Interpreting Data Using Statistical Models with Python

UNDERSTANDING INFERENTIAL STATISTICS

Janani Ravi
CO-FOUNDER, LOONYCORN
www.loonycorn.com

Overview

Descriptive and inferential statistics

Hypothesis testing

Null hypothesis and alternative hypothesis

Common statistical tests: t-test, Z-test, chi2

Power, alpha and p-value of a statistical test

Understanding Analysis of Variance i.e. ANOVA

Prerequisites and Course Outline

Prerequisites

Basic Python programming

Understanding mean, median, mode, distributions

Prerequisite Courses

Python Fundamentals

Course Outline

Understanding inferential statistics

Performing hypothesis testing in Python

Implementing predictive models for continuous data

Implementing predictive models for categorical data

Statistics in Understanding Data

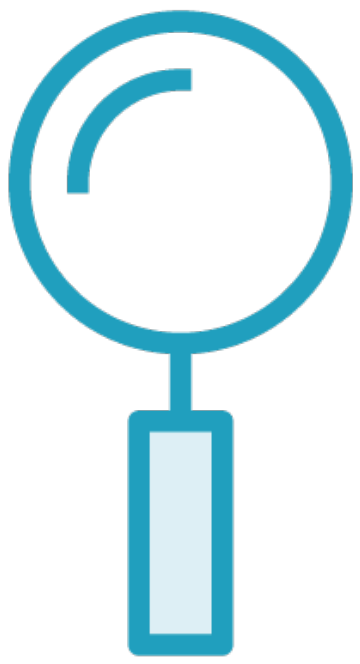
“There are two kinds of statistics, the kind you look up and the kind you make up”

Rex Stout

Statistics

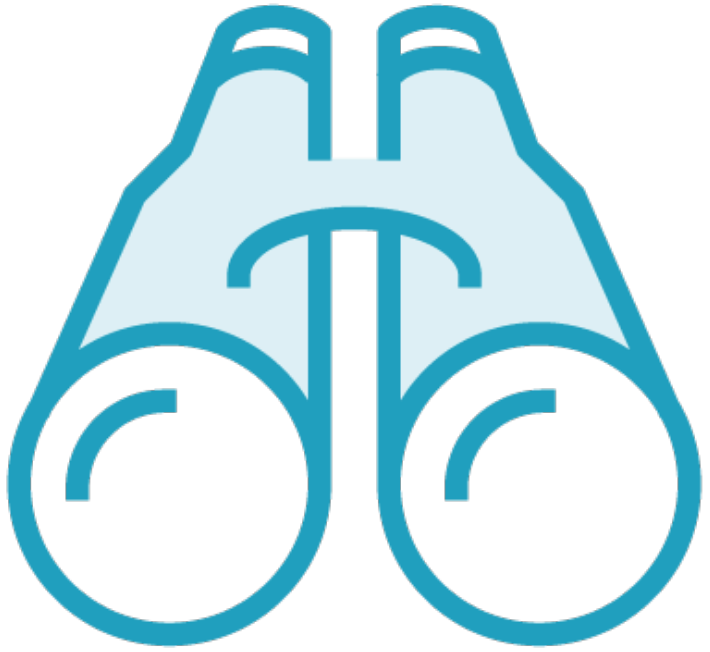
A branch of mathematics that deals with collecting, organizing, analyzing, and interpreting data

Two Sets of Statistical Tools



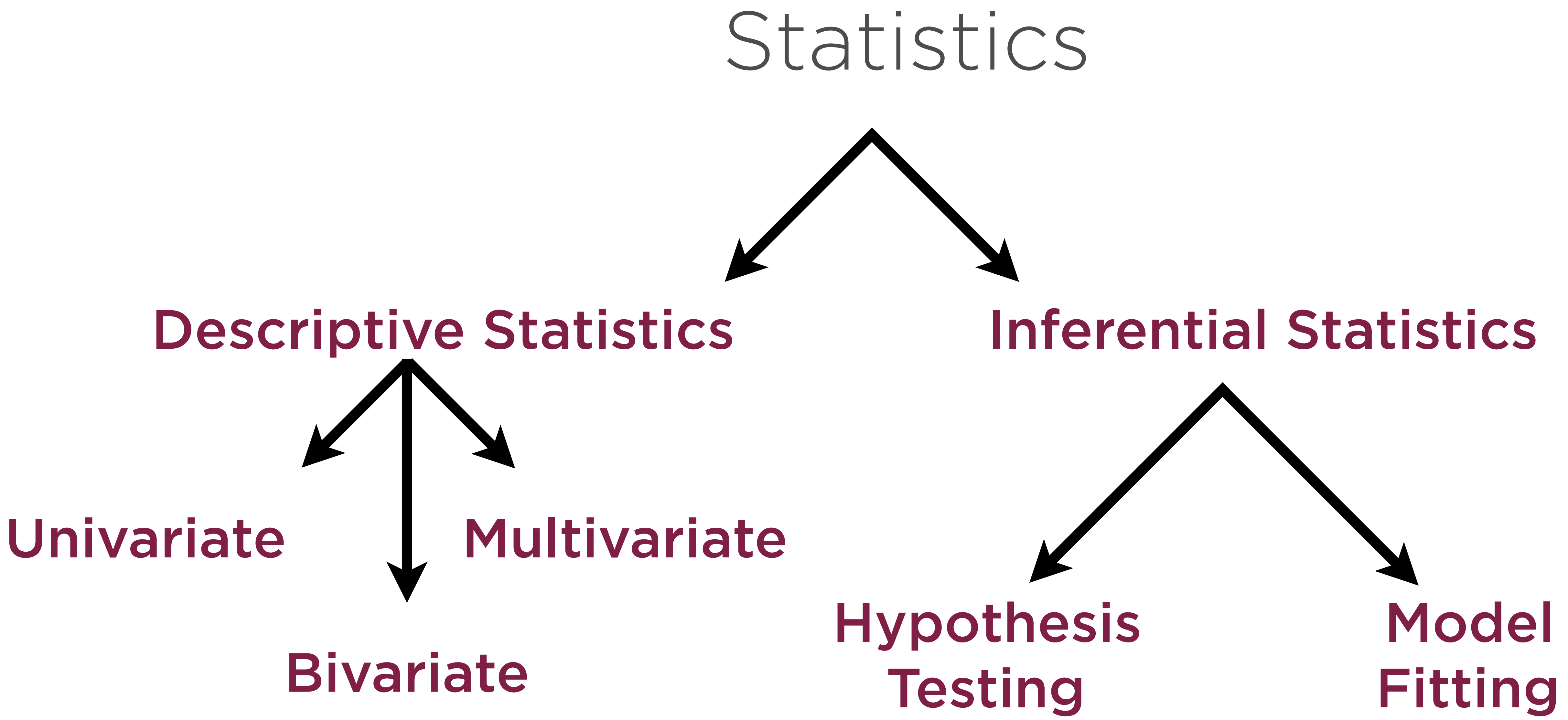
Descriptive Statistics

Identify important elements in a dataset



Inferential Statistics

Explain those elements via relationships with other elements


Statistics
Descriptive Statistics
Inferential Statistics
Univariate
Bivariate
Multivariate
Hypothesis Testing
Model Fitting

