Probability and Statistics Review

Introduction to Bioinformatics Course

Niloufar Razani
Department of Computer Engineering
Sharif University of Technology

Fall 2023

Random Variables

- A **random variable** is a variable whose value, when measured during or following an experiment, may be one or two or more possible numeric values.

There are two key parts to this definition:

1. the value measured for a random variable must be numeric. Often, the measured value is restricted to some domain.

2. The value of the random variable that results from an experiment is not known in advance.

Random Variables

Discrete

X = Number of matching bases when comparing two strands of DNA each of length N.

X = Number of LEU amino acids in a randomly selected human protein product.

Continuous

X = Time until a certain cellular molecule degrades.

X = Molecular weight of a randomly selected RNA molecule. X = (0, M) where M is some constrained upper bound for such a quantity.

X = The expression of a particular gene

Probability Function

Discrete

Probability Mass Function or pmf:

A function that assigns a probability to each possible value of a discrete random variable.

Cumulative Distribution Function:

The probability that the random variable will be less than or equal to x.

$$F_X(x) = P(X \le x)$$

Continuous

Probability Distribution Function or pdf:

A function that describes the probability distribution of a continuous random variable (f(x)).
Area = Probability

Cumulative Distribution Function:

$$F_X(x) = \int_{t=-\infty}^{t=x} f_X(t)dt$$

In real applications, the probability distribution we use for a random variable often falls into one of certain special cases (Binomial, Normal , …)

Statistics

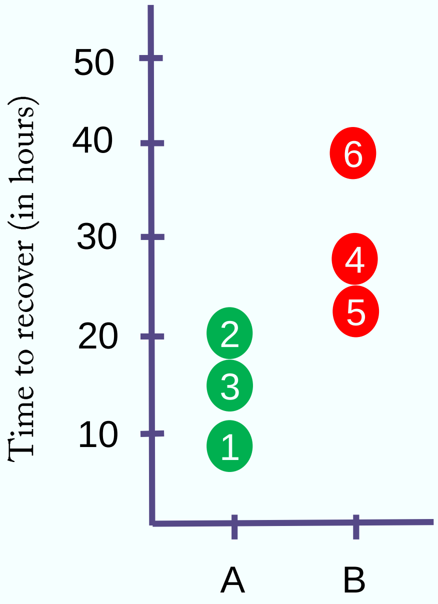
- A statistic (or statistics) is typically a way to summarize the result of a random sample with one (or a few) values.

- $\bar{X} = \frac{\sum_i x_i}{n}$ the average of the random variables from the random sample.
- $S = \sqrt{\frac{\sum_i x_i^2 - \frac{\sum_i x_i{}^2}{n}}{n-1}}$ the standard deviation of the values of the random variables. This is measure of how spread out the values were.
- The median of the random variables (the middle value).
- The smallest and largest of the random variables.

This is just a small list of statistics that may be used in any particular situation.

Hypothesis Testing

- Hypothesis testing, is an inferential procedure in statistics. The situation is that we have a probability model with unknown parameters.
- Imagine that there is a virus and we had two drugs that we could use to treat it.
- We give drug A to 3 people and drug B to 3 other people and measure how long it takes for each person to recover.
- Based on the results, it looks like people taking drug A took less time to recover than people taking drug B!

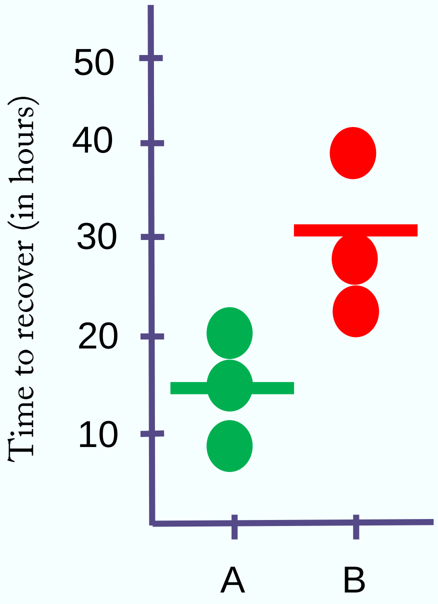


Hypothesis Testing

- When we calculate the average value for Drug A and Drug B we see that on average, there is a 15 hours difference between Drug A and Drug B.

- So after seeing this preliminary data, it might seem reasonable to form the following hypothesis:

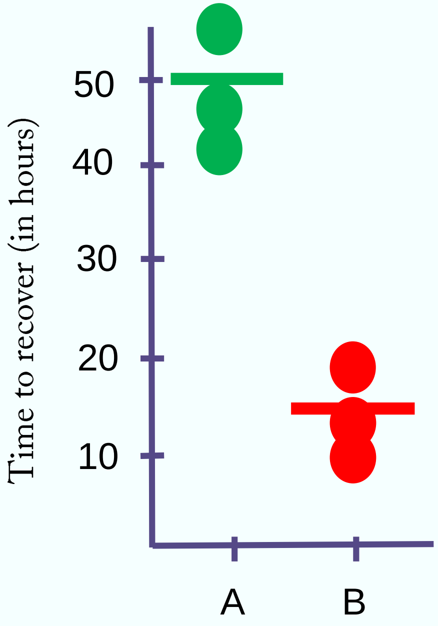
People taking Drug A need on average 15 fewer hours to recover than people taking Drug B.



- Now that we have this hypothesis, we can test it by repeating the experiment.

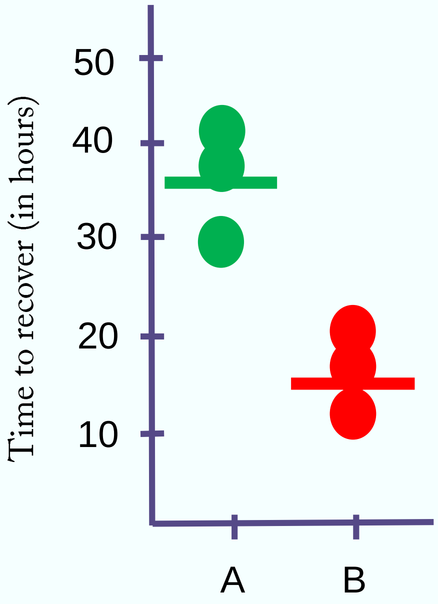
Hypothesis Testing

- This time we see that, on average people taking drug A need 35 more hours than people taking Drug B.
- Compared to our preliminary data, this result is very unexpected! It is the opposite of original hypothesis.
- It is possible that all 3 people that took Drug A in the second experiment have super stressful jobs and unhealthy lifestyle and everyone taking Drug B was well rested and super healthy!
- But it is also possible that we mislabeled Drug A and Drug B and did the experiment wrong. **So we repeat the experiment again.**



