Building Statistical Summaries with R

UNDERSTANDING STATISTICAL SUMMARIES

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

Overview

Descriptive vs. inferential Statistics

Hypothesis testing

Interpreting p-values, power and alpha of a test

Understanding t-tests and z-tests

Type-I and Type-II errors in hypothesis testing

Understanding the chi2 test

Prerequisites and Course Outline

Prerequisites

Some exposure to statistics at the level of mean, median, and standard deviation

Comfortable programming in R

Familiar with Jupyter notebooks

Course Outline

Hypothesis testing - t-tests, one-way and two-way ANOVA, chi2 test

Implementing and interpreting statistical tests

Building predictive models such as linear regression and logistic regression

A/B testing and Bayesian A/B testing

Statistics in Understanding Data

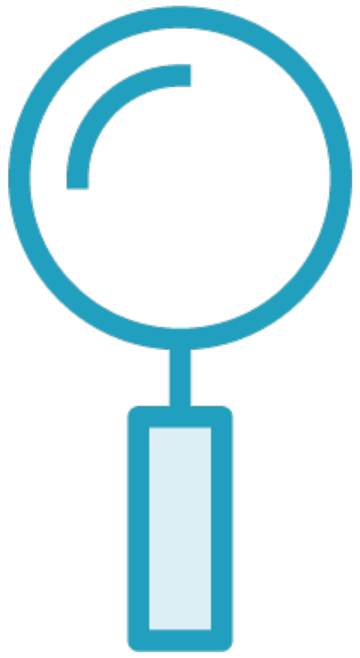
"There are two kinds of statistics, the kind you look up and the kind you make up"

Rex Stout

Statistics

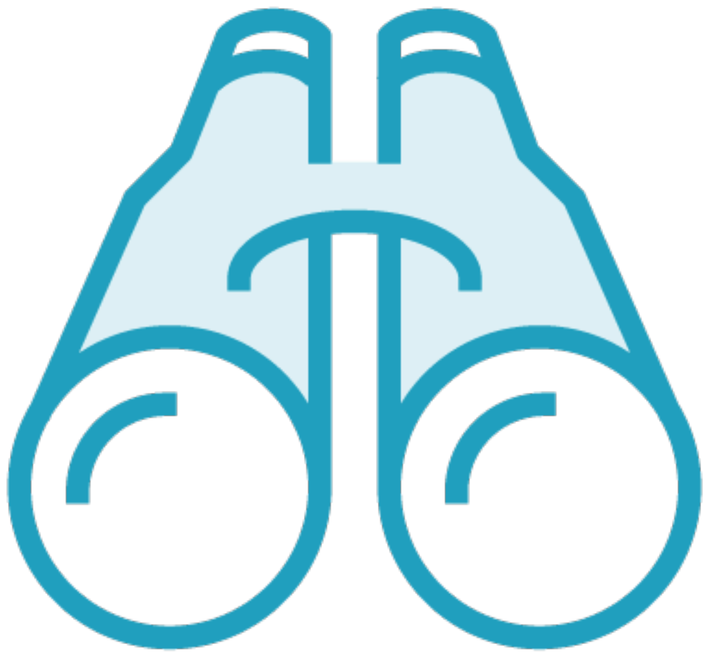
A branch of mathematics that deals with collecting, organizing, analyzing, and interpreting data

