Analysis of Variance

Data Science | COM 221-ML

The **t-test** tells you how significant the differences between group means are.
It lets you know if those differences in means could have **happened by chance**.
The **t-test** is usually used when data sets follow a normal distribution but you don't know the population variance.

Analysis of Variance

Analysis of Variance or **ANOVA** is a statistical test to find out if a survey or experiment results are **significant**.
ANOVA helps you to figure out if you need to **reject the null hypothesis** or accept the **alternate hypothesis**.

Examples:
Basically, you're testing groups to see if there's a difference between them. Examples of when you might want to test different groups:

- A group of psychiatric patients are trying three different therapies: **counseling**, **medication** and **biofeedback**. You want to see if one therapy is better than the others.
- A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other.
- Students from different colleges take the same exam. You want to see if if students who study alone outperforms students who studied in a group.

ANOVA VS. T-test

Descriptive	Inferential
ANOVA can compare three or more groups while	T tests are only useful for comparing two groups at one time.
ANOVA compares variances between populations	T test will tell you if there is a significant variation between groups
ANOVA will give you a single number (f-statistic) and one (p-value) to help you support or reject the null hypothesis	In T test, as the groups grow in number, you may end up with a *lot* of pair comparisons that you need to run

Analysis of Variance Examples

Water	Juice	Coffee
10	11	12
12	14	13
18	19	17
24	23	25
36	38	37

For example, we have a dataset that shows the reaction time of different people when taking either water, juice or coffee.

You gathered this data because you want to compare how effective drinking a coffee is when it comes to reaction time.

Analysis of Variance Examples

Water	Juice	Coffee
10	11	12
12	14	13
18	19	17
24	23	25
36	38	37

There is a lot of variation in each group. Some people have faster reaction times and other are slower

Water	Juice	Coffee
10	11	12
12	14	13
18	19	17
24	23	25
36	38	37

But each group looks pretty much the same. There is not much variation between each group.
You would conclude that most of the difference is due to the people and the type of drink did not make much of a difference
You would **accept** the **null hypothesis** that the type of drink does not an effect on a person's reaction time.

Analysis of Variance Examples

Water	Juice	Coffee
29	17	10
29	18	11
30	19	12
31	19	12
31	20	13

In this case, all the scores within each group are close to one another. There is not much variation in each group

Water	Juice	Coffee
29	17	10
29	18	11
30	19	12
31	19	12
31	20	13

But each group are very different from one another . There is a lot of variation between the groups .
In this case, you would **reject** the **null hypothesis** and that the type of drink makes a difference on a person's reaction time.

In ANOVA, we figure out how much of the total variance comes from:
1. The variance between the groups
2. The variance within the groups
Calculate the ratio:

$$F = \frac{variance\ between\ groups}{variance\ within\ groups}$$

The larger the ratio, the more likely it is that the groups have different means (**reject the null hypothesis**)

$$F = \frac{variance\ between\ groups}{variance\ within\ groups}$$

Water	Juice	Coffee
10	11	12
12	14	13
18	19	17
24	23	25
36	38	37

In this case, the variance between and within groups is **not so obvious**. You probably cannot tell if there is a significant effect because its not clear whether there is more variance between or within groups or how much.

F= 4.27

p=0.4

The calculation for this example shows that the ration is 4.27 with a p-value of 0.04. In this case, we **cannot reject the null hypothesis**

With this example, the drink you give does have an effect on reaction time.

Multiple Variables

	Water	Juice	Coffee
Morning	30,31,31, 32,32	28,30,27, 29,32	25,26,25, 28,29
Evening	31,31,33, 35,30	29,30,28, 29,31	28,30,27, 26,27

Summary

The main idea of ANOVA is to figure out how much of the total variance comes from:

The variance ***between*** the groups

The variance ***within*** the groups

Water	Juice	Coffee
29	17	10
29	18	11
30	19	12
31	19	12
31	20	13

If most of the variation is between groups, there is **probably a significant effect.**

Water	Juice	Coffee
10	11	12
12	14	13
18	19	17
24	23	25
36	38	37

If most of the variation is within groups, there is **probably not a significant effect**.