Hypothesis Testing

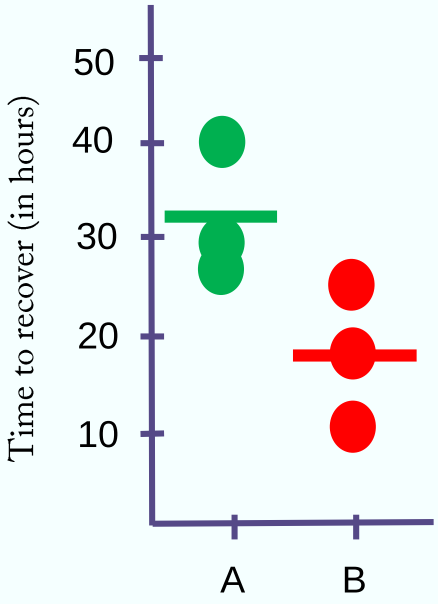
- Again the results are totally backwards from preliminary experiment and the hypothesis that we made.
- So again just to make sure we didn't mislabel things, **we redo the experiment.**



Hypothesis Testing

- Again the results are the opposite of the original hypothesis.
- So we just keep repeating the experiment, each time double checking every little detail, and every time we do the experiment we get the opposite result of the original hypothesis.
- After doing all of these repeats, we can confidently reject preliminary hypothesis.

~~People taking Drug A need on average 15 fewer hours to recover than people taking Drug B.~~

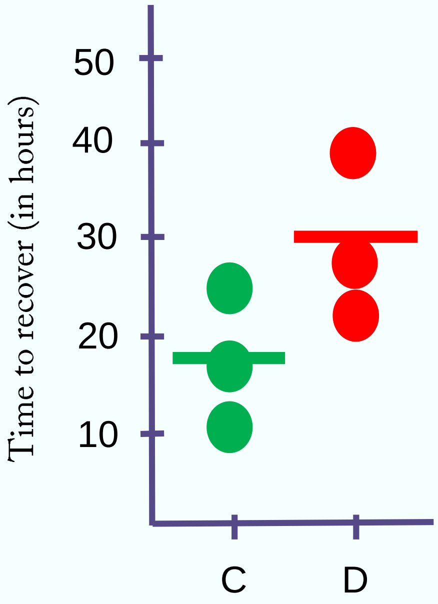


Hypothesis Testing

- Let's imagine we had two more drugs C and D.
- Now based on the data we can create a hypothesis about Drug C and Drug D

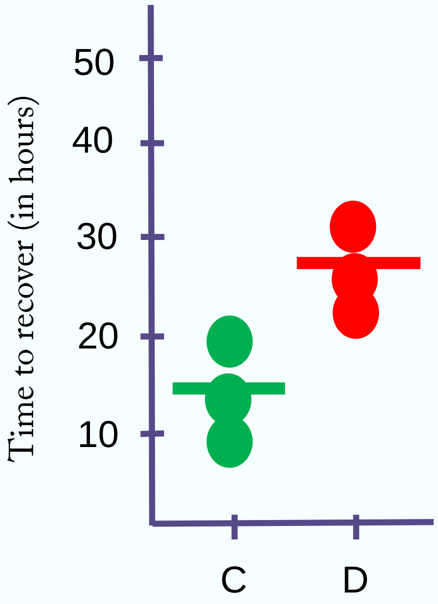
People taking Drug C need on average 13 fewer hours to recover than people taking Drug D.

- Just like before we decide to test this hypothesis by repeating the experiment.



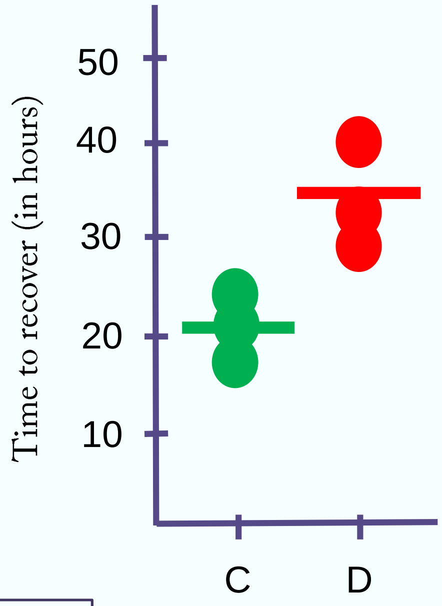
Hypothesis Testing

- This time instead of getting something that's exact opposite of what we expected we get something that is only slightly different.
- In this case the difference is in the same direction, but it is only **12 hours.**
- We repeat the experiment again.



Hypothesis Testing

- Again we get something slightly different from the preliminary experiment and hypothesis.
- The difference is in the same direction, but this time it is **13.5 hours.**
- So good news is that we probably didn't mislabel the drug like we did last time.
- But these experiments don't make us super confident that the hypothesis 13 fewer hours is correct.



So the best we can do is **fail to reject the hypothesis**

Hypothesis Testing

- In summary we can create a hypothesis, and if data give us strong evidence that the hypothesis is wrong then we can reject the hypothesis.
- And When we have data that is similar to the hypothesis, but not exactly the same then the best we can do is fail to reject the hypothesis. Because it is unclear if the hypothesis should be based on which one of the result.

Let's take a closer look at the hypothesis itself

- The only reason the hypothesis is 13 fewer hours is that it was the first result!
- So there are a lot of reasonable hypotheses. How do we know which one to test?

Null Hypothesis

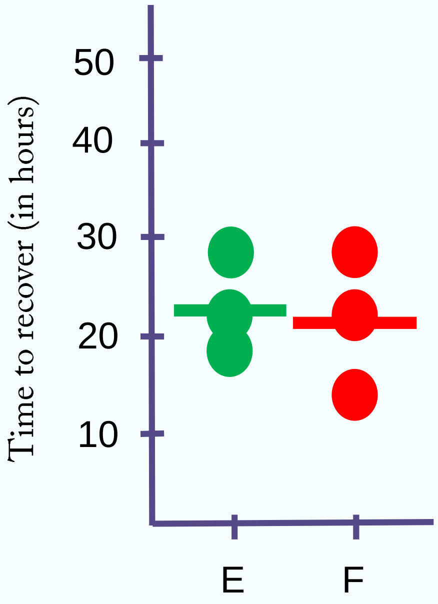
- Since the goal is to see if Drug C is different from Drug D, we simply test to see if there is no difference between the drugs.

H_0 : There is **no difference** in recovery time between Drug C and Drug D.

