

# Lecture #2: Statistical Analysis

Hypothesis Testing  
Effect Size Analysis

---



Michael Fu

<https://michaelfu1998-create.github.io/>



Dr. Kla Tantithamthavorn

Senior Lecturer in Software Engineering

<http://chakkrit.com> @klainfo



# Schedule <http://chakkrit.com/teaching/quantitative-research-methods>

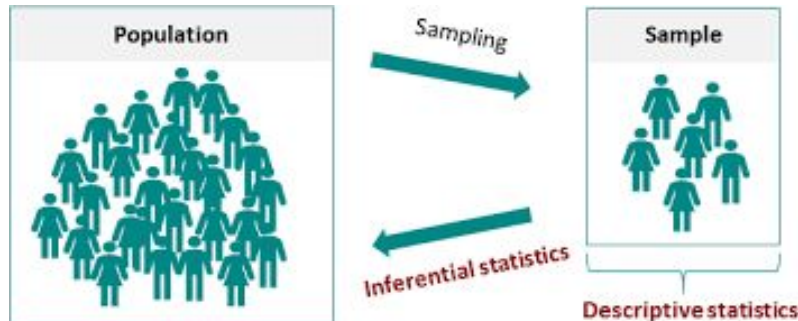
Date	Location	Topics
<del>11 November 2024</del>	<del>G.13 Woodside Building (20 Exhibition Walk)</del>	<del>Design Science Paradigm</del>
<b>15 November 2024</b>	G.13 Woodside Building (20 Exhibition Walk)	Statistical Analysis
25 November 2024	G.13 Woodside Building (20 Exhibition Walk)	Modern Regression Analysis
29 November 2024	G.13 Woodside Building (20 Exhibition Walk)	ML Quality Assurance

To participate, go to  
flux.qa/F3HR6K



# Statistical Analysis

- **Definition:** The process of collecting and analyzing data to identify patterns, identify trends, and inform decision-making.
- **Descriptive statistics (describes data)** explains and visualises the data you have for easier comprehension (e.g., using mode, median, mean, or charts)
- **Inferential statistics** extrapolate the data of representative samples onto a larger population (e.g., using confidence interval to support the claim) ...
  - **infer conclusions from samples statistically drawn from a population**



# Examples of questions that require statistical methods

1. What is the relationship between education level and income in a specific population?
2. How does age impact the likelihood of developing a specific medical condition?
3. What is the effect of a new drug on reducing symptoms of a particular disease?
4. What is the correlation between exercise and heart health in a particular group of people?
5. Is there a significant difference in job satisfaction levels between two groups of employees with different salaries?
6. What factors influence customer satisfaction in a particular industry?
7. How does the use of a particular teaching method impact student performance in a specific subject?
8. What is the relationship between the amount of sleep and academic performance among college students?
9. How effective is a particular marketing strategy in increasing sales for a specific product?
10. What is the correlation between environmental factors and the prevalence of a particular disease in a certain geographic region?

# Examples of questions that require statistical methods (in SE)

1. What is the impact of code reviews on the quality of software products?
2. What factors contribute to the success of agile software development methodologies?
3. What is the relationship between test coverage and defect density in software development projects?
4. How does team size affect software development productivity and quality?
5. What is the effect of using different software development frameworks on project outcomes?
6. What are the most common causes of software project failure and how can they be prevented?
7. How does the use of automated testing tools impact software development efficiency and quality?
8. What is the relationship between software complexity and the occurrence of defects in software products?
9. How do different software development methodologies impact the accuracy of project estimates?
10. What factors contribute to the adoption of new software development technologies and tools in the industry?

# Examples of questions that require statistical methods (in ML)

1. How does the use of a specific algorithm affect the accuracy of a natural language processing system?
2. Is there a significant difference in the performance of different deep learning models for image classification tasks?
3. What is the impact of varying amounts of training data on the performance of a machine learning model for predicting customer churn?
4. How does the choice of feature extraction method affect the performance of a sentiment analysis system?
5. What is the relationship between the complexity of a neural network and its accuracy on a given task?
6. Is there a statistically significant difference in the performance of different clustering algorithms for unsupervised learning?
7. How does the use of different regularization techniques affect the performance of a deep learning model for image segmentation?
8. What is the impact of hyperparameter tuning on the performance of a machine learning model for fraud detection?
9. Is there a significant correlation between the amount of data used for training and the generalization ability of a reinforcement learning algorithm?
10. What is the relationship between the size of a neural network and its training time for a specific task?

# Choosing the Right Statistical Test + Effect Size | Cheat Sheet

	Interval/Ratio (Normality assumed) Called "Parametric tests"	Interval/Ratio (Normality not assumed), Ordinal Called "non-parametric tests"	Binomial
<b>Compare 2 paired groups</b>	Paired t test	Wilcoxon test	McNemar's test
<b>Compare 2 unpaired groups</b>	Unpaired t test	Mann-Whitney test	Fisher's test
<b>Compare &gt;2 matched groups</b>	Repeated-measures ANOVA	Friedman test	Cochran's Q test
<b>Compare &gt;2 unmatched groups</b>	ANOVA	Kruskal-Wallis test	Chi-square test
<b>Find relationship between 2 variables</b>	Pearson correlation	Spearman correlation	Cramer's V

# Choosing the Right Statistical Test + Effect Size | Cheat Sheet

	<b>Interval/Ratio (Normality assumed) Called "Parametric tests"</b>	<b>Interval/Ratio (Normality not assumed), Ordinal Called "non-parametric tests"</b>	<b>Binomial</b>
<b>Compare 2 paired groups</b>	Paired t test	Wilcoxon test	McNemar's test
<b>Compare 2 unpaired groups</b>	Unpaired t test	Mann-Whitney test	Fisher's test
<b>Compare &gt;2 matched groups</b>	Repeated-measures ANOVA	Friedman test	Cochran's Q test
<b>Compare &gt;2 unmatched groups</b>	ANOVA	Kruskal-Wallis test	Chi-square test
<b>Find relationship between 2 variables</b>	Pearson correlation	Spearman correlation	Cramer's V



# Hypothesis Testing | Using (inferential) statistical analysis

- The goal of research is usually to investigate a relationship between two or more variables within a population. We start with a **hypothesis** about a population, and use a statistical test to test that **hypothesis** (i.e., **hypothesis testing**).
- First, we formulate two types of hypotheses.  
**Null Hypothesis (H0)** proposes that no statistical significance exists in a set of observations.  
**Alternative Hypothesis (H1)** proposes that there is a statistical significance exist in a set of observations.
- **Hypothesis testing** provides a method to reject null hypothesis with a certain confidence level
  - If you can reject the null hypothesis, it provides support for the alternative hypothesis.

## An Example: Do smart children tend to come from rich family?

- **Null hypothesis (H0):** Parental income and GPA of children have no relationship with each other in college students.
- **Alternative hypothesis (H1):** Parental income and GPA are positively correlated in college student.

