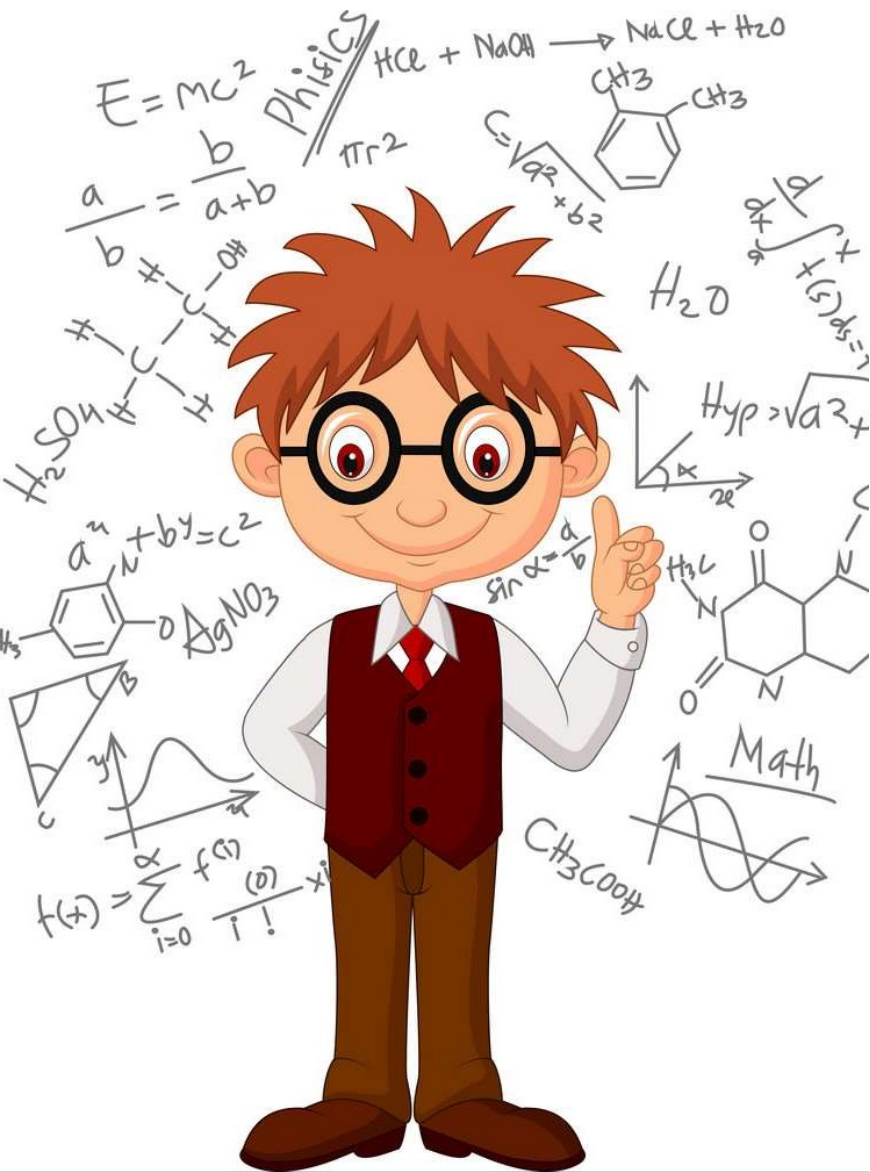


# Statistics for Data Analytics

Lecturer: Marina Iantorno

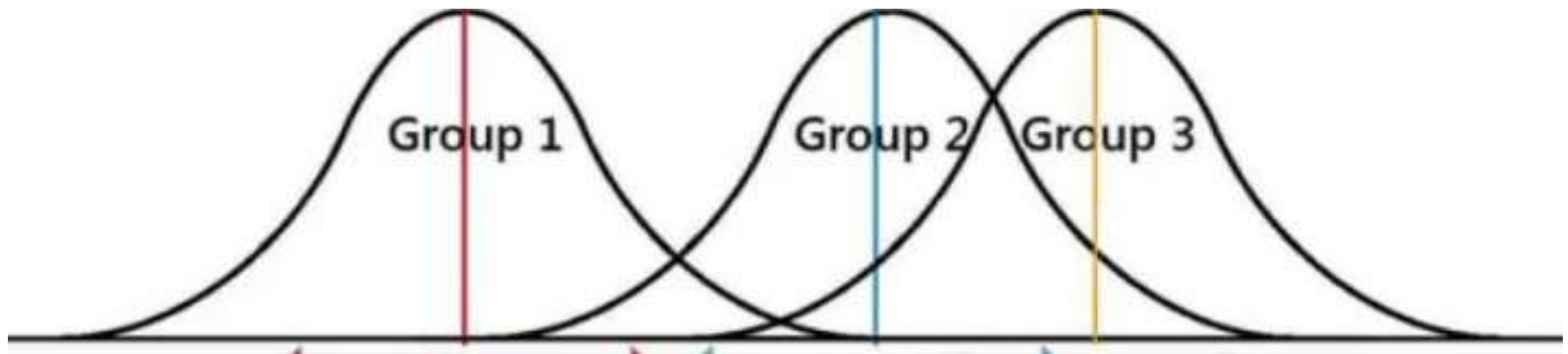
E-mail: [miantorno@cct.ie](mailto:miantorno@cct.ie)





In today's class we will cover:

- ☐ Two-Way ANOVA
- ☐ Introduction to non-parametric tests
- ☐ Sign test



## Two-Way ANOVA

---

# ANOVA

Analysis of Variance is often used in experiments to see whether different levels of an explanatory variable ( $x$ ) get different results on some quantitative variable  $y$ . In these cases, the  $x$  variable is called “a factor”.

For example, suppose we want to compare the average change in blood pressure on certain doses of a drug. Let's say that the doses could be 10mg, 20mg or 30mg per day. The factor would be the drug doses.

# ANOVA

Now, let's suppose that someone else studies the response to that same drug and examines whether the times taken per day (one or two times) has any effect on the blood pressure. In this case the factor would be the number of times per day, and it will have two levels: once and twice.



# ANOVA

Imagine that we want to study the effects of dosage **AND** number of times taken together because we believe that both may have an effect on the response. This analysis will be called a two-way ANOVA, which will use two factors together to compare the average response. It is an extension of one-way ANOVA with a twist, because the two factors you use may operate on the response differently together than separately.