- The hypothesis that there is no difference between things is called the **null hypothesis.**

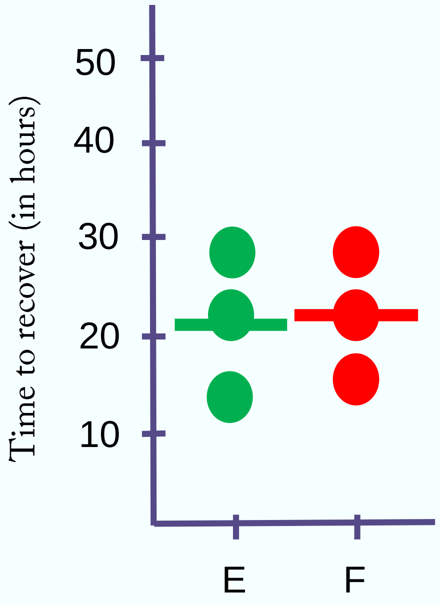
Null Hypothesis

- Now imagine we are testing two new drugs, E and F.
- This time we only get a **0.5 hour difference.**
- But a small random differences give us a slightly different result



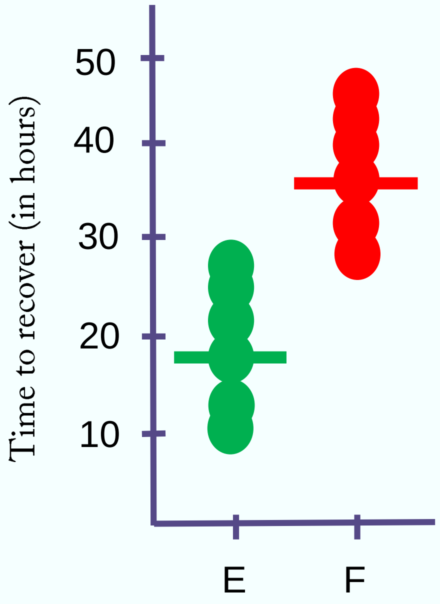
Null Hypothesis

- In this situation instead of Drug F being slightly better by 0.5 hour, Drug E is slightly better by **0.25 hours.**
- Because these small, random difference give us slightly different results we can use the null hypotheses so **we don't have to worry about whether or not the difference is exactly 0.25 or 0.5 hours.**
- Instead **we simply see if the data convince us to reject the hypothesis that there in no difference between Drug E and Drug F.**
- Thus the data does not overwhelmingly convince us to reject the null hypothesis. So **we fail to reject the null hypothesis** that there is no difference between the drugs.



Null Hypothesis

- In contrast, if we tested the drugs on a lot of people and little random things would not change the result very much, then we could confidently reject the null hypothesis that there is no difference between Drug E and Drug F.
- So without the null hypothesis we need preliminary data in order to make a statement that we can test in follow up experiments.
- This is because we don't know if we should test if the difference is 13 hours or 13,000 hours until we get some data.
- In contrast the null hypothesis does not require preliminary data because the only value that represent no difference is 0.



Hypothesis Testing (summary)

- Rather than get stressed out over a large number of possible hypothesis that we could test if Drug C is different from Drug D, we use the null hypothesis to determine if there is a difference.
- If we do an experiment with a bunch of people and a lot more people taking Drug C had shorter recovery time than people taking Drug D, it would be hard to imagine that the results were due to random things, like everyone taking Drug C had better diet or got more exercise than the people taking Drug D.
- Then we could reject the null hypothesis and we know that there is difference between Drug C and Drug D.
- Alternatively, if little random things could easily shift the result from one drug to another and then back again, then we would fail to reject the null hypothesis

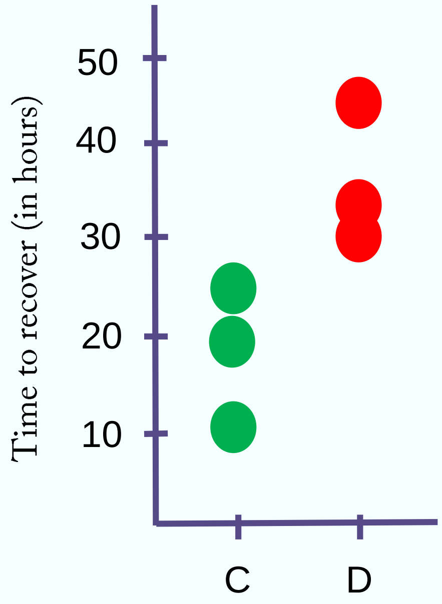
Alternative Hypothesis

- Here are some data that shows how quickly people taking Drug C and D recovered form a virus.

- The goal of collecting all of this data is to determine if we should reject the null hypothesis or fail to reject it.

- In order to decide about the null hypothesis we run the data through a **Statistical Test**.

- The output of a statistical test is a Decision about whether or not to reject or fail to reject the null hypothesis.



Alternative Hypothesis

A statistical test need 3 things:

1. It needs Data

2. It needs a Null or primary hypothesis (i.e. it needs something to reject or fail to reject)

3. It needs an Alternative Hypothesis.

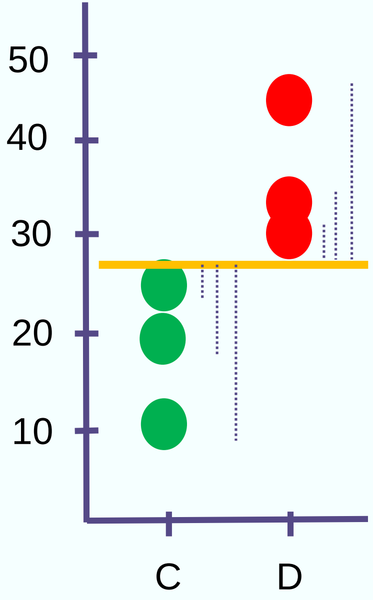
➢ In our case the alternative hypothesis is simply the opposite of the null hypothesis

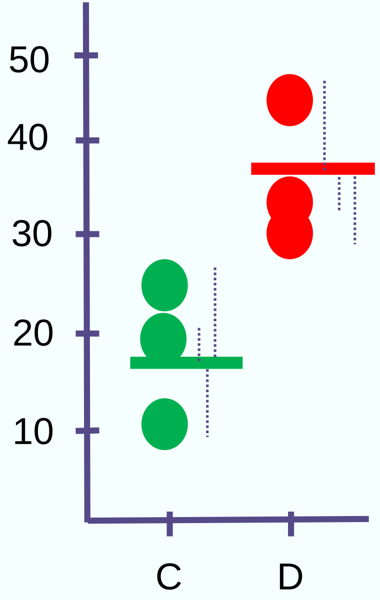
H_a : Alternative Hypothesis
There is a difference in recovery times between Drug C and Drug D

Alternative Hypothesis

- One way to test the null hypothesis is to calculate the mean value of all the data from both drugs and calculate the distances between each observation and the mean.