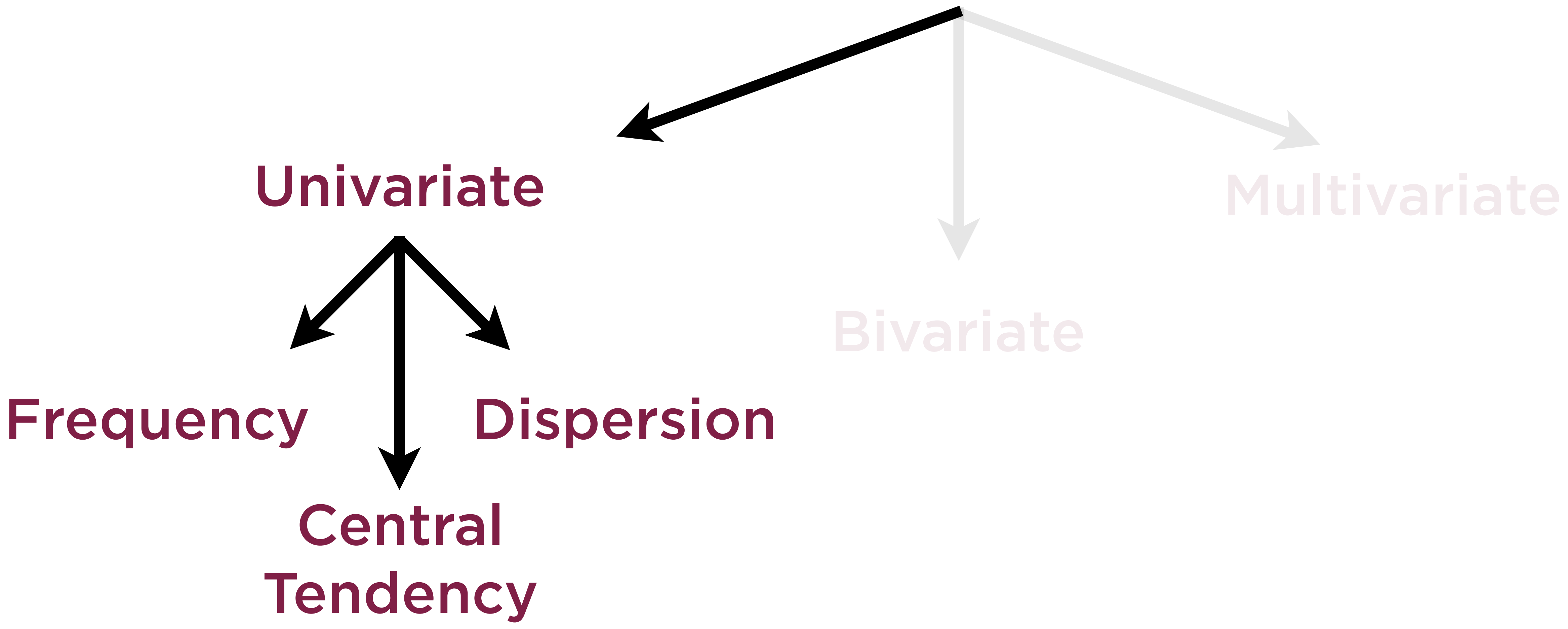
Descriptive Statistics

Univariate

Bivariate

Multivariate

Descriptive Statistics



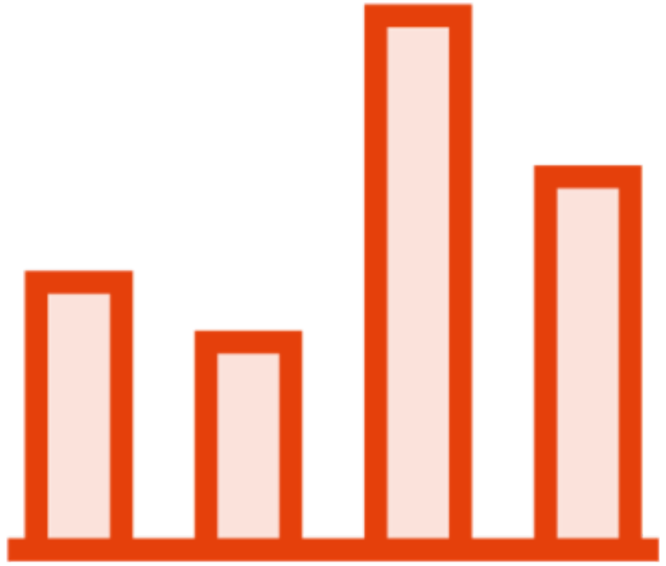
Univariate Descriptive Statistics

Measures of Frequency

Measures of Central Tendency

Measures of Dispersion

Measures of Frequency



Frequency tables

Histograms

Measures of Central Tendency

Average (Mean)