# Hypothesis Testing | More Examples

- Example 1: Does eating breakfast everyday improve academic performance in college students?
  - Hypothesis: Eating breakfast every day will improve academic performance in college students.
  - Null Hypothesis: Eating breakfast every day will not improve academic performance in college students.
- Example 2: Does drinking coffee before a workout increase endurance?
  - Hypothesis: Drinking coffee before a workout will increase endurance.
  - Null Hypothesis: Drinking coffee before a workout will not increase endurance.
- Example 3: Are people who get 8 hours of sleep each night more productive than at work?
  - Hypothesis: People who get 8 hours of sleep each night will be more productive at work.
  - Null Hypothesis: People who get 8 hours of sleep each night will not be more productive at work.

# Hypothesis Testing | An example of house price dataset



A larger house (**GrLivArea**)  
should be more expensive (**SalePrice**)

**H0 (Null hypothesis):** **Size of living area** have no relationship with the **sale prices**

**H1 (Alternative hypothesis):** **Size of living area** have a positive relationship with the **sale prices**

SalePrice	GrLivArea
208500	1710
181500	1262
223500	1786
140000	1717
250000	2198
143000	1362
307000	1694
200000	2090

Let's define the size of house based on the living area  
(lower half = smaller\_house, upper half = larger\_house)

```
df = df[order(df$GrLivArea),]
smaller_houses = df %>% slice(0:as.integer(nrow(df)/2))
larger_houses = df %>% slice(as.integer(nrow(df)/2):nrow(df))
```

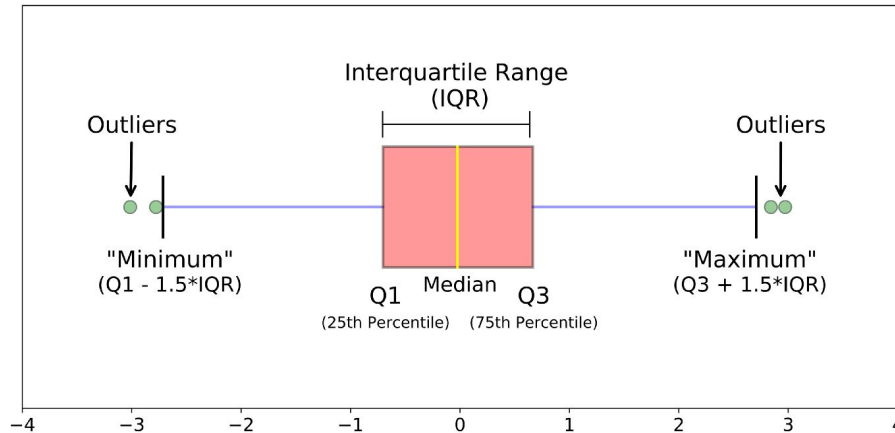
**Which statistical test should be used for this scenario?**

# Outlier Detection | Having outliers in the dataset may interfere subsequent analyses

An outlier is a data point that lies an abnormal distance from others in a dataset, i.e., values lower than  $Q1 - 1.5IQR$  or higher than  $Q3 + 1.5IQR$ .

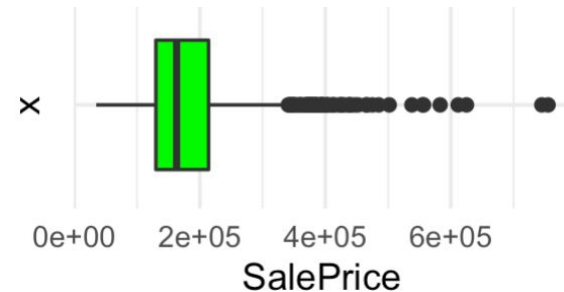
Lets visualize outliers using the house prices dataset:

<https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data>



It is a common practice to exclude outliers from a study. You can calculate Q1, Q3, and IQR to find an outlier and remove them or use this method: `df = df[!df$SalePrice %in% boxplot.stats(df$SalePrice)$out,]`

## Visualizing outliers using box plot



```
library(ggplot2)
ggplot(df) +
  aes(x = "", y = SalePrice) +
  geom_boxplot(fill = "green") +
  theme_minimal() +
  rotate()
```

# Data Preparation

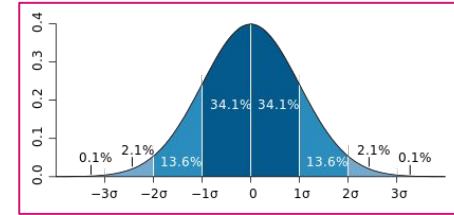
Source: <https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data>

```
df <- read.csv("train.csv", header=TRUE)
# remove outliers
df = df[!df$SalePrice %in% boxplot.stats(df$SalePrice)$out,]
df = df[!df$GrLivArea %in% boxplot.stats(df$GrLivArea)$out,]
df$has_fireplace = df$FireplaceQu > 0 # define whether a house has fireplace

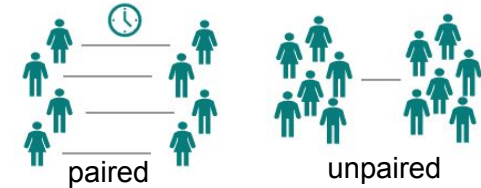
library(dplyr)
# define smaller house as the first half,
# and larger house is second half based on the living area
df = df[order(df$GrLivArea),]
smaller_houses = df %>% slice(0:as.integer(nrow(df)/2))
larger_houses = df %>% slice(as.integer(nrow(df)/2):nrow(df))
```

# Choosing a Right Statistical Test | There are three aspects to consider:

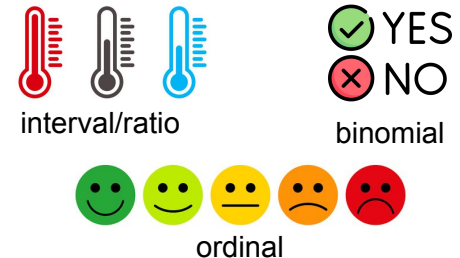
1) **Parametric vs Non-Parametric Tests? Is the data follows a normal distribution (or a Gaussian distribution)?**



2) **Comparing paired or unpaired samples?**



3) **Comparing interval/ratio, ordinal, or binomial data?**

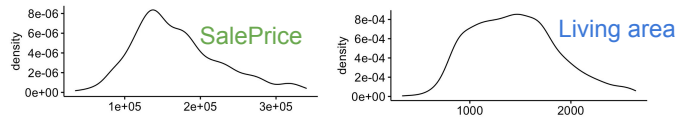


# Q1) Parametric vs Non-Parametric Tests?

## 1.1) Visual judgment based on density plot and quantile-quantile plot

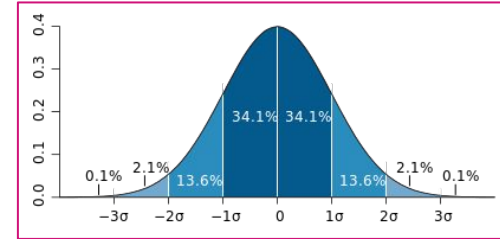
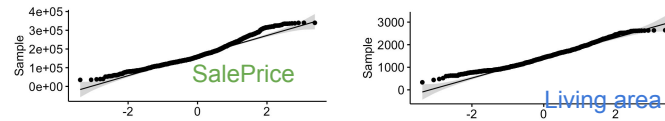
```
library(ggpubr)
ggdensity(df$SalePrice)
ggdensity(df$GrLivArea)
```

If the distribution is normal,  
distribution should be bell-shaped



```
ggqqplot(df$SalePrice)
ggqqplot(df$GrLivArea)
```

If the distribution is normal,  
the dots should form a straight line.



