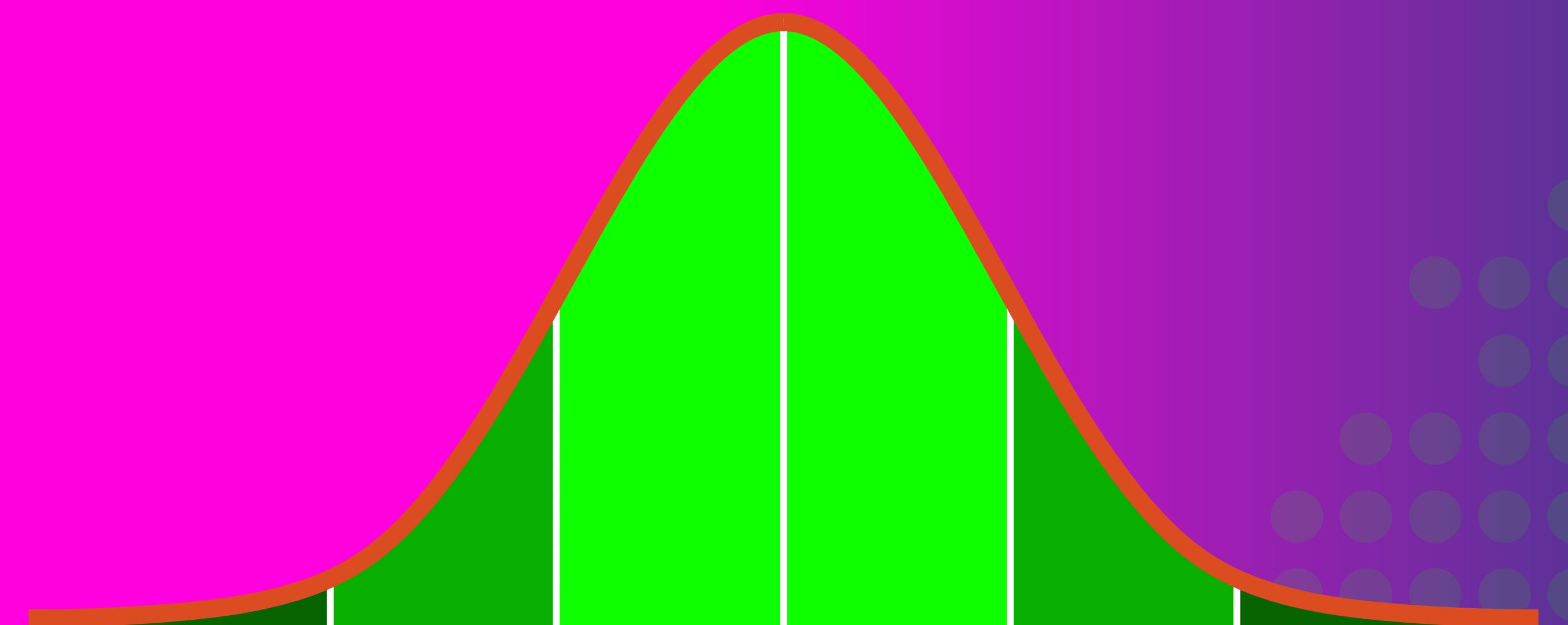


WHICH TEST WHEN?

The Data Analyst's Guide to
Understanding Data Types, Distributions,
and Statistical Tests.



ANDREW MADSON

WHAT WILL YOU LEARN?

DATA TYPES

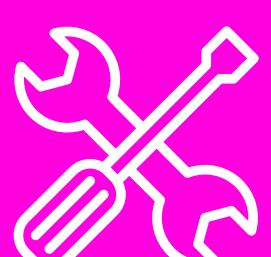
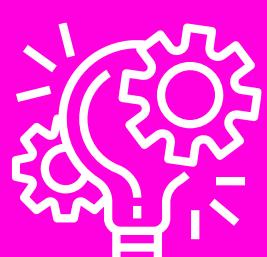


DISTRIBUTIONS

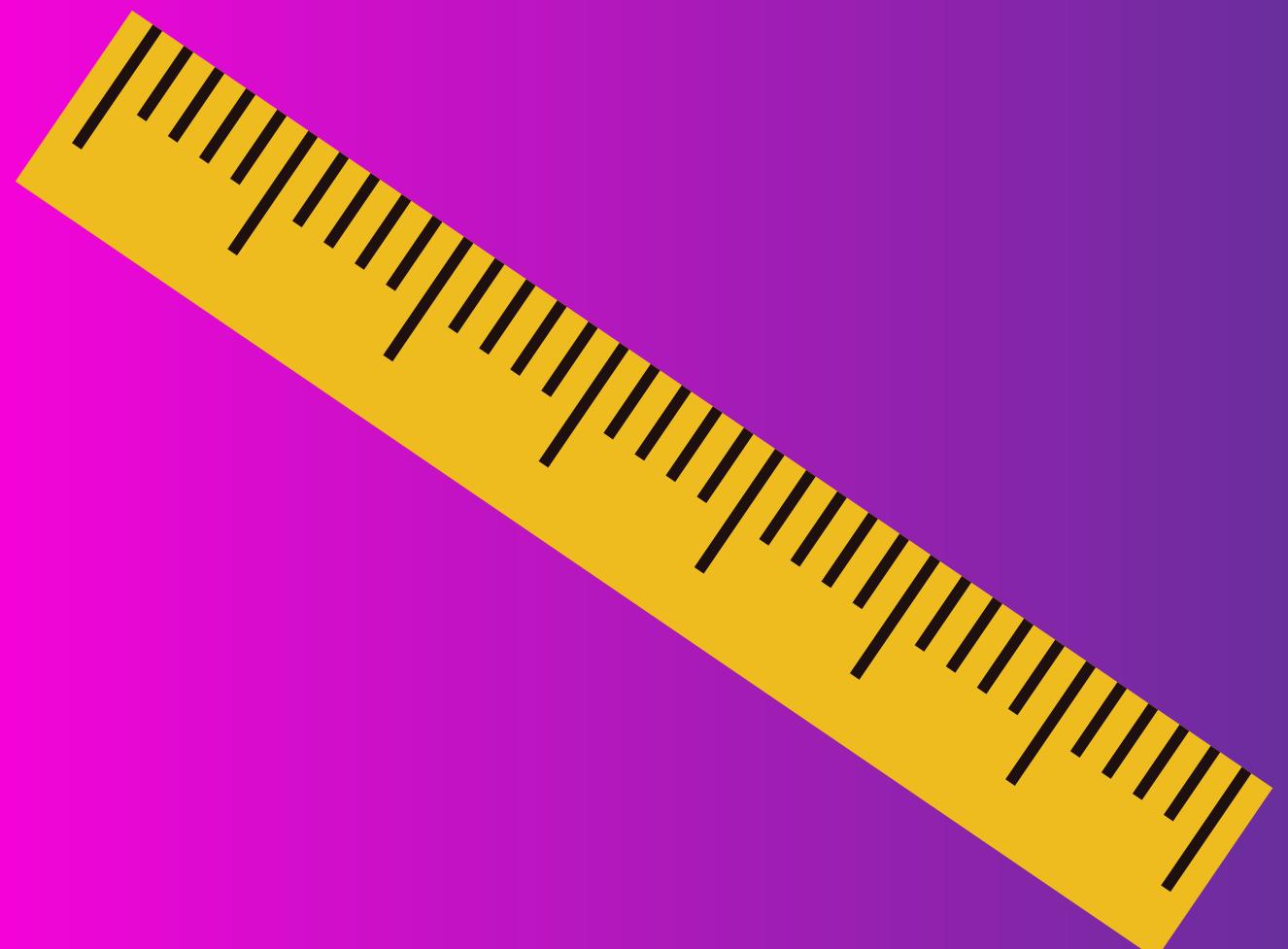


STATISTICAL

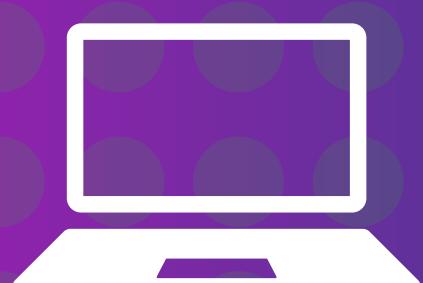
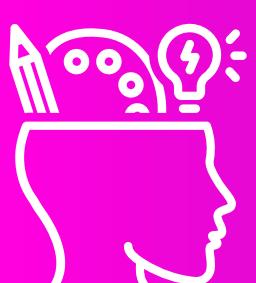
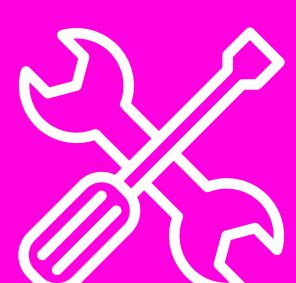
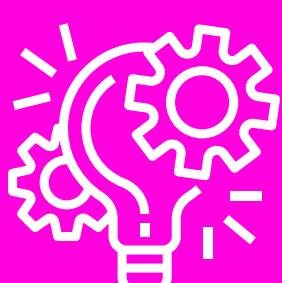
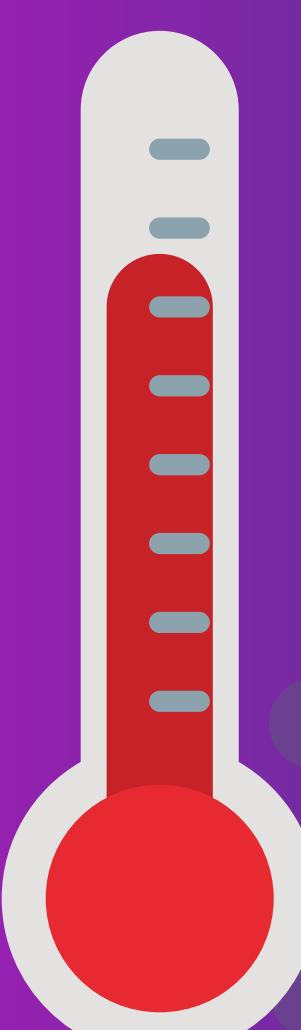
TESTS



DATA TYPES



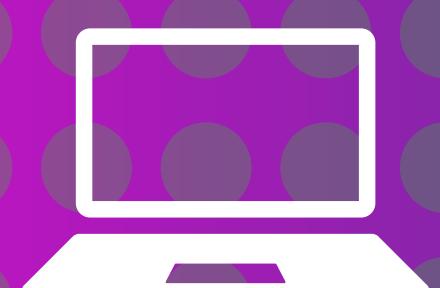
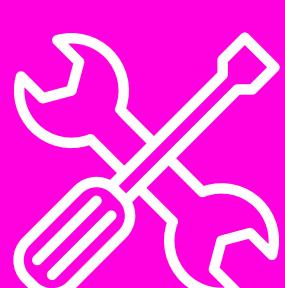
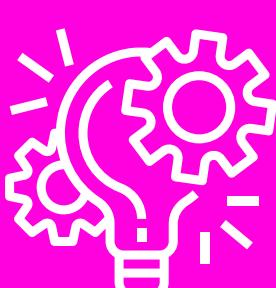
101010
010101
101010
010101



DATA TYPES

WHY IT MATTERS

- 1. Appropriate Analysis:** Different types of data require different statistical tests. For example, nominal data can be analyzed using a Chi-square test, while interval data can be analyzed using a t-test or ANOVA. Using the wrong test can lead to incorrect conclusions.
- 2. Data Visualization:** Your data type determines the best way to visualize it. For instance, categorical data might be best represented in a bar chart, while continuous data might be better suited for a histogram or scatter plot.
- 3. Data Transformation:** Understanding your data type can guide you in transforming your data, if necessary. For example, ordinal data might be converted into interval data under certain conditions, or continuous data might be categorized into ordinal data.
- 4. Data Quality:** Knowing your data type can help you identify potential errors or inconsistencies in your data. For instance, if you expect a variable to be continuous and find string values, this could indicate a data quality issue.
- 5. Interpretation of Results:** The type of data you have influenced how you interpret your results. For example, if you have ordinal data, you can make statements about the order of values but not the difference between values.