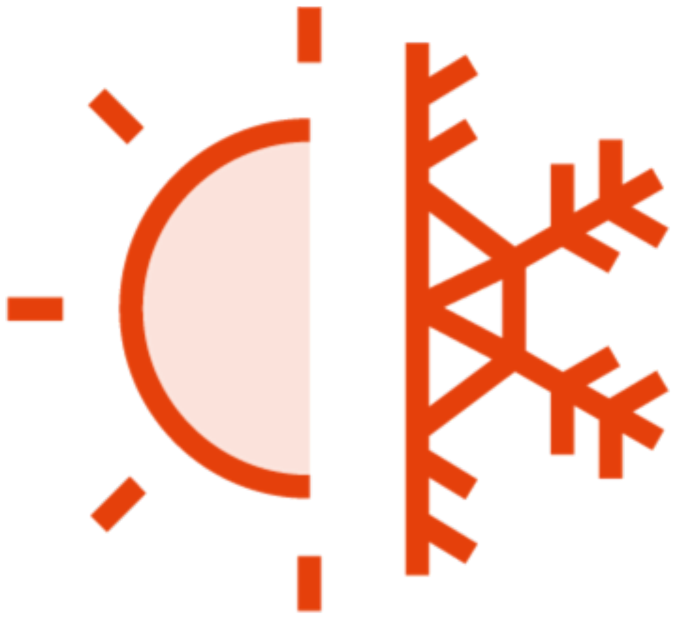
Median

Mode

Other infrequently used measures

- Geometric Mean
- Harmonic Mean

Measures of Dispersion

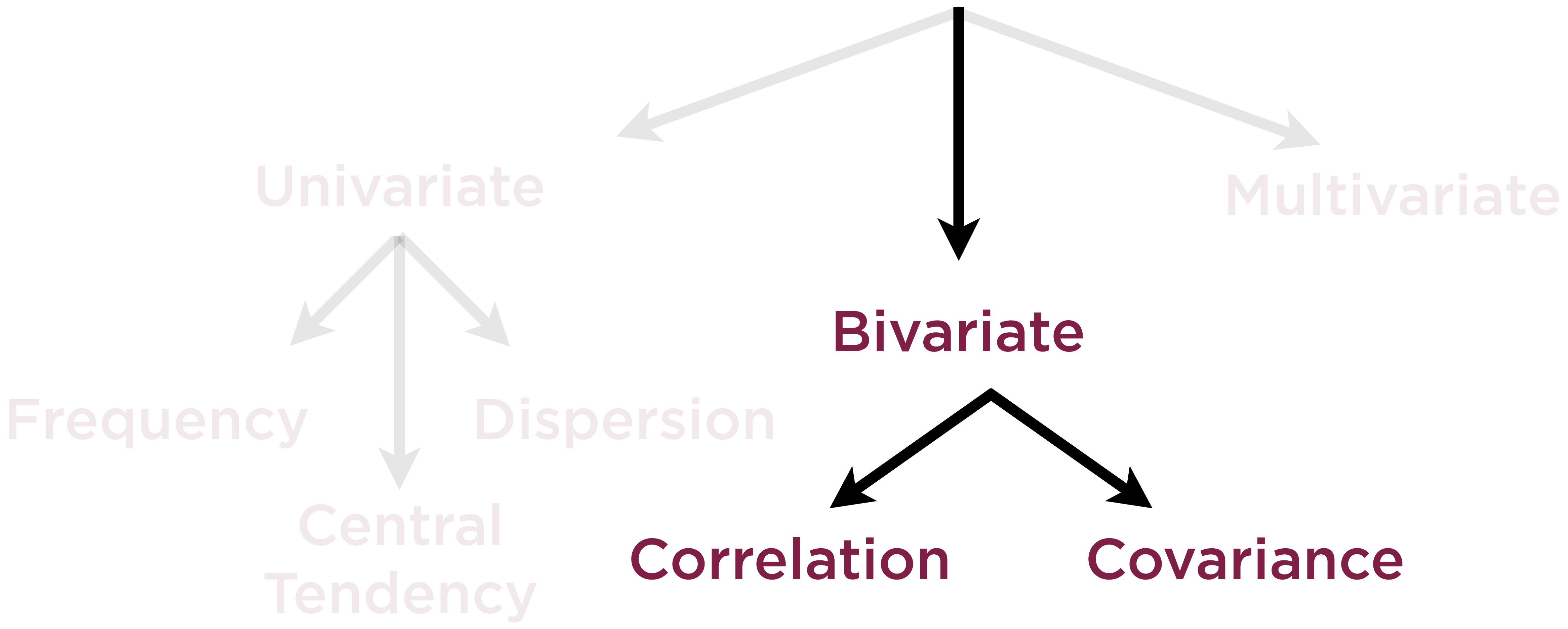


Range (max - min)

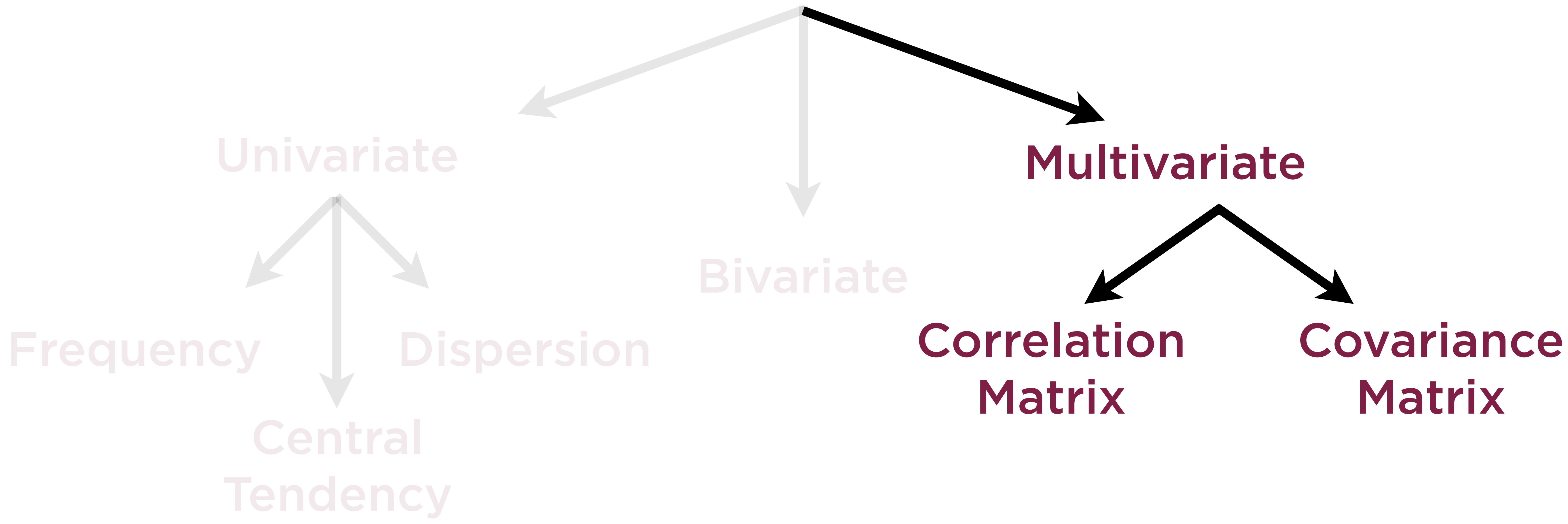
Inter-quartile range (IQR)