Two Sets of Statistical Tools



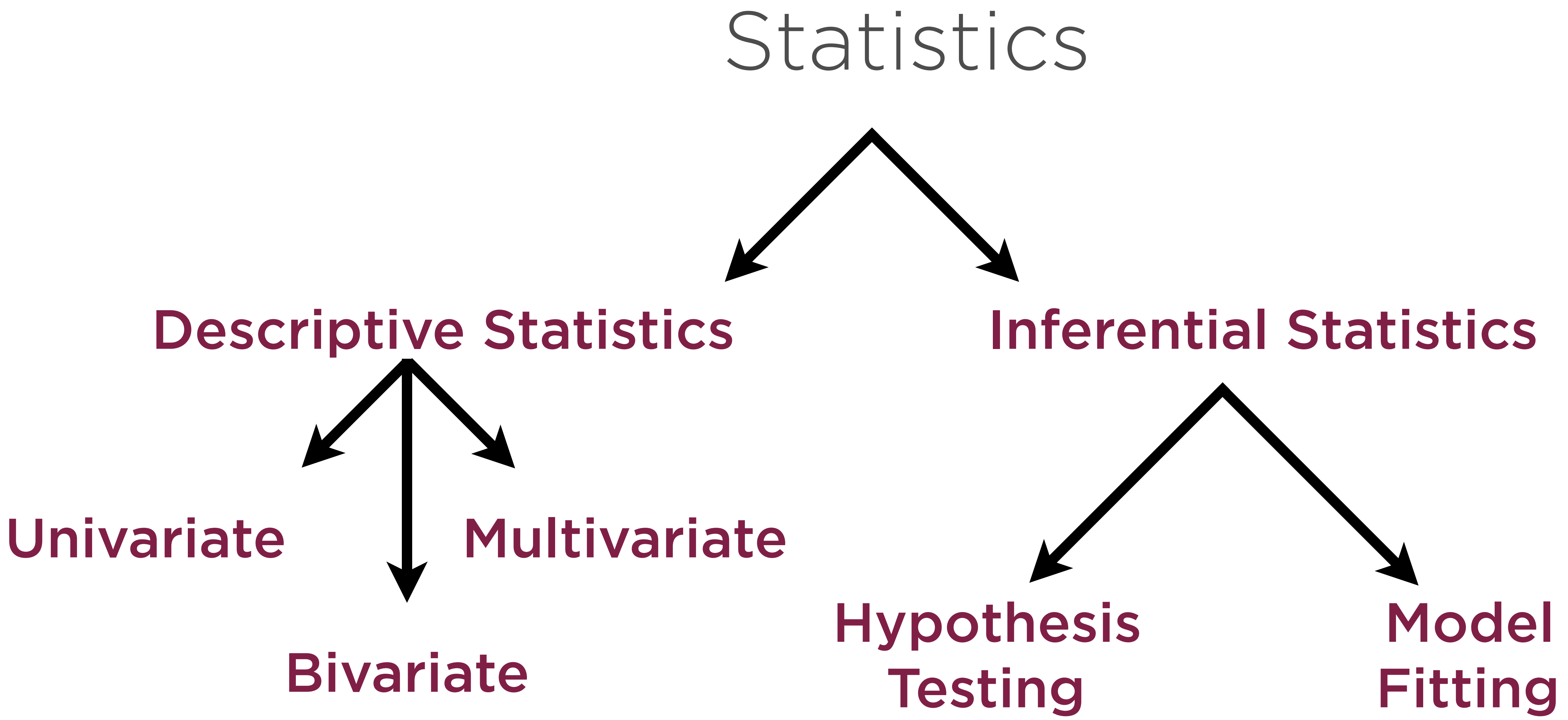
Descriptive Statistics

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

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 Testing

Hypothesis

Proposed explanation for a phenomenon.

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)	Alternate Hypothesis (H_1)
The lady **cannot** tell if milk was poured first	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%

