Lesson 5:
Hypothesis testing (two samples)

Victoria Blanes-Vidal
Manuella Lech Cantuaria
The Maersk Mc-Kinney Møller Institute
Applied AI and Data Science

Lesson	Week	Date	TOPICS	Teacher
1	35	1/Sep	Introduction to the course Descriptive statistics –Part I	MLC
2	36	8/sep	Descriptive statistics –Part II	MLC
3	37	15/Sep	Probability distributions	MLC
4	38	22/Sep	Hypothesis testing (one sample)	VBV
5	39	29/Sep	Hypothesis testing (two samples)	VBV
6	40	6/Oct	ANOVA one-way	VBV
7	41	13/Oct	R class (Introduction to R and descriptive statistics) Point giving activity (in class)	MLC
-	42	20/Oct	NO CLASS (Autum holidays)	
8	43	27/Oct	R class (hypothesis testing + ANOVA)	MLC
9	44	3/Nov	ANOVA two-way	VBV
-	45	10/Nov	NO CLASS	
10	46	17/Nov	Regression analysis	VBV
11	47	24/Nov	Notions of experimental design and questions Point giving activity (in class)	MLC
12	48	1/Dec	Multiple regression	VBV+MLC

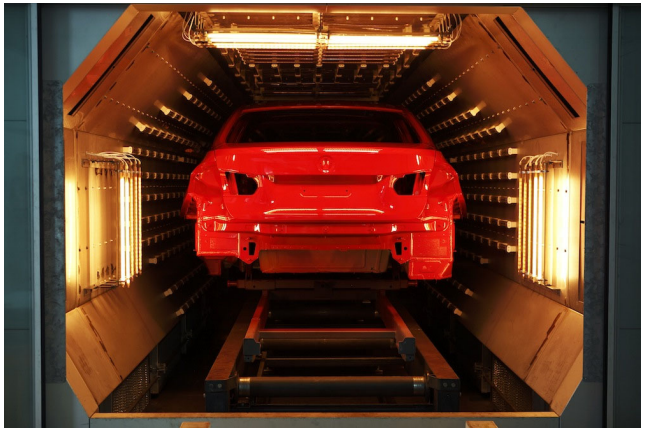
VBV = Victoria Blanes-Vidal
MLC = Manuella Lech Cantuaria

Hypothesis testing (two samples)

1. **Comparison of means (independent samples)**
2. **Paired data t-test (dependent samples)**
3. **Selecting the right statistical test**
4. **Three in-class exercises** 

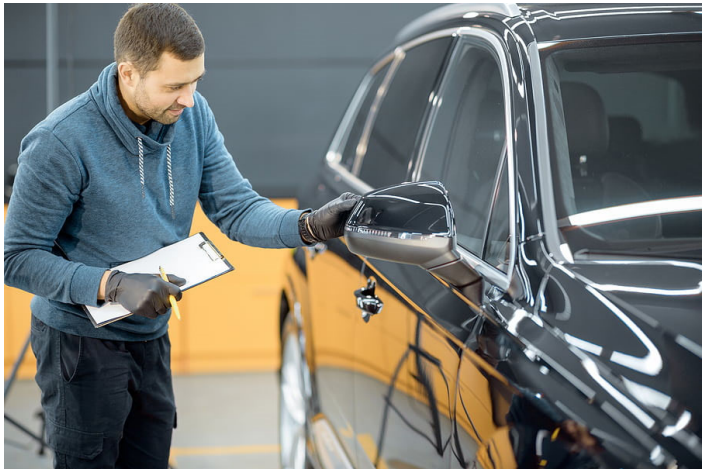
An example: Car painting oven

- The cars produced in a factory are dried in an oven after being painted.



- At the factory, they notice a problem:

  After this process (drying in the oven), there are dirt stains on the car body.



- One of the experienced engineers at the factory proposes that **keeping the car hood open in the oven might solve the problem.**

Is this measure effective?

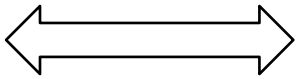
<u>We need to carry out an experiment, and test this hypothesis.</u>

- **What do we want to study?**

We want to find out whether this measure (“keeping the car hood open”) has an effect on the number of dirt stains on the cars manufactured by the factory.