Standard deviation and variance

Descriptive Statistics



Descriptive Statistics



Bivariate Descriptive Statistics

Correlation

Covariance

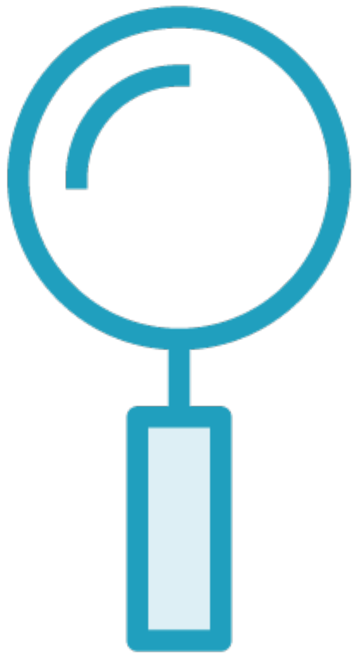
Multivariate Descriptive Statistics

Correlation Matrices

Covariance Matrices

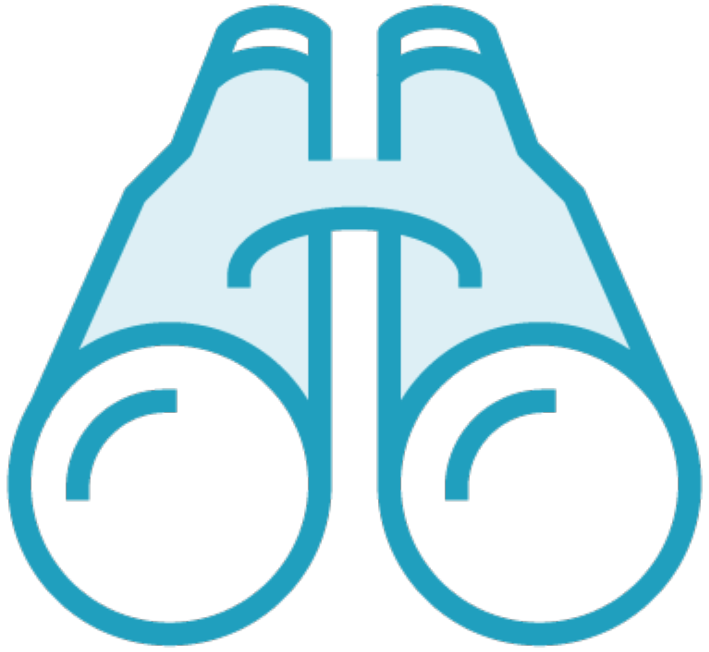
Hypothesis Testing

Two Sets of Statistical Tools



Descriptive Statistics

Identify important elements in a dataset



Inferential Statistics

Explain those elements via relationships with other elements

From Statistics to ML

Descriptive Statistics
Explore the data

No points-of-view yet

Inferential Statistics
Frame hypotheses and test them

Tentatively evaluating many points-of-view

Rule-based Learning Models
Frame rules based on the data

Performed by experts - risk of too much certainty

Machine Learning Models
Build models that change with the data

Full circle - back to no points-of-view

From Statistics to ML

Descriptive Statistics
Explore the data

No points-of-view yet

Rule-based Learning Models
Frame rules based on the data

Performed by experts - risk of too much certainty

Inferential Statistics
Frame hypotheses and test them

Tentatively evaluating many points-of-view

Machine Learning Models
Build models that change with the data

Full circle - back to no points-of-view

Hypothesis

Proposed explanation for a phenomenon.

Hypothesis

Proposed explanation

Objectively testable

Singular - hypothesis

Plural - hypotheses

Hypothesis Testing

Null Hypothesis H_0
True until proven false

Usually posits no relationship

Alternative Hypothesis
Negation of null hypothesis

Usually asserts specific relationship

Select Test
Pick from vast library

Know which one to choose

Test Statistic
Convert to p-value

How likely it was just luck?

Significance Level
Usually 1% or 5%

What threshold for luck?

Accept or Reject
Small p-value? Reject H_0

Small: Below significance level

Lady Tasting Tea

Lady tasting tea: famous experiment

Was tea added before or after milk?

Muriel Bristol claimed she could tell

Lady Tasting Tea

Null Hypothesis

(H_0)

The lady cannot tell if milk was poured first

Alternate Hypothesis

(H_1)

The lady can tell if milk was poured first

Lady Tasting Tea

Null Hypothesis

The lady **cannot** tell if the milk was poured first

Alternate Hypothesis

The lady **can** tell if the milk was poured first

It is good practice to assume that the null hypothesis is correct unless proven otherwise

Lady Tasting Tea

Null Hypothesis

The lady cannot tell if the milk was poured first

Alternate Hypothesis

The lady can tell if the milk was poured first

It is good practice to assume that the null hypothesis is correct unless proven otherwise

Lady Tasting Tea

Null Hypothesis H_0

"Lady cannot tell difference"

Can't tell if milk poured first

Alternative Hypothesis

"Lady can tell difference"

Can indeed discern if milk poured first

Select Test

8 cups, 4 of each type

Lady got all 8 correct

Test Statistic

p-value = 1/70 = 1.4%

${}^{8}C_{4}$ = 70 combinations

Significance Level

Choose 5% significance level

Part of design of experiment

Accept or Reject

1.4% < 5% => Reject H_0

Lady can indeed tell difference

Lady Tasting Tea

Experiment proved that she could

Conducted by Sir Ronald Fisher

(considered founder of modern statistics)

Errors in Hypothesis Testing

		Decision about Null Hypothesis	
		REJECT	**DON'T REJECT**
Null Hypothesis is actually	**TRUE**	**Type I error**	**Correct Inference**
	FALSE	**Correct Inference**	**Type II error**

Errors in Hypothesis Testing

		Decision about Null Hypothesis	
		REJECT	DON'T REJECT
Null Hypothesis is actually	TRUE	Type I error	
	FALSE		

Claim the lady can tell the difference based on spurious test results which are not statistically significant

Errors in Hypothesis Testing

		Decision about Null Hypothesis	
		REJECT	DON'T REJECT
Null Hypothesis is actually	TRUE		
	FALSE		Type II error

Fail to realize that the test for the alternative hypothesis was statistically significant

Power of a Statistical Test

Probability of rejecting H_0 when H_1 is true

Ranges from 0 to 1

High power is good

High statistical power implies low probability of Type-II error

Power of a binary classifier is also known as recall

α of a Statistical Test

α is probability of rejecting H_0 when H_0 is true

α = Probability of Type-I error

Ranges from 0 to 1

High α is not good

p-value of a Statistical Test

Same as statistical significance

P-value is compared to α to decide whether to accept H_0

P-value should be as small as possible (i.e. below α-threshold)

Typical cut-off values for statistical significance are 1% and 5%

The t-test and Z-test

Hypothesis Testing

Null Hypothesis H_0

True until proven false

Usually posits no relationship

Select Test

Pick from vast library

Know which one to choose

Significance Level

Usually 1% or 5%

What threshold for luck?

Alternative Hypothesis

Negation of null hypothesis

Usually asserts specific relationship

Test Statistic

Convert to p-value

How likely it was just luck?

Accept or Reject

Small p-value? Reject H_0

Small: Below significance level

Hypothesis Testing

Null Hypothesis H_0

True until proven false

Usually posits no relationship

Select Test

Pick from vast library

Know which one to choose

Significance Level

Usually 1% or 5%

What threshold for luck?

Alternative Hypothesis

Negation of null hypothesis

Usually asserts specific relationship

Test Statistic

Convert to p-value

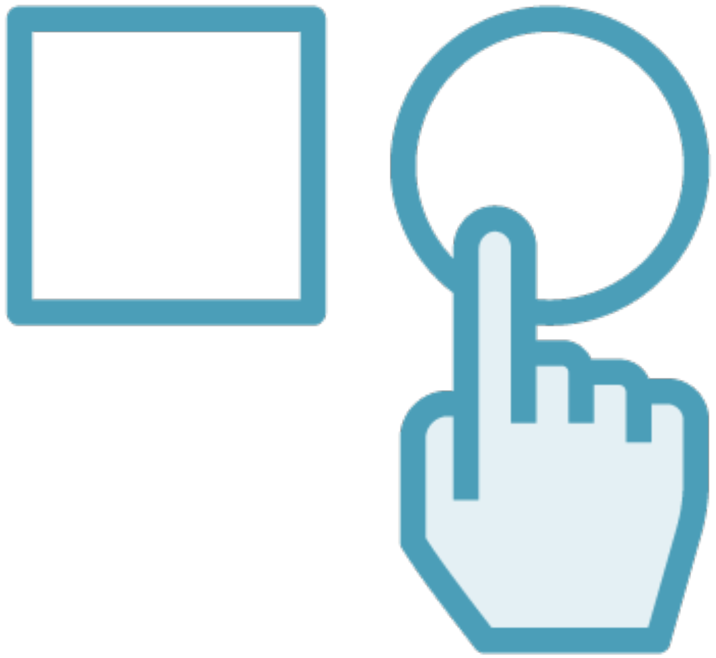
How likely it was just luck?