## 1.2) Using Shapiro-Wilk's normality test (H0=normal distribution)

```
shapiro.test(df$SalePrice)
shapiro.test(df$GrLivArea)
```

```
> shapiro.test(df$SalePrice)

Shapiro-Wilk normality test

data:  df$SalePrice
W = 0.86967, p-value < 2.2e-16

> shapiro.test(df$GrLivArea)

Shapiro-Wilk normality test

data:  df$GrLivArea
W = 0.92798, p-value < 2.2e-16
```

Even though the density plots above indicate a bell-shaped distribution, the Shapiro-Wilk tests yield P-value < 0.05.

This **P-value** helps us determine the significance of the test results in relation to the hypothesis. A p-value less than 0.05 indicates strong evidence against the null hypothesis (H0).

Therefore, we **reject H0 (data is normally distributed)** and **accept H1 (data is not normally distributed)**

# Q1) Parametric vs Non-Parametric Tests?

## Parametric test

- Assume that the data are drawn from a population with a normal distribution.
- Assume that the variables are measured based on an interval or ratio scale.

## Non-parametric test

- Not assume that the population has a normal distribution
- Can be used when the sample size is very small

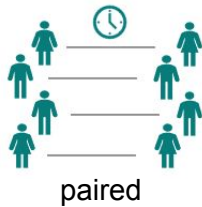
## Suggestion!

- Since this dataset does not have a normal distribution, we must to use **non-parametric tests!**
- It's generally safe to use **non-parametric tests** to **avoid any normal distribution assumptions**



## Q2) Comparing paired or unpaired samples?

**Paired samples (dependent)** are the sample in which natural or matched couplings occur. The data point in one sample is uniquely paired to a data point in the second sample.

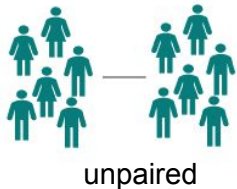


Student	Test1	Test2
ID1	100	100
ID2	80	90
ID3	60	80

Example Research Questions:

- Is there any difference between the pre-test and post-test exam scores?

**Unpaired samples (independent)** are the sample of unrelated groups.

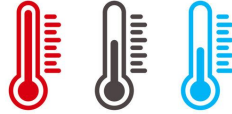


Price of small houses	Price of large houses
500,000	1,200,000
600,000	1,500,000
700,000	

Example Research Questions:

- Is there any difference between the price of small houses and the price of large houses?

## Q3) Comparing interval/ratio, ordinal, or binomial data?



**Interval/ratio:** numbers in interval, difference is meaningful,

e.g., **Interval** -> temperature in Celsius or Fahrenheit

**Ratio** -> temperature in Kelvin (has a clear definition of zero,  
0 Kelvin = lowest temp possible)



**Ordinal:** ordinal scale where the order matters

E.g., likert scale: Extremely dislike, dislike, neutral, like, extremely like



**Binomial:** having only two possible values

E.g., Diagnosed as having COVID or not

# Choosing the Right Statistical Test + Effect Size | Cheat Sheet

	<b>Interval/Ratio (Normality assumed) Called “Parametric tests”</b>	<b>Interval/Ratio (Normality not assumed), Ordinal Called “non-parametric tests”</b>	<b>Binomial</b>
<b>Compare 2 paired groups</b>	Paired t test	Wilcoxon test	McNemar’s test
<b>Compare 2 unpaired groups</b>	Unpaired t test	Mann-Whitney test	Fisher’s test
<b>Compare &gt;2 matched groups</b>	Repeated-measures ANOVA	Friedman test	Cochran’s Q test
<b>Compare &gt;2 unmatched groups</b>	ANOVA	Kruskal-Wallis test	Chi-square test
<b>Find relationship between 2 variables</b>	Pearson correlation	Spearman correlation	Cramer’s V

# Choosing the Right Statistical Test + Effect Size | Cheat Sheet

	Interval/Ratio (Normality assumed) Called “Parametric tests”	Interval/Ratio (Normality not assumed), Ordinal Called “non-parametric tests”	Binomial
<b>Compare 2 paired groups</b>	Paired t test	Wilcoxon test	McNemar’s test
<b>Compare 2 unpaired groups</b>	Unpaired t test	Mann-Whitney test	Fisher’s test
<b>Compare &gt;2 matched groups</b>	Repeated-measures ANOVA	Friedman test	Cochran’s Q test
<b>Compare &gt;2 unmatched groups</b>	ANOVA	Kruskal-Wallis test	Chi-square test
<b>Find relationship between 2 variables</b>	Pearson correlation	Spearman correlation	Cramer’s V

# Paired T-Test

Comparing the means of two paired groups (parametric test)

## Requirements

- dependent variable is interval or ratio
- samples are drawn from a normally distributed population
- the comparing data must have the same size (i.e., paired)

## Interpretation

- $H_0$  (accept if  $p \geq 0.05$ ): There is no significant difference in the means between the two groups
- $H_1$  (accept if  $p < 0.05$ ): There is a significant difference in the means between the two groups

## Case study

A group of students took pre and post lecture exams. Do the students achieved a higher score in post-exam than the pre-exam or not?

## $H_0$ (Null hypothesis)

The pre- and post-lecture exam scores are not statistically different.

## $H_1$ (alternative hypothesis)

The post-lecture exam scores are significantly statistically higher than the pre-lecture exam scores.