That is, we want to compare the number of dirt stains in two situations (Hood closed vs. Hood open), and find out whether the number of dirt stains is significantly different or not.

Hood closed



Hood open

An example: Car painting oven

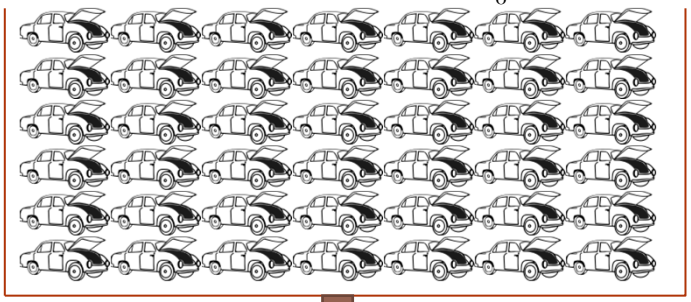
- In order to study if “keeping the car hood open” really makes a difference in the number of dirt spots on the cars, we carry out a statistical experiment:
  - We will dry 10 cars with the hood open and 10 cars with the hood closed.
  - We will then count the number of dirt stains on each of these cars.

An example: Car painting oven

Population O: All cars open hood

Mean = μ_o

Standard deviation = σ_o



Sample O: 10 cars open hood

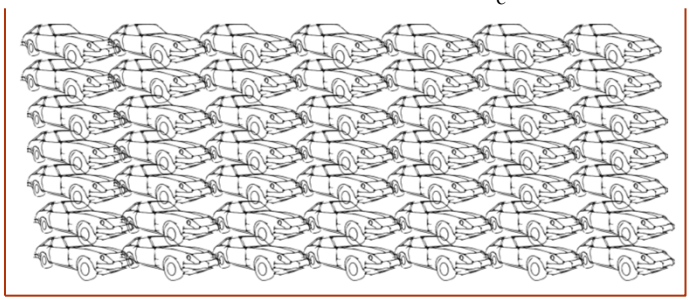
Mean = $\bar{x}_o$

Standard deviation = s_0

Population C: All cars closed hood

Mean = μ_c

Standard deviation = σ_c



Sample C: 10 cars closed hood

Mean = $\bar{x}_c$

Standard deviation = s_c

An example: Car painting oven

- The measurements we got were:

Open	Closed
3	3
3	4
2	3
2	2
1	2
2	3
3	4
5	6
4	5
3	4

Open	Closed
$\bar{x}_o = 2.80$	$\bar{x}_c = 3.60$
$s_o = 1.14$	$s_c = 1.26$

We can observe that, in the samples, the cars with open hood (sample O) had less dirt stains than the cars with closed hood (sample C).

$$\bar{x}_o < \bar{x}_c$$

But, what about, in the population?

$$\mu_o < \mu_c \ ?$$

Is there evidence that keeping the hood open really has an effect on the average number of dirty stains on the cars in the population, and so we have to change the way the cars are dried in the factory?

Hypothesis testing for comparison of means of two populations: 5 Steps

Step 1: Formulate the null hypothesis and the alternative hypothesis

Step 2: Select the statistical test

Step 3: Calculate the test statistic

Step 4: Find the critical value(s) of the Student-t distribution

Step 5: Make the decision to reject or accept the null hypothesis

Step 1:
Formulate the null hypothesis and the alternative hypothesis

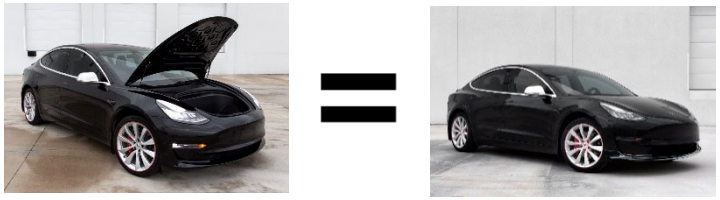
In hypothesis testing (comparison of means in two samples)…

The **null hypothesis (H0)** is a statistical hypothesis that states that there is no statistical difference between the means of the two populations.