Accept or Reject

Small p-value? Reject H_0

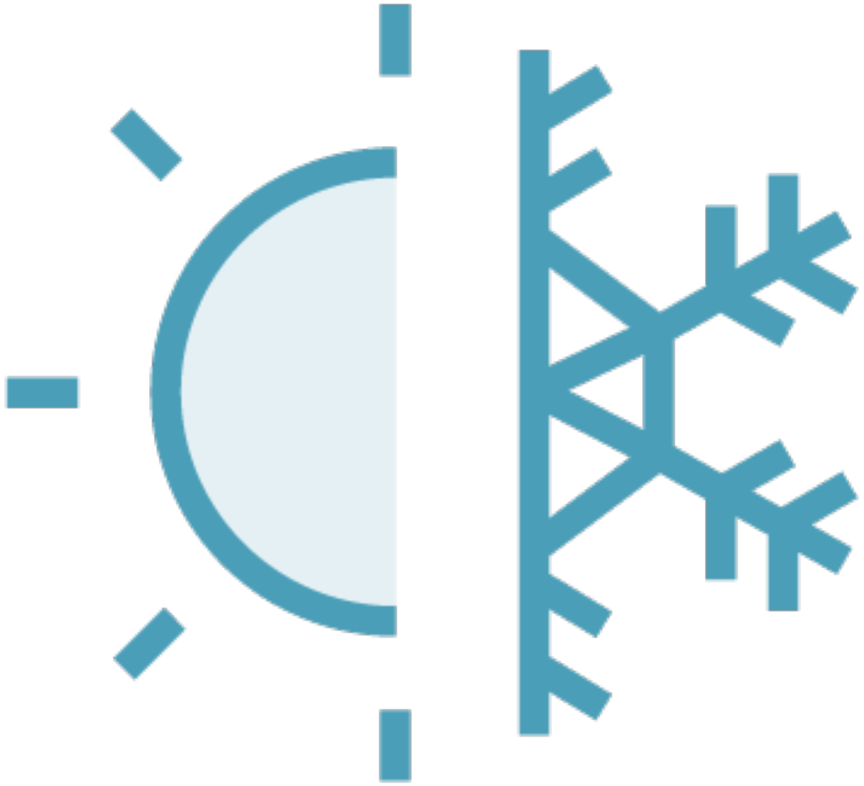
Small: Below significance level

Statistical Test Selection



- There are tests for pretty much everything
- Developed by statisticians to be sound
- Knowing which one to use is hard
- Actually using them is relatively easy

t-tests



Most common, simple statistical tests out there

Used to learn about averages across two categories

Also tells whether the differences are significant

t-tests

Average male baby birth weight =
Average female baby birth weight?

Is the difference statistically significant?

t-tests

t-statistic

- Score which indicates the difference in means

P-value

- Whether the t-statistic is significant
- Low p-values of <5% mean the result cannot be due to chance

Assumptions of t-tests

Sample mean(s) are normally distributed

Sample variance(s) follow chi^2 distribution

Sample mean and variance are independent

Some more mathematical fine print around degrees of freedom etc.

Types of t-tests

One sample location test

Two sample location test

Paired difference test

Regression coefficient test

One-sample Location Test

One sample location test

What is the average weight of babies born in a certain town?

Is it different from the average of the general population?

One-sample Location Test

One sample location test

Null hypothesis of form

"Population mean is equal to specified value"

$$H_0: \mu = \mu_0$$

One-sample Location Test

One sample location test

Test statistic

$$t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$$

Related Test: Z-test

Test statistic of one sample t-test follows Student's t-distribution

The same test statistic can be used for the simpler Z-test if

- Number of samples is large (>>30)
- Population variance is known

Z-test assumes test statistic follows normal distribution

Related Test: Z-test

Z-test is simpler to interpret as compared with the t-test

Need not take into account the degrees of freedom

Related Test: Z-test

However, population variance is rarely known in practice

So, t-test is usually preferred to Z-test

Two-sample Location Test

Two sample location test

Is the average weight of babies in Town A different from that in Town B?

Two-sample Location Test

Two sample location test

Null hypothesis of form

“Population means of two samples are equal”

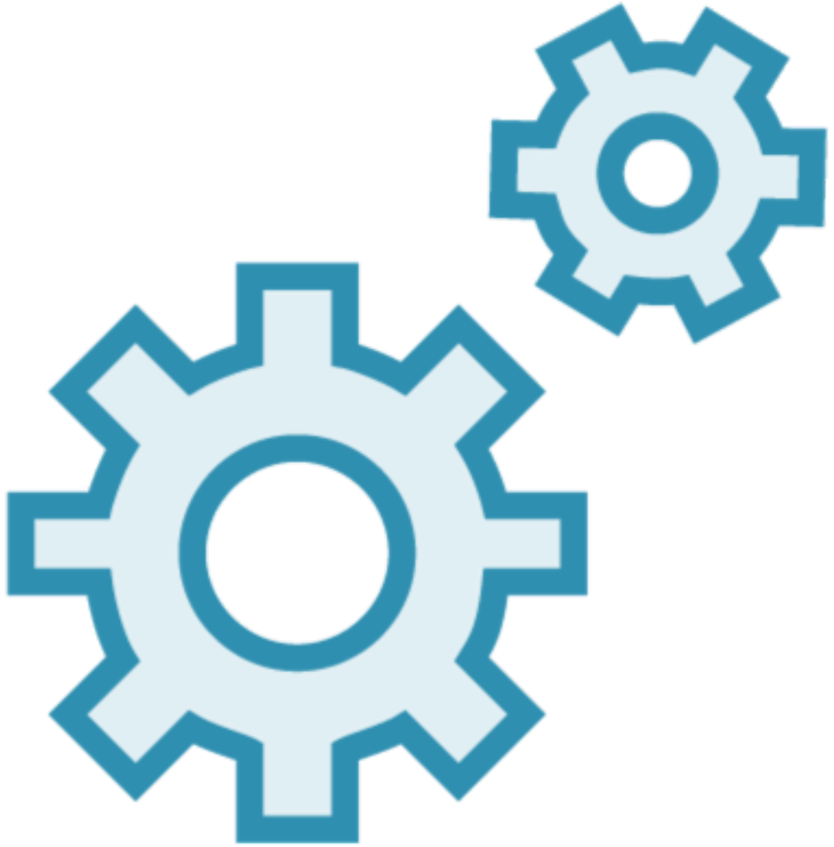
Two-sample Location Test

Two sample location test

Slightly different test statistics for

- Equal sample sizes, equal variance
- Unequal sample sizes, equal variance
- Equal or unequal sample sizes, unequal variances (Welch’s t-test)