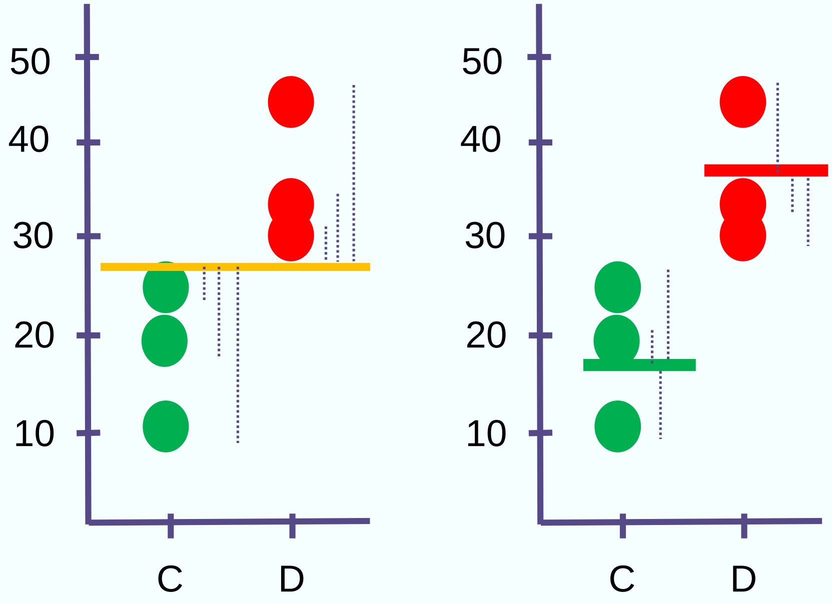
- And Then compare those to distances calculated from individual means for Drug C and Drug D.





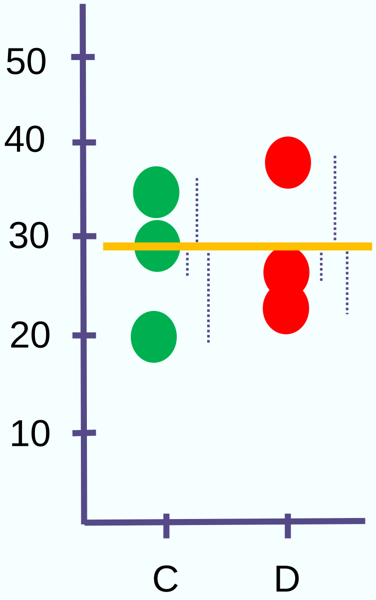
Alternative Hypothesis

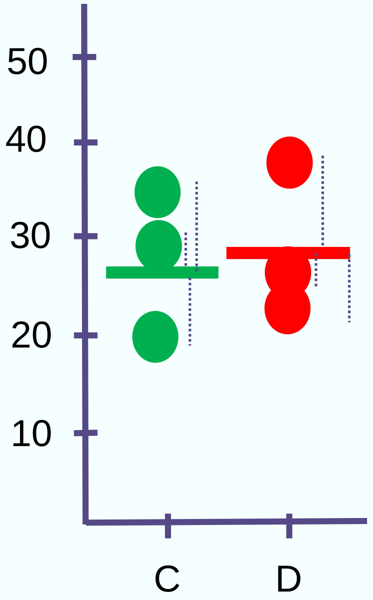
- The distance around the single mean represents the null hypothesis that there is no difference.
- The distances around the two separate means represent the alternative hypothesis.
- If the distances around two means are much shorter than the distances around the single mean then that suggests that using two means to summarize the data makes more sense than using one.
- So we will reject the null hypothesis.



Alternative Hypothesis

- But If the data look like this and the distance from the single mean were not dramatically different from the distances around the separate means.
- Then that would suggest that the difference between two means only reflects little random things that we can't account for.
- So in this case we fail to reject the Null Hypothesis





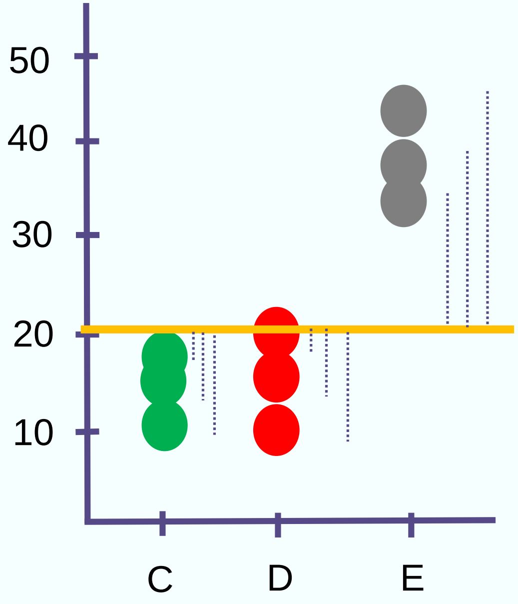
Alternative Hypothesis

- When we only have two groups of data the Alternative Hypothesis is pretty obvious because it is simply the opposite of the Null Hypothesis.

- However when we have 3 or more groups the alternative hypothesis become more interesting.

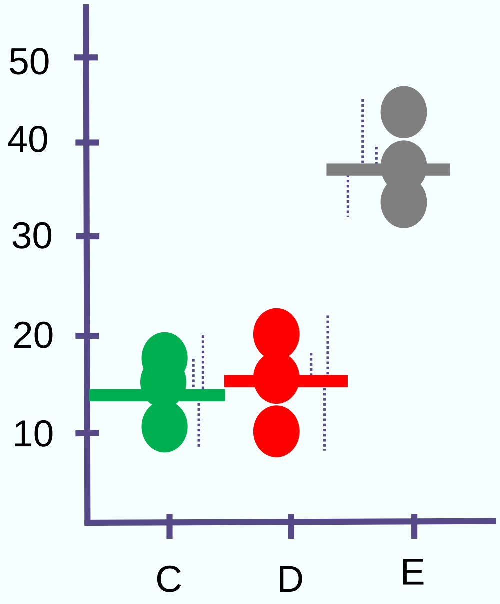
- In this case the null hypothesis is that there is no difference between Drugs C,D and E.

- Like before we can represent the null hypothesis by measuring the distances from the data to a single mean value.



Alternative Hypothesis

- However, now we have choices for the Alternative Hypothesis.
- **One Alternative Hypothesis could be that all 3 drugs are different.**
- So in this case we would measure the distances from a separate mean for each drug

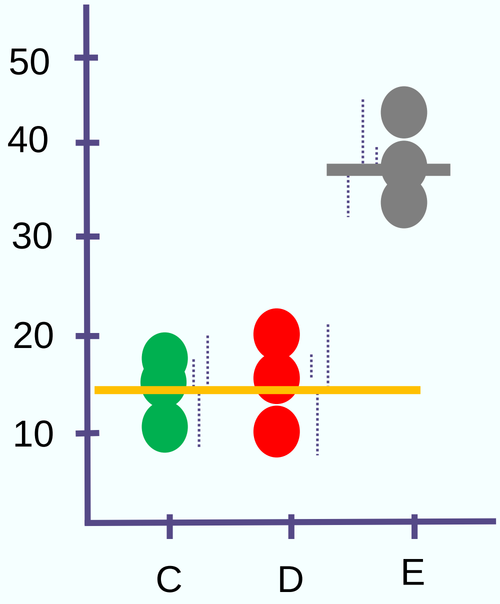


Alternative Hypothesis

- Or the Alternative Hypothesis could be that there is no difference between Drugs C and D, but Drug E is doing its own thing.