The **alternative hypothesis (H1)**, is a statistical hypothesis that states that there is a statistical difference between the means of the two populations

Null hypothesis:

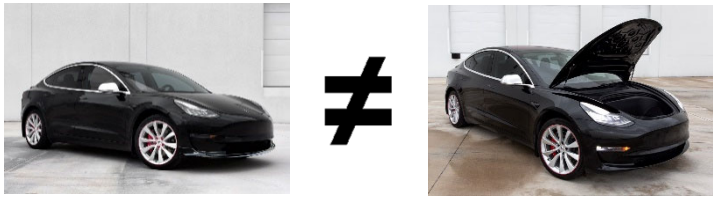
$$\mu_o = \mu_c$$



"The mean of the number of dirt stains on the cars dried with the hood open and the mean of the number of dirt stains on the cars dried with the hood closed, are NOT significantly different"

Alternative hypothesis:

$$\mu_o \neq \mu_c$$

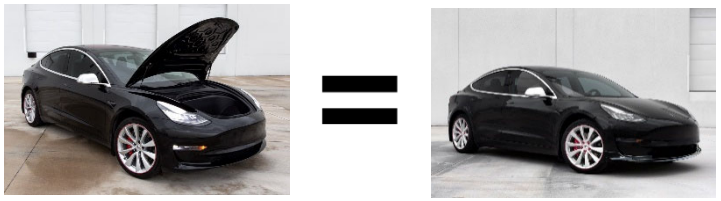


"The mean of the number of dirt stains on the cars dried with the hood open and the mean of the number of dirt stains on the cars dried with the hood closed, are significantly different"

Step 1:
Formulate the null hypothesis and the alternative hypothesis

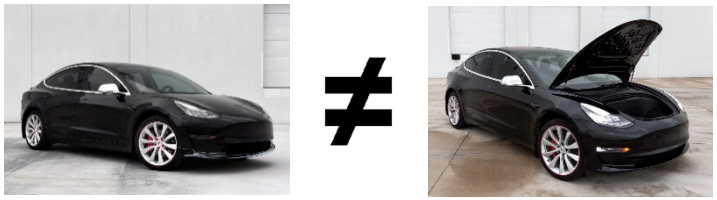
Null hypothesis:

$$\mu_o = \mu_c$$



Alternative hypothesis:

$$\mu_o \neq \mu_c$$



We will perform a **statistical test, that uses the data obtained from two samples, to make a decision about** whether the null hypothesis should be rejected or accepted.

If the test shows that the null hypothesis should be rejected, then we will accept the alternative hypothesis.

Step 2: Select the statistical test

Select the statistical test:
Student t-test for comparison of means of two independent samples

If the null hypothesis (H0) is true (if $\mu_o = \mu_c$), then,

the test statistic: $\frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_{\overline{x_1} - \overline{x_2}}}$ "follows" a Student t distribution ($t_{N1+N2-2}$).

$$\frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_{\overline{x_1} - \overline{x_2}}} \approx t_{N1+N2-2}$$

To make a decision, we compare the test statistic, with the Student-t probability distribution.

Step 3: Calculate the test statistic

Let's calculate the test statistic:

1= Closed 2= Open

$$\frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_{\overline{x_1}-\overline{x_2}}}$$

This is the difference between the means of the two populations in the null hypothesis. In this case, as the null hypothesis states: $\mu_o = \mu_c$ Then $\mu_o - \mu_c = 0$

This is the difference between the means of the two samples.