$^8C_4$ = 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

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

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

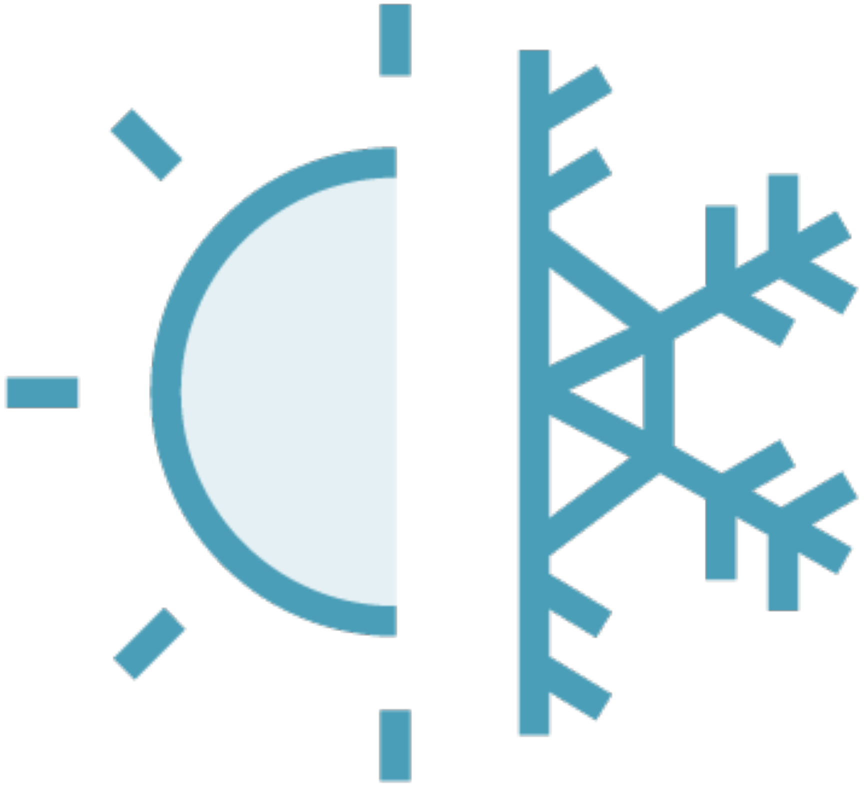
How likely it was just luck?

Accept or Reject

Small p-value? Reject H_0

Small: Below significance level

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

(Samples, populations need not be normal)

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”

$$H_0: \mu = \mu_0$$

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”

