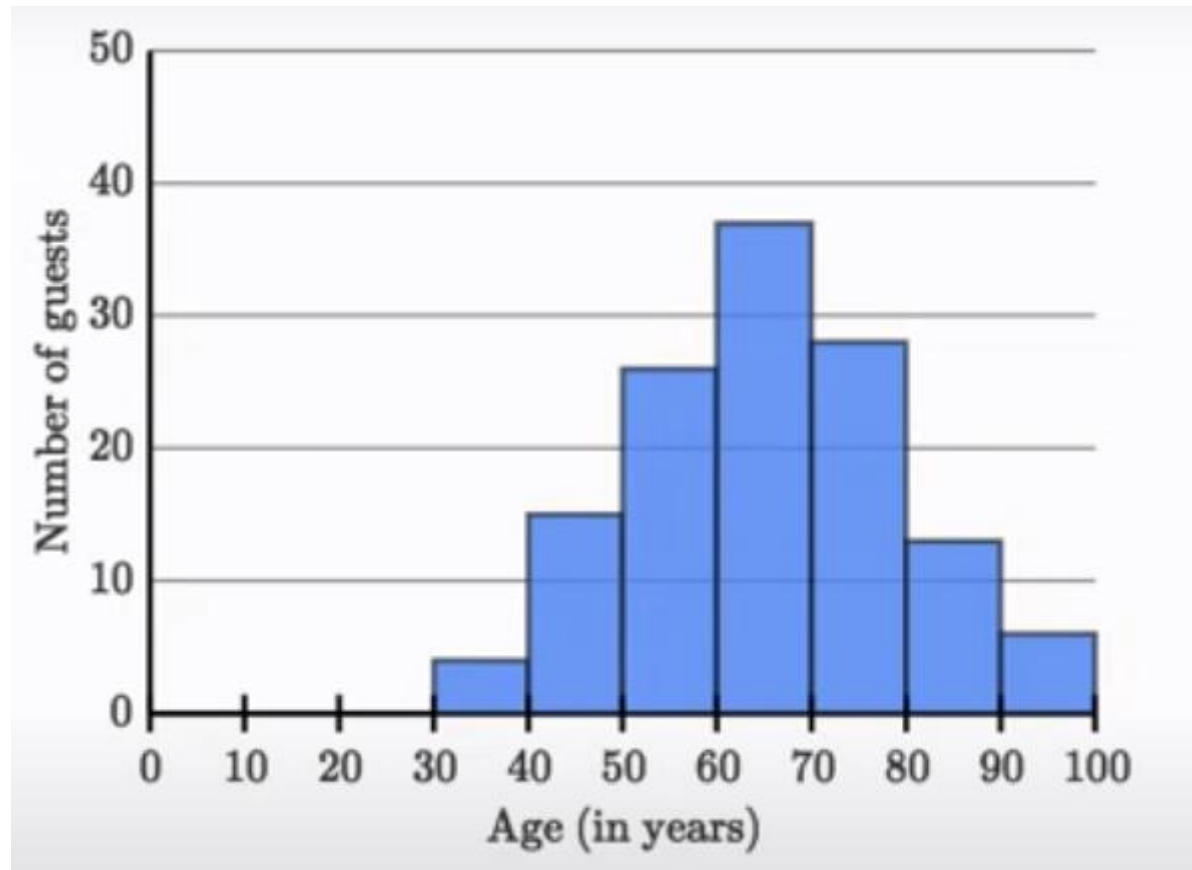
Task 3

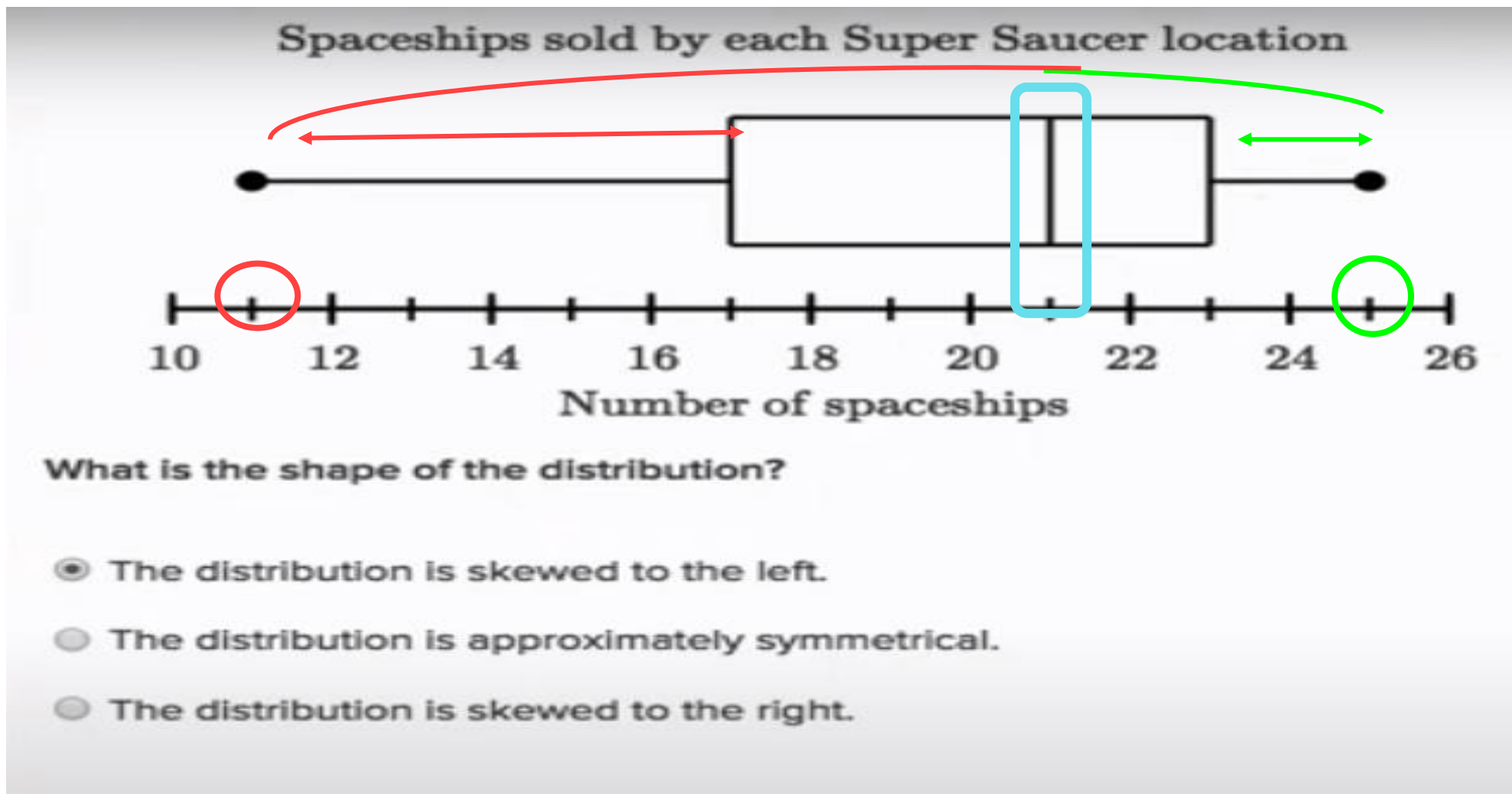
types of data distribution
What is famous in each field?