- In this case we would calculate the distance from a single mean value for Drug C and D and a separate mean for Drug E.

- Depending on which Alternative Hypothesis we use in the statistical test, we can end up making a different decision about the Null Hypothesis.



Alternative Hypothesis

- So it is important to clearly state which Alternative Hypothesis we want to use.
- However regardless of the alternative hypothesis we use in the test, we only reject or fail to reject the primary or null hypothesis.
- We would still not say that we accept the alternative hypothesis, because other alternative might be better.
- In other words there are too many possibilities to test to know if we have accepted the correct one.

P-Value

- Imagine we have two drugs, Drug A and Drug B. And we want to know if Drug A is different from Drug B.
- One person using Drug A is cured and the one person used Drug B is not cured!
- **Can we conclude that Drug A is better than Drug B?**
- We need to try each drug on more than just one person each.
- So we test the drug on a lot of different people and the following are the results:

Drug A

Cured	Not Cured
1043	3

99.7 % are cured

Drug B

Cured	Not Cured
2	1432

0.1 % are cured

P-Value

- What if the results were like this:

Drug A

Cured	Not Cured
73	125

37 % are cured

Drug B

Cured	Not Cured
59	131

31 % are cured

- As we see, No study is perfect and there are always a few random things that happen.
- **How confident can we be that the Drug A is superior?**
- **That's where the P-Value comes in.**

P-Value

- P-value is a number between 0 and 1, that in our example **it quantifies how confident we should be that Drug A is different from Drug B.**

- The closer a P-Value is to 0, the more confidence we have that Drug A and Drug B are different.

- **Significance Level(α):** How small does a P-value have to be before we are sufficiently confident that Drug A is different from Drug B? In other words what threshold can we use to make a good decision?

- In practice, a commonly used **threshold or Significance level** is **0.05**. It means **that if there is no difference between Drug A and Drug B, and if we did this exact same experiment a bunch of times, the only 5 % of those experiments would result in the wrong decision.**

P-Value

- So a 0.05 threshold for P-value means that 5% of the experiments, where the only difference comes from weird random things, will generate a P-value smaller that 0.05.
- In other words if there is no difference between drug A and drug B, 5% time we do the experiment, we'll get a P-Value less than 0.05, aka a False Positive.
- Note: if it is extremely important that we are correct when we say that drugs are different then we can use a smaller threshold like 0.000001. it means we would only get a false positive every 100,000 experiment.
- The **Null Hypothesis** is that the drugs are the same.
- **The P-Value helps us decide if we should reject the null hypothesis or not.**

P-Value

P-Value = **P(observed or more extreme outcome** | H_0 **true)**

- If the P-value is low(lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the Null Hypothesis were true and hence **reject H_0.**

- If the P-value is high(higher than) we say that it is likely to observe the data even if the Null Hypothesis were true, and hence **do not reject H_0.**

Calculating P-Value

1) **Identify the correct test statistic**: All hypothesis tests boil your sample data down to a single number known as a test statistic. T-tests use t-values, F-tests use F-values, Chi-square tests use chi-square values. Choosing the correct one depends on the type of data you have and how you want to analyze it.

2) **Calculate the test statistic**: How you calculate the test statistic depends on which one you're using. The method for calculating test statistics varies by test type.

3) **Specify the characteristics of the test statistic's sampling distribution**: Test statistics are unitless, making them tricky to interpret on their own. You need to place them in a larger context to understand how extreme they are. The sampling distribution for the test statistic provides that context.

4) **Place your test statistic in the sampling distribution to find the p-value**: This step involves calculating the probability of observing a test statistic as extreme as the one calculated from your data, assuming the null hypothesis is true

Calculating P-Value

- Different types of tests have different methods for calculating the p-value. For example:

  Z-test: Used for hypotheses concerning the mean of a normal distribution with known variance.

  T-test: Used for hypotheses concerning the mean of a normal distribution when the variance is unknown.

  F-test: Used for hypotheses concerning the variance.

Types of Errors

		Truth	
		Null Hypothesis is TRUE	Null Hypothesis is FALSE
Decision	Reject null hypothesis	Type I Error (False positive)	Correct Outcome! (True positive)
	Fail to reject null hypothesis	Correct Outcome! (True negative)	Type II Error (False negative)

Type I Error Rate

- We reject H_0 when the p-value is less than 0.05 (α = 0.05).
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

$$P(\text{Type 1 error}) = P(\text{reject } H_0 \mid H_0 \text{ true}) = \alpha$$

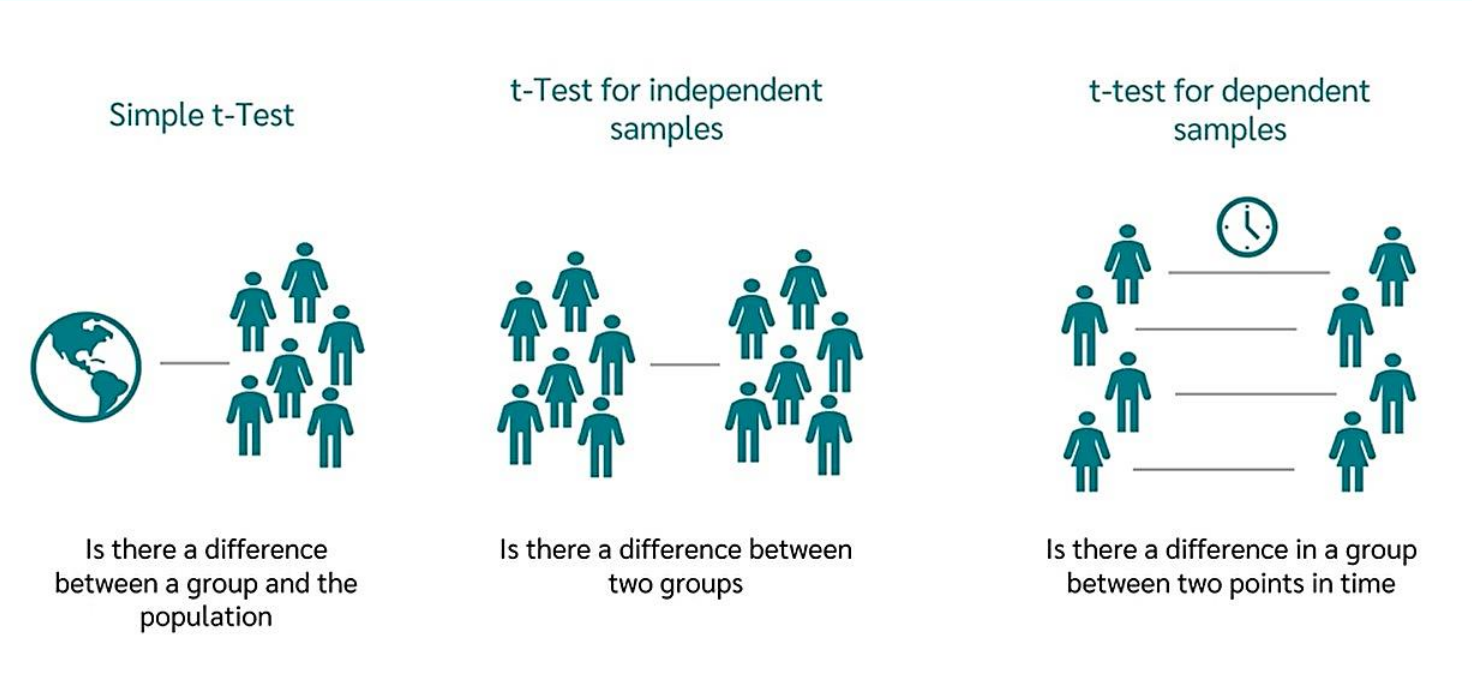
- This is why we prefer small values of α: increasing α increases the Type 1 error rate.