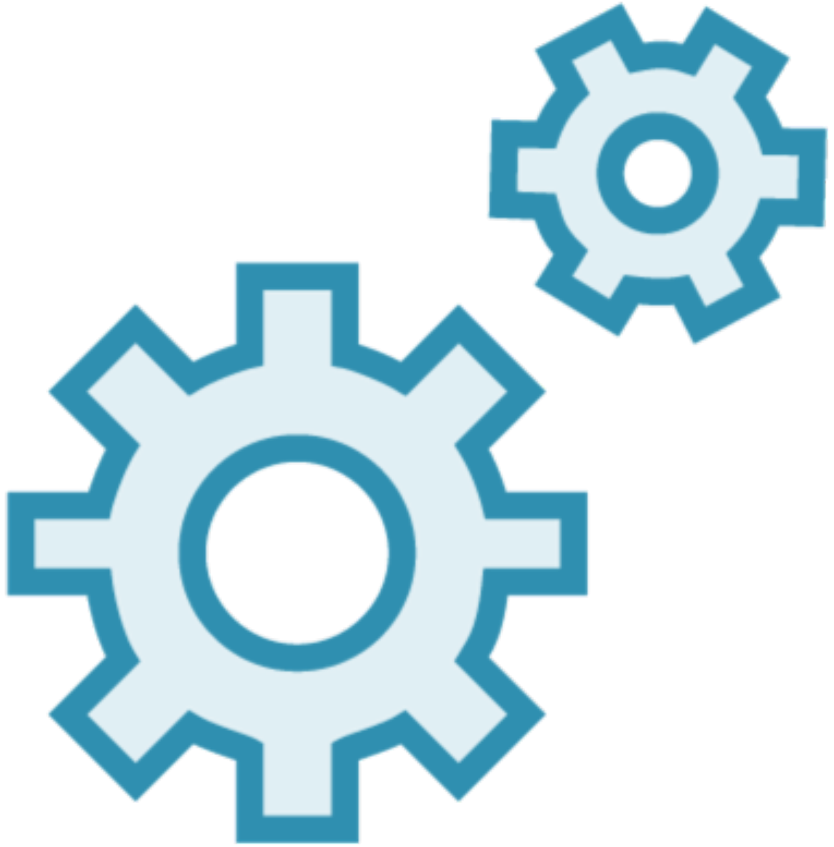
Two-sample Location Test

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

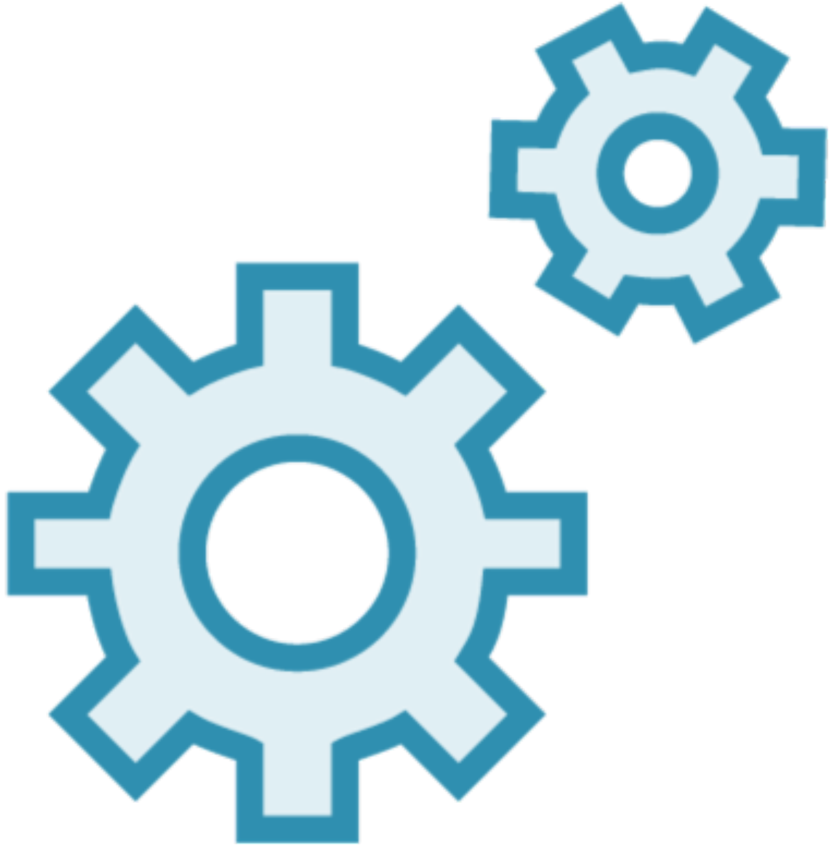


Different forms of t-test based on whether variances are equal or not

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

Regression Coefficient Test

Regression coefficient test

Is the coefficient of any of the independent variables > 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

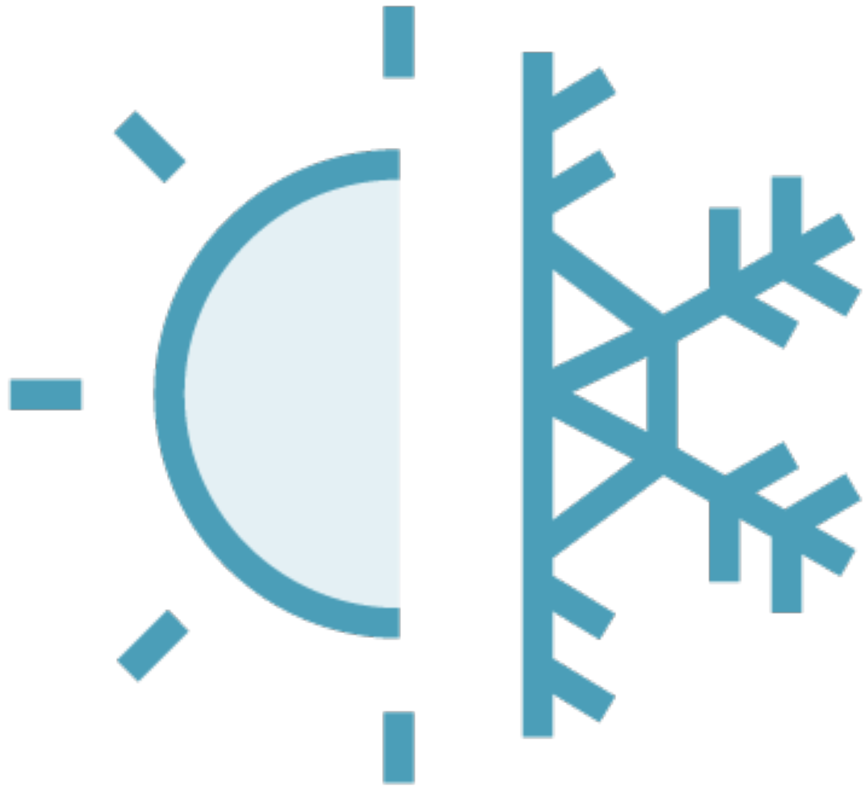
Need not take into account the degrees of freedom