the distribution is right-tailed



the distribution is Approximately symmetrical

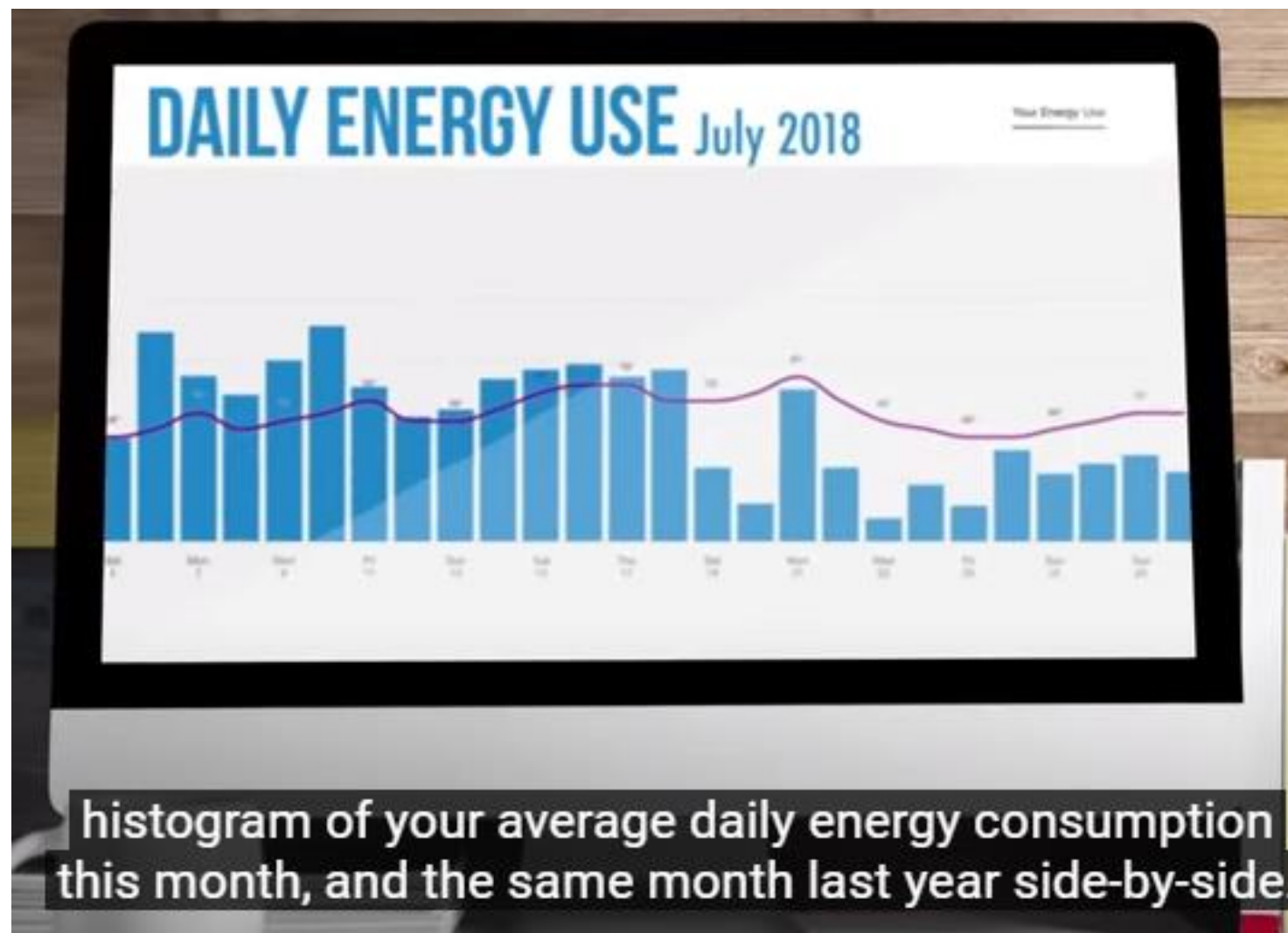




Uniform distribution

LIKE A HISTOGRAM, THE DISTRIBUTION TELLS US
ABOUT THE SHAPE AND SPREAD OF DATA.