# ANOVA

The two-way ANOVA model contains two factors, namely A and B, and each factor has a certain number of levels (let's say "i" levels of factor A and "j" levels of factor B).

In our drug study example, we could see

A = drug dosage, with  $i = 1, 2$  or  $3$ .

B = number of times taking the medicine per day, with  $j = 1$  or  $2$ .

Each person involved in the study is subject to one of the three different drug dosages and will take the drug in one of the two methods given.

# ANOVA

Each person involved in the study is subject to one of the three different drug dosages and will take the drug in one of the two methods given. That means that we will have  $3 * 2 = 6$  different combinations of Factors A and B that you can apply to the subjects, and you can study these combinations and their effects on blood pressure changes in the two-way ANOVA model.

Each different combination of levels of Factors A and B is called “treatment” in the model.





# ANOVA

Dosage Amount	One Dose Per Day	Two Doses Per Day
10mg	Treatment 1	Treatment 2
20mg	Treatment 3	Treatment 4
30mg	Treatment 5	Treatment 6

For example, Treatment 4 is the combination of 20mg of the drug taken in two doses of 10mg per day.

If Factor A has  $i$  levels and Factor B has  $j$  levels, you have  $i * j$  different combinations of treatments in the two-way ANOVA model.

# ANOVA

While the purpose of one-way ANOVA model is to test to see whether the different levels of Factor A produce any different response in the y variable, the two-way ANOVA model include another factor (B) plus an interaction term between these factors (AB).