Related Test: Z-test

However, population variance is rarely known in practice

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

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

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

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

Continuous y	1 categorical, independent X	Independent observations
No outliers	Normally distributed y for each x-value	Equal variances of y for each combination

One-way ANOVA Assumptions

Coping with violations of assumptions

- If y is ordinal: Use Kruskal-Wallis ANOVA
- If variances are unequal, use
    - Welch's t-test (2 groups) or
    - Welch's ANOVA (>2 groups)

Kruskal-Wallis ANOVA

One-way ANOVA on samples of ordinal data examines whether those samples originate from the same distribution.

Kruskal-Wallis ANOVA

Non-parametric test

Does not assume normal distribution of residuals

Works with ordinal y-variables

Can be used with ranks

Kruskal-Wallis ANOVA

Does assume that different groups have same variance

Do not use if data is heteroscedastic

- Use Welch's ANOVA instead

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

Kruskal-Wallis ANOVA

Null hypothesis is that mean ranks of all groups are equal

- For n observations, mean rank is (n+1)/2

Works out equivalent to medians of all groups being equal only if

- Each group has same distribution

Null hypothesis is not that means are the same

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

Males

Females

Employees <= 40

Males

Females

Two-way ANOVA Hypotheses

Null Hypothesis (H_{01})	Null Hypothesis (H_{02})	Null Hypothesis (H_{03})
H_{01}: All genders have equal levels of stress	H_{02}: All ages have equal levels of stress	H_{03}: There is no interaction between age and gender

Assumptions of Two-way ANOVA

Continuous y	2 categorical, independent X variables	Independent observations
No outliers	Normally distributed y for each combination	Equal variances of y for each combination