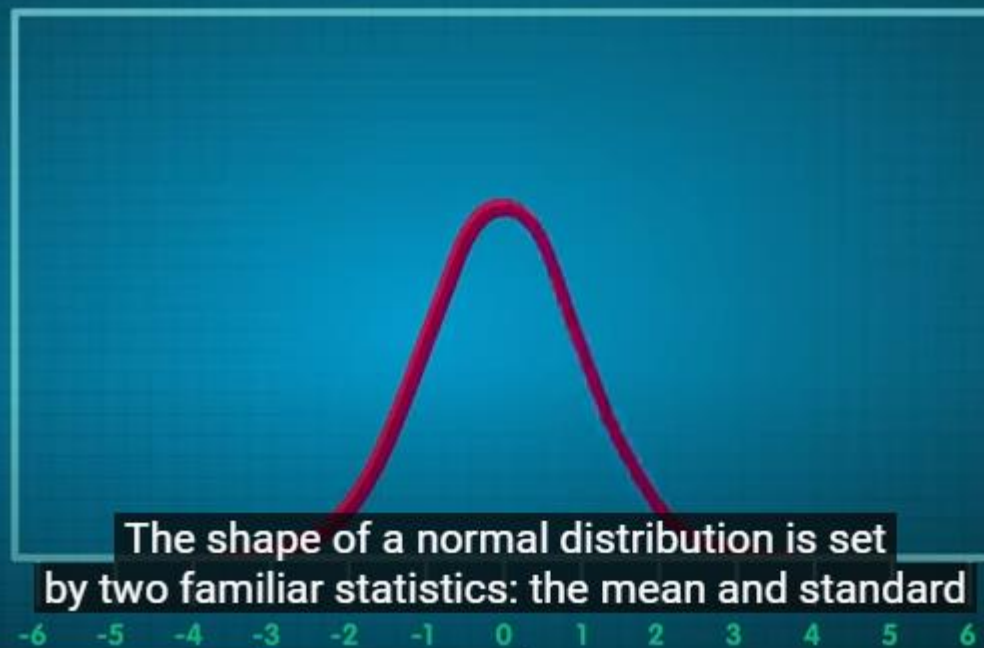




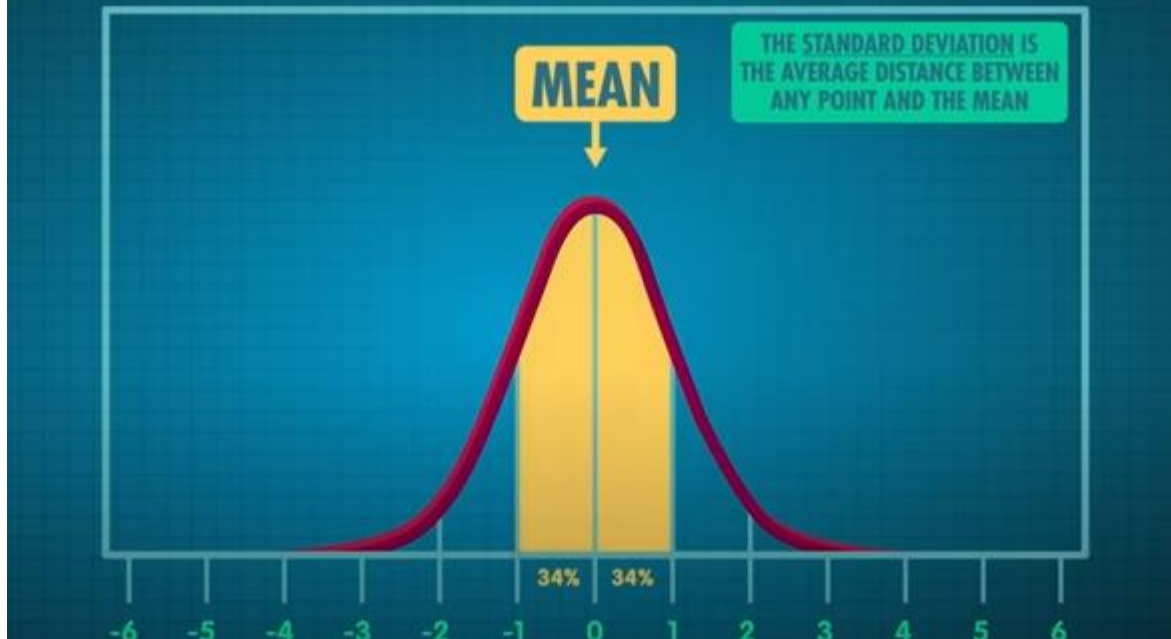
CrashCourse ✓

14.8M subscribers

NORMAL DISTRIBUTION CURVE



NORMAL DISTRIBUTION CURVE



BOXPLOT

NORMALLY DISTRIBUTED DATA



CrashCourse ✓

14.8M subscribers

BOXPLOT

NORMALLY DISTRIBUTED DATA



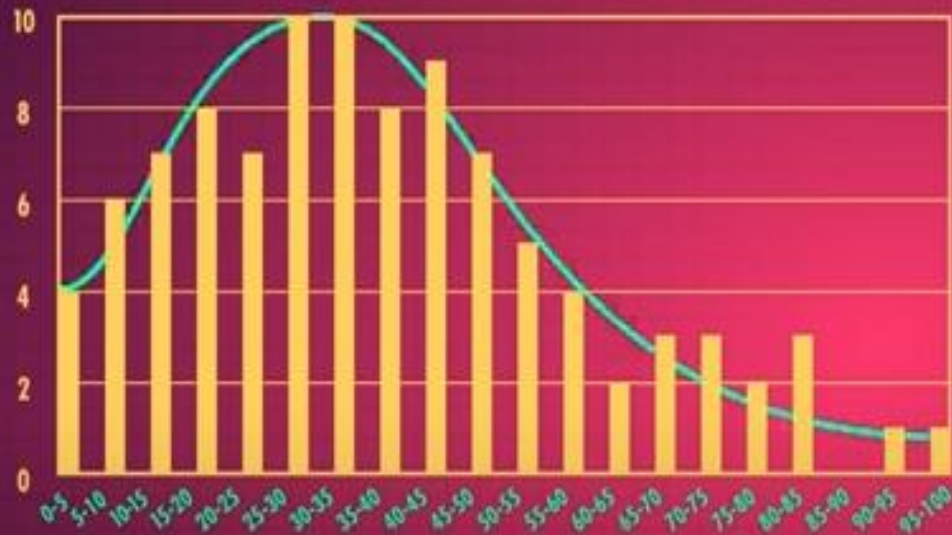
POSITIVE SKEW



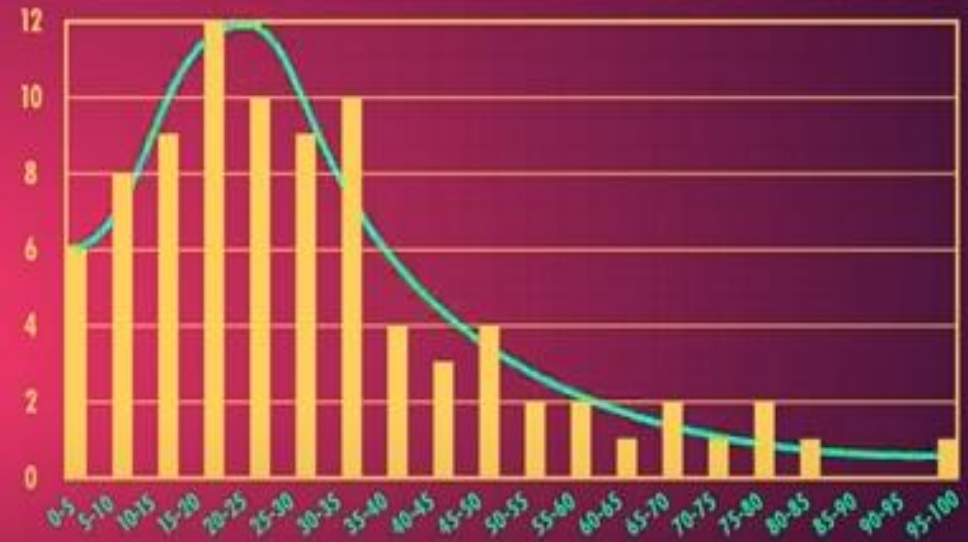
NEGATIVE SKEW



"WHO'S THAT POKÉMON?" TEST SCORES



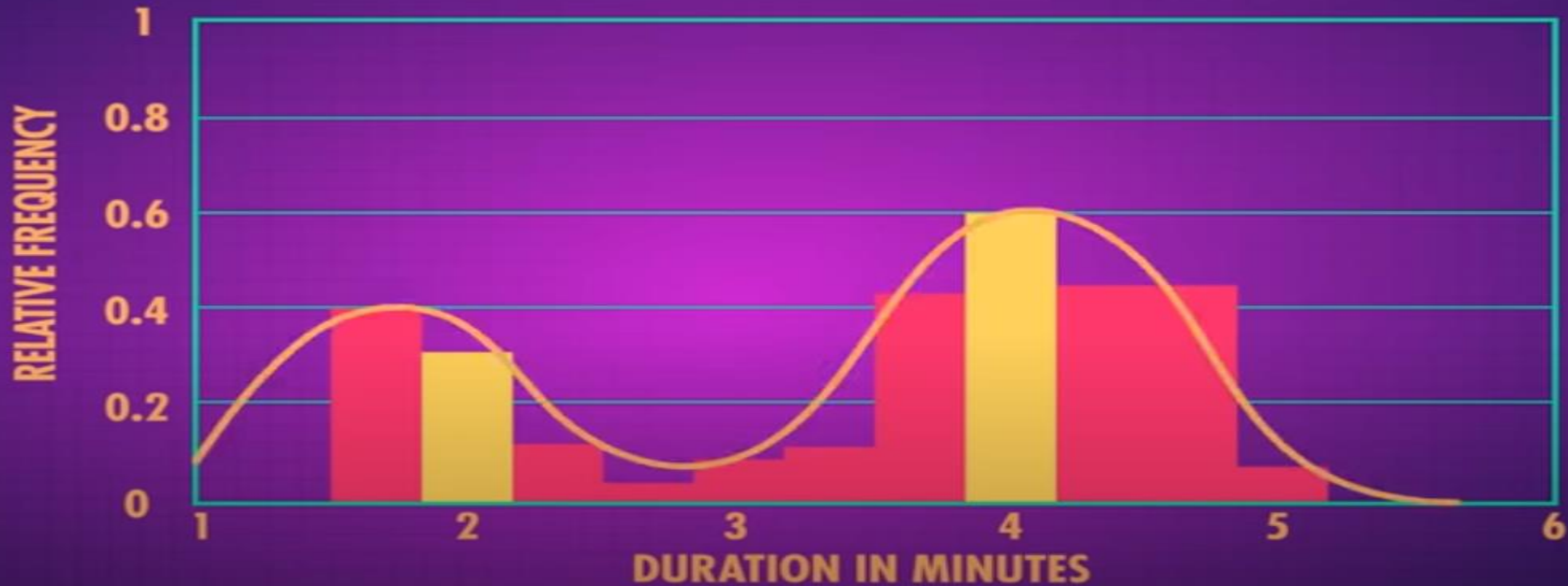
"NAME ALL COUNTRIES" TEST SCORES



of all these two samples look pretty similar both have a right skew



ERUPTIONS OF OLD FAITHFUL GEYSER



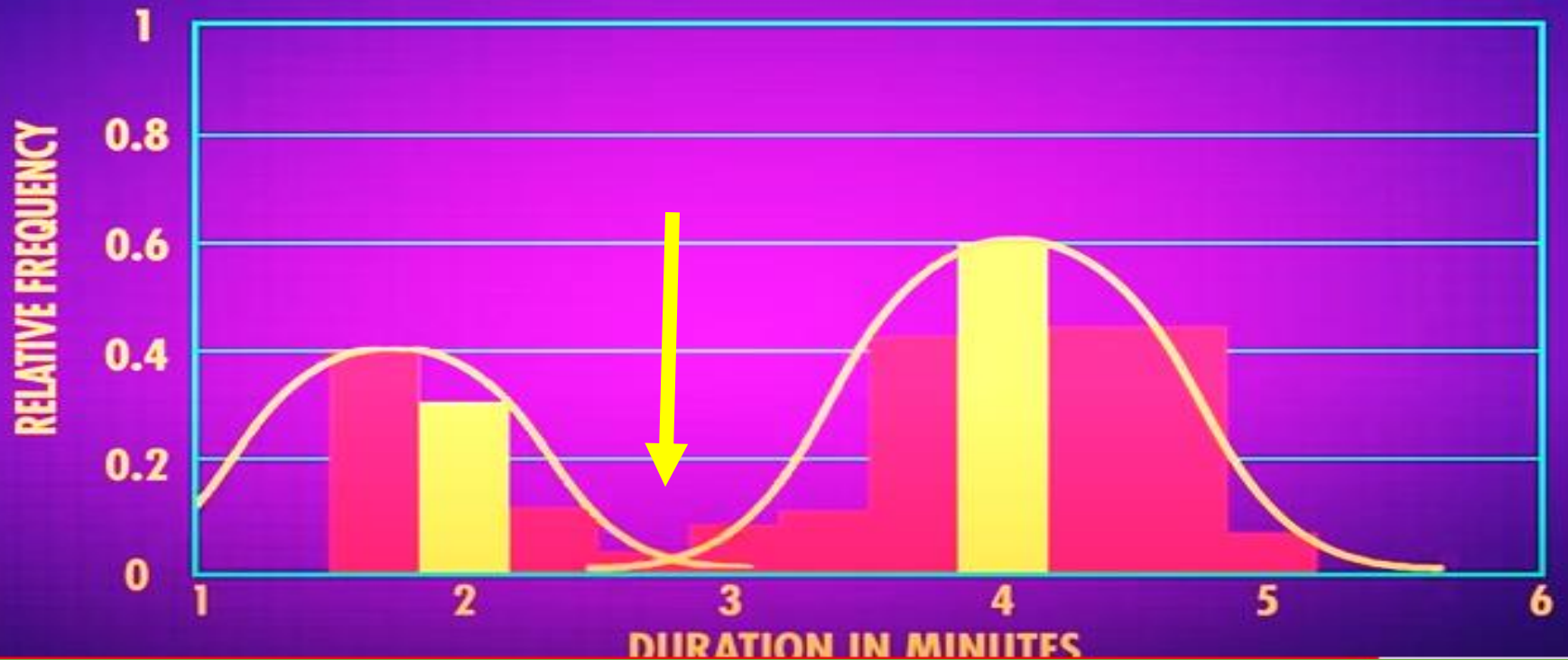
it looks like come from one distribution with two pumps



CrashCourse ✓

14.8M subscribers

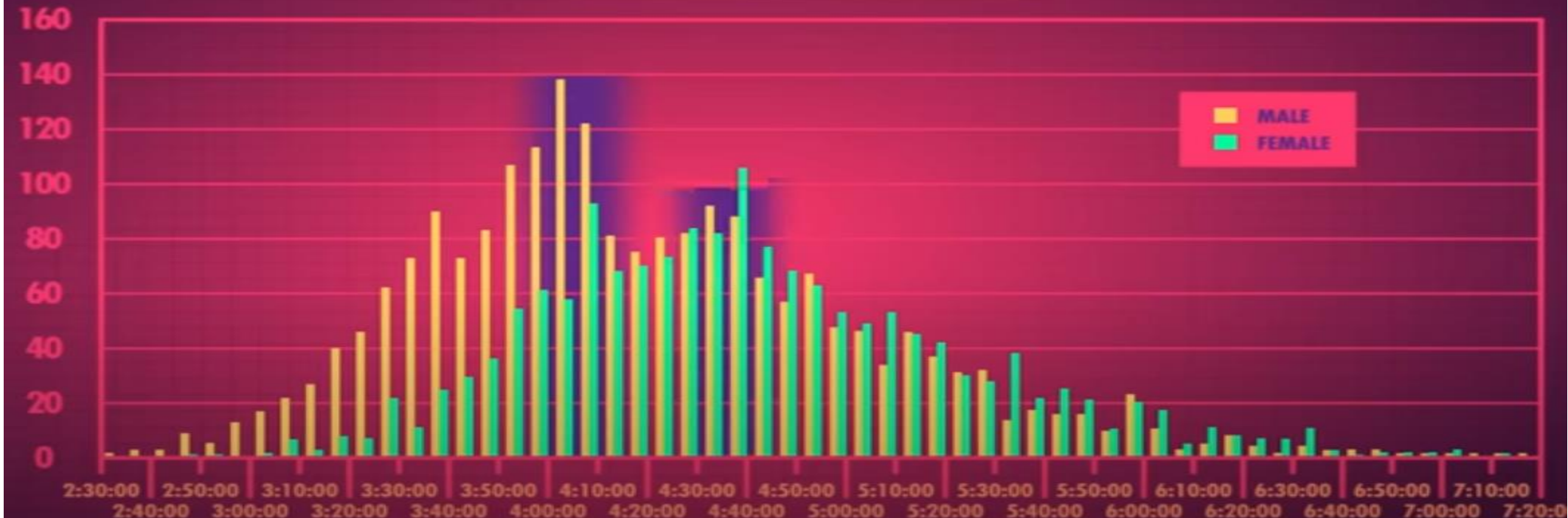
ERUPTIONS OF OLD FAITHFUL GEYSER



They look like two unimodal distributions being measured in the same time

2014 BMO VANCOUVER MARATHON

FINISHING TIMES



They are two unimodal distributions being measured in the same time, but are very close