Related Test: Levene’s Test

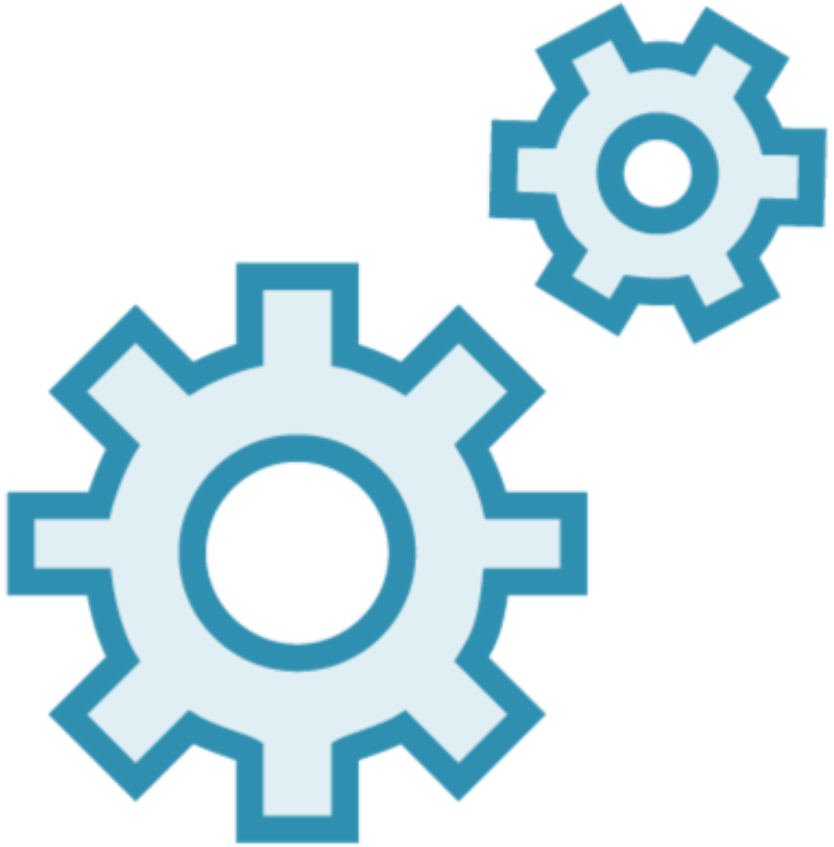


Different forms of t-test based on whether variances are equal or not

So need a way to test for equality of variances

Levene’s test serves this purpose

Related Test: Levene's Test



Null hypothesis: Populations from which two samples are drawn have equal variance

If Levene's test shows that null hypothesis needs to be rejected

- Use two sample t-test for unequal variances (Welch's t-test)
- Else can use two sample t-test for equal variances

Paired Difference Test

Paired difference test

Is the average weight of babies born in winter different from babies born in summer?

Paired Difference Test

Paired difference test

In the one sample and two sample tests, samples are assumed to be independent

Those forms of tests are not suitable for matched samples

In such cases, use paired difference t-test instead

Wilcoxon Signed-rank Test

Alternative to paired difference t-test

Used to compare median of matched samples

Prefer to paired difference t-test when

- Sample size is small
- Population is not normally distributed

Regression Coefficient Test

Regression coefficient test

Is the coefficient of any of the independent variables > 0?

From Sample to Population

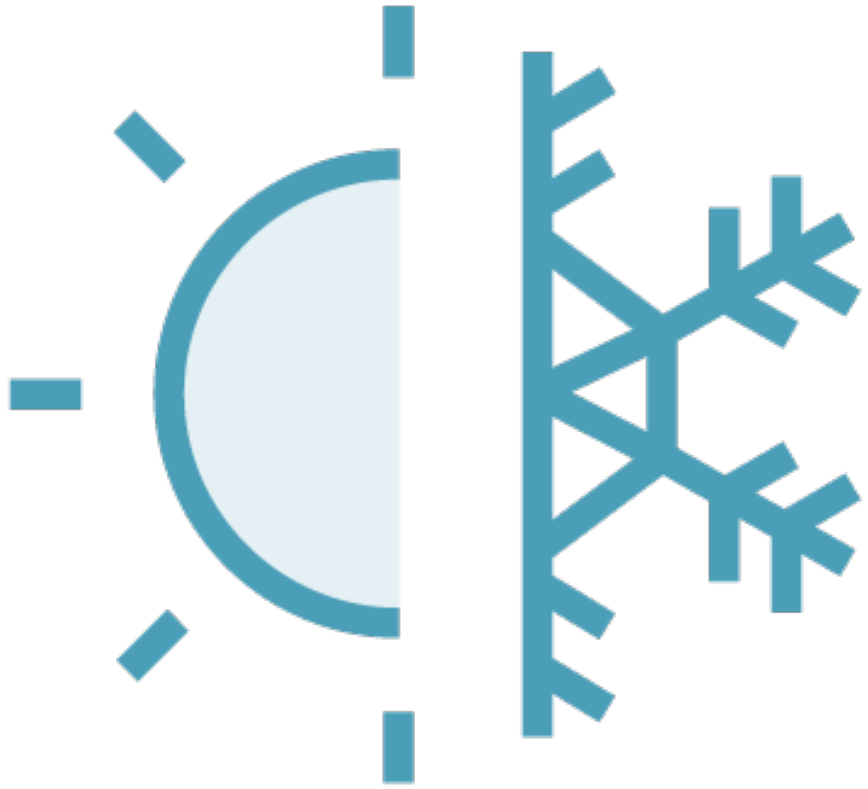
Population

All the data out there in the universe

Sample

A subset - hopefully representative - of the population

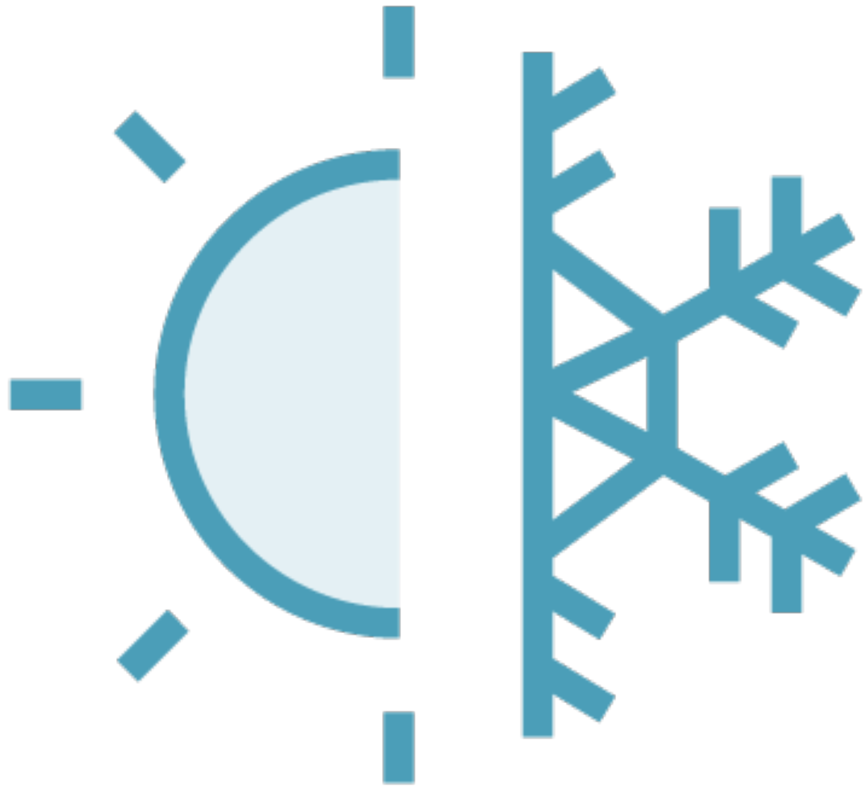
t-tests Assumptions



Notably, that

- populations are normal
- samples are representative
- samples are randomly drawn

t-tests



Work best for two group comparisons

Comparing multiple groups gets tricky

- need many pairwise tests
- increases likelihood of Type 1 error (alpha inflation)