Student ID	Pre-lecture exam scores	Post-lecture exam scores
1	4	7
2	3	5
3	8	9
4	2	7
5	3	8

# Paired T-Test

Comparing the means of two paired groups (parametric test)

## R Code

```
preExam = c(4,3,8,2,3)
postExam = c(7,5,9,7,8)
t.test(x=postExam, y=preExam, alternative = "greater",
       var.equal = FALSE, paired = TRUE)  ** "greater" = test whether x is greater than y
```

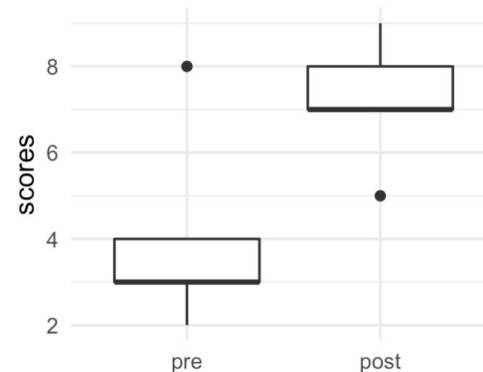
## Results

```
Paired t-test
data:  postExam and preExam
t = 4, df = 4, p-value = 0.008065
alternative hypothesis: true difference
in means is greater than 0
95 percent confidence interval:
 1.494523      Inf
sample estimates:
mean of the differences
      3.2
```

## Effect size

```
library(effectsiz)
effect_size = cohens_d(x=postExam, y=preExam, var.equal = FALSE)
interpret_cohens_d(effect_size)
```

Cohen's d	95% CI	Interpretation
1.63	[0.09, 3.10]	large



## Interpretation

Rejecting H0, accepting H1.

The post-lecture test scores are statistically significantly higher than the pre-lecture exam scores.  
(with a large effect size)

**Conclusion:** Students learn well during the lectures.

## Caveat Is this conclusion statistically sound?

Answer: **NO**, because we did not test whether the population where the data was drawn from has normal distribution or not.

# Wilcoxon (Signed-Rank) Test

Comparing the means of two paired groups (non-parametric test)



Student ID	Pre-lecture exam scores	Post-lecture exam scores
1	4	7
2	3	5
3	8	9
4	2	7
5	3	8

## Requirements

- dependent variable is ordinal, interval/ratio
- the data must have the same size (i.e., paired)

## Interpretation

- $H_0$  (accept if  $p \geq 0.05$ ): There is no significant difference in the means between the two groups
- $H_1$  (accept if  $p < 0.05$ ): There is a significant difference in the means between the two groups

## Case study

A group of students took pre and post lecture exams. Do the students achieved a higher score in post-exam than the pre-exam or not?

## $H_0$ (Null hypothesis)

The pre- and post-lecture exam scores are not statistically different.

## $H_1$ (alternative hypothesis)

The post-lecture exam scores are significantly statistically higher than the pre-lecture exam scores.

# Wilcoxon (Signed-Rank) Test

Comparing the means of two paired groups (non-parametric test)

## R Code

```
preExam = c(4,3,8,2,3)
postExam = c(7,5,9,7,8)
library(coin)
wilcox.test(postExam, preExam, paired=T, alternative = "greater")
```

## Results

```
Wilcoxon signed rank test with continuity correction
data:  postTest and preTest
V = 15, p-value = 0.02895
alternative hypothesis: true location shift is greater than 0
```

## Effect size

```
scores = c(4,3,8,2,3,7,5,9,7,8) # re-organize the data to calculate effect size
type = c("pre", "pre", "pre", "pre", "pre", "post", "post", "post", "post", "post")
df = data.frame(scores=scores, type=type)
library(rcompanion)
cliffDelta(scores~type, data = df)
```

Cliff.delta 0.72

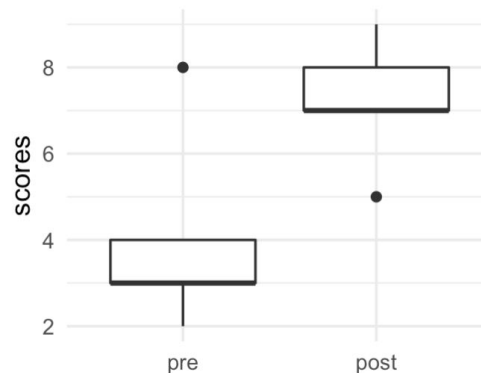
Rule of thumb	Small	Medium	Large
Effect size	0.10	0.30	0.50

## Interpretation

Rejecting H0, accepting H1.

The post-lecture exam scores are significantly statistically higher than the pre-lecture exam scores (with a large effect size).

**Conclusion:** Students learn well during the lectures.





# Mcnemar's test

A test that measures an association between two (paired) categorical variables.

## Requirements

- dependent variable is binomial
- the data must have the same size (i.e., paired)
- the data is in “before & after” table format (see the right tables carefully)

## Interpretation

- $H_0$  (accept if  $p \geq 0.05$ ): The occurrences of the outcomes for the two groups are equal.
- $H_1$  (accept if  $p < 0.05$ ): The occurrences of the outcomes for the two groups are not equal.

## Case study

A group of students took pre- and post-lecture exam. Do the students that passed the pre-exam will also pass the post-exams or not?

## $H_0$ (Null hypothesis)

The number of students that passed the pre- and post-lecture exams are not statistically different.

## $H_1$ (alternative hypothesis)

The number of students that passed the pre- and post-lecture exams are statistically different.

	Pre-lecture exam	Post-lecture exam
Passed	30	90
Not passed	70	10



For this test,  
we have to  
re-format the table

	Pre-lecture Passed	Pre-lecture Not passed
Post-lecture Passed	5	5
Post-lecture Not passed	25	65

# Mcnemar's test

A test that measures an association between two (paired) categorical variables.

## R Code

```
data <- matrix(c(5,5,25,65), ncol=2, byrow=T)
mcnemar.test(data)
```

## Results

McNemar's Chi-squared test with continuity correction

```
data: data
McNemar's chi-squared = 12.033, df = 1, p-value = 0.0005226
```

## Interpretation

Rejecting H0, accepting H1.

The number of students that passed the pre- and post-lecture exams are statistically different.

Implication: The number of students that passed the exam is significantly changed after the lecture (in this case, decreased).

**Conclusion:** Students did not learn well in the lecture.


	Pre-lecture exam	Post-lecture exam
Passed	30	90
Not passed	70	10




For this test,  
we have to  
re-format the table

	Pre-lecture Passed	Pre-lecture Not passed
Post-lecture Passed	5	5
Post-lecture Not passed	25	65

# Choosing the Right Statistical Test + Effect Size | Cheat Sheet

	Interval/Ratio (Normality assumed) Called “Parametric tests”	Interval/Ratio (Normality not assumed), Ordinal Called “non-parametric tests”	Binomial
<b>Compare 2 paired groups</b>	Paired t test	Wilcoxon test 	McNemar’s test
<b>Compare 2 unpaired groups</b>	Unpaired t test	Mann-Whitney test	Fisher’s test
<b>Compare &gt;2 matched groups</b>	Repeated-measures ANOVA	Friedman test	Cochran’s Q test
<b>Compare &gt;2 unmatched groups</b>	ANOVA	Kruskal-Wallis test	Chi-square test
<b>Find relationship between 2 variables</b>	Pearson correlation	Spearman correlation	Cramer’s V