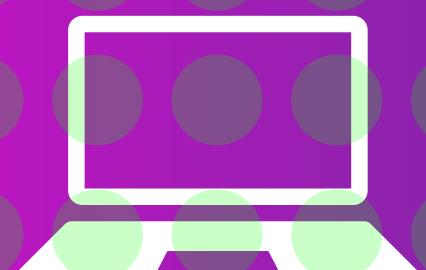
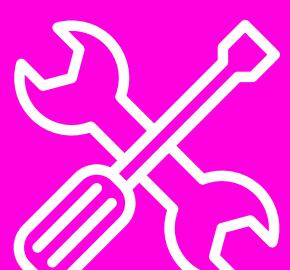
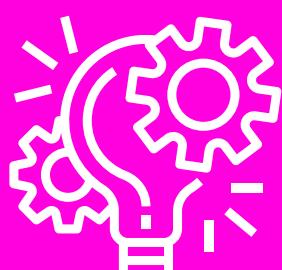
DATA TYPES

QUANTITATIVE

Numerical data that can be measured or counted and can be represented numerically, such as height, weight, or temperature.

QUALITATIVE

Non-numerical data that consists of descriptive information, such as colors, tastes, textures, or any other characteristics that cannot be counted or measured.



DATA TYPES

QUANTITATIVE

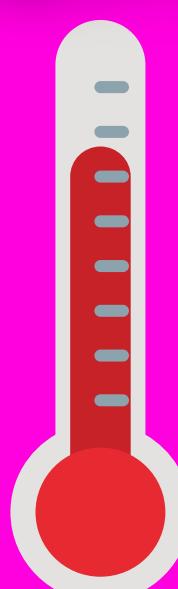
DISCRETE



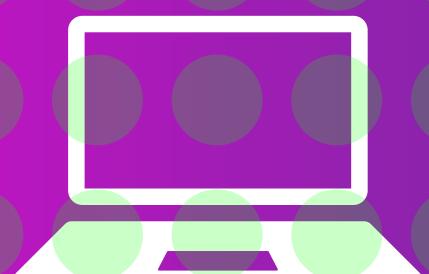
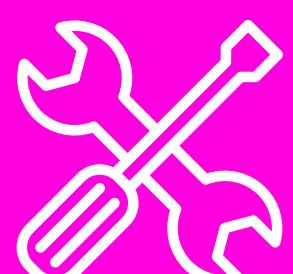
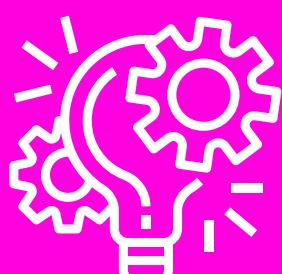
CONTINUOUS



INTERVAL



RATIO



QUANTITATIVE DATA TYPES

DISCRETE

Distinct and separate values with no intermediate values in between.

CONTINUOUS

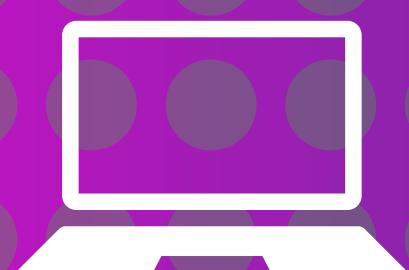
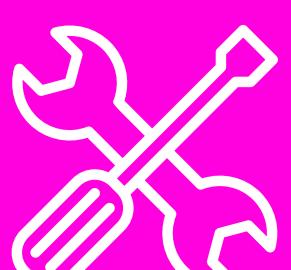
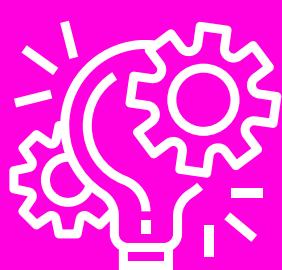
Infinitely divisible and can take on any value within a certain range or interval. Encompasses both INTERVAL and RATIO data.

INTERVAL

Continuous Data Type - numerical data where the intervals between values are equal but no true zero point exists.

RATIO

Continuous Data Type - numerical data with a true zero point, allowing for meaningful ratios and comparisons between values.



DATA TYPES

QUALITATIVE

CATEGORICAL

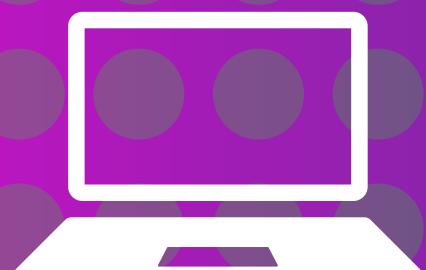
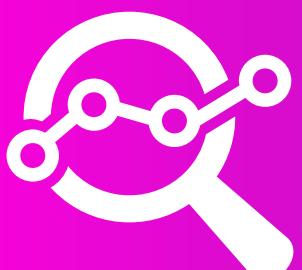
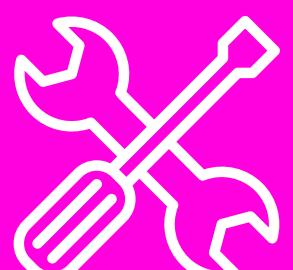
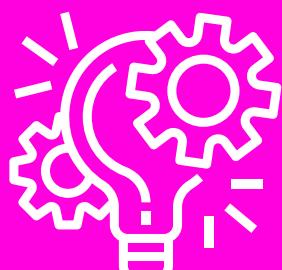


ORDINAL



BINARY

101010
010101
101010
010101



QUALITATIVE DATA TYPES

CATEGORICAL

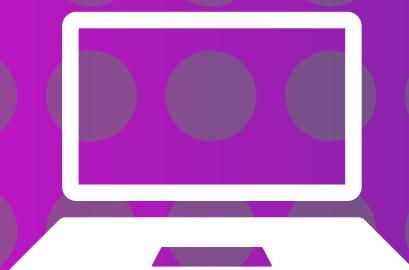
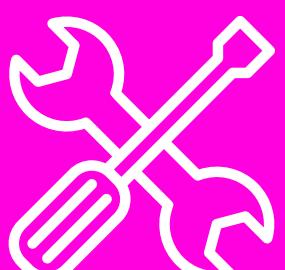
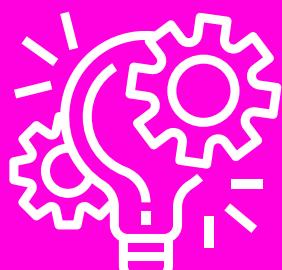
Distinct categories or groups with no inherent order or numerical significance.