The sums of squares equation for the two-way ANOVA model is:

$$SSTO = SSA + SSB + SSAB + SSE$$

Let's analyse it!

# ANOVA

$$SSTO = SSA + SSB + SSAB + SSE$$

SSTO = this is the total variability in the y-values.

SSA = this is the sums of squares due to the Factor A (variability in the y-values explained by the factor A).

SSB = this is the sums of squares due to the Factor B (variability in the y-values explained by the factor B).

SSAB = this is the sums of squares due to the interaction of Factors A and B.

SSE is the amount of variability left unexplained by the model (the error).

# ANOVA

The main point of this model is the interaction between the variables, because we know that the two factors may act together in a different way than they would separately.

“Interaction” is when two factors meet or interact with each other. In any two-way ANOVA we must check out the interaction term first. If A and B interact with each other and the interaction is statistically significant, we cannot examine the effects of either factor separately. Their effects are intertwined and cannot be separated.

# ANOVA

We have 4 conditions to proceed with the two-way ANOVA model:

1. The samples must be independent.
2. The groups must have the same sample size.
3. The populations from which the samples were obtained must be normally or approximately normally distributed.
4. The variances of the populations must be equal.

# ANOVA

Let's see this with an example!

Suppose that you are working at a laboratory that studies the performance of two different brands of detergent: Best and Super. In order to evaluate the efficiency, the dishes were washed with cold, warm and hot water, respectively. You are required to use the data from the file “detergent.csv” that is in Moodle to confirm whether the performance score is related to the brand and the temperature of the water.



# ANOVA

$H_0$  = The performance of the detergent is not affected by the brand and the temperature of the water.

$H_1$  = The performance of the detergent is affected by the brand and the temperature of the water

*Remember that  $H_0$  states that nothing happens, while  $H_1$  will create a statement that “fights” against it.*

## 1. Samples must be independent.

We will follow the same principle as we did with one-way ANOVA: as the samples are randomly taken there is no reason to presume dependence from one variable to another one.



# ANOVA

## 2. The groups must have the same sample size.

We should have the same number of observations for Best than for Super, as well as the same number of observations of the same temperature of the water. Let's have a look at our data.

	Cold	Warm	Hot
<b>Super</b>	4	7	10
	5	9	12
	6	8	11
	5	12	9
<b>Best</b>	6	13	12
	6	15	13
	4	12	10
	4	12	13





# ANOVA

3. The populations from which the samples were obtained must be normally or approximately normally distributed.

We need to test the normality of the sample. We will use the Shapiro Wilk test to evaluate in function of our y variable. In this case the y variable is the performance score.

Our  $H_0$  says that the population data follows a normal distribution and  $H_1$  states that the population data does not follow a normal distribution.

# ANOVA

3. The populations from which the samples were obtained must be normally or approximately normally distributed.

```
#Shapiro wilk test  
  
stats.shapiro(dataset.Perf)  
  
ShapiroResult(statistic=0.9176760315895081, pvalue=0.05190473049879074)
```

As the p-value is greater than 0.05, we accept the null hypothesis and therefore, we can assume that the data comes from a normal distribution.



# ANOVA

## 4. The variances of the populations are equal.

There are different tests to measure the homogeneity of the variances. Today we will use the Levene Test.

We will perform the test using  $y$  as the performance and  $x$  as the brands and temperatures that are part of this analysis.

# ANOVA

## 4. The variances of the populations are equal.

The null hypothesis states that the variances are equal while the alternative hypothesis states the opposite. We will use “Pingouin’s homoscedasticity” to calculate it with Levene.

```
7]: pg.homoscedasticity(dataset, dv='Perf',  
                        group='Detergent',  
                        method='levene')
```

7]:

	W	pval	equal_var
levene	0.452442	0.508182	True

```
5]: pg.homoscedasticity(dataset, dv='Perf',  
                        group='Temp',  
                        method='levene')
```

5]:

	W	pval	equal_var
levene	2.015152	0.158274	True

As p-value is greater than alpha, we accept the null hypothesis, and therefore we have no reason to think that the variances are not equal.