# Choosing the Right Statistical Test + Effect Size | Cheat Sheet

	Interval/Ratio (Normality assumed) Called "Parametric tests"	Interval/Ratio (Normality not assumed), Ordinal Called "non-parametric tests"	Binomial
Compare 2 paired groups	Paired t test	Wilcoxon test 	McNemar's test
Compare 2 unpaired groups	Unpaired t test	Mann-Whitney test	Fisher's test
Compare >2 matched groups	Repeated-measures ANOVA	Friedman test	Cochran's Q test
Compare >2 unmatched groups	ANOVA	Kruskal-Wallis test	Chi-square test
Find relationship between 2 variables	Pearson correlation	Spearman correlation	Cramer's V

# Unpaired T-Test

Comparing the means of two unpaired groups (parametric test)

## Requirements

- dependent variable is interval or ratio
- samples are drawn from a normally distributed population
- the data is unpaired (i.e., all columns in the dataset may not have the same size)

## Interpretation

- $H_0$  (accept if  $p \geq 0.05$ ): There is no significant difference in the means between the two groups
- $H_1$  (accept if  $p < 0.05$ ): There is a significant difference in the means between the two groups

## Case study

A group of students are randomly sampled to take a final exam. The students can choose to take the exam in the morning or the afternoon (i.e., the number of students in the exam can be different). Do the scores achieved in the morning and the afternoon exams are different or not?

## $H_0$ (Null hypothesis)

The scores achieved in the morning exam and the afternoon exams are not statistically different.

## $H_1$ (alternative hypothesis)

The scores achieved in the morning exam and the afternoon exams are statistically different.

Students' exam score	
Morning exam	Afternoon exam
8	7
4	5
2	3
9	-
5	-

# Unpaired T-Test

Comparing the means of two unpaired groups (parametric test)

## R Code

```
morning = c(8,4,2,9,5)
afternoon = c(7,5,3)
t.test(x=morning, y=afternoon, alternative = "two.sided",
       var.equal = FALSE, paired = FALSE)
```

"two.sided" = test whether the  
two samples are different

## Results

```
Welch Two Sample t-test
data: morning and afternoon
t = 0.3468, df = 5.6789, p-value = 0.7412
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.692178  4.892178
sample estimates:
mean of x mean of y
  5.6      5.0
```

## Effect size

```
effectsize::cohens_d(x=morning, y=afternoon, var.equal = FALSE).
```

## Interpretation

Accepting H0

The scores achieved in the morning exam and the afternoon exams are not statistically different.

**Conclusion:** Time of the day is not affecting the ability of the students to take the exam.

Students' exam score	
Morning exam	Afternoon exam
8	7
4	5
2	3
9	-
5	-

# Mann-Whitney Test

or Mann-Whitney's U test,  
or Wilcoxon Rank sum test,  
or non-parametric t test



Comparing the means of two unpaired groups (non-parametric test)

## Requirements

- dependent variable is ordinal, interval/ratio
- if interval or ratio, the population must not be normally distributed
- the data is unpaired (i.e., all columns in the dataset may not have the same size)

## Interpretation

- $H_0$  (accept if  $p \geq 0.05$ ): There is no significant difference in the means between the two groups
- $H_1$  (accept if  $p < 0.05$ ): There is a significant difference in the means between the two groups

## Case study

A group of students are randomly sampled to take a final exam. The students can choose to take the exam in the morning or the afternoon (i.e., the number of students in the exam can be different). Do the scores achieved in the morning exam are higher than the scores achieved in the afternoon exams or not?

## $H_0$ (Null hypothesis)

The scores achieved in the morning exam and the afternoon exams are not statistically different

## $H_1$ (alternative hypothesis)

The scores achieved in the morning exam are higher than the scores achieved in the afternoon exams

Students' exam score	
Morning exam	Afternoon exam
8	7
4	5
2	3
9	-
5	-

# Mann-Whitney Test

Comparing the means of two unpaired groups (non-parametric test)

## R Code

```
morning = c(8,4,2,3,5)
afternoon = c(9,9,9)
examTime = factor(c(rep("morning", length(morning)), rep("afternoon", length(afternoon))))
scores = c(morning, afternoon)
wilcox.test(scores ~ examTime, distribution="exact", alternative="greater")
```

## Results

```
Wilcoxon rank sum test with continuity correction
data:  v by g
W = 15, p-value = 0.01624
alternative hypothesis: true location shift is greater than 0
```

## Interpretation

Rejecting  $H_0$ , accepting  $H_1$ .

The scores achieved in the morning exam are higher than the scores achieved in the afternoon exams

**Conclusion:** Taking the exam in the morning lead to a better test scores.

Students' exam score	
Morning exam	Afternoon exam
8	7
4	5
2	3
9	-
5	-



# Fisher's Test

A test that measures an association between two (unpaired) categorical variables that define a contingency table.

## Requirements

- dependent variable is binomial (in form of a contingency table)
- the data is unpaired (i.e., all columns in the dataset may not have the same size)
- the data is in contingency table format

	Morning exam	Afternoon exam
Passed	10	100
Not passed	30	30

## Interpretation

- $H_0$  (accept if  $p \geq 0.05$ ): The occurrences of the outcomes for the two groups are equal.
- $H_1$  (accept if  $p < 0.05$ ): The occurrences of the outcomes for the two groups are not equal.

## Case study

A group of students are randomly sampled to take a final exam. The students can choose to take the exam in the morning or the afternoon (i.e., the number of students in the exam can be different). Do the students that took the morning exam are less likely to passed the exam?

## $H_0$ (Null hypothesis)

The odds that the students passed the morning exam and the afternoon exam are not statistically different.

## $H_1$ (alternative hypothesis)

The odd that the students passed the morning exam is statistically less than that of the morning exam.

# Fisher's Test

A test that measures an association between two (unpaired) categorical variables that define a contingency table.

## R Code

```
m <- matrix(c(10,100,30,30), ncol=2, byrow=T)
fisher.test(m, alternative = "less")
```

## Results

`EFisher's Exact Test for Count Data`

```
data:  m
p-value = 4.452e-09
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.000000 0.214895
sample estimates:
odds ratio
 0.101681
```

	Morning exam	Afternoon exam
Passed	10	100
Not passed	30	30



## Interpretation

Rejecting H0, accepting H1.