For multiple groups, just use ANOVA

ANOVA

t-tests are useful to compare differences between **two** groups

Running **multiple** significance tests to compare across many groups is **risky**

ANOVA

ANalysis ***O***f ***VA***riance

ANOVA

Looks across multiple groups of populations, compares their means to produce one score and one significance value

ANOVA

Looks across multiple groups of populations, compares their means to produce one score and one significance value

Diabetes Risk

Underweight patients	Normal weight patients	Overweight patients

In order to compare across 3 groups the we'll need to perform multiple t-tests

Diabetes Risk

Underweight patients

Normal weight patients

Overweight patients

Perform a single ANOVA test to know whether the risk of diabetes is significantly different between these groups

ANOVA Hypotheses

Null Hypothesis

(H_0)

Alternate Hypothesis

(H_1)

H_0: All groups of patients are at an equal risk of diabetes

H_1: All groups of patients are NOT at an equal risk of diabetes

ANOVA

Looks across multiple groups of populations, compares their means to produce one score and one significance value

F-statistic

$$F = \frac{\text{Variance between groups}}{\text{Variance within a group}}$$

F-statistic

If the groups are similar, F ~ 1

If the groups are different, F will be large

P-value

Significance of the F-statistic

Smaller p-values indicate that the results are not due to chance

Large F-statistic and small p-value - means the null hypothesis can be rejected

ANOVA Hypotheses

Large F-statistic and small p-values < 0.05 significance level

Accept the alternative hypothesis and reject the null hypothesis

Alternate Hypothesis

(H_1)

H_1: All groups of patients are NOT at an equal risk of diabetes

ANOVA Hypotheses

Null Hypothesis

(H_0)

H_0: All groups of patients are at an equal risk of diabetes

Small F-statistic and large p-values > 0.05 significance level

Accept the null hypothesis and reject the alternative hypothesis

One-way ANOVA helps compare means across two or more groups

A **single** categorical variable is used to split the population into these groups

One-way ANOVA Assumptions

Notably, that

- populations are normal
- samples are representative
- samples are randomly drawn
- variances of the population are constant within each group

Two-way ANOVA

Examines the influence of two different independent variables on one continuous dependent variable

Two-way ANOVA

Examines the influence of two different independent variables on one continuous dependent variable

Two-way ANOVA

Employees > 40	Employees <= 40
Males	Females

Two-way ANOVA

Employees > 40		Employees <= 40	
Males	Females	Males	Females

Two-way ANOVA Hypotheses

Null Hypothesis (H_{01})

H_{01}: All genders have equal levels of stress

Null Hypothesis (H_{02})

H_{02}: All ages have equal levels of stress

Null Hypothesis (H_{03})

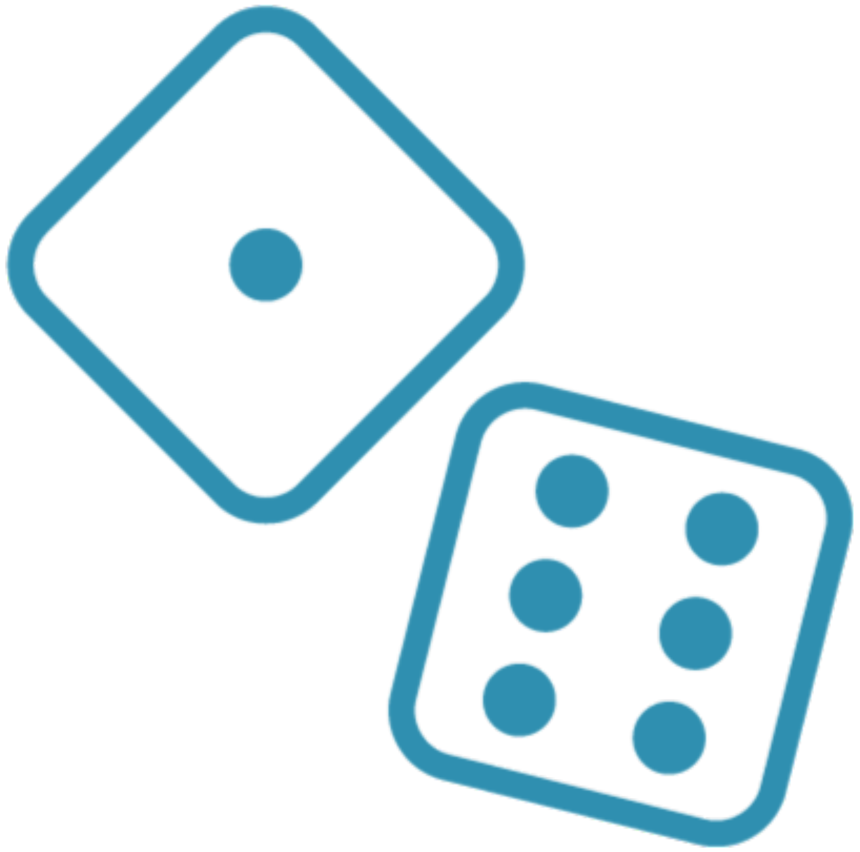
H_{03}: There is no interaction between age and gender

Pearson's χ^2 Test

Pearson's χ^2 Test

Test applied to ascertain whether frequencies of events (values of a categorical variable) follow a specific distribution.