ORDINAL

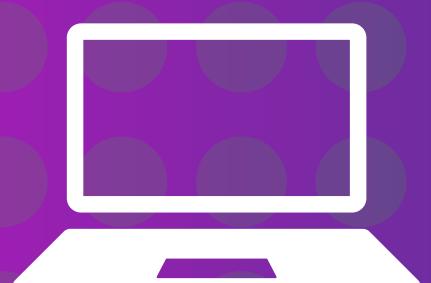
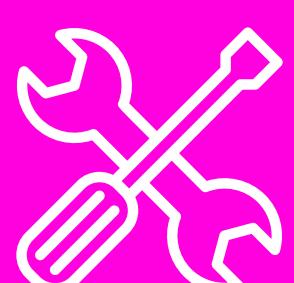
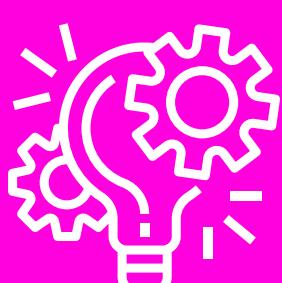
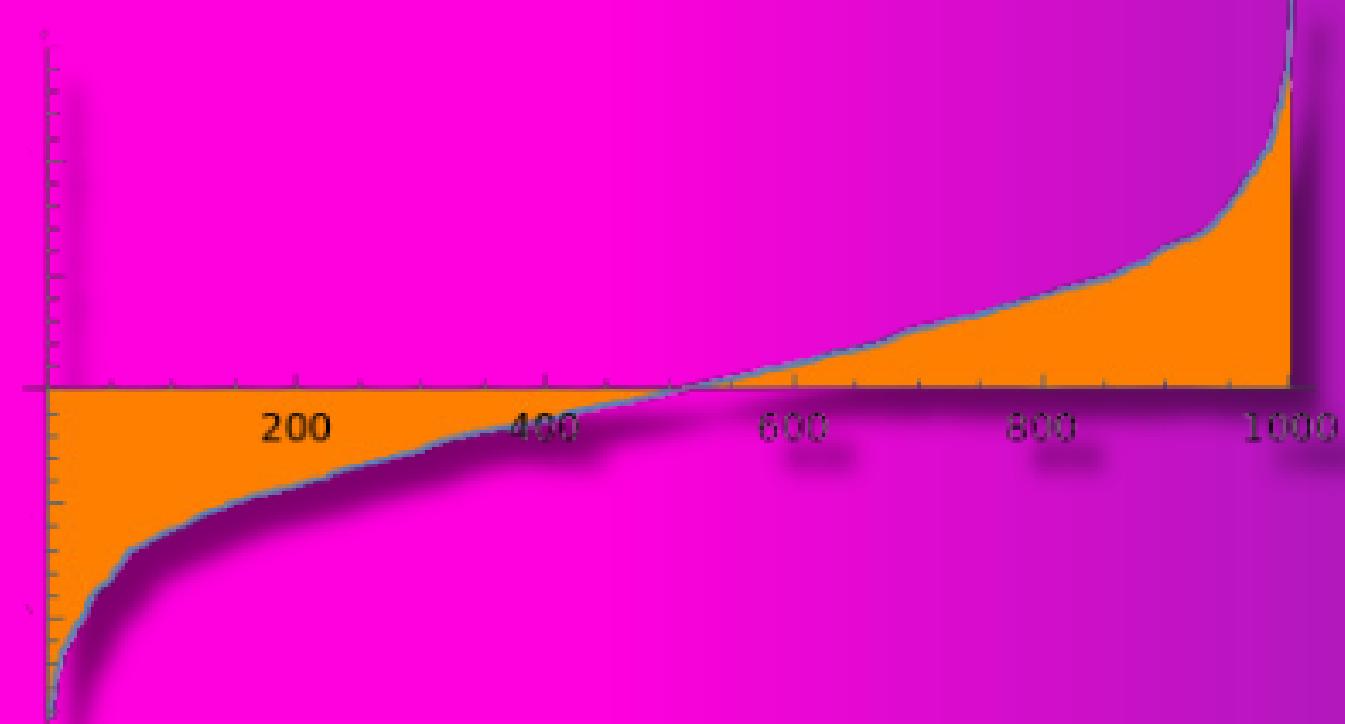
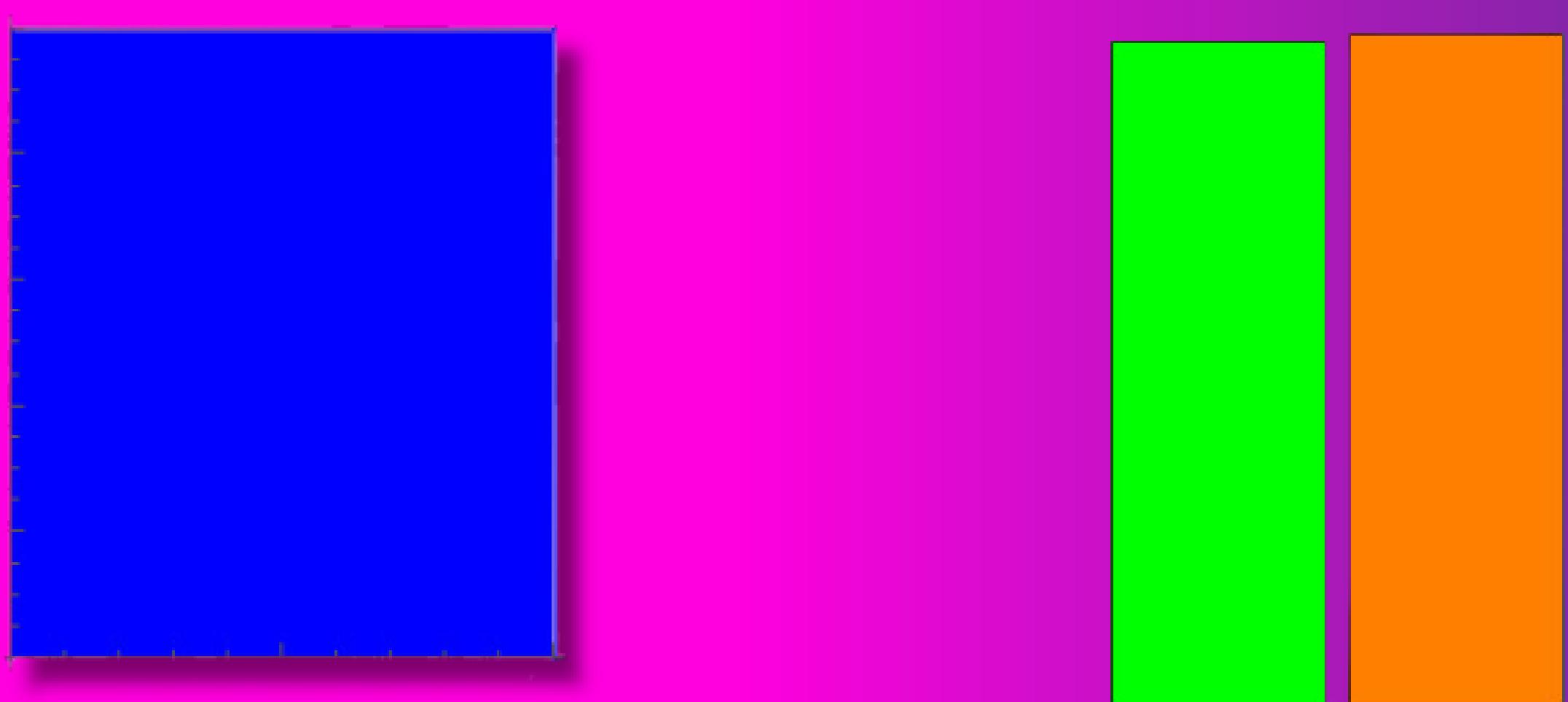
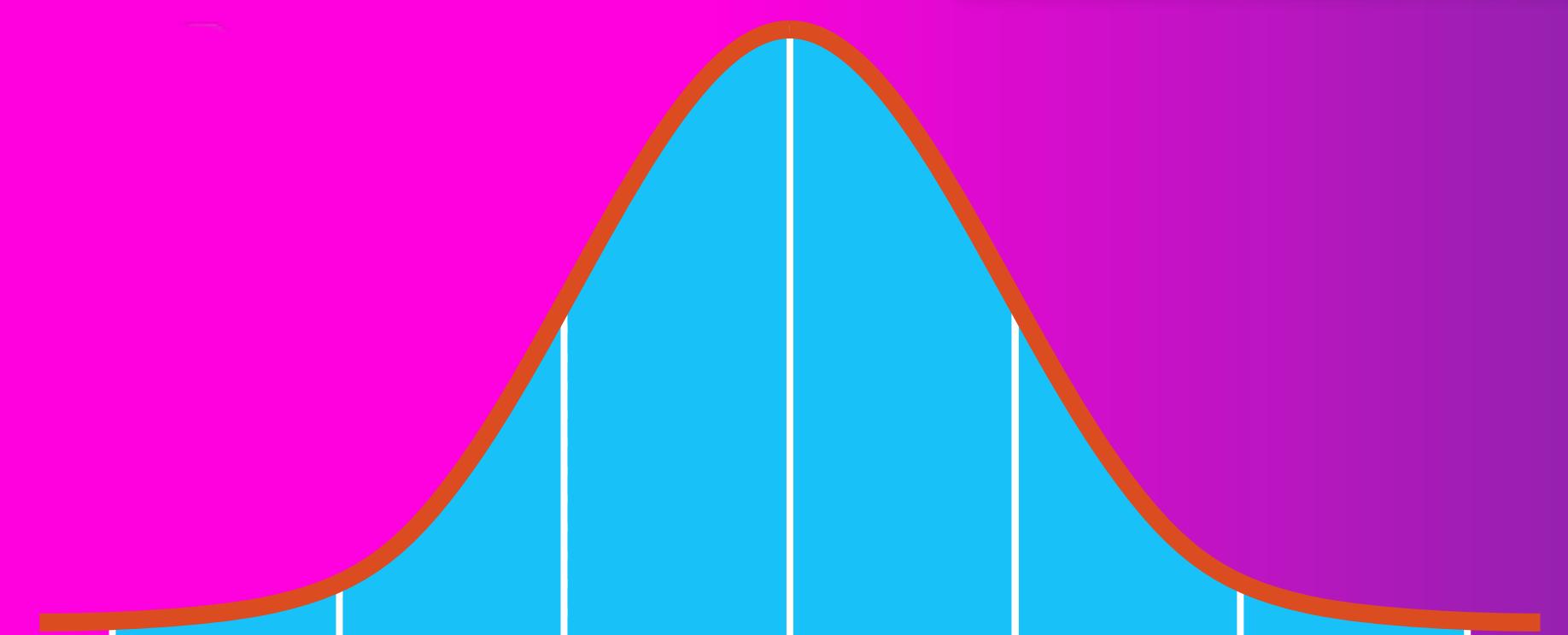
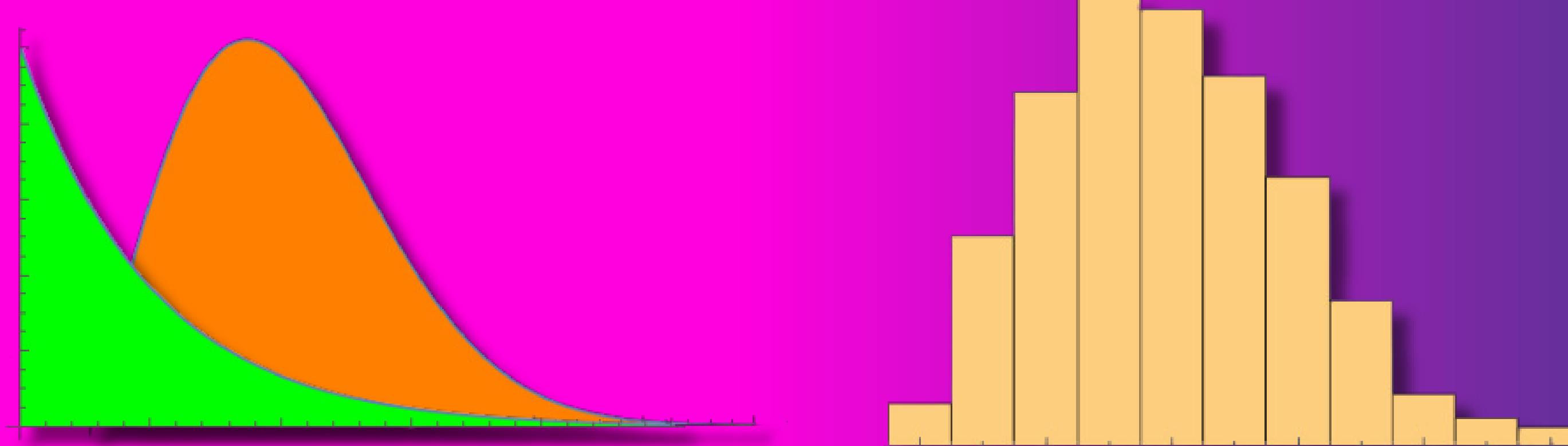
Data with a natural order or ranking among its categories, indicating relative differences or preferences.

BINARY

Categorical data that has only two possible outcomes or categories.