CrashCourse ✓

14.8M subscribers

TYPES OF PROBABILITY DISTRIBUTIONS IN **MACHINE LEARNING**



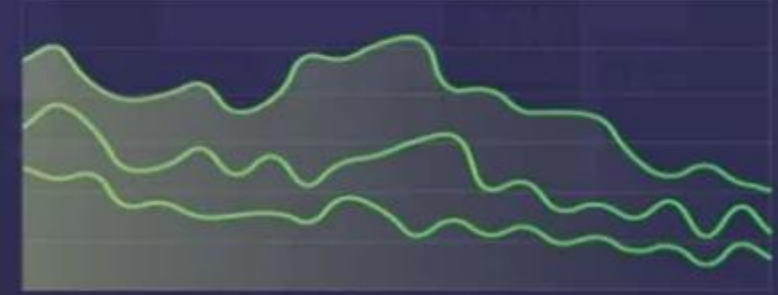
Uniform Distribution



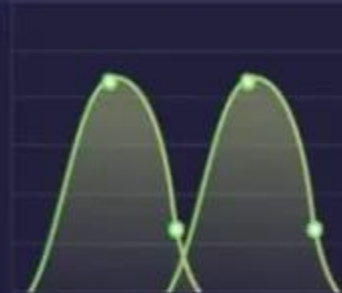
Binomial Distribution



Bernoulli Distribution



Poisson Distribution



Normal Distribution



T-test Distribution



Exponential Distribution

Whether you're guessing if it's going to rain tomorrow, betting on a sports team to win an away match, framing a policy for an insurance company, or simply trying your luck on blackjack at the casino, probability and distributions come into action in all aspects of life to determine the likelihood of events.

Having a sound **statistical** background can be incredibly beneficial in the daily life of a data scientist. Probability is one of the main building blocks of data science and machine learning. While the concept of probability gives us mathematical calculations, statistical distributions help us visualize what's happening underneath.

Having a good grip on statistical distribution makes exploring a new dataset and finding patterns within a lot easier. It helps us choose the appropriate machine learning model to fit our data on and speeds up the overall process.

Probability

A measure of the likelihood of an event's occurrence.



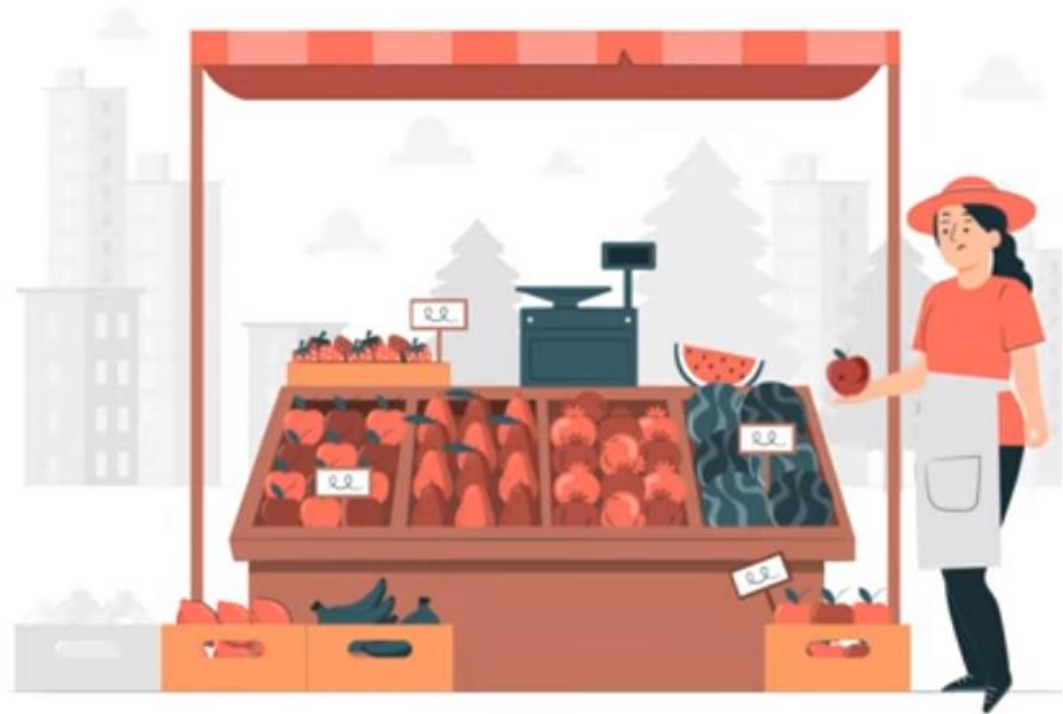
0.6 percent probability
of raining.

Note that

Probability is always
between 0 to 1.

| Days | Rotten apples | Total apples | Probability (X) |
|------|---------------|--------------|-----------------|
| 1 | 3 | 24 | 0.125 |
| 2 | 2 | 18 | 0.111 |
| 3 | 5 | 33 | 0.151 |
| 4 | 1 | 26 | 0.038 |
| 5 | 4 | 32 | 0.125 |
| 6 | 3 | 37 | 0.111 |
| 7 | 6 | 38 | 0.157 |

$$P(X) = \frac{\text{Number of rotten apples}}{\text{Total number of apples}}$$



Probability of picking a rotten apple

| Days | Rotten apples | Total apples | Probability (X) |
|------|---------------|--------------|-----------------|
| 1 | 3 | 24 | 0.125 |
| 2 | 2 | 18 | 0.111 |
| 3 | 5 | 33 | 0.151 |
| 4 | 1 | 26 | 0.038 |
| 5 | 4 | 32 | 0.125 |
| 6 | 3 | 37 | 0.111 |
| 7 | 6 | 38 | 0.157 |

$$P(X) = \frac{\text{Number of rotten apples}}{\text{Total number of apples}}$$

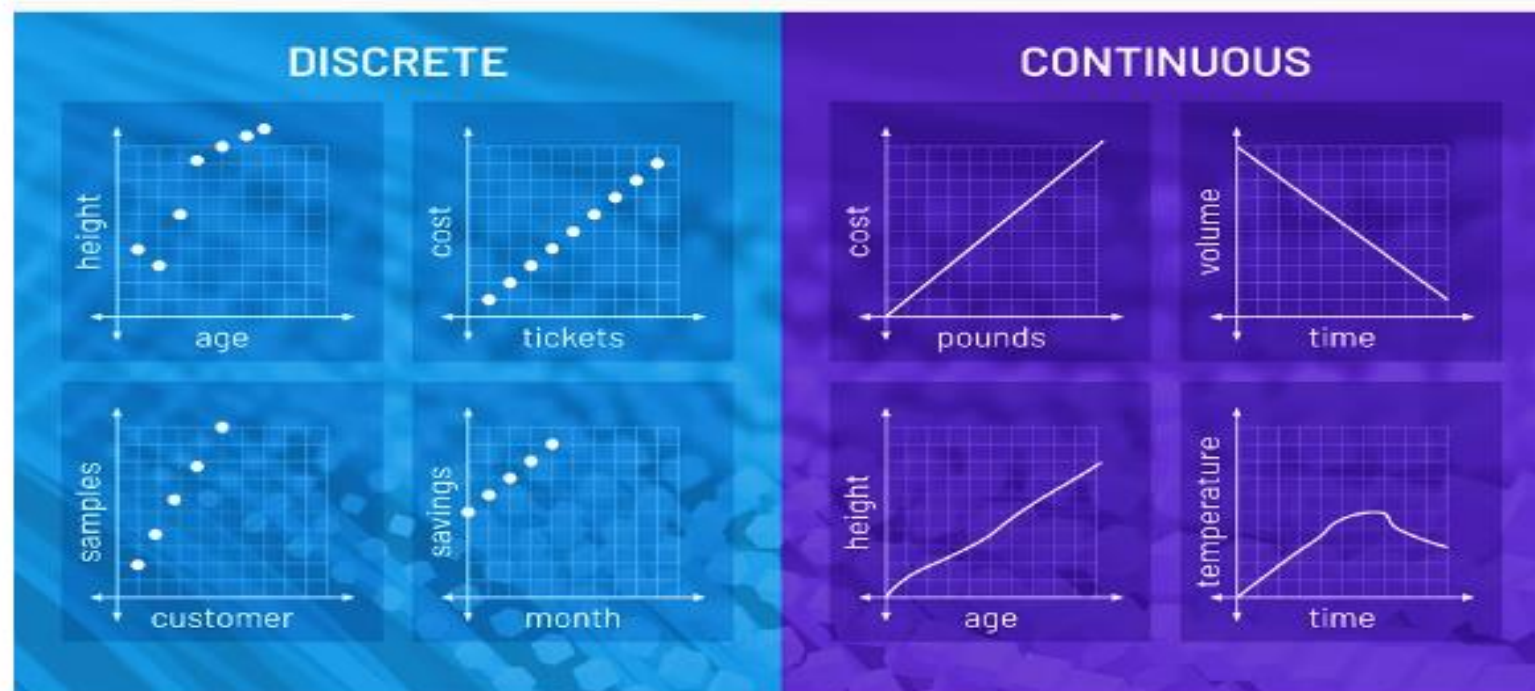


Common types of data

Explaining various distributions becomes more manageable if we are familiar with the type of data they use. We encounter two different outcomes in day-to-day experiments: finite and infinite outcomes.