Pearson's X^2 Test

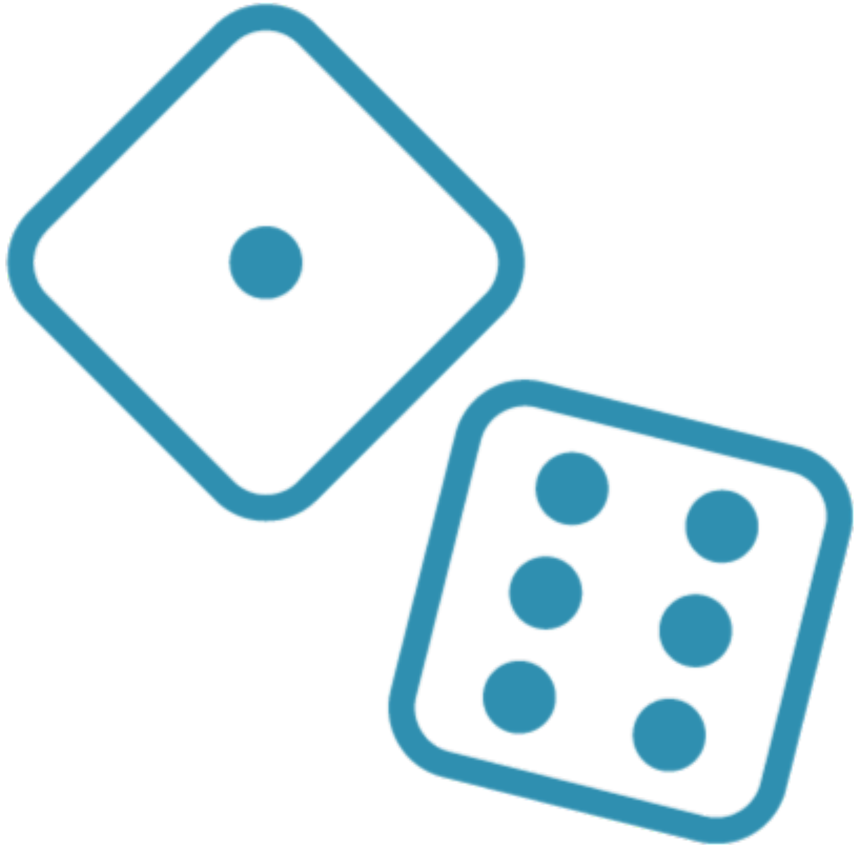


Best understood with an example

Given results of throws of a dice

Are the results consistent with the dice being fair?

Pearson's X^2 Test

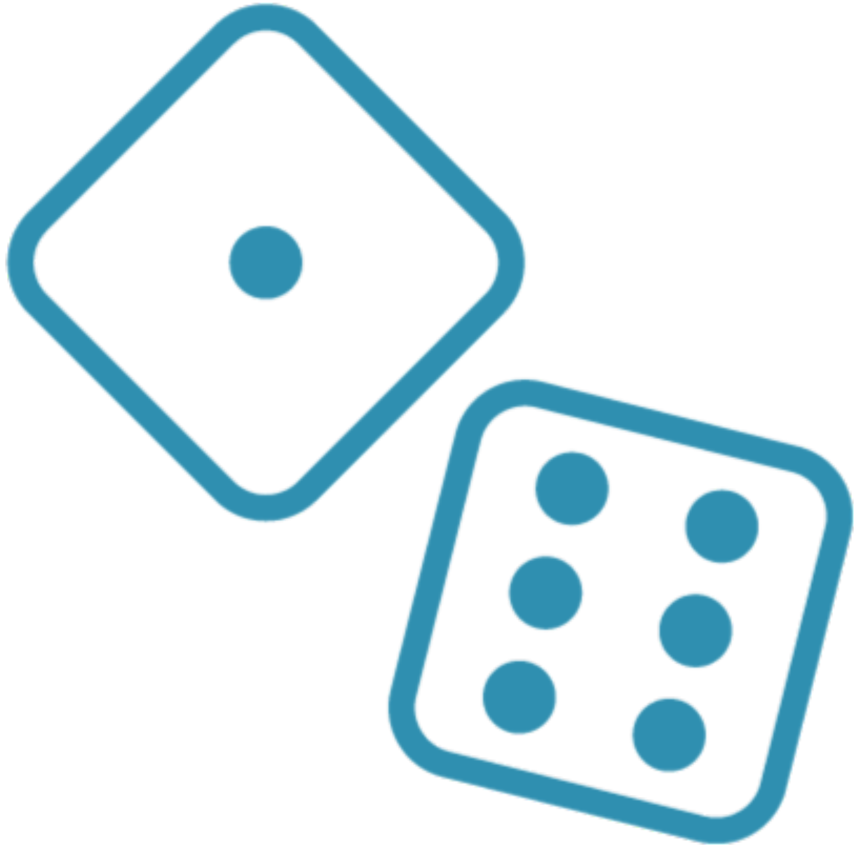


Result of each throw is a categorical variable with values between 1 and 6

If dice is fair, each outcome has equal probability of 1/6

This set of equal probabilities represent the theoretical distribution

Pearson's X^2 Test



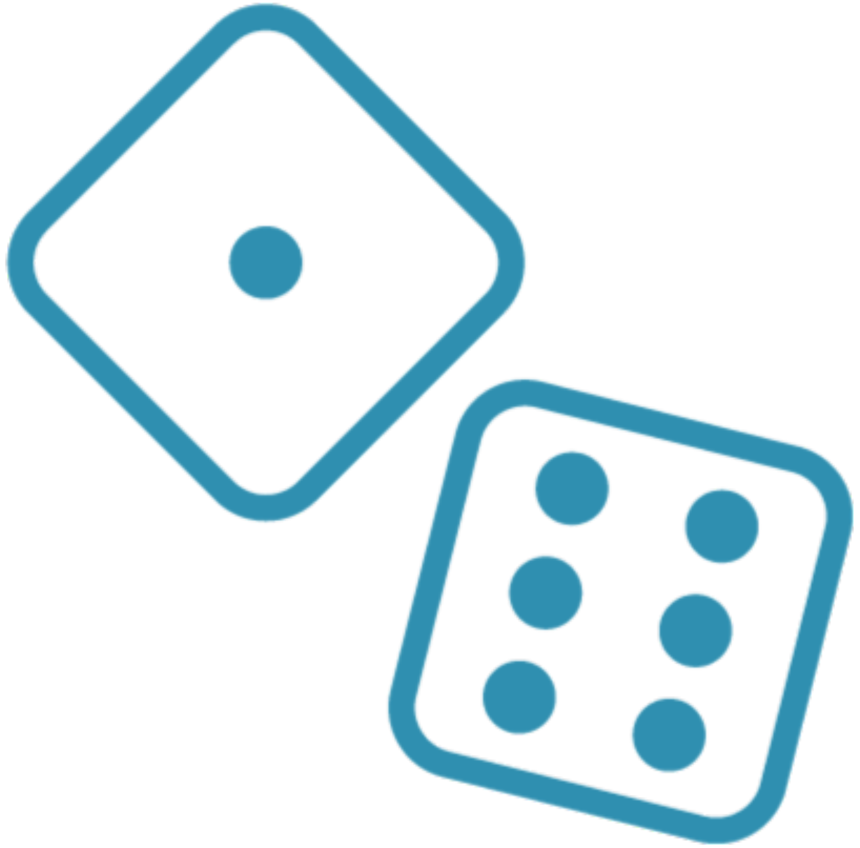
Test statistic follows X^2 distribution

Dice example uses Pearson's test to check for goodness of fit

- Does observed frequency distribution match a theoretical distribution?

Pearson's test can also be used to test for independence

Pearson's X^2 Test for Independence



Take two categorical variables to be tested for independence

Create a contingency table

- Values of 1st categorical variable as rows
- Values of 2nd categorical variable as columns
- Cells correspond to frequency of corresponding combination

Summary

Descriptive and inferential statistics

Hypothesis testing

Null hypothesis and alternative hypothesis

Common statistical tests: t-test, Z-test, chi2

Power, alpha and p-value of a statistical test

Understanding Analysis of Variance i.e. ANOVA