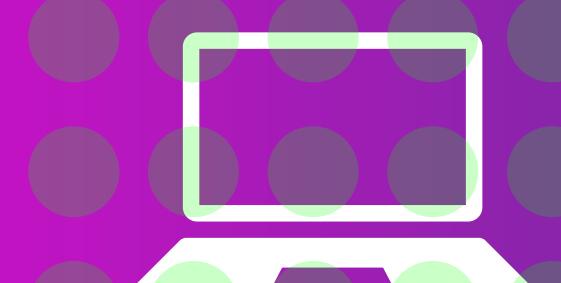
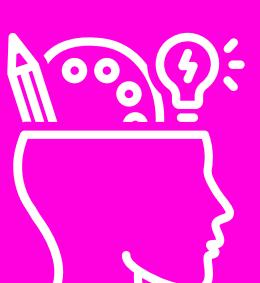
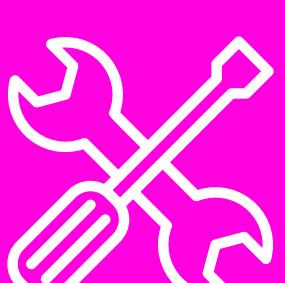
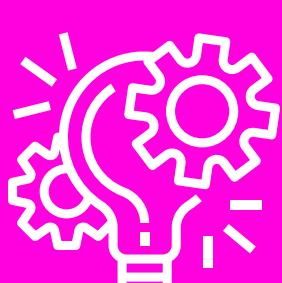
DISTRIBUTIONS



DISTRIBUTION TYPES

WHY IT MATTERS

1. **Understanding the data:** Understanding the distribution of your data gives insight into the nature and behavior of the variables you are studying. It helps you identify your data's patterns, trends, and potential outliers.
2. **Statistical assumptions:** Many statistical tests and models make assumptions about the distribution of the data. For example, the t-test assumes that the data follows a normal distribution. If these assumptions are violated, it can lead to incorrect conclusions. Knowing the distribution of your data helps you choose the appropriate statistical methods.
3. **Predictive modeling:** When building predictive models, the distribution of the data can inform the selection of algorithms or the model's configuration. Some machine learning algorithms are more suited to certain types of distributions.
4. **Data transformation:** If your data does not follow the distribution required by a particular statistical method, you may need to transform it. For example, if your data is skewed, you might apply a logarithmic transformation to make it more symmetrical. Understanding the distribution can guide these transformations.
5. **Risk management:** In fields like finance and insurance, understanding data distribution is crucial for risk assessment. For example, the distribution of returns on investment can help determine the probability of a significant loss.
6. **Data quality:** Examining data distribution can also be a way to check data quality. If the data doesn't follow expected distributions, it may indicate errors or bias in the data collection process.



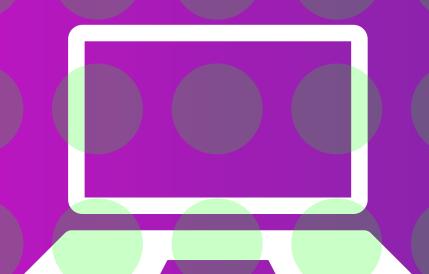
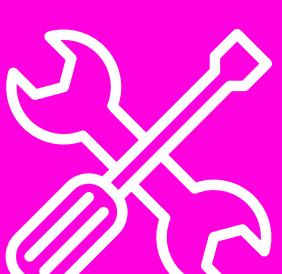
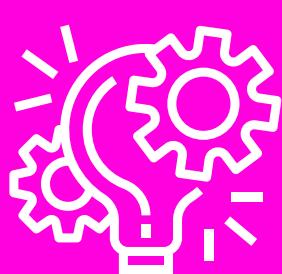
DISTRIBUTION TYPES

PARAMETRIC

Assume that the data follows a certain specific distribution pattern, and the parameters of that distribution are estimated from the data.

NON-PARAMETRIC

Do not assume that the data follow any specific distribution. They are defined without the assumption of underlying parameters

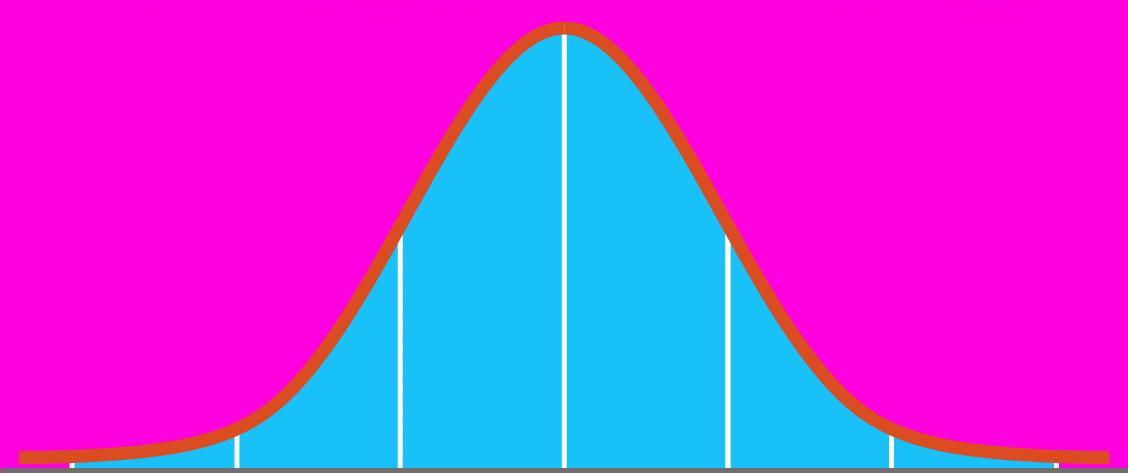


DISTRIBUTION

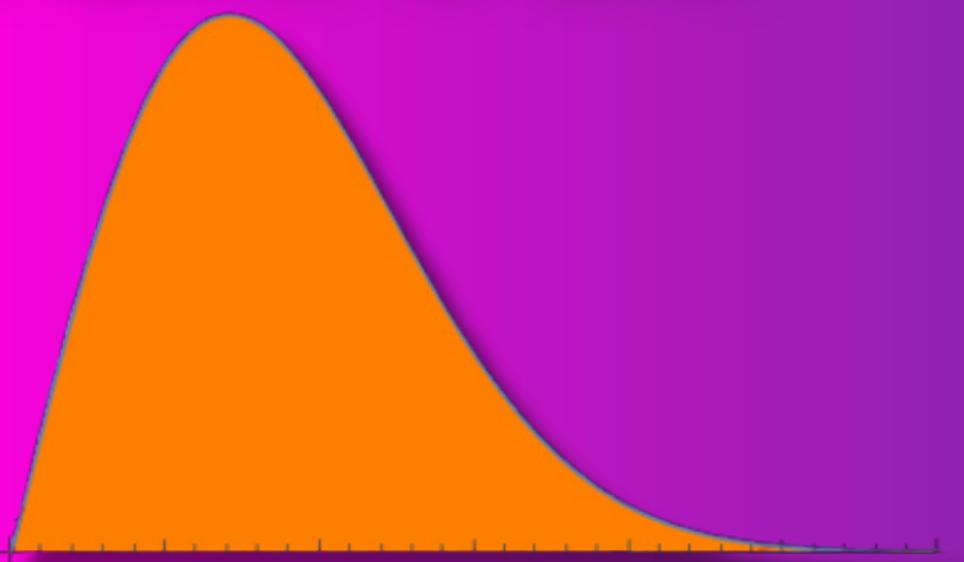
TYPES

PARAMETRIC

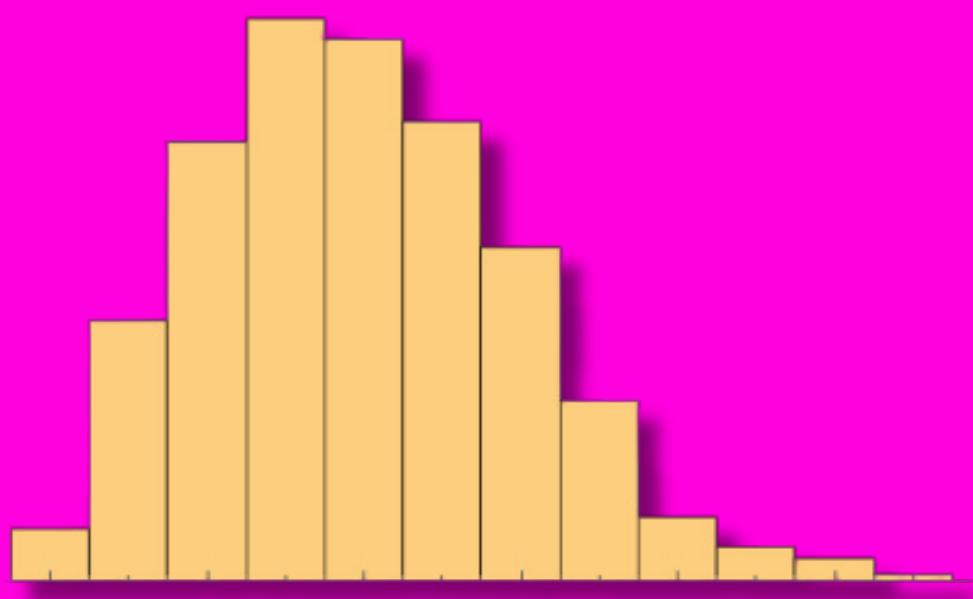
NORMAL



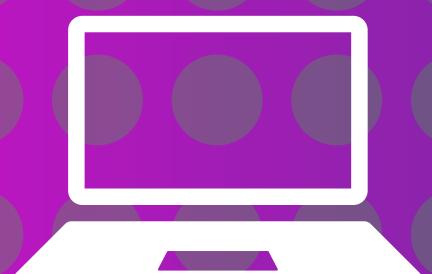
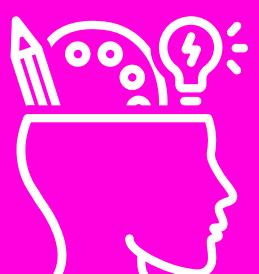
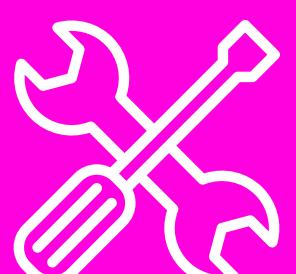
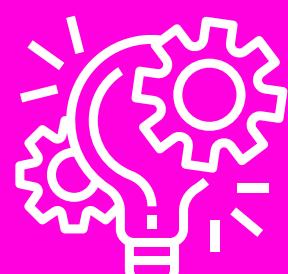
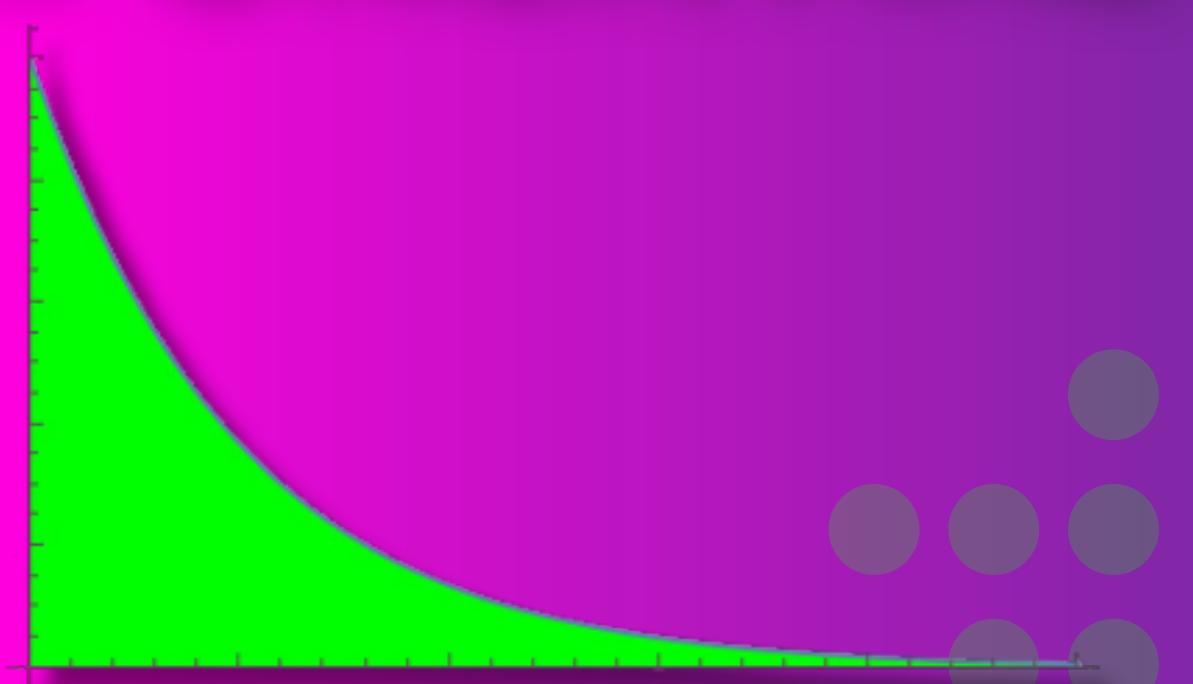
WEIBULL