Discrete vs Continuous Data



Difference between Discrete and Continuous Data (Source)

Types of statistical distributions

Depending on the type of data we use, we have grouped distributions into two categories, discrete distributions for discrete data (finite outcomes) and continuous distributions for continuous data (infinite outcomes).

Discrete distributions

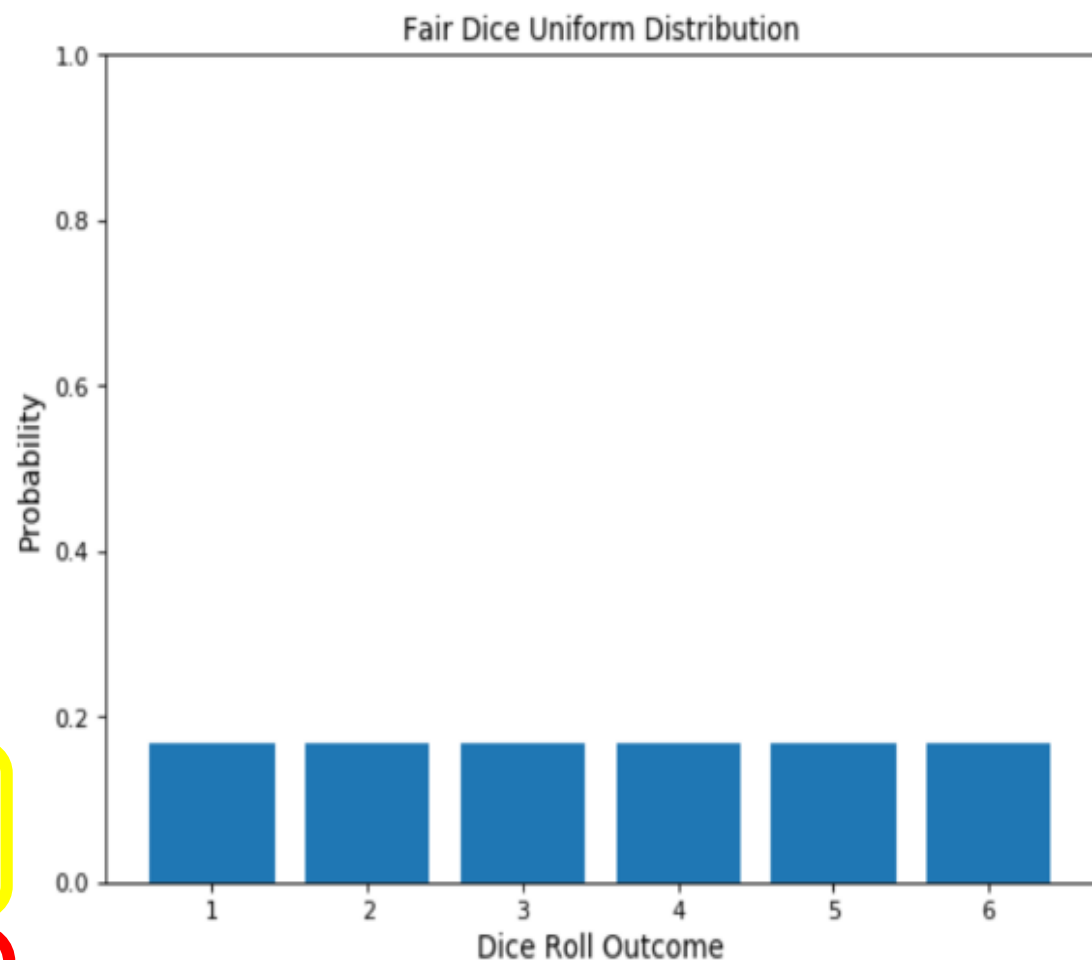
Discrete uniform distribution: All outcomes are equally likely

In statistics, uniform distribution refers to a statistical distribution in which all outcomes are equally likely. Consider rolling a six-sided die. You have an equal probability of obtaining all six numbers on your next roll, i.e., obtaining precisely one of 1, 2, 3, 4, 5, or 6, equaling a probability of $1/6$, hence an example of a discrete uniform distribution.

As a result, the uniform distribution graph contains bars of equal height representing each outcome. In our example, the height is a probability of $1/6$ (0.166667).

Uniform distribution is represented by the function $U(a, b)$, where a and b represent the starting and ending values, respectively. Similar to a discrete uniform distribution, there is a continuous uniform distribution for continuous variables.

The drawbacks of this distribution are that it often provides us with no relevant information. Using our example of a rolling die, we get the expected value of 3.5, which gives us no accurate intuition since there is no such thing as half a number on a dice. Since all values are equally likely, it gives us no real predictive power.



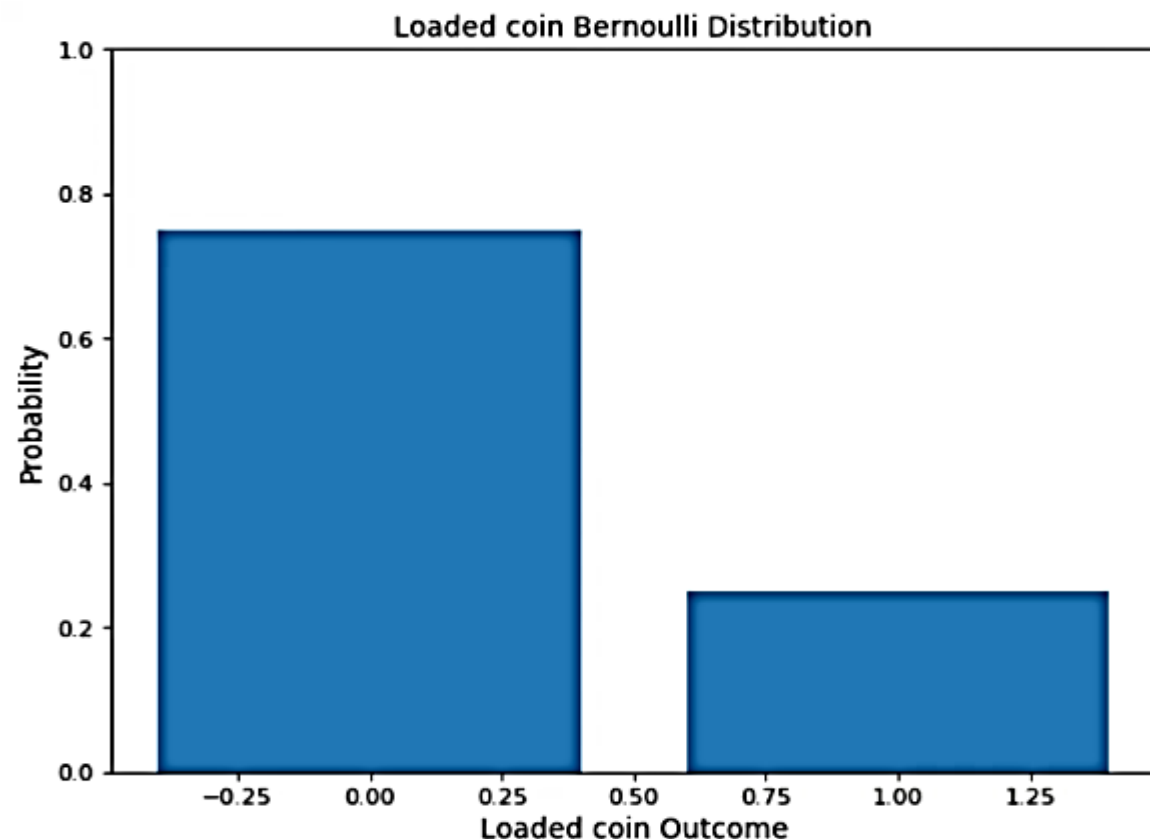
Bernoulli Distribution: Single-trial with two possible outcomes

The Bernoulli distribution is one of the easiest distributions to understand. It can be used as a starting point to derive more complex distributions. Any event with a single trial and only two outcomes follows a Bernoulli distribution. Flipping a coin or choosing between True and False in a quiz are examples of a Bernoulli distribution.

They have a single trial and only two outcomes. Let's assume you flip a coin once; this is a single trial. The only two outcomes are either heads or tails. This is an example of a Bernoulli distribution.

Usually, when following a Bernoulli distribution, we have the probability of one of the outcomes (p). From (p), we can deduce the probability of the other outcome by subtracting it from the total probability (1), represented as $(1-p)$.

It is represented by $\text{bern}(p)$, where p is the probability of success. The expected value of a Bernoulli trial ' x ' is represented as, $E(x) = p$, and similarly Bernoulli variance is, $\text{Var}(x) = p(1-p)$.



Loaded Coin Bernoulli distribution | Data Science Dojo

The graph of a Bernoulli distribution is simple to read. It consists of only two bars, one rising to the associated probability p and the other growing to $1-p$.