# ANOVA

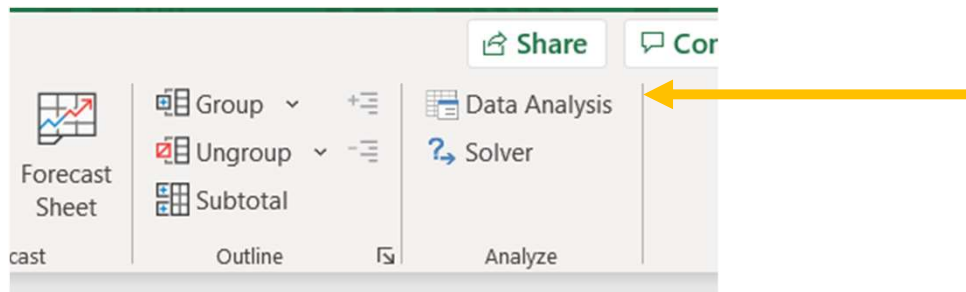
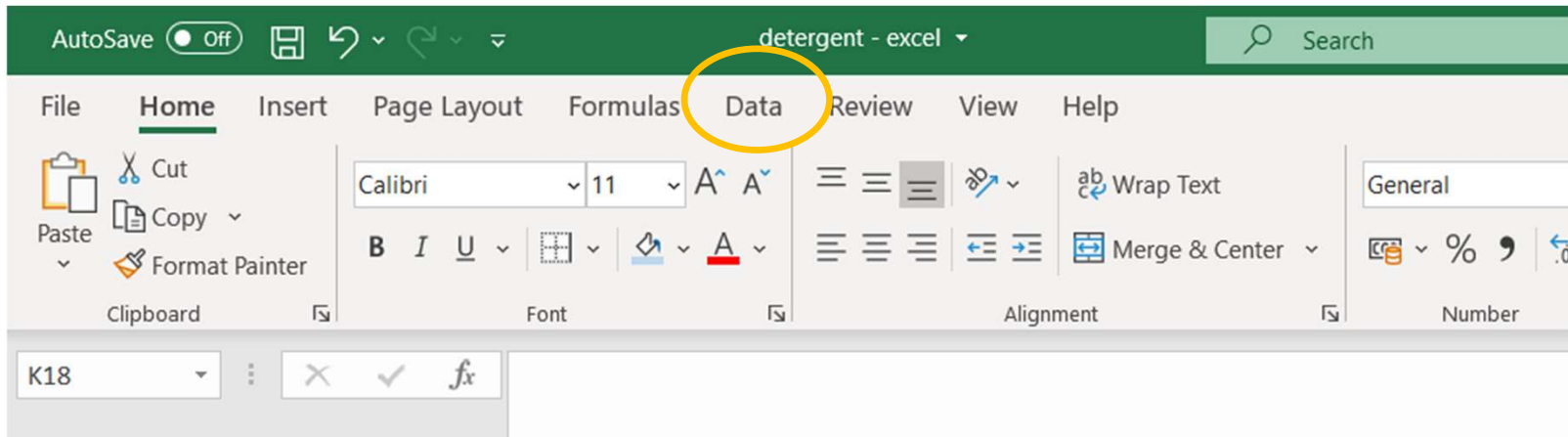
We checked all the conditions. Now we can proceed with our two-way ANOVA model!

Let's try it first in Excel!