POISSON



EXPONENTIAL



PARAMETRIC DISTRIBUTIONS

NORMAL

Symmetric around the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

WEIBULL

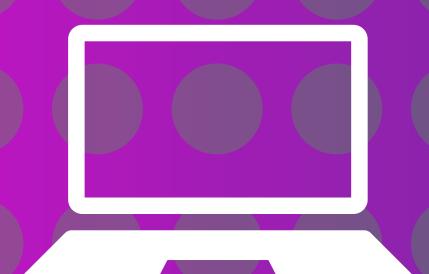
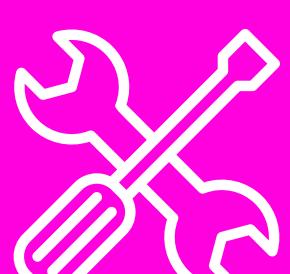
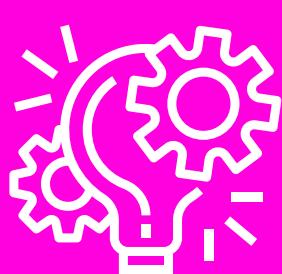
Continuous probability distribution that models the time it takes for an event to occur and is commonly used in reliability and survival analysis.

POISSON

Discrete probability distribution that models the number of events occurring in a fixed interval of time or space.

EXPONENTIAL

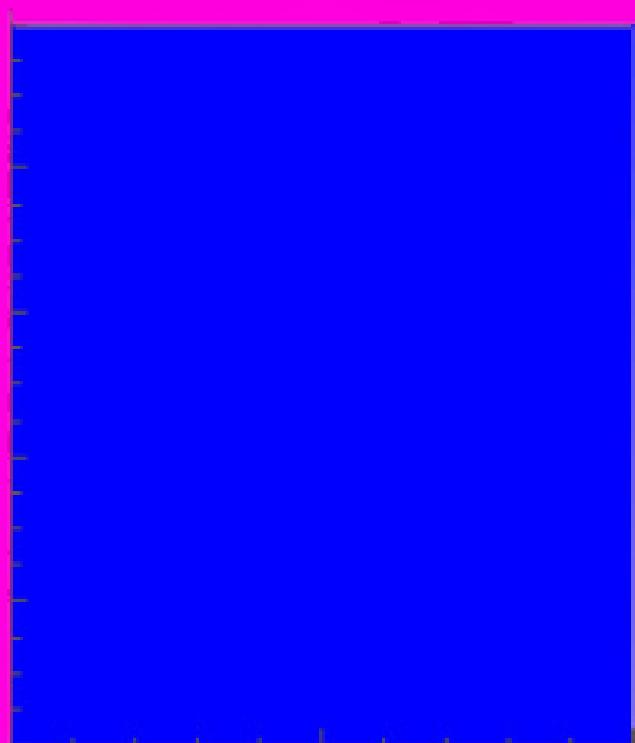
Continuous probability distribution that models the time between events in a Poisson process, where events occur independently and at a constant average rate.



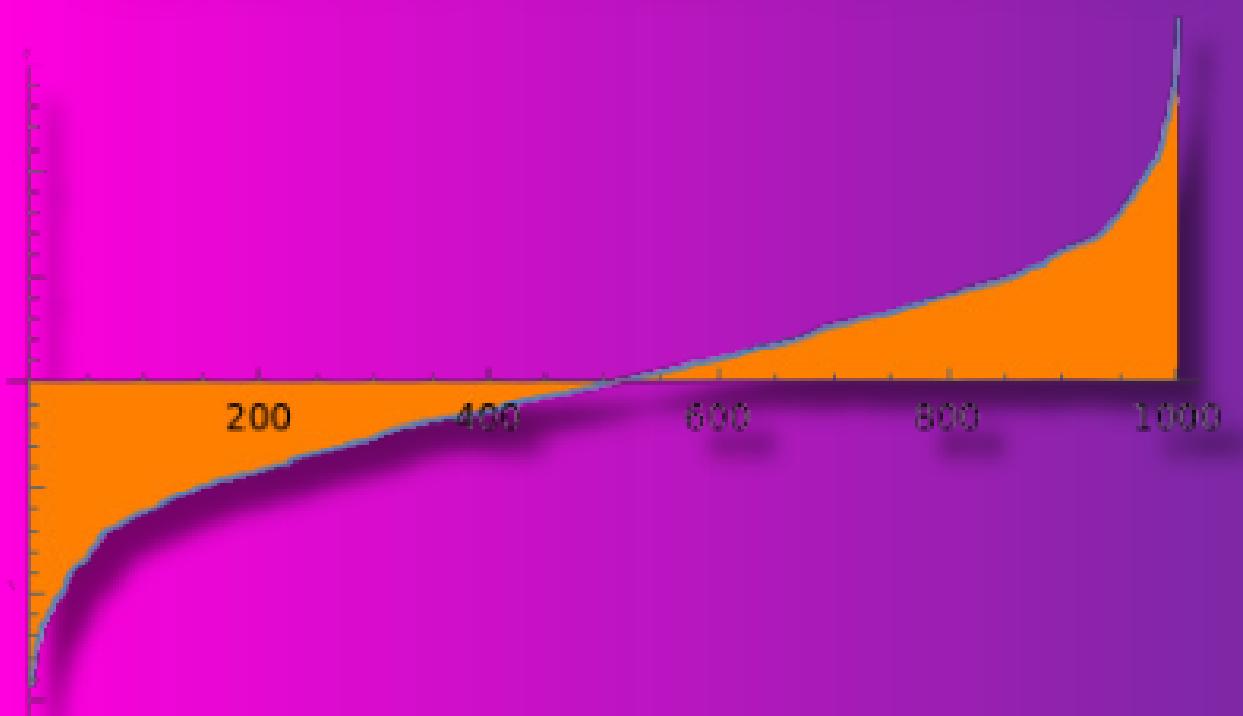
DISTRIBUTION TYPES

NON-PARAMETRIC

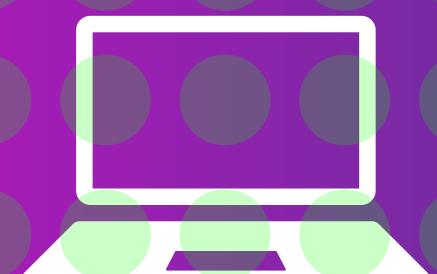
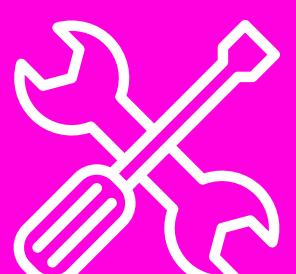
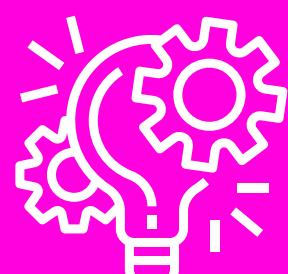
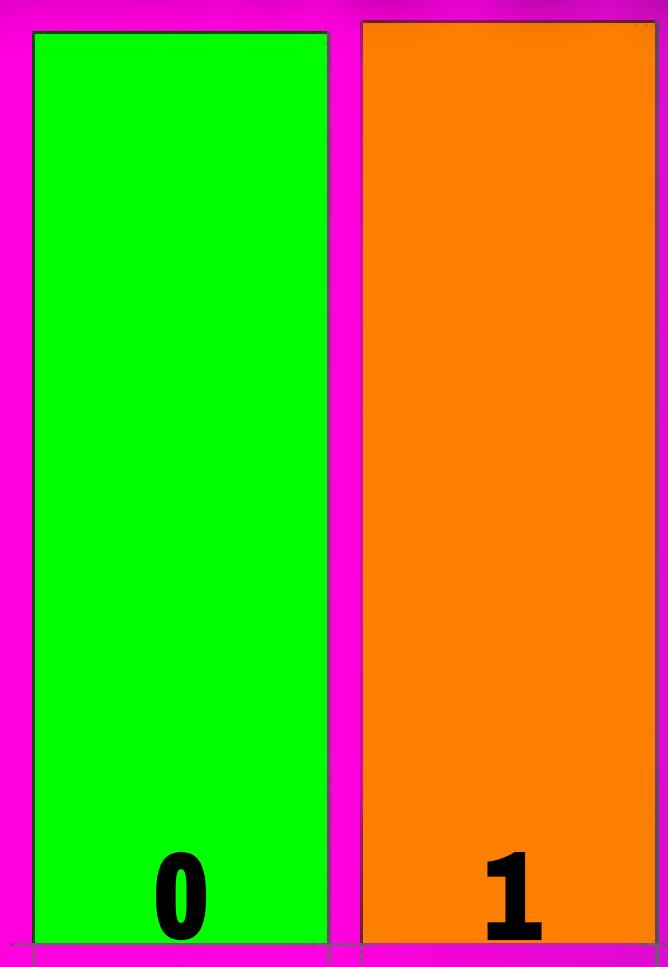
UNIFORM



EMPIRICAL



BERNOULLI



NON-PARAMETRIC DISTRIBUTIONS

UNIFORM

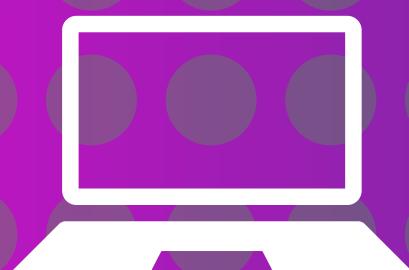
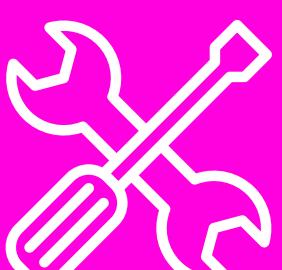
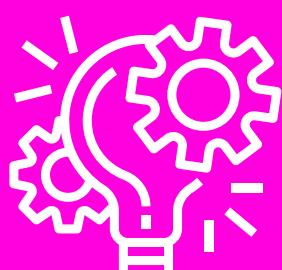
Probability distribution where all outcomes or values within a given range have an equal probability of occurring.