$3.60 - 2.80$

This is called "the pooled standard deviation", and is a weighted average of the two standard deviations from the two different groups.

$$s_{\overline{x}_1-\overline{x}_2} = \sqrt{\frac{(N_1-1)s_1^2 + (N_2-1)s_2^2}{N_1+N_2-2}} \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

$$s_{\overline{x}_1-\overline{x}_2} = \sqrt{\frac{(10-1)1.14^2 + (10-1)1.26^2}{10+10-2}} \sqrt{\frac{1}{10} + \frac{1}{10}}$$

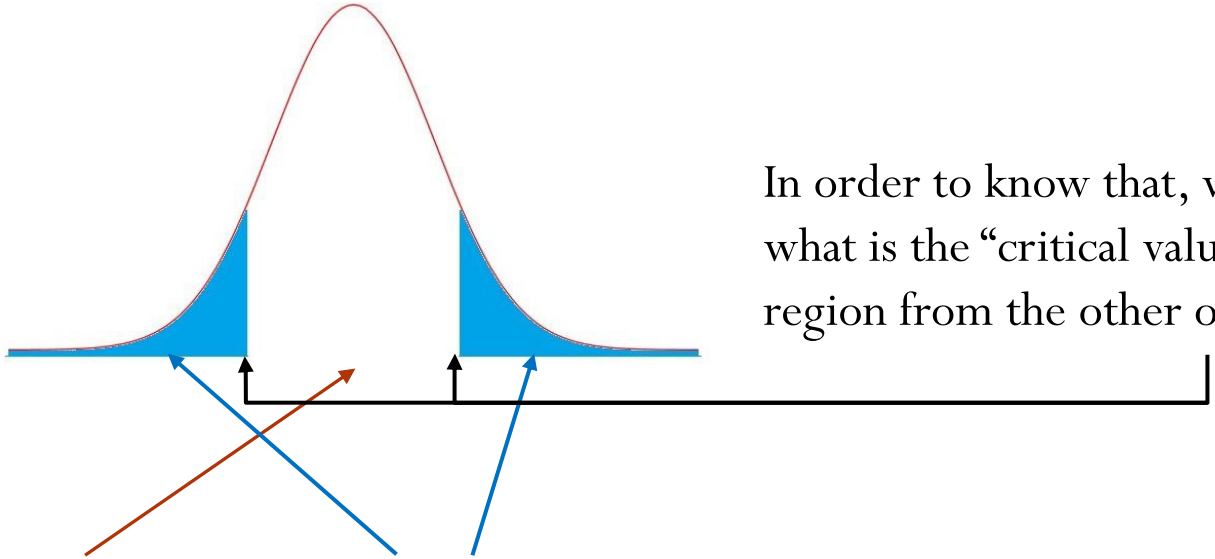
$$s_{\overline{x}_1-\overline{x}_2} = 0.56$$

Step 3: Calculate the test statistic

1= Closed 2= Open

Let's calculate the test statistic:

$$\frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_{\overline{x_1} - \overline{x_2}}} = \frac{(3.60 - 2.80) - (0)}{0.56} = 1.48$$



In order to know that, we need to find out what is the "critical value" that separates one region from the other one.

Is the number 1.48 here?
(and then we should accept the null hypothesis)

Or is it here?
(and then we should reject the null hypothesis)

Step 4: Find the critical value(s) of the Student-t distribution

In order to find the "critical value" that separates one region from the other one in the Student-t probability distribution, we need to follow 3 steps:

1. Decide on which level of significance we want to use

$$\alpha=0.05$$

2. Find out whether our test is a "two-tailed test" or a "one-tailed test":

This is a "two-tailed test".

3. Find the critical value of the Student t distribution:

$$t_{N1+N2-2}$$

Step 4: Find the critical value(s) of the Student-t distribution

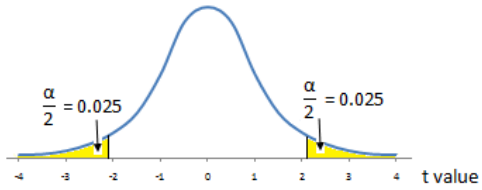
3. Find the critical value of the Student t distribution, with:

α=0.05 (confidence level = 95%)

d.f. = $N_1+N_2-2 = 10 + 10 - 2 = 18$

Student's t Distribution Table

For example, the t value for 18 degrees of freedom is 2.101 for 95% confidence interval (**2-Tail** $\alpha = 0.05$).