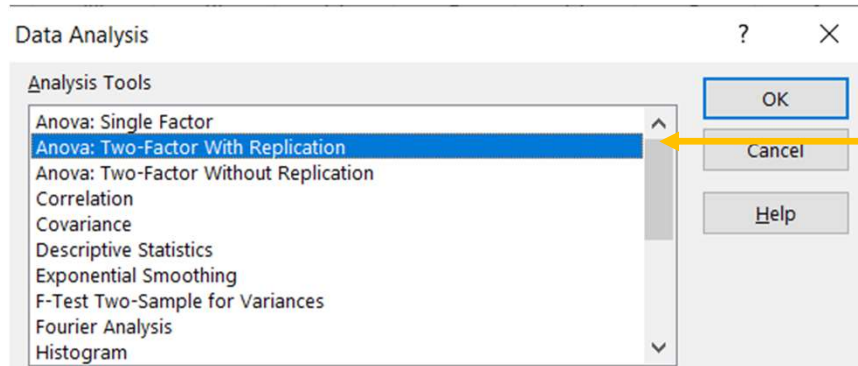
First, filter your data and allocate it on a table like this one:

	Cold	Warm	Hot
<b>Super</b>	4	7	10
	5	9	12
	6	8	11
	5	12	9
<b>Best</b>	6	13	12
	6	15	13
	4	12	10
	4	12	13

# ANOVA



# ANOVA



# ANOVA

	G	H	I	J	K	L
		Cold	Warm	Hot		
Super		4	7	10		
		5	9	12		
		6	8	11		
		5	12	9		
Best		6	13	12		
		6	15	13		
		4	12	10		
		4	12	13		

Anova: Two-Factor With Replication

Input  
Input Range:   
Rows per sample:   
Alpha:

Output options  
☐ Output Range:   
☒ New Worksheet Ply:   
☐ New Workbook

OK  
Cancel  
Help

The range will be the whole table (including labels)

This is the number of observations per category (4 cold, 4 warm, 4 hot in this example)

Alpha will always be 0.05. Just ensure this is the number of this cell.



# ANOVA

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Sample	20.16667	1	20.16667	9.810811	0.005758	4.413873
Columns	200.3333	2	100.1667	48.72973	5.44E-08	3.554557
Interaction	16.33333	2	8.166667	3.972973	0.037224	3.554557
Within	37	18	2.055556			

If our p-value is below alpha, we can say that there are differences between the results of one sample and another one (Best and Super in this case)

# ANOVA

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Sample	20.16667	1	20.16667	9.810811	0.005758	4.413873
Columns	200.3333	2	100.1667	48.72973	5.44E-08	3.554557
Interaction	16.33333	2	8.166667	3.972973	0.037204	3.554557
Within	37	18	2.055556			

When p-value of Columns is below alpha, we can say that there are differences between the y value per each factor (there are differences between the performance score of Best and Super)

# ANOVA

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Sample	20.16667	1	20.16667	9.810811	0.005758	4.413873
Columns	200.3333	2	100.1667	48.72973	5.44E-08	3.554557
Interaction	16.33333	2	8.166667	3.972973	0.037224	3.554557
Within	37	18	2.055556			




The most important is the interaction between the variables. When our p-value is below alpha, we can say that there is an interaction between the brand and the temperature of the water (interaction between independent variables)

# ANOVA

Now let's do it with Python.

```
9]: #ANOVA TWO WAYS
model = ols('Perf~Detergent+Temp', data = dataset).fit()
aov2 = sm.stats.anova_lm(model, type=2)
print(aov2)
```

	df	sum_sq	mean_sq	F	PR(>F)
Detergent	1.0	20.166667	20.166667	7.5625	1.234634e-02
Temp	2.0	200.333333	100.166667	37.5625	1.687925e-07
Residual	20.0	53.333333	2.666667	NaN	NaN



We need to look at the p-value and understand our main goal. In this case, we rejected the null hypotheses, but what were the null hypotheses?

# ANOVA

$H_0$  = The performance of the detergent is not affected by the brand and the temperature of the water.

$H_1$  = The performance of the detergent is affected by the brand and the temperature of the water

## Interpretation of the results:

We reject the null hypothesis, therefore there is enough evidence to conclude that the performance score of the detergent will be affected by the brand and the temperature of the water.