EMPIRICAL

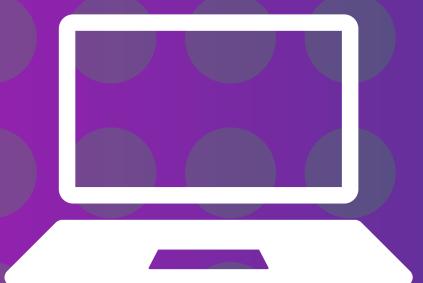
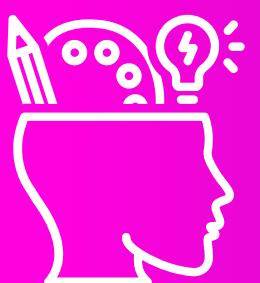
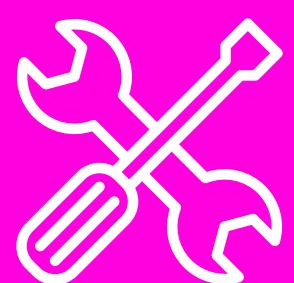
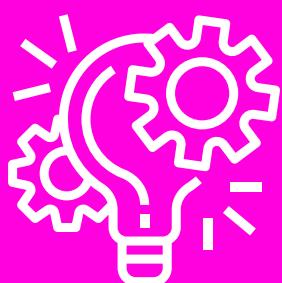
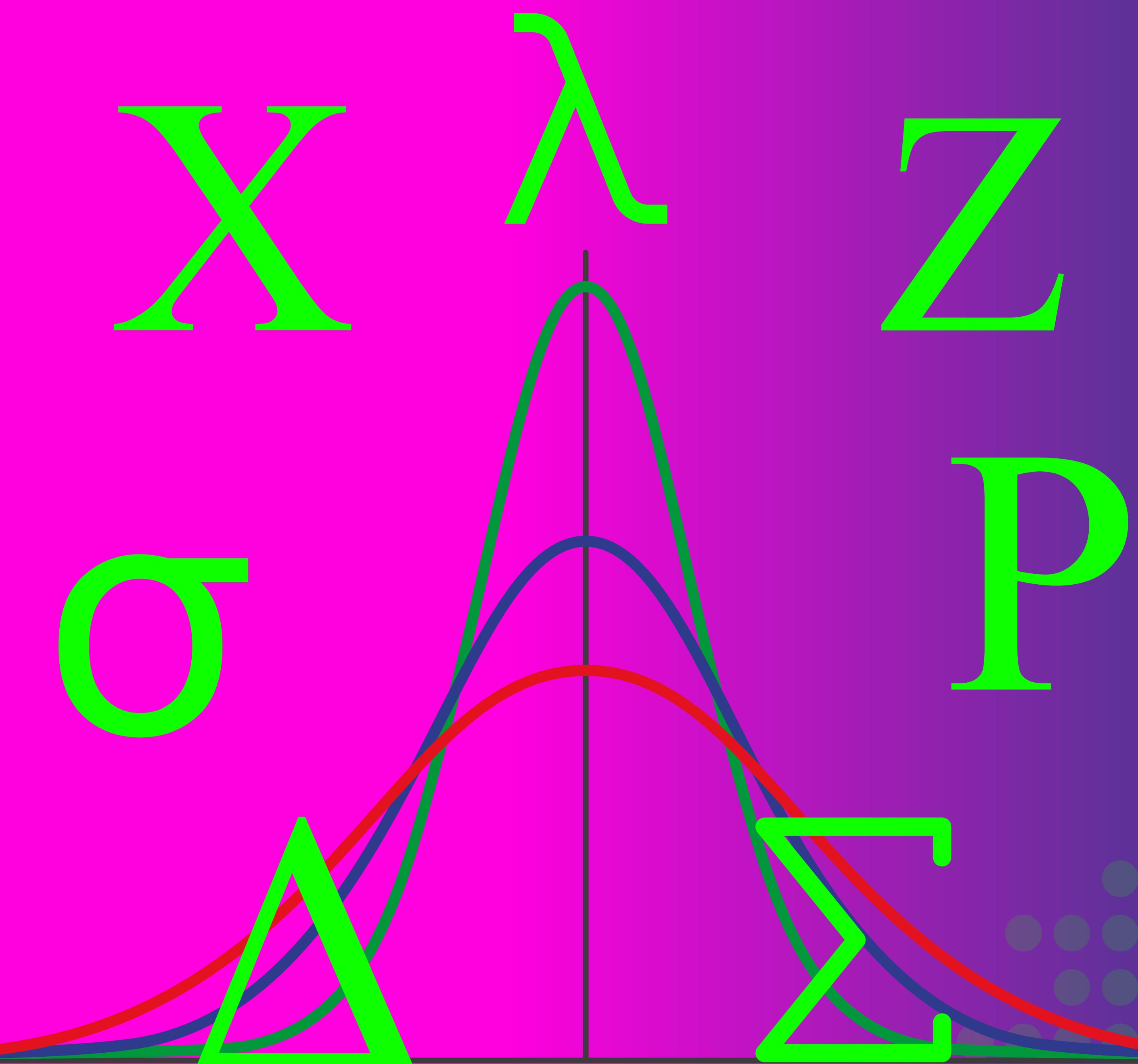
Based on observed data rather than being derived from a known mathematical formula.

BERNOULLI

Discrete probability distribution representing a random experiment with only two possible outcomes, typically denoted as success (1) or failure (0), each with a fixed probability.



STATISTICAL TESTS



T-TEST

PURPOSE

C.compares the
means of two
groups

WHEN TO USE IT

Two related groups
to compare

DISTRIBUTION

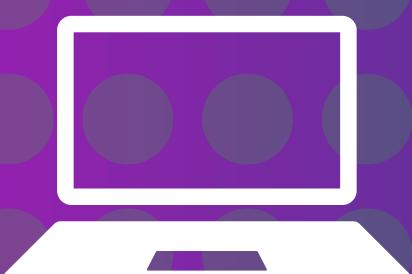
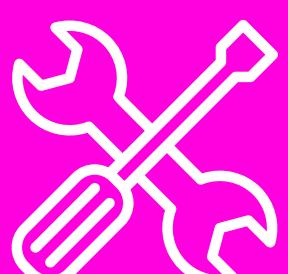
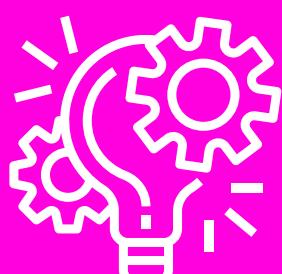
Normal

DATA TYPE

Continuous

WHAT IT SHOWS

If there is a
significant
differences between
group means



T-TEST OUTPUT



Independent t-test

Group 1: Treatment A ($n = 30$)
Group 2: Treatment B ($n = 35$)

Test Statistic: $t = -2.54$

Degrees of Freedom: $df = 63$

p-value: $p = 0.014$

Test Statistic

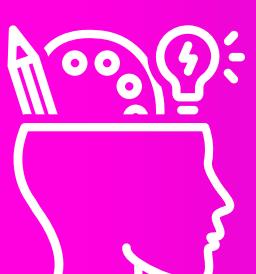
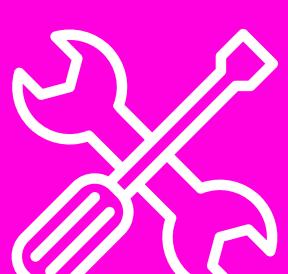
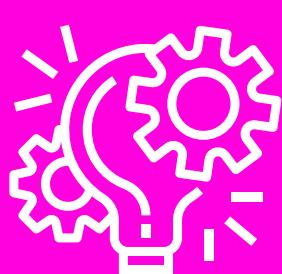
The t-value is calculated based on the difference in means between the two groups and the variability within the groups.

Degrees of Freedom

The number of independent pieces of information available to estimate the population parameter.

p-value

Probability of obtaining the observed difference (or a more extreme difference) between the groups by chance alone, assuming that the null hypothesis is true (i.e., there is no difference between the groups)



CHI-SQUARE

PURPOSE

Test for association between variables

WHEN TO USE IT

Assess relationship between categorical variables

DISTRIBUTION

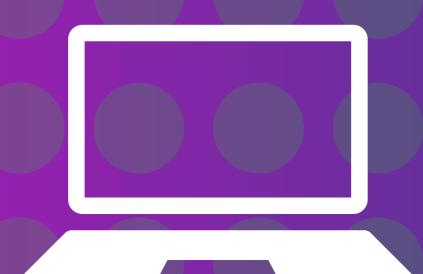
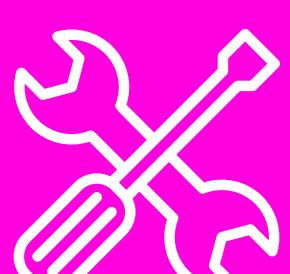
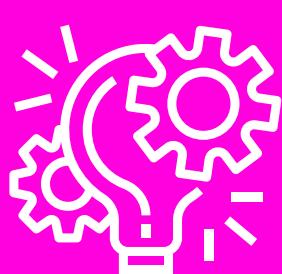
No strict distribution requirement

DATA TYPE

Categorical

WHAT IT SHOWS

Look for significant differences between observed and expected values



CHI-SQUARE

OUTPUT



Chi-Square Test of Independence

	Group A	Group B	Group C	Group D
Observed	50	45	35	60
Expected	40	50	30	60
Total	190	190	190	190

Chi-Square Test Statistics

Chi-Square Value	8.96
Degrees of Freedom	3
p-value	0.0304

Chi-Square Value

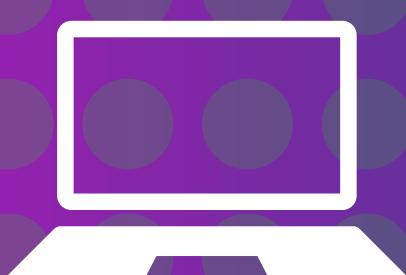
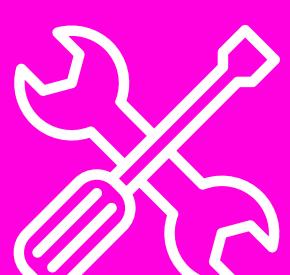
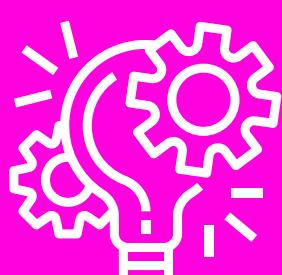
Measures the discrepancy between the observed and expected frequencies.