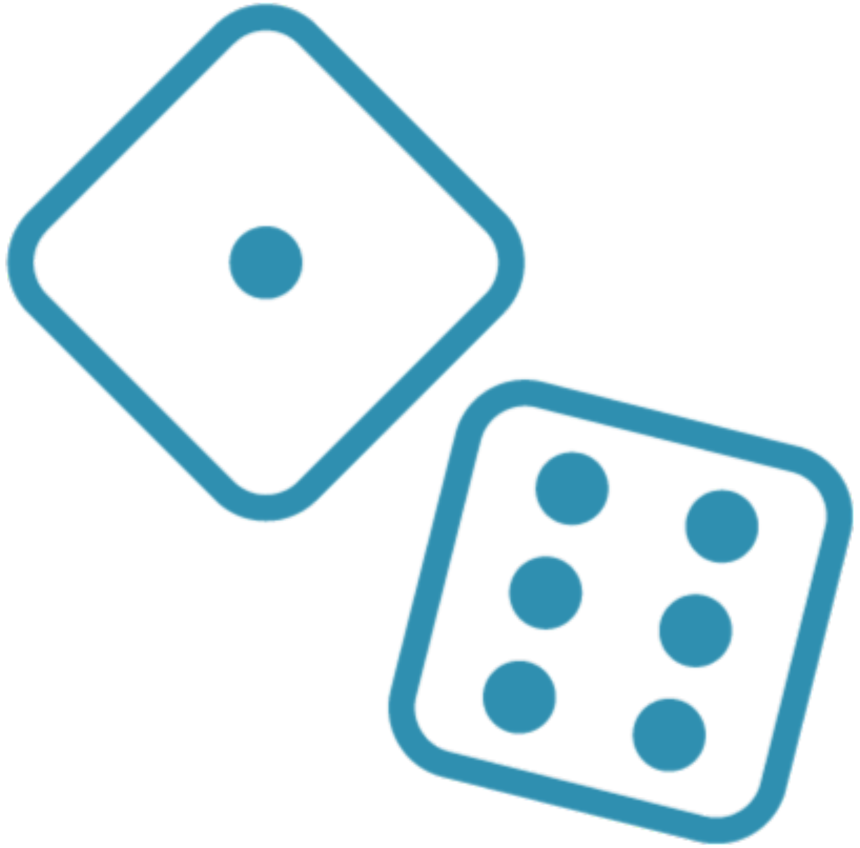
- **Assumption #1:** Your **dependent variable** should be measured at the **continuous** level (i.e., they are **interval** or **ratio** variables). Examples of **continuous variables** include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth. You can learn more about interval and ratio variables in our article: Types of Variable.
- **Assumption #2:** Your **two independent variables** should each consist of **two or more categorical**, **independent groups**. Example independent variables that meet this criterion include gender (2 groups: male or female), ethnicity (3 groups: Caucasian, African American and Hispanic), profession (5 groups: surgeon, doctor, nurse, dentist, therapist), and so forth.
- **Assumption #3:** You should have **independence of observations**, which means that there is no relationship between the observations in each group or between the groups themselves. For example, there must be different participants in each group with no participant being in more than one group. This is more of a study design issue than something you would test for, but it is an important assumption of the two-way ANOVA. If your study fails this assumption, you will need to use another statistical test instead of the two-way ANOVA (e.g., a repeated measures design). If you are unsure whether your study meets this assumption, you can use our Statistical Test Selector, which is part of our enhanced guides.
- **Assumption #4:** There should be **no significant outliers**. Outliers are data points within your data that do not follow the usual pattern (e.g., in a study of 100 students' IQ scores, where the mean score was 108 with only a small variation between students, one student had a score of 156, which is very unusual, and may even put her in the top 1% of IQ scores globally). The problem with outliers is that they can have a negative effect on the two-way ANOVA, reducing the accuracy of your results. Fortunately, when using SPSS Statistics to run a two-way ANOVA on your data, you can easily detect possible outliers. In our enhanced two-way ANOVA guide, we: (a) show you how to detect outliers using SPSS Statistics; and (b) discuss some of the options you have in order to deal with outliers.
- **Assumption #5:** Your **dependent variable** should be **approximately normally distributed for each combination of the groups of the two independent variables**. Whilst this sounds a little tricky, it is easily tested for using SPSS Statistics. Also, when we talk about the two-way ANOVA only requiring approximately normal data, this is because it is quite "robust" to violations of normality, meaning the assumption can be a little violated and still provide valid results. You can test for normality using the Shapiro-Wilk test for normality, which is easily tested for using SPSS Statistics. In addition to showing you how to do this in our enhanced two-way ANOVA guide, we also explain what you can do if your data fails this assumption (i.e., if it fails it more than a little bit).
- **Assumption #6:** There needs to be **homogeneity of variances for each combination of the groups of the two independent variables**. Again, whilst this sounds a little tricky, you can easily test this assumption in SPSS Statistics using Levene's test for homogeneity of variances. In our enhanced two-way ANOVA guide, we (a) show you how to perform Levene's test for homogeneity of variances in SPSS Statistics, (b) explain some of the things you will need to consider when interpreting your data, and (c) present possible ways to continue with your analysis if your data fails to meet this assumption.