In conclusion, ...



## Non-Parametric tests

# Non-parametric tests

The hypothesis tests we have been using require certain conditions, such as specific distribution of the population the data came from or the sample size. Those tests are called “*parametric- tests*” and are very powerful and precise; we try to use them as often as we can.

But sometimes the data do not meet the conditions for a parametric procedure. Maybe one of the conditions is not met, or our data is different than the typical quantitative data, such as ranks. When that happens, the best option is follow a “*non-parametric procedure*”.

# Non-parametric tests

In general, non-parametric procedures are not as powerful as the parametric ones, but they are still useful in Data Analysis.

A positive aspect of these procedures is that they are easy to carry out, and most importantly, non-parametric tests give accurate results compared to the parametric procedures when the conditions of parametric procedures are not met or are not appropriate.



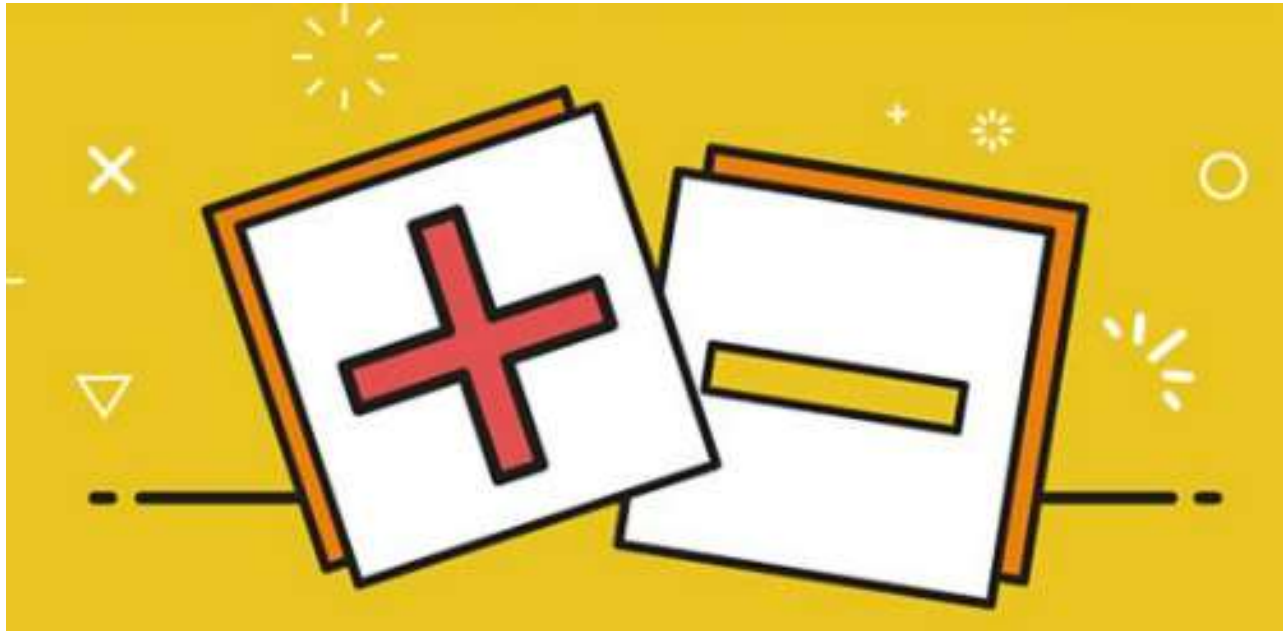
# Non-parametric tests

We have to bear in mind that some tests are more powerful than others, but at the same time there are some limitations, therefore we must be careful when we decide to place our study, because there are different factors to consider. Also, another interesting fact of these tests is that they are testing on the whole distribution and not only on a specific parameter, thus its name:

non-parametric tests.

Let's explore a bit more!





Sign Test

# Sign Test

This test, along with the Wilcoxon signed rank test are meant to test or estimate the median or mean of one population. These two non-parametric procedures are the counterparts to the one-sample and matched pairs t-tests, which require data from a normal population.