Degrees of Freedom

The number of categories minus 1

p-value

The probability associated with the test statistic. It indicates the level of statistical significance.



ANOVA

PURPOSE

Compare means of multiple groups

WHEN TO USE IT

Three or more groups

DISTRIBUTION

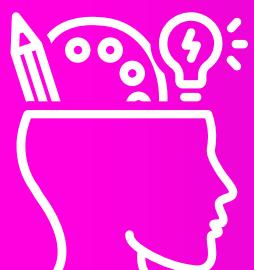
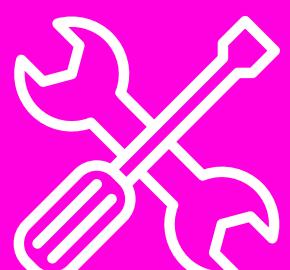
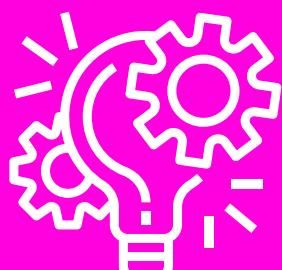
Normally distributed

DATA TYPE

Numerical

WHAT IT SHOWS

Significant differences between group means



ANOVA OUTPUT



Analysis of Variance (ANOVA)

Source	Sum of Squares	Degrees of Freedom	Mean Square	F Value	p-value
Between Groups	45.12	2	22.56	3.74	0.031
Within Groups	126.48	27	4.69		
Total	171.60	29			
Chi-Square Value	8.96				
Degrees of Freedom	3				
p-value	0.0304				

Between Groups

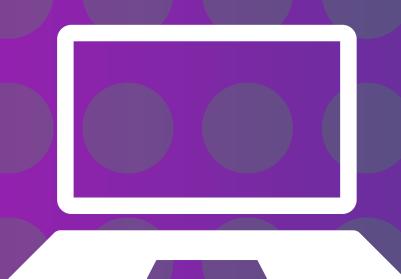
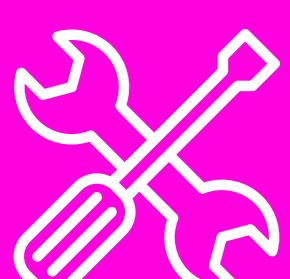
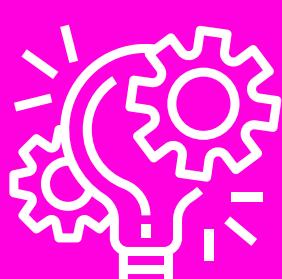
Information about the variation between the different groups being compared.

Within Groups

Information about the variation within each group.

Total

Overall sum of squares and degrees of freedom for the entire dataset, combining the between and within group variations.



REGRESSION

PURPOSE

Examine relationships between variables

WHEN TO USE IT

Predict the value of a dependent variable

DISTRIBUTION

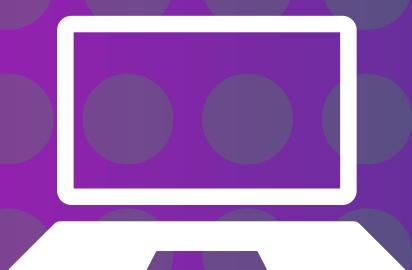
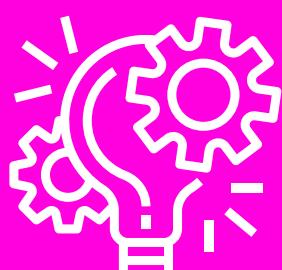
No strict distribution requirement

DATA TYPE

Numerical

WHAT IT SHOWS

Assess the strength and significance of relationships



REGRESSION

OUTPUT



	Coefficient	Standard Error	t-value	p-value
Intercept	12.345	0.678	18.154	<0.001
X_Variable	0.987	0.043	22.885	<0.001

Regression Equation

$$Y = 12.345 + 0.987 * X_{\text{Variable}}$$

Coefficients

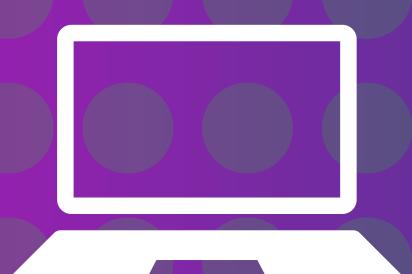
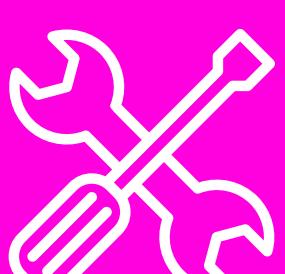
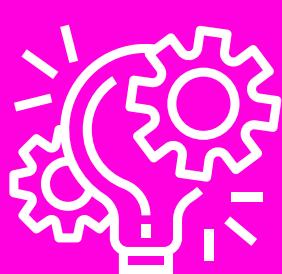
- The intercept (12.345) represents the estimated value of the dependent variable when the independent variable (X_{Variable}) is zero.
- The coefficient for X_{Variable} (0.987) represents the estimated change in the dependent variable for a one-unit increase in X_{Variable} .

R-Square

Proportion of the variance in the dependent variable that is explained by the independent variables.

p-value

Statistical significance of a coefficient.



Mann-Whitney U

Test

PURPOSE

Compare distributions of two groups

WHEN TO USE IT

Compare distributions of two independent groups

DISTRIBUTION

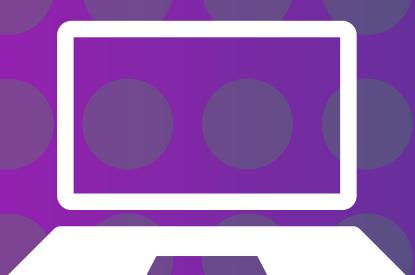
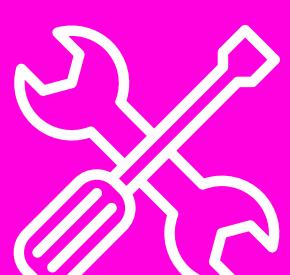
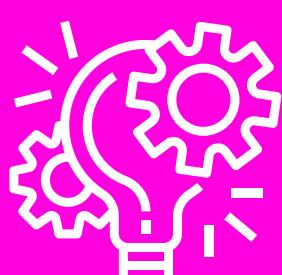
No strict distribution requirement

DATA TYPE

Numerical/Ordinal

WHAT IT SHOWS

Significant differences in rank order



MANN-WHITNEY

OUTPUT



Mann-Whitney U Test Results:

U Statistic: 1234.5

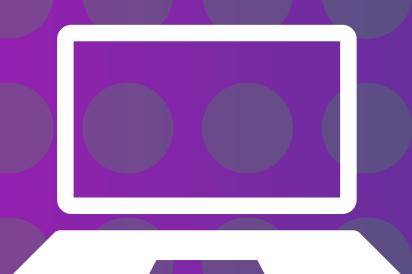
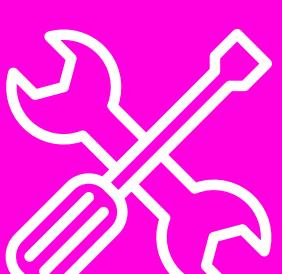
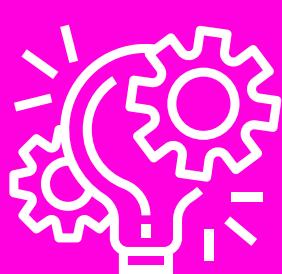
p-value: 0.0234

U Statistic

Rank-based test statistic used in the Mann-Whitney U test. It quantifies the degree of difference between the two groups.

p-value

Statistical significance of the test. It indicates the probability of obtaining the observed difference between the groups if there were no true differences in the populations from which the samples were drawn.



Kruskal-Wallis

PURPOSE

Compare distributions of multiple groups

WHEN TO USE IT

Compare distributions of three or more independent groups

DISTRIBUTION

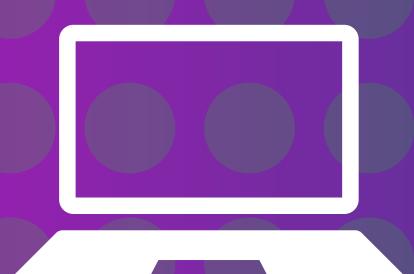
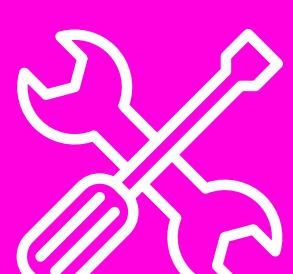
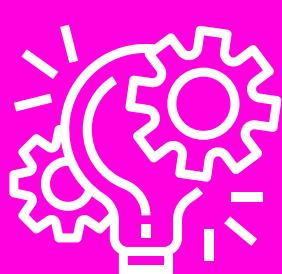
No strict distribution requirement

DATA TYPE

Numerical/Ordinal

WHAT IT SHOWS

Look for significant differences in rank order



Kruskal-Wallis

Output



Kruskal-Wallis Test

$H = 12.45$

$df = 2$

$p\text{-value} = 0.0023$

H

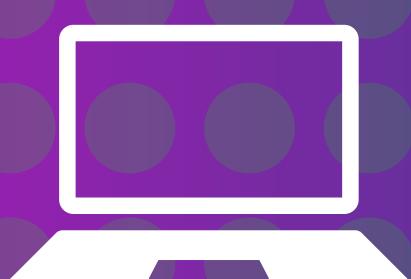
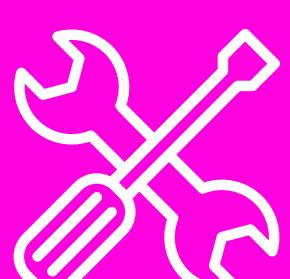
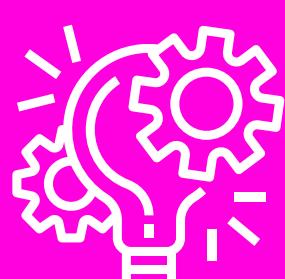
Sum of ranks across all groups and is used to assess the differences between the groups.