The odd that the students passed the morning exam is statistically less than that of the morning exam. (with the odds ratio of 0.1)

**Conclusion:** Students that took the exam in the afternoon tend to perform better.

# Choosing the Right Statistical Test + Effect Size | Cheat Sheet

	Interval/Ratio (Normality assumed) Called "Parametric tests"	Interval/Ratio (Normality not assumed), Ordinal Called "non-parametric tests"	Binomial
Compare 2 paired groups	Paired t test	Wilcoxon test 	McNemar's test
Compare 2 unpaired groups	Unpaired t test	Mann-Whitney test 	Fisher's test
Compare >2 matched groups	Repeated-measures ANOVA	Friedman test	Cochran's Q test
Compare >2 unmatched groups	ANOVA	Kruskal-Wallis test	Chi-square test
Find relationship between 2 variables	Pearson correlation	Spearman correlation	Cramer's V

# Choosing the Right Statistical Test + Effect Size | Cheat Sheet

	Interval/Ratio (Normality assumed) Called "Parametric tests"	Interval/Ratio (Normality not assumed), Ordinal Called "non-parametric tests"	Binomial
Compare 2 paired groups	Paired t test	Wilcoxon test 	McNemar's test
Compare 2 unpaired groups	Unpaired t test	Mann-Whitney test 	Fisher's test
Compare >2 matched groups	Repeated-measures ANOVA	Friedman test	Cochran's Q test
Compare >2 unmatched groups	ANOVA	Kruskal-Wallis test	Chi-square test
Find relationship between 2 variables	Pearson correlation	Spearman correlation	Cramer's V

# Analysis of Variance Test

Comparing the means of more than two groups

## Requirements

- dependent variable is interval or ratio
- samples are drawn from a normally distributed population
- the data can be both paired or unpaired

## Interpretation

- $H_0$  (accept if  $p \geq 0.05$ ): There is no significant difference in the means among all groups
- $H_1$  (accept if  $p < 0.05$ ): There is a significant difference in the means among all groups

## Case study

Three groups of students took a final exam. Do the exam scores achieved by the students in three groups are different or not?

## $H_0$ (Null hypothesis)

The exam scores achieved by the student in three groups are not statistically different.

## $H_1$ (alternative hypothesis)

The exam scores achieved by the students in three groups are statistically different.

Students' exam score		
GroupA	GroupB	GroupC
5	5	5
6	6	6
7	7	7
8	8	-
9	-	-

This library require different form of data.

# Analysis of Variance Test

Comparing the means of more than two groups

## R Code

```
score = c(5,6,7,8,9,5,6,7,8,5,6,7)
group = c("A","A","A","A","A","B","B","B","C","C","C")
df <- data.frame(score, group)

# test for homogeneity of variances
bartlett.test(score ~ group, df)
# p-value = 0.7974 (>0.5), variances are equal

# anova test
aov <- aov(score ~ group, df)
summary(aov)
```

## Results

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	2	1.917	0.9583	0.507	0.618
Residuals	9	17.000	1.8889		

## Effect size

$\eta^2 = \frac{SS_{effect}}{SS_{total}} = \text{sum sq of effect} / \text{sum sq total} = 1.917 / (1.917 + 17) = 0.1$   
# Indicating that only ~10% of total variance is accounted for by treatment effect

## Interpretation

Accepting H0

The exam scores achieved by the students in three groups are not statistically different.

**Conclusion:** The students from three groups performed similarly in the exam.

Students' exam score		
GroupA	GroupB	GroupC
5	5	5
6	6	6
7	7	7
8	8	-
9	-	-

Students' exam score	
Group	Score
A	5
A	6
A	7
A	8
A	9
B	5
B	6
B	7
B	8
C	5
C	6
C	7

# Friedman Test

Comparing the means of more than two matched groups (non-parametric test)

## Requirements

- dependent variable is interval/ratio, ordinal
- if interval or ratio, the population can be not normally distributed
- the data must have the same size (i.e., matched)

## Interpretation

- $H_0$  (accept if  $p \geq 0.05$ ): There is no significant difference in the means among all groups
- $H_1$  (accept if  $p < 0.05$ ): There is a significant difference in the means among all groups

## Case study

A group of students takes the exams in three subjects. Do the exam scores achieved by the students in three groups are different or not?

## $H_0$ (Null hypothesis)

The exam scores achieved by the student in three groups are not statistically different.

## $H_1$ (alternative hypothesis)

The exam scores achieved by the students in three groups are statistically different.

Student ID	Subject A	Subject B	Subject C
1	4	7	9
2	3	5	8
3	8	9	9
4	2	7	8
5	3	8	9

# Friedman Test

Comparing the means of more than two matched groups (non-parametric test)

## R Code

```
data <- cbind(c(4,3,8,2,3), c(7,5,9,7,8), c(9,8,9,8,9))  
friedman.test(data)
```

## Results

```
Friedman rank sum test  
data: data2  
Friedman chi-squared = 9.5789, df = 2, p-value = 0.008317
```

## Effect size

Unfortunately, there is no direct way to calculate the effect size for Friedman test.  $r = \frac{Z}{\sqrt{N}}$

You need to perform Mann-Whitney test to calculate, where Z is outcome from the Mann-Whitney test and N is the total number of samples. See <https://yatani.jp/teaching/doku.php?id=hcistats:kruskalwallis>

## Interpretation

Rejecting H0, accepting H1.

The exam scores achieved by the students in three groups are statistically different.

**Conclusion:** The students from three groups performed differently in the exam.

Student ID	Subject A	Subject B	Subject C
1	4	7	9
2	3	5	8
3	8	9	9
4	2	7	8
5	3	8	9



# Cochran's Q Test

A test that measures an association between two or more (matched) categorical variables.

## Requirements

- dependent variable is binomial
- the data must have the same size (i.e., matched)

## Interpretation

- $H_0$  (accept if  $p \geq 0.05$ ): The occurrences of the outcomes for all groups are equal.
- $H_1$  (accept if  $p < 0.05$ ): The occurrences of the outcomes for all groups are not equal.

## Case study

A group of students took the exams in three subjects. Do the odds that the students passed the exams are similar for all three subjects.

## $H_0$ (Null hypothesis)

The odds that the students passed the exams in three subjects are not statistically different.

## $H_1$ (alternative hypothesis)

The odds that the students passed the exams in three subjects are statistically different.

Student passed the exam (0=no, 1=yes)			
Student ID	Subject A	Subject B	Subject C
1	1	0	1
2	0	0	1
3	0	1	0
4	0	1	1
5	0	0	1
6	1	1	1
7	0	0	1
8	0	1	1
9	0	1	1
10	0	1	1

# Cochran's Q Test

A test that measures an association between two or more (matched) categorical variables.

## R Code

```
passed <- c(1,0,1,0,0,1,0,1,0,0,1,1,0,0,1,1,1,1,0,0,1,0,1,1,0,1,1,0,1,1)
student <- factor(c(1,1,1,2,2,2,3,3,3,4,4,4,5,5,5,6,6,6,7,7,7,8,8,8,9,9,9,10,10,10))
subject <- factor(rep(1:3, 10))
data <- data.frame(student, subject, passed)

library(coin)
symmetry_test(passed ~ factor(subject) | factor(student), data = data, teststat = "quad")
```

## Results

```
Asymptotic General Symmetry Test
data: Answer by factor(Software) (1, 2, 3)
stratified by factor(Participant)
chi-squared = 8.2222, df = 2, p-value = 0.01639
```

## Effect size

We can use McNemar's test on each pair of the group to find odds ratio.

## Interpretation




Rejecting H0, accepting H1.

The odds that the students passed the exams in three subjects are statistically different.