	90%	95%	97.5%	99%	99.5%	99.95%	1-Tail Confidence Level
	80%	90%	95%	98%	99%	99.9%	2-Tail Confidence Level
	0.100	0.050	0.025	0.010	0.005	0.0005	1-Tail Alpha
df	0.20	0.10	0.05	0.02	0.01	0.001	2-Tail Alpha
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192	
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991	
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240	
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103	
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688	
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588	
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079	
8	1.3968	1.8595	2.3060	2.8965	3.3554	5.0413	
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809	
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869	
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370	
12	1.3562	1.7823	2.1788	2.6810	3.0545	4.3178	
13	1.3502	1.7709	2.1604	2.6503	3.0123	4.2208	
14	1.3450	1.7613	2.1448	2.6245	2.9768	4.1405	
15	1.3406	1.7531	2.1314	2.6025	2.9467	4.0728	
16	1.3368	1.7459	2.1199	2.5835	2.9208	4.0150	
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.9651	
18	1.3304	1.7341	2.1009	2.5524	2.8784	3.9216	
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.8834	

$$t_{table\ N1+N2-2} = t_{table\ 20-2}(95\%) = t_{table\ 18}(95\%) = 2.10$$

Step 4: Find the critical value(s) of the Student-t distribution

Let's go back to the statistical test (Student t-test for comparison of means of two independent samples):

If the null hypothesis (H0) is true (if $\mu_o = \mu_c$), then,

the test statistic: $\frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_{\overline{x_1} - \overline{x_2}}}$ "follows" a Student t distribution ($t_{N1+N2-2}$).

$$\frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_{\overline{x_1} - \overline{x_2}}} \approx t_{N1+N2-2}$$

Test statistic = 1.48
(also called $t_{calculated}$)

Critical value = 2.10
(also called t_{table})

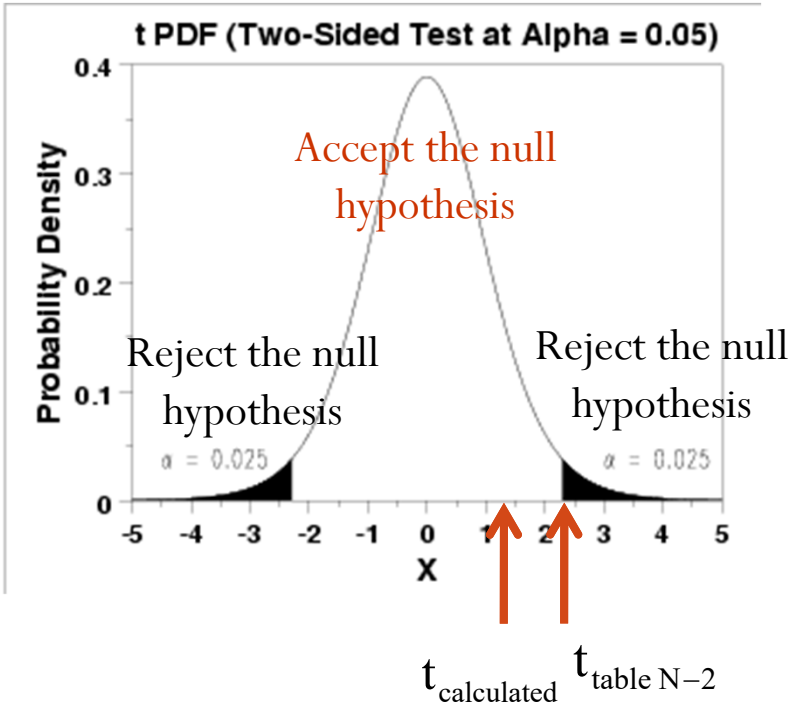
Step 5: Make the decision to reject or accept the null hypothesis

$$t_{calculated} = 1.48$$

$$t_{table\ N1+N2-2}(95\%) = 2.10$$

$$-t_{tableN-2} < t_{calculated} < t_{tableN-2}$$

$$-2.10 < 1.48 < 2.10$$



As $-t_{tableN-2} < t_{calculated} < t_{tableN-2}$,

We ACCEPT the null hipothesis: $\mu_o = \mu_c$

We have NOT demostrated that keeping the hood open decreases the number of dirt stains on the cars manufactured in the factory.