Pearson's χ^2 Test

Pearson's X^2 Test

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

Pearson's X^2 Test

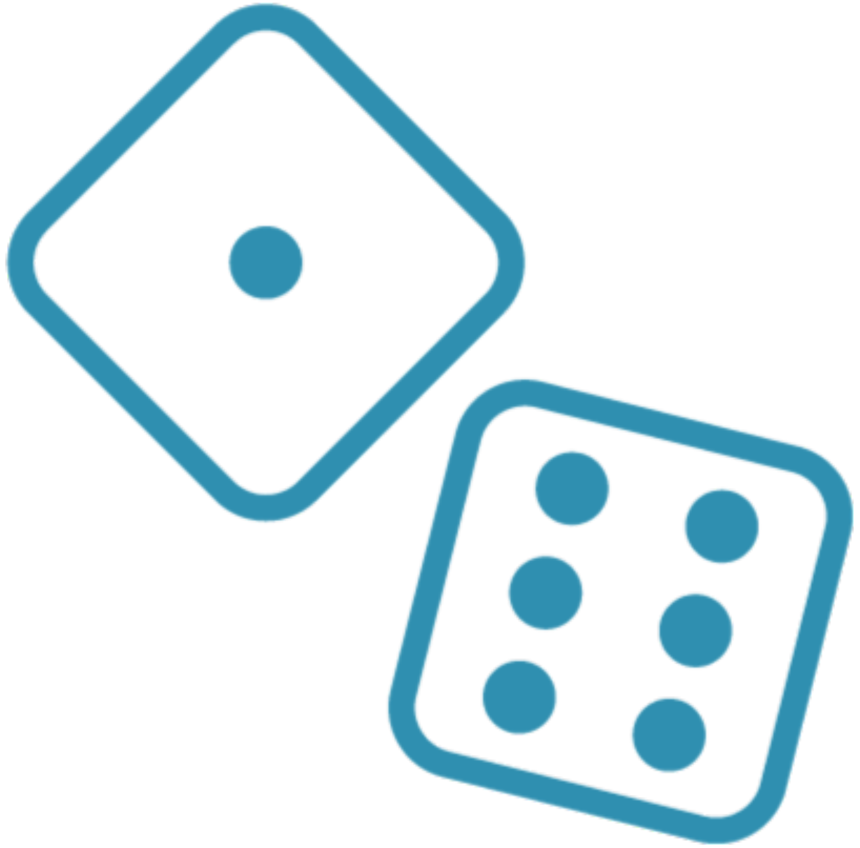


Best understood with an example

Given results of throws of a dice

Are the results consistent with the dice being fair?