**Conclusion:** The students performed differently in three exams.

Student passed the exam (0=no, 1=yes)			
Student ID	Subject A	Subject B	Subject C
1	1	0	1
2	0	0	1
3	0	1	0
4	0	1	1
5	0	0	1
6	1	1	1
7	0	0	1
8	0	1	1
9	0	1	1
10	0	1	1

# Choosing the Right Statistical Test + Effect Size | Cheat Sheet

	Interval/Ratio (Normality assumed) Called "Parametric tests"	Interval/Ratio (Normality not assumed), Ordinal Called "non-parametric tests"	Binomial
<b>Compare 2 paired groups</b>	Paired t test	Wilcoxon test 	McNemar's test
<b>Compare 2 unpaired groups</b>	Unpaired t test	Mann-Whitney test 	Fisher's test
<b>Compare &gt;2 matched groups</b>	Repeated-measures ANOVA	Friedman test 	Cochran's Q test
<b>Compare &gt;2 unmatched groups</b>	ANOVA	Kruskal-Wallis test	Chi-square test
<b>Find relationship between 2 variables</b>	Pearson correlation	Spearman correlation	Cramer's V

# Kruskal-Wallis Test

Comparing the means of more than two unmatched groups (non-parametric test)

## Requirements

- dependent variable is interval/ratio, ordinal
- not required the population to be normally distributed
- the data is unpaired (i.e., all columns in the dataset may not have the same size)

## Interpretation

- $H_0$  (accept if  $p \geq 0.05$ ): There is no significant difference in the means among all groups
- $H_1$  (accept if  $p < 0.05$ ): There is a significant difference in the means among all groups

Exam scores of the students		
Morning	Afternoon	Evening
4	7	9
3	5	8
8	9	9
2	7	8
3	8	

## Case study

A group of students takes a final exam. The students can choose to take the exam in the morning, afternoon, or evening (i.e., the number of students in the exam can be different). Do the exam scores achieved in different time of the day are different or not?

## $H_0$ (Null hypothesis)

The exam scores achieved in different time of the day are not statistically different.

## $H_1$ (alternative hypothesis)

The exam scores achieved in different time of the day are statistically different.

# Kruskal-Wallis Test

Comparing the means of more than two unmatched groups (non-parametric test)

## R Code

```
data <- cbind(c(4,3,8,2,3), c(7,5,9,7,8), c(9,8,9,8,9))  
kruskal.test(data)
```

## Results

```
Kruskal-Wallis rank sum test  
data: data  
Kruskal-Wallis chi-squared = 60.049, df = 2, p-value = 9.132e-14
```

## Effect size

$$r = \frac{Z}{\sqrt{N}}$$

Unfortunately, there is no direct way to calculate the effect size for Kruskal-Wallis test.

You need to perform Mann-Whitney test to calculate, where Z is outcome from the Mann-Whitney test and N is the total number of samples. See <https://yatani.jp/teaching/doku.php?id=hcistats:kruskalwallis>

## Interpretation

Rejecting H0, accepting H1.

- The exam scores achieved in different time of the day are statistically different.

**Conclusion:** The time chosen to take the exam can affect the exam scores.

Exam scores of the students		
Morning	Afternoon	Evening
4	7	9
3	5	8
8	9	9
2	7	8
3	8	

# Chi-square Test

A test that measures an association between two or more (unmatched) categorical variables that define a contingency table.

## Requirements

- dependent variable is binomial
- data is in contingency table format

	Morning	Afternoon	Evening
Passed	16	11	3
Not passed	21	8	15

## Interpretation

- $H_0$  (accept if  $p \geq 0.05$ ): The occurrences of the outcomes for all groups are equal.
- $H_1$  (accept if  $p < 0.05$ ): The occurrences of the outcomes for all groups are not equal.

## Case study

A group of students are randomly sampled to take a final exam. The students can choose to take the exam in the morning, afternoon, or evening (i.e., the number of students in the exam can be different). Do the odds that the students passed the exams in different time of the day are similar?

## $H_0$ (Null hypothesis)

The odds that the students passed the exams in different time of the day are not statistically different.

## $H_1$ (alternative hypothesis)

The odds that the students passed the exams in different time of the day are statistically different.

# Chi-square Test

A test that measures an association between two or more (unmatched) categorical variables that define a contingency table.

## R Code

```
data <- matrix(c(16, 11, 3, 21, 8, 15), ncol=3, byrow=T)
chisq.test(data)
```

## Results

```
Pearson's Chi-squared test
data: data
X-squared = 6.742, df = 2, p-value = 0.03435
```

## Effect size

```
library(vcd)
assocstats(data)
```

```
X^2 df P(> X^2)
Likelihood Ratio 7.2218 2 0.027027
Pearson          6.7420 2 0.034355
```

```
Phi-Coefficient      : NA
Contingency Coeff.: 0.289
Cramer's V           : 0.302
```

Rule of thumb	small size	medium size	large size
Cramer's phi or V	0.10	0.30	0.50

	Morning	Afternoon	Evening
Passed	16	11	3
Not passed	21	8	15




## Interpretation

Rejecting H0, accepting H1.

The odds that the students passed the exams in different time of the day are statistically different.

**Conclusion:** Taking the exam in different time of the day could affect the exam outcome.

# Choosing the Right Statistical Test + Effect Size | Cheat Sheet

	Interval/Ratio (Normality assumed) Called "Parametric tests"	Interval/Ratio (Normality not assumed), Ordinal Called "non-parametric tests"	Binomial
Compare 2 paired groups	Paired t test	Wilcoxon test 	McNemar's test
Compare 2 unpaired groups	Unpaired t test	Mann-Whitney test 	Fisher's test
Compare >2 matched groups	Repeated-measures ANOVA	Friedman test 	Cochran's Q test
Compare >2 unmatched groups	ANOVA	Kruskal-Wallis test	Chi-square test
Find relationship between 2 variables	Pearson correlation	Spearman correlation	Cramer's V



# Pearson Correlation

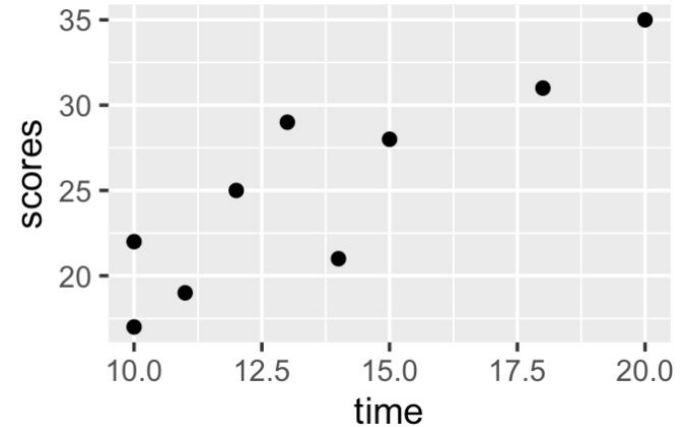
A test for correlation (i.e., the strength of the relationship) between two interval/ratio variables

## Requirements

- dependent variable is interval/ratio
- assume normal distribution

## Interpretation

- $H_0$  (accept if  $r = 0$ ): There is no correlation between the two variables.
- $H_1$  (accept if  $r \neq 0$ ): There is a correlation between the two variables.