Binomial Distribution: A sequence of Bernoulli events

The Binomial Distribution can be thought of as the sum of outcomes of an event following a Bernoulli distribution. Therefore, Binomial Distribution is used in binary outcome events, and the probability of success and failure is the same in all successive trials. An example of a binomial event would be flipping a coin multiple times to count the number of heads and tails.

Binomial vs Bernoulli distribution.

The difference between these distributions can be explained through an example. Consider you're attempting a quiz that contains 10 True/False questions. Trying a single T/F question would be considered a Bernoulli trial, whereas attempting the entire quiz of 10 T/F questions would be categorized as a Binomial trial. **The main characteristics of Binomial Distribution are:**

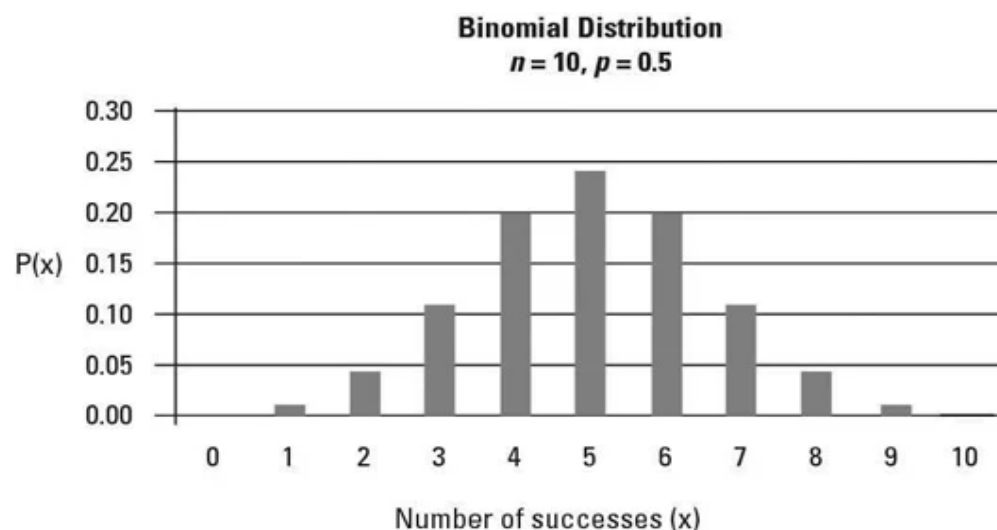
- Given multiple trials, each of them is independent of the other. That is, the outcome of one trial doesn't affect another one.
- Each trial can lead to just two possible results (e.g., winning or losing), with probabilities p and $(1 - p)$.

A binomial distribution is represented by $B(n, p)$, where n is the number of trials and p is the probability of success in a single trial. A Bernoulli distribution can be shaped as a binomial trial as $B(1, p)$ since it has only one trial. The expected value of a binomial trial " x " is the number of times a success occurs, represented as $E(x) = np$. Similarly, variance is represented as $\text{Var}(x) = np(1-p)$.

Let's consider the probability of success (p) and the number of trials (n). We can then calculate the likelihood of success (x) for these n trials using the formula below:

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x (1 - p)^{n-x}$$

For example, suppose that a candy company produces both milk chocolate and dark chocolate candy bars. The total products contain half milk chocolate bars and half dark chocolate bars. Say you choose ten candy bars at random and choosing milk chocolate is defined as a success. The probability distribution of the number of successes during these ten trials with $p = 0.5$ is shown here in the binomial distribution graph:



Binomial Distribution Graph

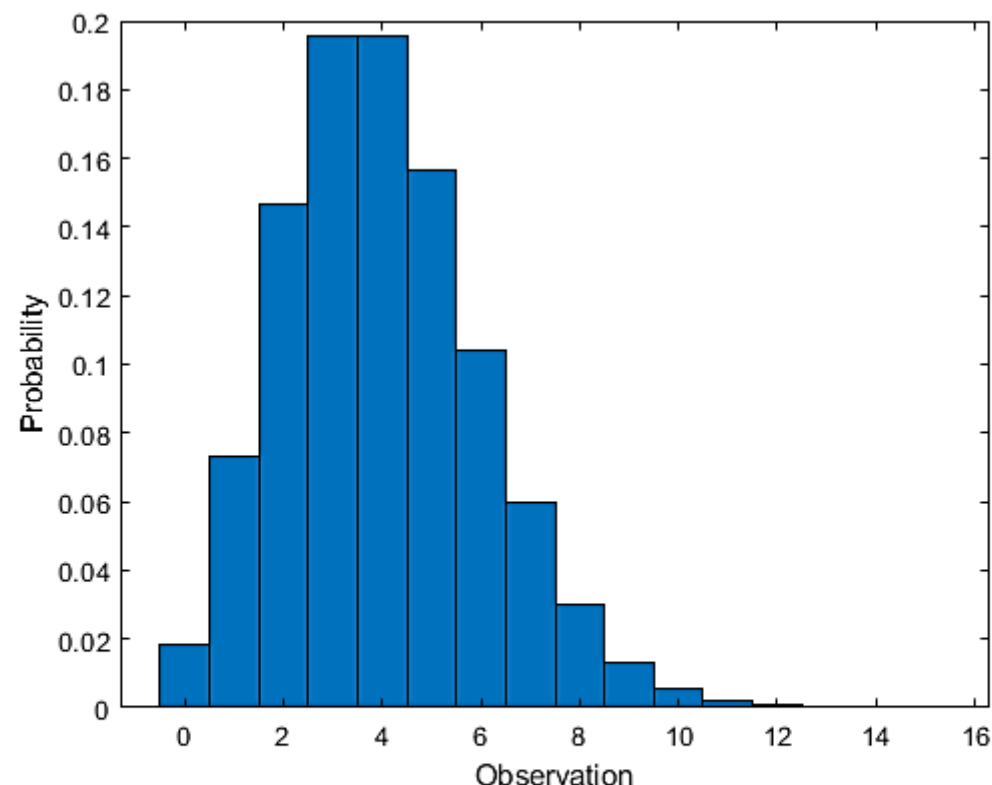
Poisson Distribution: The probability that an event may or may not occur

Poisson distribution deals with the frequency with which an event occurs within a specific interval. Instead of the probability of an event, Poisson distribution requires knowing how often it happens in a particular period or distance. For example, a cricket chirps two times in 7 seconds on average. We can use the Poisson distribution to determine the likelihood of it chirping five times in 15 seconds.

A Poisson process is represented with the notation $Po(\lambda)$, where λ represents the expected number of events that can take place in a period. The expected value and variance of a Poisson process is λ . X represents the discrete random variable. A Poisson Distribution can be modeled using the following formula.

The main characteristics which describe the Poisson Processes are:

- The events are independent of each other.
- An event can occur any number of times (within the defined period).
- Two events can't take place simultaneously.



Poisson Distribution Graph

The graph of Poisson distribution plots the number of instances an event occurs in the standard interval of time and the probability of each one.

Continuous distributions

Normal Distribution: Symmetric distribution of values around the mean

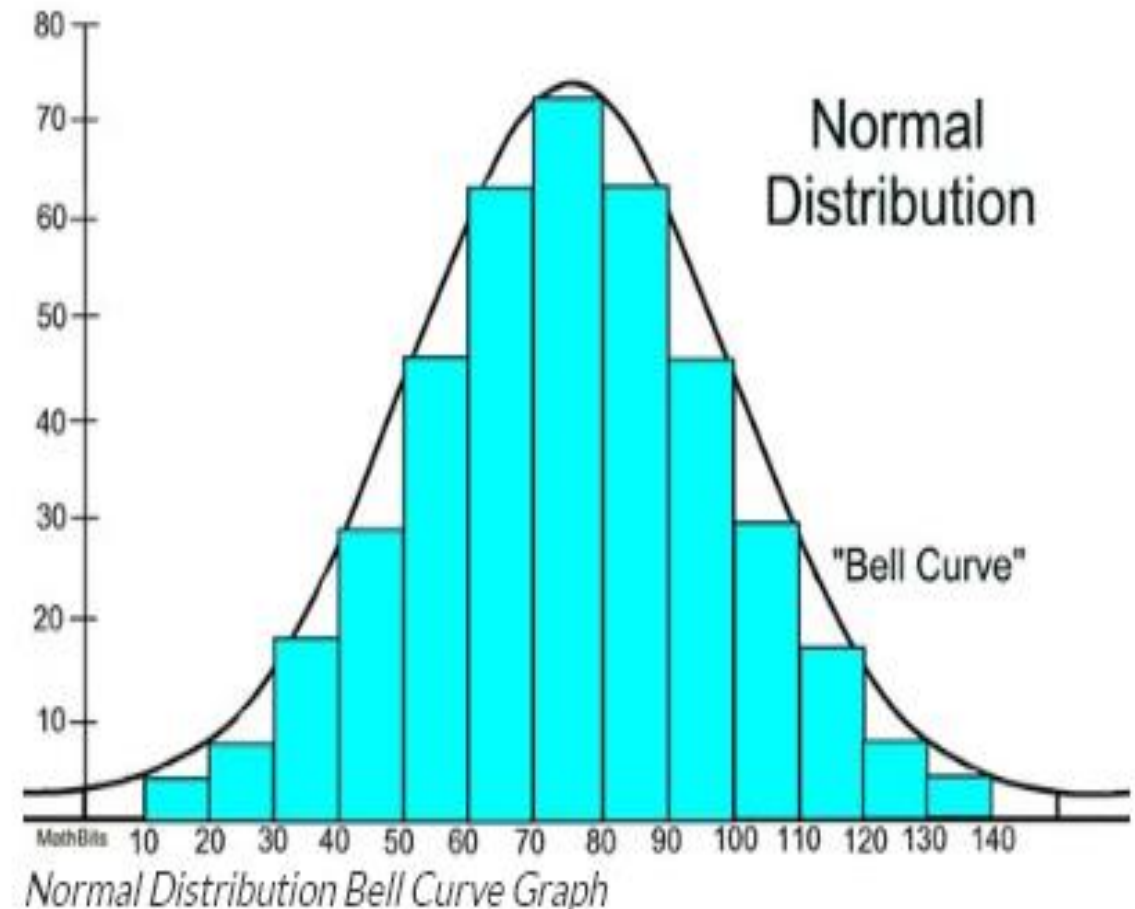
Normal distribution is the most used distribution in data science. In a normal distribution graph, data is symmetrically distributed with no skew. When plotted, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the center.