Pearson's X^2 Test

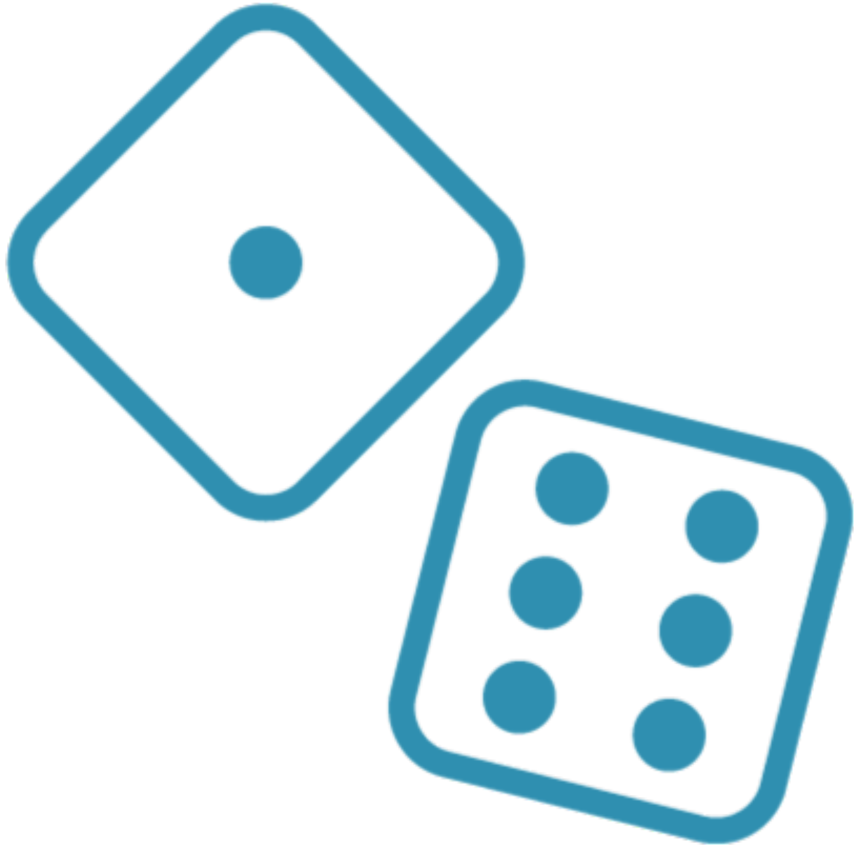


Result of each throw is a categorical variable with values between 1 and 6

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



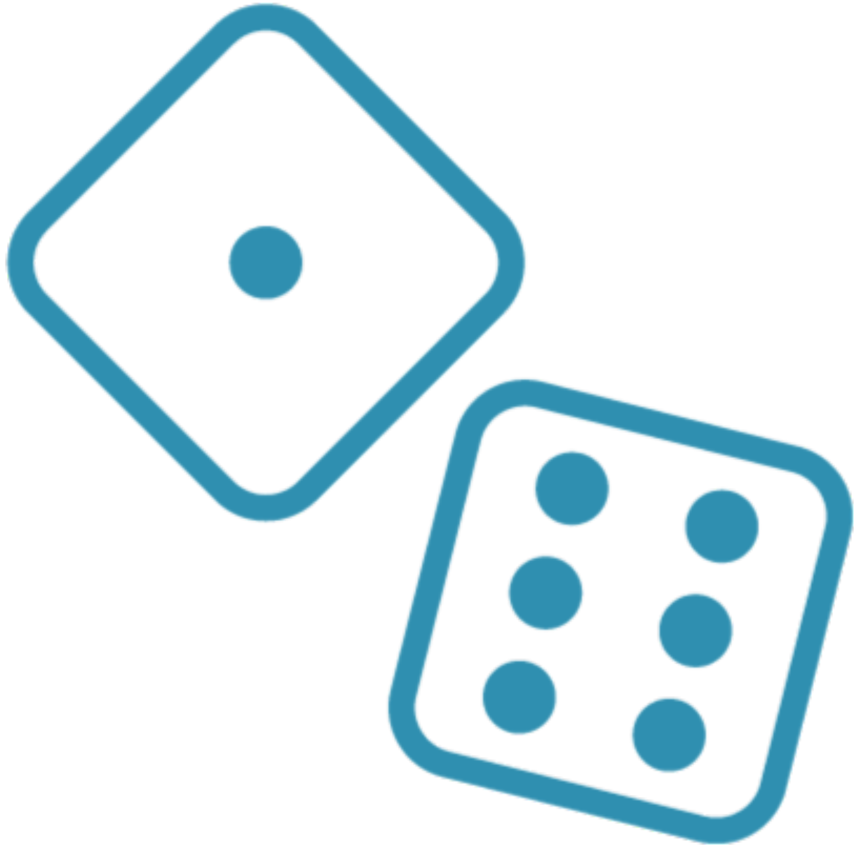
Test statistic follows X^2 distribution

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 vs. inferential Statistics

Hypothesis testing

Interpreting p-values, power and alpha of a test

Understanding t-tests and z-tests

Type-I and Type-II errors in hypothesis testing

Understanding the chi2 test