Hypothesis testing (two samples)

1. Comparison of means (independent samples)
2. **Paired data t-test (dependent samples)**
3. **Selecting the right statistical test**
4. **Three in-class exercises** 

Paired data t-test (Student t- test for dependent samples)

- In the previous example, the Student t test was used to compare two sample means when the samples were **independent**.
- In this section, a different version of the t test is explained. This version is used when the samples are dependent.
- Samples are considered to be **dependent samples** when the subjects are paired or matched in some way:
  - When the same subjects are tested in a pre/post situation
  - When we can match the subjects using a variable relevant to the study.

Paired data test (Student t-test for dependent samples)

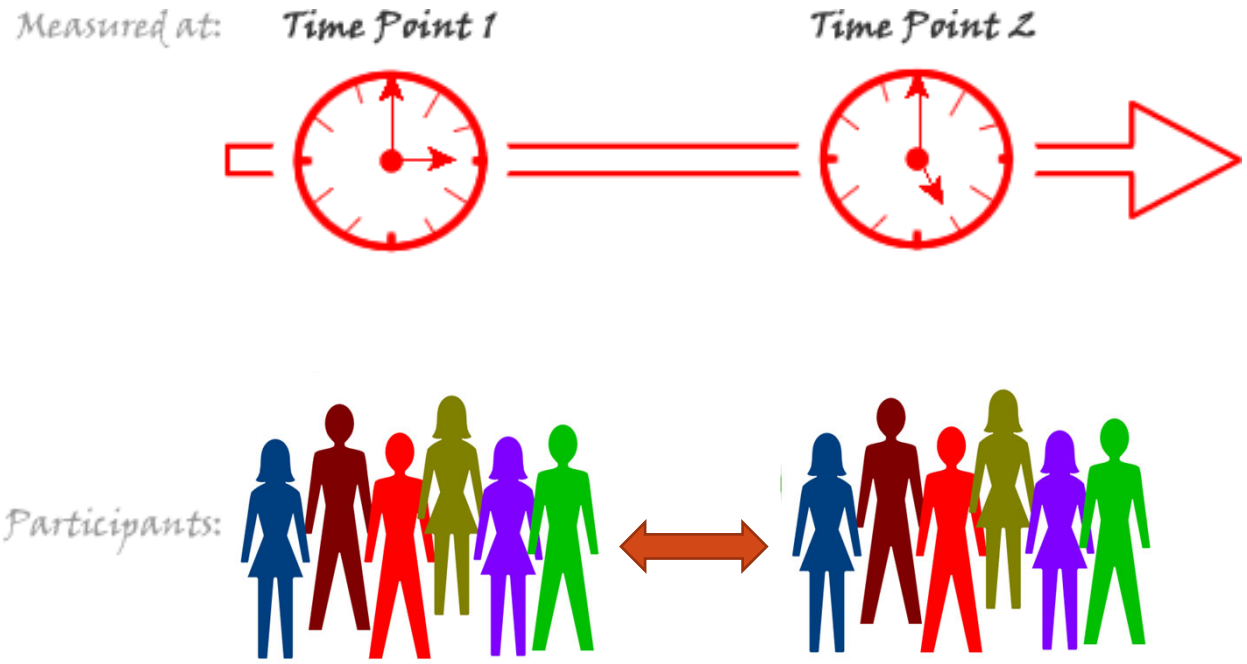
An example

A researcher wants to see if a vitamin included in the diet changes the cholesterol.

Six subjects were pretested at Week 0, and then they took the vitamin during 6-weeks. After the 6-weeks period, their cholesterol level was measured again.

Measured cholesterol level in mg/dL						
Subject	**1**	**2**	**3**	**4**	**5**	**6**
Week 0	210	235	208	190	172	244
Week 6	190	170	210	188	173	228

Can we conclude (with 99% confidence level) that the cholesterol level has been changed?



Since the same subjects are used in both cases, the samples are related; subjects scoring high on the pretest will generally score high on the posttest, even after consuming the vitamin.

Likewise, those scoring lower on the pretest will tend to score lower on the posttest.