The normal distribution frequently appears in nature and life in various forms. For example, the scores of a quiz follow a normal distribution. Many of the students scored between 60 and 80 as illustrated in the graph below. Of course, students with scores that fall outside this range are deviating from the center.

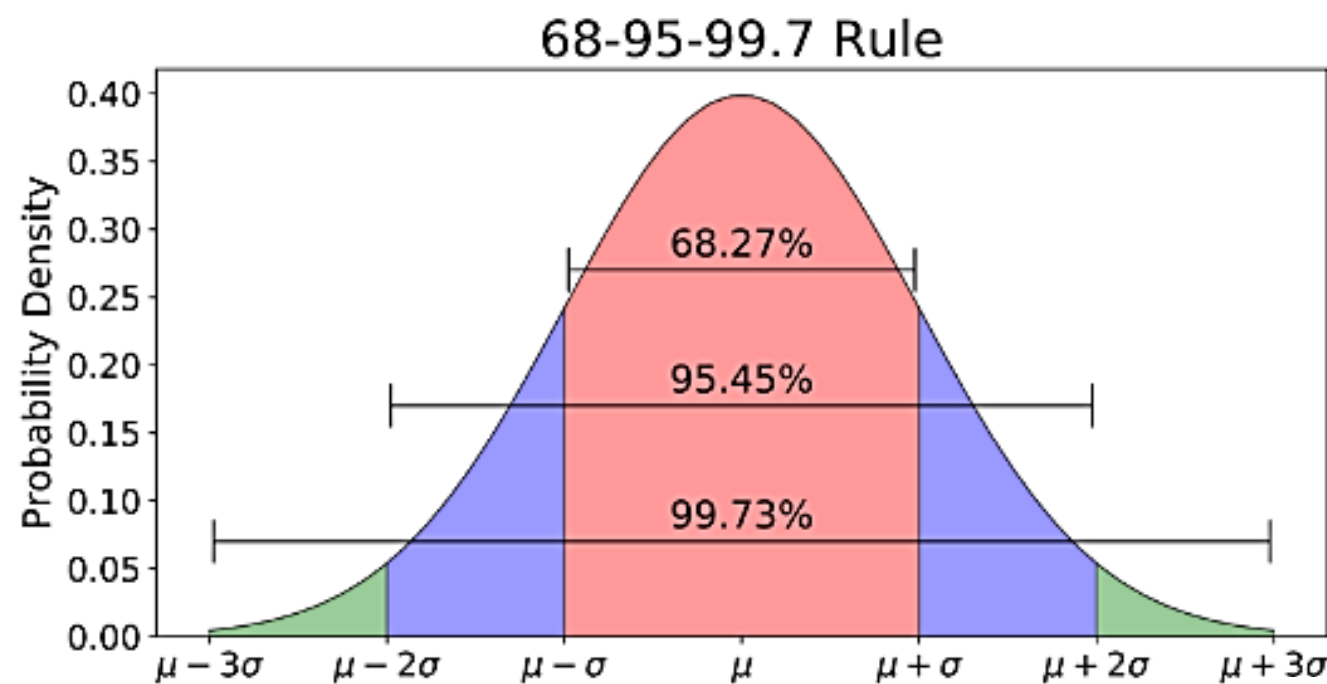
Here, you can witness the "bell-shaped" curve around the central region, indicating that most data points exist there. The normal distribution is represented as $N(\mu, \sigma^2)$ here, μ represents the mean, and σ^2 represents the variance, one of which is mostly provided. The expected value of a normal distribution is equal to its mean. Some of the characteristics which can help us to recognize a normal distribution are:

- The curve is symmetric at the center. Therefore mean, mode, and median are equal to the same value, distributing all the values symmetrically around the mean.
- The area under the distribution curve equals 1 (all the probabilities must sum up to 1).

68-95-99.7 Rule



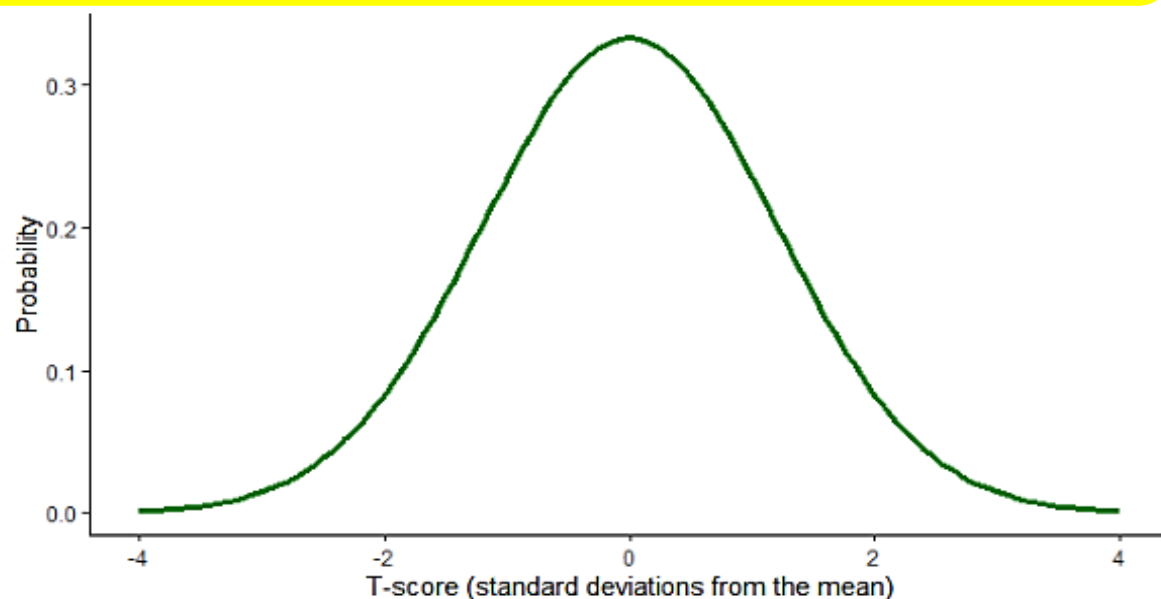
While plotting a graph for a normal distribution, 68% of all values lie within one standard deviation from the mean. In the example above, if the mean is 70 and the standard deviation is 10, 68% of the values will lie between 60 and 80. Similarly, 95% of the values lie within two standard deviations from the mean, and 99.7% lie within three standard deviations from the mean. This last interval captures almost all matters. If a data point is not included, it is most likely an outlier.



Probability Density and 68-95-99.7 Rule

Student t-Test Distribution: Small sample size approximation of a normal distribution

The student's t-distribution, also known as the t distribution, is a type of statistical distribution similar to the normal distribution with its bell shape but has heavier tails. The t distribution is used instead of the normal distribution when you have small sample sizes.



Student t-Test Distribution Curve

For example, suppose we deal with the total apples sold by a shopkeeper in a month. In that case, we will use the normal distribution. Whereas, if we are dealing with the total amount of apples sold in a day, i.e., a smaller sample, we can use the t distribution.

Another critical difference between the students' t distribution and the Normal one is that apart from the mean and variance, we must also define the degrees of freedom for the distribution. In statistics, the number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary. A Student's t distribution is represented as $t(k)$, where k represents the number of degrees of freedom. For $k=2$, i.e., 2 degrees of freedom, the expected value is the same as the mean.

T-Distribution Table

| df | $\alpha = 0.1$ | 0.05 | 0.025 |
|----------|----------------------|-------|--------|
| ∞ | $t_{\alpha} = 1.282$ | 1.645 | 1.960 |
| 1 | 3.078 | 6.314 | 12.706 |
| 2 | 1.886 | 2.920 | 4.303 |
| 3 | 1.638 | 2.353 | 3.182 |
| 4 | 1.533 | 2.132 | 2.776 |
| 5 | 1.476 | 2.015 | 2.571 |

T-Distribution Table

Degrees of freedom are in the left column of the t-distribution table.

Overall, the student t distribution is frequently used when conducting statistical analysis and plays a significant role in performing hypothesis testing with limited data.