Degrees of Freedom

Number of groups minus 1

p-value

Strength of evidence against the null hypothesis (the assumption that there are no differences between the groups).



Pearson's Correlation

PURPOSE

Measure the strength of linear relationship

WHEN TO USE IT

Assess the strength and direction of a linear relationship

DISTRIBUTION

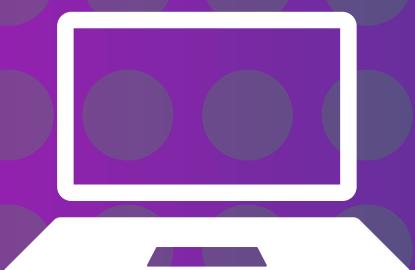
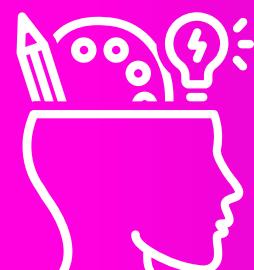
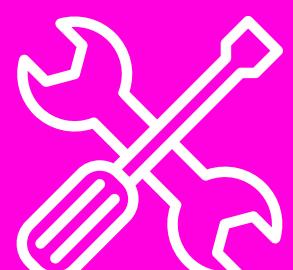
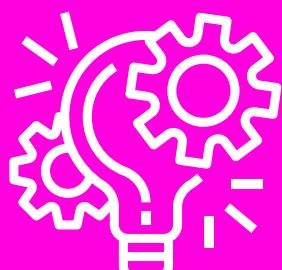
Normally distributed

DATA TYPE

Numerical

WHAT IT SHOWS

Look for correlation coefficient and its significance



Pearson's Correlation Output



Correlation Coefficient (r): 0.85

p-value: < 0.001

Sample Size (n): 100

p-value = 0.0023

Correlation Coefficient (r)

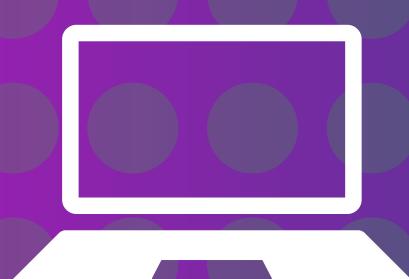
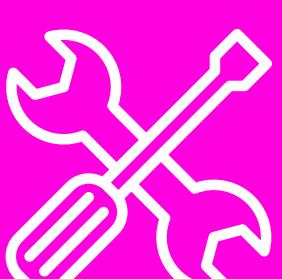
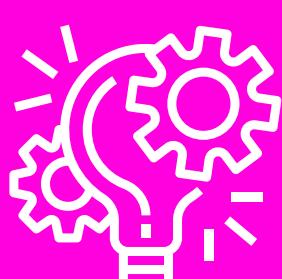
Strength and direction of the linear relationship between the variables. It ranges from -1 to +1. A positive value indicates a positive correlation, a negative value indicates a negative correlation, and a value close to zero indicates a weak or no correlation.

p-value

Probability of observing the given correlation coefficient by chance.

Sample Size (n)

Number of data points used to calculate the correlation coefficient.



Spearman's Correlation

PURPOSE

Measure the strength of monotonic relationship

WHEN TO USE IT

Assess the strength and direction of a monotonic relationship

DISTRIBUTION

No strict distribution requirement

DATA TYPE

Numerical/Ordinal

WHAT IT SHOWS

Look for correlation coefficient and its significance

Spearman's Correlation Output



Spearman's Rank-Order Correlation Test:

Correlation-0.732

p-0.002

Sample-50

**Correlation
Coefficient (r)**

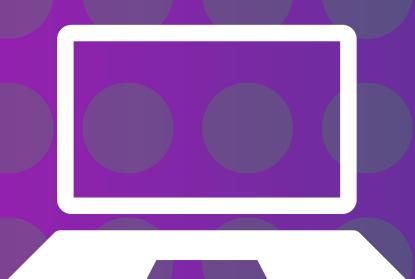
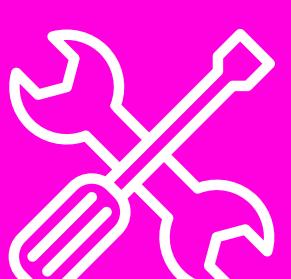
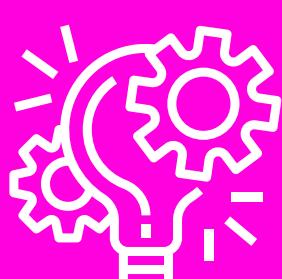
Strength and direction of the linear relationship between the variables.
It ranges from -1 to +1. A positive value indicates a positive correlation, a negative value indicates a negative correlation, and a value close to zero indicates a weak or no correlation.

p-value

Probability of observing the given correlation coefficient by chance.

**Sample Size
(n)**

Number of data points used to calculate the correlation coefficient.



One-Sample T-Test

PURPOSE

Compare sample mean to a known population mean

WHEN TO USE IT

Compare a sample mean to a known value

DISTRIBUTION

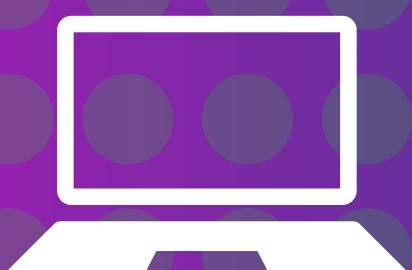
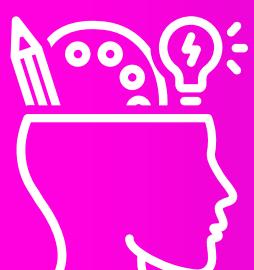
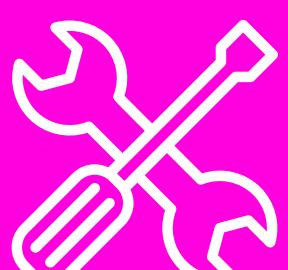
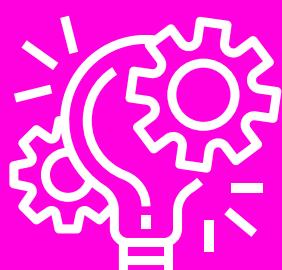
Normally distributed

DATA TYPE

Numerical

WHAT IT SHOWS

Look for significant differences between the sample mean and the known population mean



One Sample T-Test

Output



One-Sample t-Test

Variable: SampleData

Sample Mean: 25.6

Population Mean: 30.0

Sample Size: 50

Degrees of Freedom: 49

t-Statistic: -2.345

p-value: 0.023

t-statistic

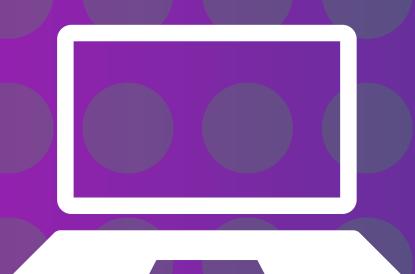
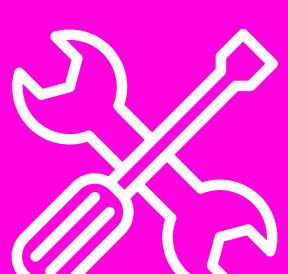
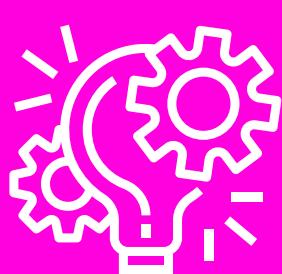
Difference between the sample mean and the hypothesized population mean in terms of standard errors

p-value

Probability of obtaining the observed difference (or a more extreme difference) between the sample and the hypothesized population by chance alone.

Sample Size (n)

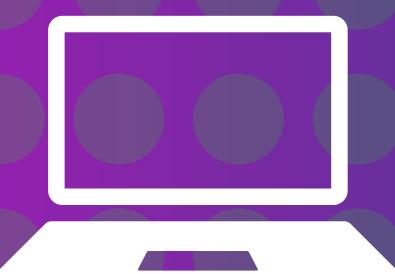
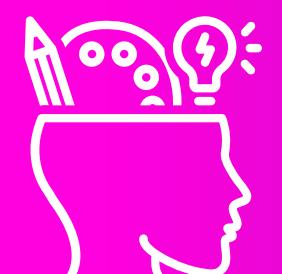
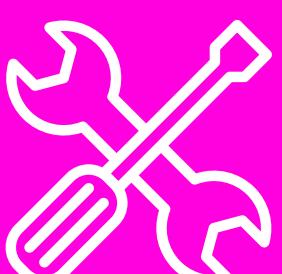
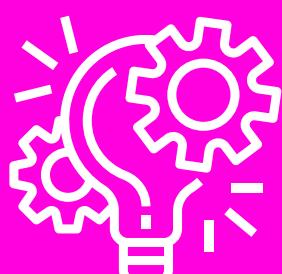
Number of data points used to calculate the correlation coefficient.



Wilcoxon

Signed-Rank

PURPOSE	Compare paired samples
WHEN TO USE IT	Compare paired observations
DISTRIBUTION	No strict distribution requirement
DATA TYPE	Numerical/Ordinal
WHAT IT SHOWS	Look for significant differences between paired observations



Wilcoxon Signed-Rank Output



Wilcoxon Signed-Rank Test

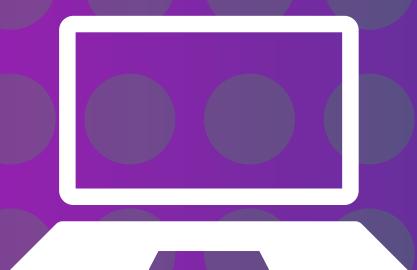
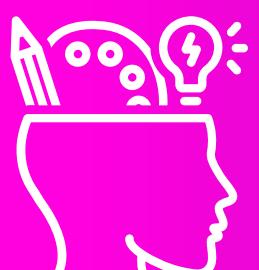
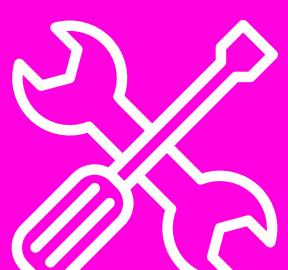
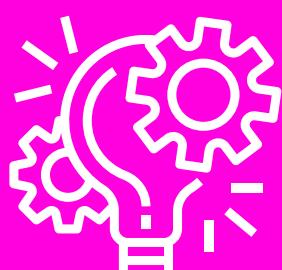
```
data: x  
V = 45, p-value = 0.028  
alternative hypothesis: true location is not equal to 0
```

V

Summarizes the data and is used to assess the statistical significance of the test.

p-value

Statistical significance of the test



YOU DID IT!

HOORAY!

Save this post, and tag me
as you develop these data
analytics core skills.

HAPPY LEARNING!