Choosing α

- If Type 1 Error is dangerous or especially costly, choose a small significance level (e.g. 0.01)
- **Goal:** we want to be very cautious about rejecting H_0, so we demand very strong evidence favoring H_a before we would do so.

- If Type 2 Error is relatively more dangerous or more costly, choose a higher significance level (e.g. 0.1)
- **Goal:** we want to be cautious about failing to reject H_0 when the null is actually false.

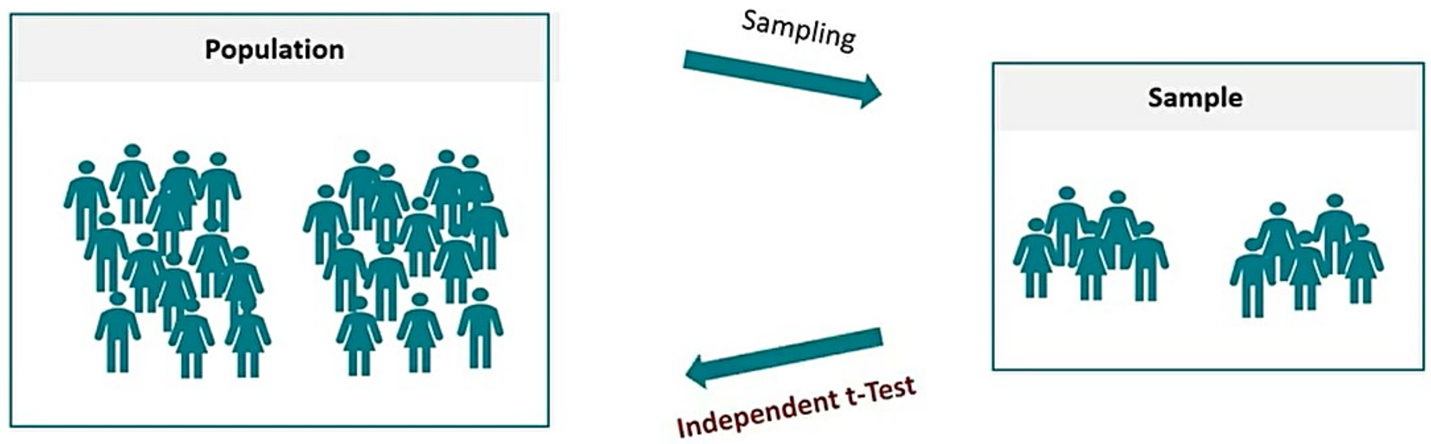
t-Test



Simple t-Test
t-Test for independent samples
t-test for dependent samples
Is there a difference between a group and the population
Is there a difference between two groups
Is there a difference in a group between two points in time

Why an Independent t-Test

- We need the independent t-test to make a statement about the population based on the sample.



Questions and Hypotheses

- **Is there a statistically significant difference between the mean value of two groups**

1. Is there a difference between people with and without studies with regard to their health?

2. Do smokers have a higher risk of heart attack than non-smokers?

3. Is there difference between the recovery time Drugs C and D?

Hypotheses

Null hypothesis H0

The null hypothesis assumes that there is no difference between two groups with respect to a characteristic.

Example:

The salary of men and women does not differ in Germany.

Alternative hypothesis H1

Alternativhypothesen hingegen gehen davon aus, dass ein Unterschied zwischen zwei Gruppen vorliegt.

Example:

The salary of men and women differs in Germany.

Directed and Undirected Hypotheses

Undirected Hypothesis :

Check if there is a difference, it doesn't matter in which direction the connection or the difference goes.

1. There is a difference between the salary of men and woman
2. There is a difference in heart attack risk between smokers and non-smokers

Directed Hypothesis:

Also indicates the direction of the difference

1. Men earn more than women
2. Smokers have a higher risk of heart attack than non smokers

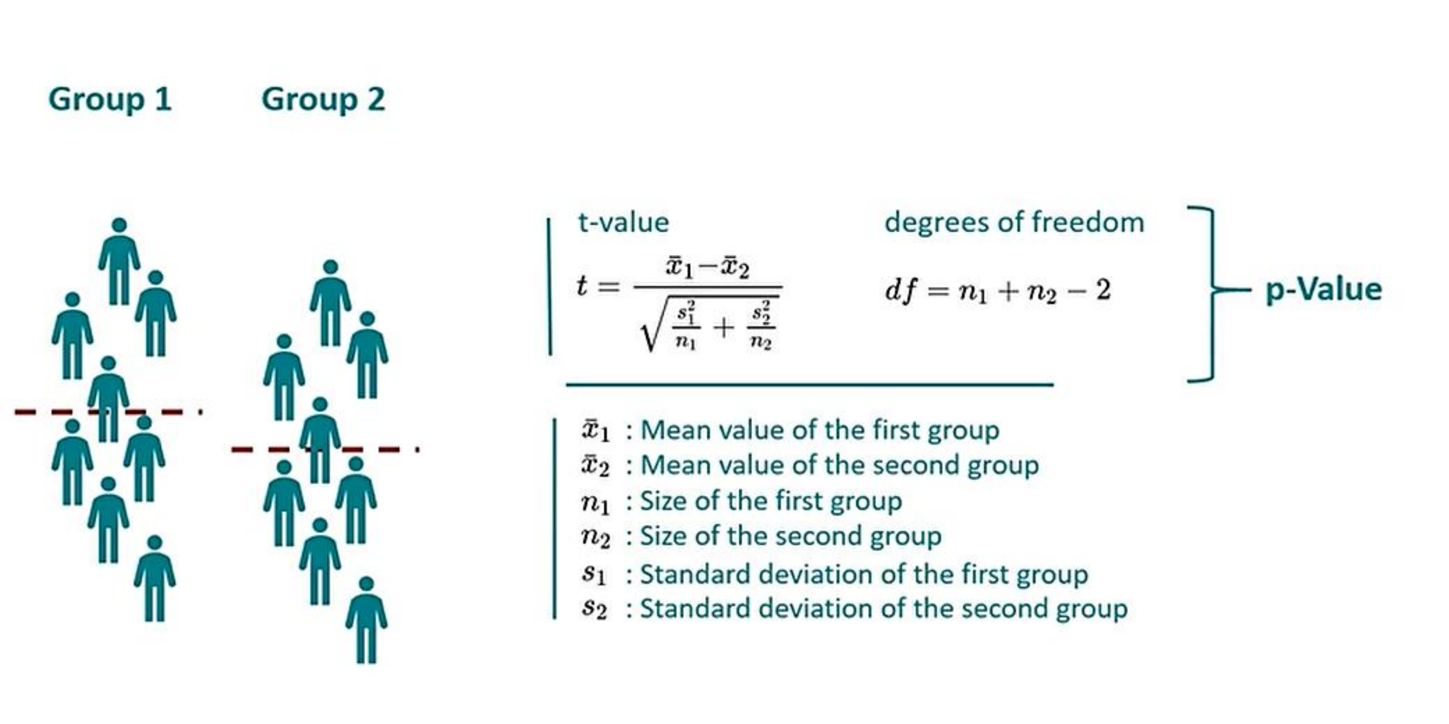
Requirements for the Independent t-test

- There is an independent variable(e.g. gender), which has two characteristics or groups(e.g. male and female).

- These two groups should be compared in the analysis. The question is thus is there a difference between the two groups regarding the dependent variable(e.g. income)

1) **The two groups or samples must be independent.**
2) **The variables must be scaled in intervals**
3) **The variables must be normally distributes**
4) **The variance within the groups should be similar**

Calculate t-Test



Chi^2-Test

- It is used when we have two categorical variables that you want to examine.

Categorical variables

Gender	Preferred newspaper	Frequency of television	Highest educational level
1 = male	1 = The Washington Post	1 = daily	1 = Without graduation
2 = female	2 = The New York Times	2 = several times per week	2 = College
	3 = USA Today	3 = more rarely	3 = Bachelor's degree
	4 = ...	4 = never	4 = Master's degree

Gender ⟷ **Preferred newspaper**

Preferred newspaper ⟷ **Frequency of television**

Gender ⟷ **Frequency of television**

Frequency of television ⟷ **Highest educational level**

Fall	Gender	Highest educational level
1	Male	College
2	Female	Without graduation
3	Male	Without graduation
4	Male	Bachelor's degree
5	Female	Master's degree
6	Male	Bachelor's degree
7	Female	Master's degree
...	...	...

	Female	Male
Without graduation	6	7
College	13	16
Bachelor's degree	16	15
Master's degree	8	11
Total	**43**	**49**

Is there a correlation between gender and the highest level of education?

Chi^2-Test

Calculate Chi^2-Test

Observed frequencies

	Female	Male
Without graduation	6	7
College	13	16
Bachelor's degree	16	15
Master's degree	8	11
Total	**43**	**49**

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}$$

Expected frequencies for perfectly independent variables

	Female	Male
Without graduation	6.08	6.92
College	13.55	15.45
Bachelor's degree	14.49	16.51
Master's degree	8.88	10.12
Total	**43**	**49**

Expected Frequency = (Row Total x Column Total) / Grand Total

Reading from the Chi^2 Table

$Chi^2 = 0{,}504$

df = (number rows - 1) (number columns - 1) = 3

	Female	Male
Without graduation	6	7
College	13	16
Bachelor's degree	16	15
Master's degree	8	11
Total	**43**	**49**

Chi-Quadrat Tabelle

Signifikanzniveau Alpha	0,995	0,975	0,2	0,1	0,05	0,025	0,02	0,01	0,005	0,002	0,001
Freiheitsgrade											
1	0	0,001	1,642	2,706	3,841	5,024	5,412	6,635	7,879	9,55	10,828
2	0,01	0,051	3,219	4,605	5,991	7,378	7,824	9,21	10,597	12,429	13,816
3	0,072	0,216	4,642	6,251	7,815	9,348	9,837	11,345	12,838	14,796	16,266
4	0,207	0,484	5,989	7,779	9,488	11,143	11,668	13,277	14,86	16,924	18,467
5	0,412	0,831	7,289	9,236	11,07	12,833	13,388	15,086	16,75	18,907	20,515
6	0,676	1,237	8,558	10,645	12,592	14,449	15,033	16,812	18,548	20,791	22,458
7	0,989	1,69	9,803	12,017	14,067	16,013	16,622	18,475	20,278	22,601	24,322
8	1,344	2,18	11,03	13,362	15,507	17,535	18,168	20,09	21,955	24,352	26,124
9	1,735	2,7	12,242	14,684	16,919	19,023	19,679	21,666	23,589	26,056	27,877
10	2,156	3,247	13,442	15,987	18,307	20,483	21,161	23,209	25,188	27,722	29,588