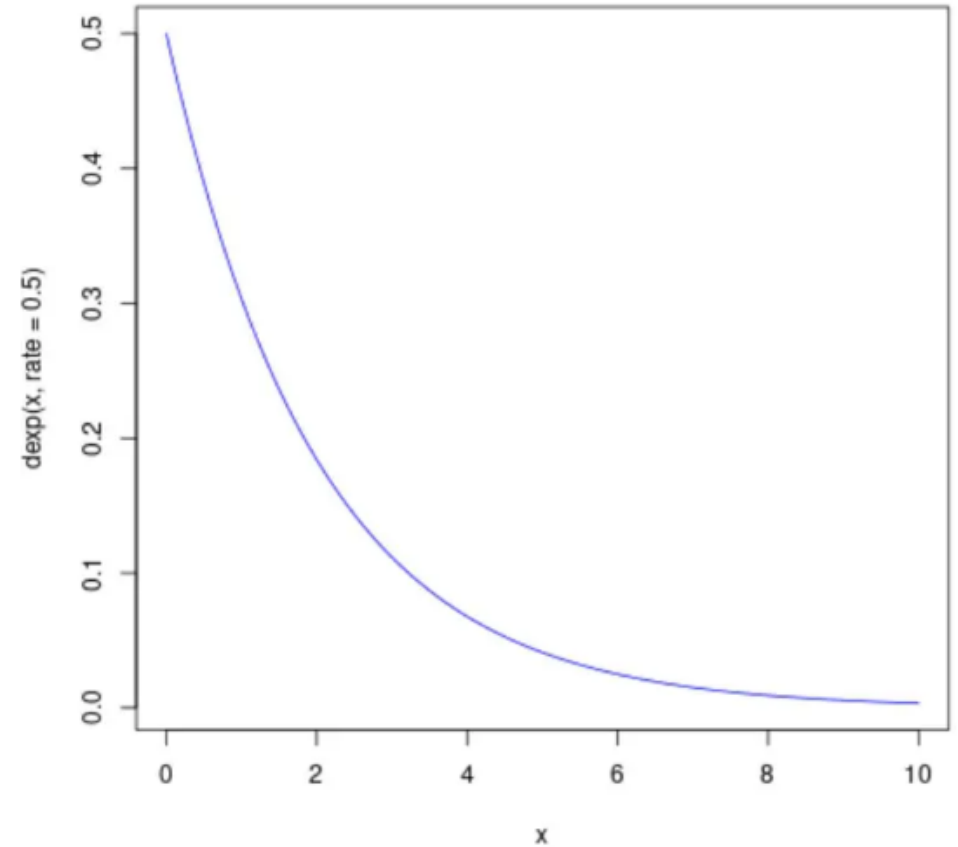
Exponential distribution: Model elapsed time between two events

Exponential distribution is one of the widely used continuous distributions. It is used to model the time taken between different events. For example, in physics, it is often used to measure radioactive decay; in engineering, to measure the time associated with receiving a defective part on an assembly line; and in finance, to measure the likelihood of the next default for a portfolio of financial assets. Another common application of Exponential distributions is in survival analysis (e.g., expected life of a device/machine).

Read the [top 10 Statistics books](#) to learn about Statistics

The exponential distribution is commonly represented as $\text{Exp}(\lambda)$, where λ is the distribution parameter, often called the rate parameter. We can find the value of λ by the formula $\lambda = 1/\mu$, where μ is the mean. Here standard deviation is the same as the mean. $\text{Var}(x)$ gives the variance $= 1/\lambda^2$.

An exponential graph is a curved line representing how the probability changes exponentially. Exponential distributions are commonly used in calculations of product reliability or the length of time a product lasts.



Exponential Distribution Curve

Conclusion

Data is an essential component of the data exploration and model development process. The first thing that springs to mind when working with continuous variables is looking at the data distribution. We can adjust our Machine Learning models to best match the problem if we can identify the pattern in the data distribution which reduces the time to get to an accurate outcome.

Indeed, specific Machine Learning models are built to perform best when certain distribution assumptions are met. Knowing which distributions, we're dealing with may thus assist us in determining which models to apply.

Task 4

What are famous statistical tests in data science?
How to choose the right statistical test for your research?

7 Ways to Choose the Right Statistical Test for Your Research Study



By Shrutika Sirisilla



(average: 5 out of 5. Total: 1)



Jan 4, 2023 4 mins read



Listen



What are Statistical Tests?

Statistical tests are a way of mathematically determining whether two sets of data are significantly different from each other. To do this, statistical tests use several statistical measures, such as the mean, standard deviation, and coefficient of variation. Once the statistical measures are calculated, the statistical test will then compare them to a set of predetermined criteria. If the data meet the criteria, the statistical test will conclude that there is a significant difference between the two sets of data.

There are various statistical tests that can be used, depending on the type of data being analyzed. However, some of the most common statistical tests are t-tests, chi-squared tests, and ANOVA tests.

Types of statistical tests

1- parametric statistical test:

1-1 regression tests

1-2 Comparison tests

1-3 Correlation tests

2- Non parametric tests

Table 3 Parametric and Non-parametric tests for comparing two or more groups

| Parametric test | Non-Parametric equivalent |
|------------------------------|---------------------------|
| Paired t-test | Wilcoxon Rank sum Test |
| Unpaired t-test | Mann-Whitney U test |
| Pearson correlation | Spearman correlation |
| One way Analysis of variance | Kruskal Wallis Test |



1.1. Regression Tests

Regression tests determine cause-and-effect relationships. They can be used to estimate the effect of one or more continuous variables on another variable.

- **Simple linear regression** is a type of test that describes the relationship between a dependent and an independent variable using a straight line. This test determines the relationship between two quantitative variables.
- **Multiple linear regression** measures the relationship between a quantitative dependent variable and two or more independent variables, again using a straight line.
- **Logistic regression** predicts and classifies the research problem. Logistic regression helps identify data anomalies, which could be predictive fraud.

1.2. Comparison Tests

Comparison tests determine the differences among the group means. They can be used to test the effect of a categorical variable on the mean value of other characteristics.

- T-test

One of the most common statistical tests is the t-test, which is used to compare the means of two groups (e.g. the average heights of men and women). You can use the t-test when you are not aware of the population parameters (mean and standard deviation).

- Paired T-test

It tests the difference between two variables from the same population (pre-and post-test scores). For example, measuring the performance score of the trainee before and after the completion of the training program.

- Independent T-test

The independent t-test is also called the two-sample t-test. It is a statistical test that determines whether there is a statistically significant difference between the means in two unrelated groups. For example, comparing cancer patients and pregnant women in a population

- One Sample T-test

In this test, the mean of a single group is compared with the given mean. For example, determining the increase and decrease in sales in the given average sales.

- ANOVA

ANOVA (Analysis of Variance) analyzes the difference between the means of more than two groups. One-way ANOVAs determine how one factor impacts another, whereas two-way analyses compare samples with [different variables](#). It determines the impact of one or more factors by comparing the means of different samples.

- MANOVA

MANOVA, which stands for Multivariate Analysis of Variance, provides regression analysis and analysis of variance for multiple dependent variables by one or more factor variables or covariates. Also, it examines the statistical difference between one continuous dependent variable and an independent grouping variable.

- Z-test

It is a statistical test that determines whether two population means are different, provided the variances are known and the sample size is large.

1.3. Correlation Tests

Correlation tests check if the variables are related without hypothesizing a cause-and-effect relationship. These tests can be used to check if the two variables you want to use in a multiple regression test are correlated.

- Pearson Correlation Coefficient

It is a common way of measuring the linear correlation. The coefficient is a number between -1 and 1 and determines the strength and direction of the relationship between two variables. The change in one variable changes the course of another variable change in the same direction.

2. Non-parametric Statistical Tests

Non-parametric tests do not make as many assumptions about the data compared to parametric tests. They are useful when one or more of the common statistical assumptions are violated. However, these inferences are not as accurate as with parametric tests.

- Chi-square test

The chi-square test compares two categorical variables. Furthermore, calculating the chi-square statistic value and comparing it with a critical value from the chi-square distribution allows you to assess whether the observed frequency is significantly different from the expected frequency.

7 Essential Ways to Choose the Right Statistical Test

1. Research Question

The decision for a statistical test depends on the [research question](#) that needs to be answered. Additionally, the research questions will help you formulate the data structure and [research design](#).

2. Formulation of Null Hypothesis

After defining the research question, you could develop a null hypothesis. A [null hypothesis](#) suggests that no statistical significance exists in the expected observations.

3. Level of Significance in Study Protocol

Before performing the study protocol, a level of significance is specified. The level of significance determines the statistical importance, which defines the acceptance or rejection of the null hypothesis.

4. The Decision Between One-tailed and Two-tailed

You must decide if your study should be a one-tailed or two-tailed test. If you have clear evidence where the statistics are leading in one direction, you must perform one-tailed tests. However, if there is no particular direction of the expected difference, you must perform a two-tailed test.

5. The Number of Variables to Be Analyzed

Statistical tests and procedures are divided according to the number of variables that are designed to analyze. Therefore, while choosing the test, you must consider how many variables you want to analyze.

6. Type of Data

It is important to define whether your data is continuous, categorical, or binary. In the case of continuous data, you must also check if the data are normally distributed or skewed, to further define which statistical test to consider.

7. Paired and Unpaired Study Designs

A paired design includes comparison studies where the two population means are compared when the two samples depend on each other. In an unpaired [or independent study design](#), the results of the two samples are grouped and then compared.

References

<https://youtu.be/4SivdTLIwHc>

https://imbalanced-learn.org/stable//references/under_sampling.htm

<https://youtu.be/5Ddf7v1M6wY>

<https://youtu.be/2oJldeE4JcU>

<https://youtu.be/bPFNxD3Yg6U>

<https://datasciencedojo.com/blog/types-of-statistical-distributions-in-ml/>

<https://youtu.be/PMs76lrqiA4>

<https://www.stratascratch.com/blog/basic-types-of-statistical-tests-in-data-science/>

<https://www.enago.com/academy/right-statistical-test/>

<https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1b-statistical-methods/parametric-nonparametric-tests>