## Case study

A group of students took different length of time (in minutes) to prepare for the exam. Do the length of the exam preparation time and the test scores are correlated.

## $H_0$ (Null hypothesis)

There is no correlation between the length of the exam preparation time and the test scores.

## $H_1$ (alternative hypothesis)

There is a correlation between the length of exam preparation time and the test scores.

# Pearson Correlation

A test for correlation (i.e., the strength of the relationship) between two interval/ratio variables

## R Code

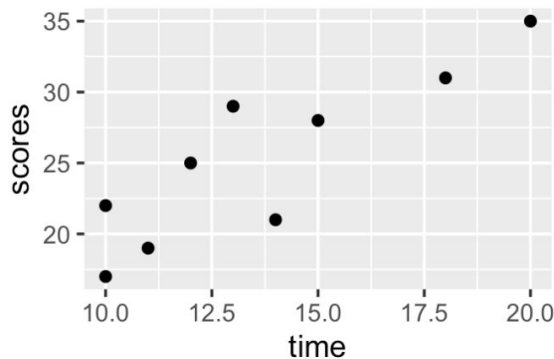
```
time <- c(10,14,12,20,15,13,18,11,10)
scores <- c(22,21,25,35,28,29,31,19,17)
cor.test(time,scores,method="pearson")
```

## Results

Pearson's product-moment correlation

```
data: time and scores
t = 4.6855, df = 7, p-value = 0.002246
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4900229 0.9724978
sample estimates:
      cor
0.8707668
```

Rule of thumb	small size	medium size	large size
Pearson's $r$	0.1	0.3	0.5



## Interpretation

Rejecting  $H_0$ , accepting  $H_1$ .

There is a large positive correlation between the length of exam preparation time and the test scores.

**Conclusion:** Spending more time in preparing can lead to a higher exam scores.

# Spearman Correlation

A test for correlation (i.e., the strength of the relationship) between two interval/ratio variables

## Requirements

- dependent variable is interval/ratio
- not assume normal distribution

## Interpretation

- $H_0$  (accept if  $r = 0$ ): There is no correlation between the two variables.
- $H_1$  (accept if  $r \neq 0$ ): There is a correlation between the two variables.

## Case study

A group of students took different length of time (in minutes) to prepare for the exam. Do the length of the exam preparation time and the test scores are correlated.

## $H_0$ (Null hypothesis)

There is no correlation between the length of the exam preparation time and the test scores.

## $H_1$ (alternative hypothesis)

There is a correlation between the length of exam preparation time and the test scores.

# Spearman Correlation

A test for correlation (i.e., the strength of the relationship) between two interval/ratio variables

## R Code

```
time <- c(10,14,12,20,15,13,18,11,10)
scores <- c(22,21,25,35,28,29,31,19,17)
cor.test(time,scores,method="spearman")
```

## Results

Spearman's rank correlation rho

data: time and scores

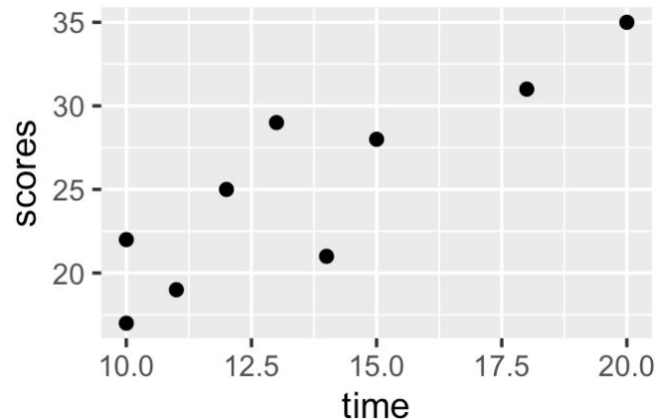
S = 22.593, p-value = 0.007889

alternative hypothesis: true rho is not equal to 0

sample estimates:

rho

0.8117226



Rule of thumb	small size	medium size	large size
Spearman's rho	0.1	0.3	0.5

## Interpretation

Rejecting H0, accepting H1.

There is a large positive correlation between the length of exam preparation time and the test scores.

**Conclusion:** Spending more time in preparing can lead to a higher exam scores.

# Cramer's V

A test that measures a coefficients of an association (i.e., correlation for categorical data) between two binomial variables in a crosstab table. In other words, it represents how the distribution of the data are changing depending on one variable.

## Requirements

- dependent variable is binomial
- the data is in a crosstab table format

## Interpretation

- $H_0$  (accept if  $r = 0$ ): There is no association between the two variables.
- $H_1$  (accept if  $r \neq 0$ ): There is an association between the two variables.

## Case study

Students are separated in group A and group B in an exam. In this exam, the students can choose to use either pen or pencil as their writing tool. Do the students in two groups choose the writing tools differently?

## $H_0$ (Null hypothesis)

There is no association between the student groups and the writing tools used.

## $H_1$ (alternative hypothesis)

There is an association between the student groups and the writing tools used.

	Writing tool	
	Pen	Pencil
Students group A	20	10
Students group B	3	27

# Cramer's V

A test that measures a coefficients of an association (i.e., correlation for categorical data) between two binomial variables in a crosstab table. In other words, it represents how the distribution of the data are changing depending on one variable.

## R Code

```
data <- matrix(c(20, 10, 3, 27), ncol=2, byrow=T)
library(vcd)
assocstats(data)
```

## Results

```
      X^2 df      P(> X^2)
Likelihood Ratio 22.185  1 2.4762e-06
Pearson          20.376  1 6.3622e-06

Phi-Coefficient   : 0.583
Contingency Coeff.: 0.503
Cramer's V       : 0.583
```

	Writing tool	
	Pen	Pencil
Students group A	20	10
Students group B	3	27

Rule of thumb	small size	medium size	large size
Cramer's phi or V	0.10	0.30	0.50

## Interpretation

Rejecting H0, accepting H1.

There is a large association between the student groups and the writing tools used.

**Conclusion:** The students in group A and group B chose the writing tools differently.

## Exercise: Which statistical methods would you use?

1. What is the relationship between education level and income in a specific population?
2. How does age impact the likelihood of developing a specific medical condition?
3. What is the effect of a new drug on reducing symptoms of a particular disease?
4. What is the correlation between exercise and heart health in a particular group of people?
5. Is there a significant difference in job satisfaction levels between two groups of employees with different salaries?
6. What factors influence customer satisfaction in a particular industry?
7. How does the use of a particular teaching method impact student performance in a specific subject?
8. What is the relationship between the amount of sleep and academic performance among college students?
9. How effective is a particular marketing strategy in increasing sales for a specific product?
10. What is the correlation between environmental factors and the prevalence of a particular disease in a certain geographic region?