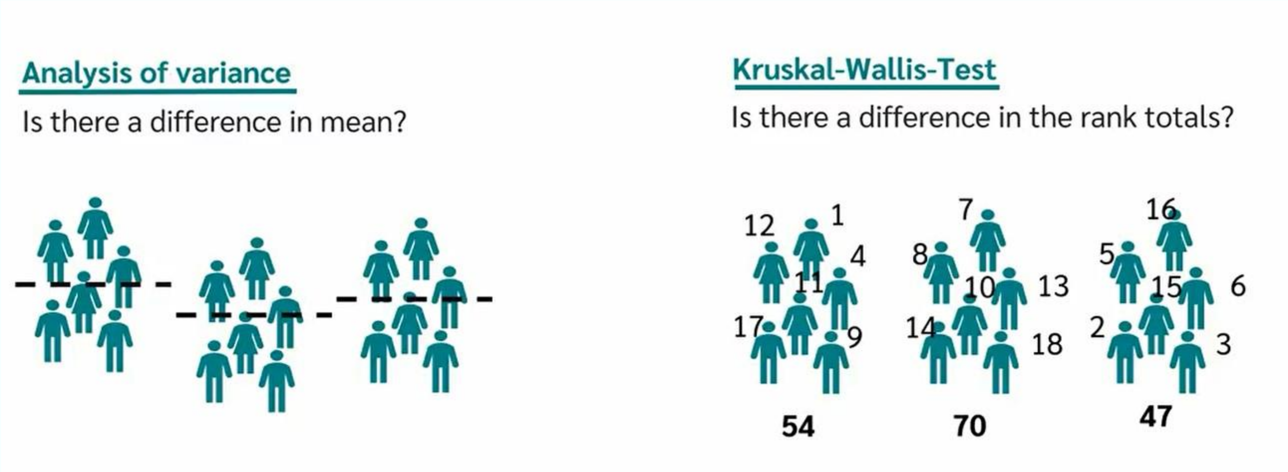
So fail to reject the null hypothesis

Kruskal-Wallis test

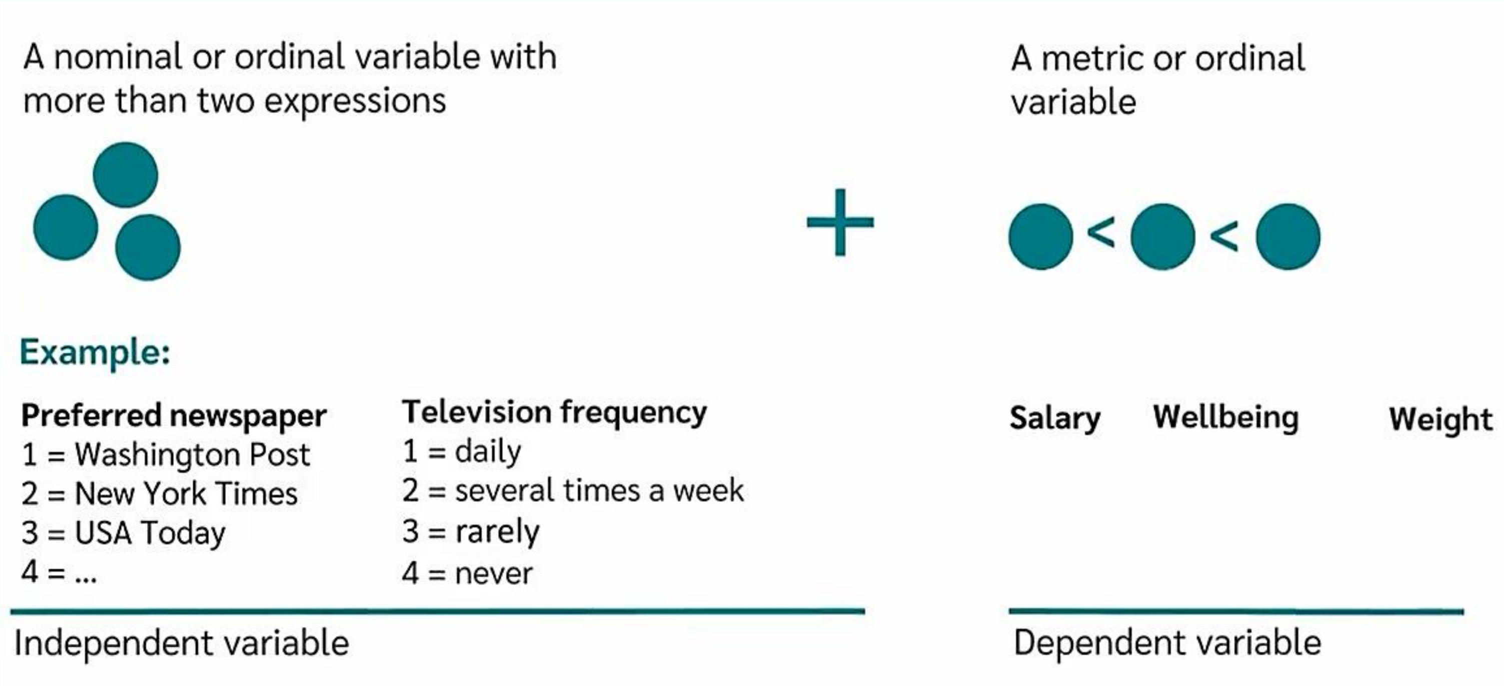
- It is a hypothesis test that is used when you want to test whether there is a difference between several independent groups.

- If your data are not normally distributed and the assumption for the analysis of variance are not met this test is used.

- If we do not look at the mean difference but at the rank sum the data does not have to be normally distributed. When using this test our data does not have to satisfy any distributional form.



When do we use Kruskal-Wallis test?



Assumptions:

Only several independent random samples with at least ordinally scaled characteristics must be available.

The variables do not have to satisfy a distribution curve.

Null hypothesis

The independent samples all have the same central tendency
and therefore come from the same population.

Alternative hypothesis

At least one of the independent samples does not have the same central tendency
as the other samples and therefore come from a different population.

Group	Response time	Rank
A	34	**2**
A	36	**4**
A	41	**7**
A	43	**9**
B	44	**10**
B	37	**5**
B	45	**11**
B	33	**1**
C	35	**3**
C	39	**6**
C	42	**8**
C	46	**12**

Rank sums:

$R_A = 2 + 4 + 7 + 9 = 22$

$R_B = 10 + 5 + 11 + 1 = 27$

$R_C = 3 + 6 + 8 + 12 = 29$

Mean Rank Sum:

$\bar{R}_A = 22 / 4 = 5.5$

$\bar{R}_B = 27 / 4 = 6.75$

$\bar{R}_C = 29 / 4 = 7.25$

$$E_R = \frac{n+1}{2} = \frac{12+1}{2} = 6.5$$

Group	Response time	Rank
A	34	2
A	36	4
A	41	7
A	43	9
B	44	10
B	37	5
B	45	11
B	33	1
C	35	3
C	39	6
C	42	8
C	46	12

Number of cases

$n = 12$

Expected value of the rankings

$E_R = 6.5$

Mean Rank Totals:

$\bar{R}_A = 22 / 4 = 5.5$

$\bar{R}_B = 27 / 4 = 6.75$

$\bar{R}_C = 29 / 4 = 7.25$

Degrees of freedom

$df = 2$

Rank variance

$$\sigma_R^2 = \frac{n^2 - 1}{12} = \frac{12^2 - 1}{12} = 11.92$$

Number of cases

$n = 12$

Expected value of the rankings

$E_R = 6.5$

Mean Rank Totals:

$\bar{R}_A = 22 / 4 = 5.5$

$\bar{R}_B = 27 / 4 = 6.75$

$\bar{R}_C = 29 / 4 = 7.25$

Degrees of freedom:

$df = 2$

Rank variance

$$\sigma_R^2 = \frac{n^2 - 1}{12} = \frac{12^2 - 1}{12} = 11.92$$

Test value H → equivalent to χ^2

$$H = \frac{n - 1}{12} \cdot \sum_{i=1}^{k} \frac{n_i (\bar{R}_i - E_R)^2}{\sigma_R^2}$$

$$H = \frac{12 - 1}{12} \cdot 4 \frac{(5.5 - 6.5)^2 + (6.75 - 6.5)^2 + (7.25 - 6.5)^2}{11.92}$$

$$= 0.5$$

Table of chi-squared distribution

Significance level Alpha	0.995	0.975	0.2	0.1	0.05	0.025	0.02	0.01
Degrees of freedom								
1	0	0.001	1.642	2.706	3.841	5.024	5.412	6.635
2	0.01	0.051	3.219	4.605	5.991	7.378	7.824	9.21
3	0.072	0.216	4.642	6.251	7.815	9.348	9.837	11.345
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277
5	0.412	0.831	7.289	9.236	11.07	12.833	13.388	15.086

So fail to reject the null hypothesis

RESOURCES

- George Casella, Roger L. Berger - Statistical Inference-Duxbury Press (2001)
- David M Diez, Christopher D Barr, Mine Çetinkaya-Rundel - OpenIntro Statistics-OpenIntro, Inc. (2015)