Therefore, if we want to compare these two populations, we have to take into account that these samples are not independent, they are **dependent samples**.

Paired data test (Student t- test for dependent samples)

- To take this effect into account, we should use another type of t test (Student t- test for dependent samples, also called Paired t-test).

- The Paired t-test uses the differences between the pre-test values and the post-test values. Thus only the gain or loss in values is compared.

- Matching helps to reduce type II error by eliminating the effect of other variables that may affect the outcome.

(Remider: Type II error occurs when accept the null hypothesis, when it is false)

		Truth	
		H_o True	H_o False
Your Findings	H_o True	Correct	Type II Error **(β)**
	H_o False	**Type I Error (α)**	Correct

Step 1:
Formulate the null hypothesis and the alternative hypothesis

Null hypothesis: $\overline{(\mu_1 - \mu_2)} = 0$ **Alternative hypothesis:** $\overline{(\mu_1 - \mu_2)} \neq 0$

Step 2: Select the statistical test

Paired t-Test for Comparing means of two populations (dependent samples)

If the null hypothesis (H0) is true (if $\overline{(\mu_1 - \mu_2)} = 0$), then,

the test statistic: $\dfrac{\overline{(x_1 - x_2)} - \overline{(\mu_1 - \mu_2)}}{s_{x_1 - x_2} / \sqrt{N}}$ "follows" a Student t distribution (t_{N-1}),

Where:

$$s_{x_1 - x_2} = \sqrt{\frac{\sum (x_1 - x_2)^2 - \dfrac{(\sum (x_1 - x_2))^2}{N}}{N - 1}}$$

Step 3: Calculate the test statistic

Let's calculate the test statistic:

x_1	x_2	x_1-x_2	$(x_1-x_2)^2$
210	190	20	400
235	170	65	4225
208	210	-2	4
190	188	2	4
172	173	-1	1
244	228	16	256
		$\sum(x_1-x_2)=100$	$\sum(x_1-x_2)^2=4890$

$$\overline{(x_1 - x_2)} = \frac{100}{6} = 16.7$$

In the null hypothesis:

$$\overline{(\mu_1 - \mu_2)} = 0$$

16.7 0

$$\frac{\overline{(x_1 - x_2)} - \overline{(\mu_1 - \mu_2)}}{s_{x_1 - x_2} / \sqrt{N}} = \frac{16.7 - 0}{25.4 / \sqrt{6}} = 1.610$$

6

4890 100

$$s_{x_1 - x_2} = \sqrt{\frac{\sum(x_1 - x_2)^2 - \frac{(\sum(x_1 - x_2))^2}{N}}{N - 1}} = \sqrt{\frac{4890 - \frac{(100)^2}{6}}{6 - 1}} = 25.4$$

Step 4: Find the critical value(s) of the Student t distribution

$$t_{N-1}(99\%) = t_5(99\%) = 4.03$$

Step 5: Make the decision to reject or accept the null hypothesis

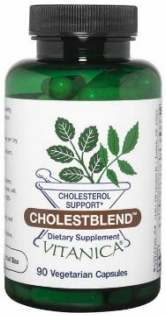
$$\frac{\overline{(x_1 - x_2)} - (\overline{\mu_1 - \mu_2})}{s_{x_1 - x_2} / \sqrt{N}} = 1.61$$

$$t_{N-1}(99\%) = t_5(99\%) = 4.03$$

} -4.03<1.610<4.03

As $-t_{tableN-1} < t_{calculated} < t_{tableN-1}$,

We will accept the null hipothesis: $\overline{(\mu_1 - \mu_2)} = 0$



We have NOT demostrated that including this vitamin in the diet changes the cholesterol level.

Hypothesis testing (two samples)

1. Comparison of means (independent samples)
2. Paired data t-test (dependent samples)
3. **Selecting the right statistical test**
4. **Three in-class exercises** 

Selecting the right statistical test

In general, the use of statistics, typically involves 3 main steps:

1. To understand the problem you are trying to solve, and find out which statistical calculation/test is appropriate to your problem.