We use the one sample t-test to find out whether or not the population mean is equal to certain value, but it requires the data to come from a normal distribution, and when this condition is not met, the sign test is a non-parametric option.

# Sign Test

There are some aspects that we must consider:

- This test is used to determine whether a post-procedure presents improvements regarding the previous one.
- The scale of measurement is ordinal (check class 8).
- It does not assume any distribution for the original variable, but we generate a binominal distribution to proceed the test.
- Our variable will be “ $x$  = number of elements that improve/get worse”.
- It could be a one side or a bilateral test.
- This test is also known as a Cox-Stuart test.



# Sign Test

You work at a factory and you want to know if, in general, the employees are working more over time (extra hours) than they did last year. You randomly pick 10 employees, and these are your results:

	A	B	C	D	E	F	G	H	I	J
Hours 2019	10	15	22	5	18	18	15	16	0	4
Hours 2020	12	16	25	4	20	18	10	9	2	4

With a 5% of significance determine if there is enough evidence to say that in 2020 the employees worked more hours than in 2019.

# Sign Test

We have to define our variable:

$X$  = number of employees who worked more hours in 2020 than in 2019 (within 8 observations)

	A	B	C	D	E	F	G	H	I	J
Hours 2019	10	15	22	5	18	18	15	16	0	4
Hours 2020	12	16	25	4	20	18	10	9	2	4

We remove those observations that have the same value, because they will not be greater or less, they will be neutral

$n = 8$

# Sign Test

## Step 1: Statement of the hypothesis test

$$H_0 : p = 0.50$$

$$H_1 : p > 0.50$$



There is no change

The employees worked more

We state this hypothesis because  $H_0$  states that half of the employees worked more, and the other half worked less, therefore there was no change between one year and another one

The alternative hypothesis says that most of the employees (more than the half of them) worked more hours than the previous year.

# Sign Test

## Step 2: Formula

$X$  = number of positive signs within my sample



# Sign Test

## Step 3: Critical values

➤ Table: binomial

➤ Sign of  $H_1$ : >

➤  $\alpha = 0,05$

r	0	1	2	3	4	5	6	7	8
P( x = r)	0.00391	0.03125	0.10938	0.21875	0.27344	0.21875	0.10938	0.03125	0.00391

# Sign Test

## Step 3: Critical values

r	0	1	2	3	4	5	6	7	8
P( x = r)	0.00391	0.03125	0.10938	0.21875	0.27344	0.21875	0.10938	0.03125	0.00391

I start counting probabilities from the right to the left until the first time I pass the value of alpha.

$$P(r = 8) = 0.00391 \rightarrow \text{less than } 0.05$$

$$P(r = 8) + P(r = 7) = 0.00391 + 0.03125 = 0.03516 \rightarrow \text{less than } 0.05$$

$$P(r = 8) + P(r = 7) + P(r = 6) = 0.00391 + 0.03125 + 0.10938 = 0.14454 \rightarrow \text{greater than } 0.05$$

The critical value is 7.

# Sign Test

Step 4: Decision rule

We reject  $H_0$  if  $x \geq 7$

I accept  $H_0$  if  $x \leq 6$



# Sign Test

## Step 5: Calculating the empiric value

Remember step 2. Here we need to count the number of positive signs. In other words, we consider positive to those observations that resulted greater and negative those that resulted less than the previous year.

	A	B	C	D	E	G	H	I
Hours 2019	10	15	22	5	18	15	16	0
Hours 2020	12	16	25	4	20	10	9	2
Sign	+	+	+	-	+	-	-	+

$$X = 5$$

# Sign Test

### Step 6: Result of the HT



# Sign Test

## Step 7: Interpretation

*According to the sample and at a 5% significance level, there is no evidence to say that the employees increased the number of extra hours.*

**THAT'S ALL FOR TODAY**

**THANK YOU**