2. To carry out the statistical calculation/test.

3. To correctly interpret and communicate the results of your statistical test.

Hypothesis testing (two samples)

1. Comparison of means (independent samples)
2. Paired data t-test (dependent samples)
3. **Selecting the right statistical test**
4. **Three in-class exercises** 

In-class exercises

In-class exercise:
Selecting the right statistical test

You have learned, so far:

1. Descriptive statistics (Lesson 1 and 2)
2. Hypothesis testing: comparison of mean of one sample (Lesson 4)
3. Hypothesis testing: comparison of means of two independent samples (Lesson 5)
4. Hypothesis testing: comparison of means of two dependent samples (Lesson 5)

You will find next, the description of 6 different problems (A, B, C, D, E and F), that can be studied using statistics. Which type of statistical method (1, 2, 3, or 4) would you use for each of them?

A

Pulse Rates of Smokers and Nonsmokers A medical researcher wishes to see whether the pulse rates of smokers are higher than the pulse rates of nonsmokers. Samples of 100 smokers and 100 nonsmokers are selected. The results are shown here. Can the researcher conclude, at $\alpha = 0.05$, that smokers have higher pulse rates than nonsmokers?

Smokers	**Nonsmokers**
$\bar{X}_1 = 90$	$\bar{X}_2 = 88$
$s_1 = 5$	$s_2 = 6$
$n_1 = 100$	$n_2 = 100$

B

Traffic Fatalities These data represent the number of traffic fatalities for two specific years for 27 selected states. Find the (*a*) mean, (*b*) median, (*c*) mode, and (*d*) midrange for each data set.

Year 1			Year 2		
1113	1488	868	1100	260	205
1031	262	1109	970	1430	300
4192	1586	215	4040	460	350
645	527	254	620	480	485
121	442	313	125	405	85
2805	444	485	2805	690	1430
900	653	170	1555	1160	70
74	1480	69	180	3360	325
158	3181	326	875	705	145

C

Hotel Room Cost A survey claims that the average cost of a hotel room in Atlanta is $69.21. To test the claim, a researcher selects a sample of 30 hotel rooms and finds that the average cost is $68.43. The standard deviation of the population is $3.72. At $\alpha = 0.05$, is there enough evidence to reject the claim?

Source: *USA TODAY.*

D

Hurricane Damage Find the percentile rank for each value in the data set. The data represent the values in billions of dollars of the damage of 10 hurricanes.

1.1, 1.7, 1.9, 2.1, 2.2, 2.5, 3.3, 6.2, 6.8, 20.3

Source: Insurance Services Office.

E

Tennis Fans The Tennis Industry Association stated that the average age of a tennis fan is 32 years. To test the claim, a researcher selected a random sample of 18 tennis fans and found that the mean of their ages was 31.3 years and the standard deviation was 2.8 years. At $\alpha = 0.05$ does it appear that the average age is lower than what was stated by the Tennis Industry Association?

F

Improving Study Habits As an aid for improving students' study habits, nine students were randomly selected to attend a seminar on the importance of education in life. The table shows the number of hours each student studied per week before and after the seminar. At $\alpha = 0.10$, did attending the seminar increase the number of hours the students studied per week?

Before	9	12	6	15	3	18	10	13	7
After	9	17	9	20	2	21	15	22	6

Problem	Statistical method (1, 2, 3 or 4)
A. Pulse rate of smokers and non smokers	
B. Traffic fatalities	
C. Hotel room cost	
D. Hurricane damage	
E. Tennis fans	
F. Improving studying habits	

In-class exercise:

Question 13

In a hypothesis test for comparing the means of two populations, the null hypothesis (H0) is:

(2 points)

Your answer:

- A statistical hypothesis that states that there is no statistical difference between the variances of the two populations.
- A statistical hypothesis that states that there is a statistical difference between the means of the two populations.
- A statistical hypothesis that states that there is no statistical difference between the means of the two populations.
- A statistical hypothesis that states that there is no statistical difference between the means of any of these two populations, and a certain value.

In-class exercise:

A researcher claims that people's right and left eyes equally good at reading letters displayed in a computer monitor.

Which kind of experiment could you perform to test this hypothesis and how would you analyze the data you collected?