**HYPOTHESIS TESTING.**

**Problem Statement.**

Our dataset contains data regarding the number of Autolib electric cars or “Blue cars” that are used daily in various cities in France over the month of April.

For our analysis this time, we are interested in finding out whether the mean numbers of blue cars used in two different postal codes are identical. For a more organized approach, we singled out the weekends which were represented by the values ‘6’ and ‘7’ according to the order of the days in a week. Also, we randomly picked two postal codes ‘75014’ and ‘75016’ for comparison.

We got the mean for one of the postal codes(75014) as ‘2’ to use as a comparison for the second mean that we are testing in our hypothesis. The objective of this test is to figure out whether the mean number of blue cars used in postal code 75016 is identical to that of postal code 75014 which is 2.

To clearly state our hypotheses:

1. **Null** hypothesis (H0): ***μ = 2*** (here, we are assuming that the mean number of the cars in postal code 75016 is equal to that of postal code 75014 which is 2)
2. **Alternative** hypothesis (Ha): ***μ ≠ 2***  (here, we assume that the mean number of the cars in postal code 75016 is not equal to that of postal code 75014 hence opposing the null hypothesis)

* This hypothesis is just to figure out whether Autolib, as an electric car-sharing company, needs to conduct extensive research about their clientele according to their regions or postal codes so as to know whether they should prioritize certain areas over others in regards to the provision of their blue cars.

**Data Description.**

Our original dataset contained data concerning the number of blue cars used over the month of April, for both weekdays and weekends, in various cities and postal codes.

However, our hypothesis only concerns two specific postal codes(75014 and 75016) and data collected over the weekends(represented by the values 6 and 7 according to the order of the days in a week. Hence, we had to derive data from the general dataset to create a more specific area of focus. At this point, we had around 1000 records.

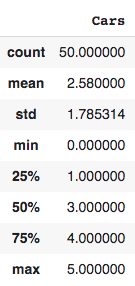
Next, we randomly sampled 50% or half of the data for our analysis, using the strata “Postal code” to minimize bias in our sampling process. The ratio of distribution of the values across both specified postal codes was generally 1:1

We grouped the values according to the two postal codes, and got the mean number of blue cars used in postal code 75014 as 2. This was in order to use it as a reference for the hypothesis testing procedure.

Now, we focus on the second mean value in postal code 75016. For our hypothesis test, we want to figure out whether the mean values of blue cars used in the two postal codes are the same.

Hence, our null hypothesis is: ***μ = 2*** ; to assume that the mean values are equal and our alternative hypothesis is: ***μ ≠ 2*** ; to oppose the null hypothesis by saying that the mean values are not equal.

Regarding the general descriptive statistics of the field ‘Cars’ for both postal codes:



**Hypothesis Testing Procedure.**

In order to conduct our hypothesis calculations, we need to find out the sample mean, sample standard deviation, and sample size of the number of blue cars used in postal code 75016.

* Sample mean(μ) = 2
* Sample standard deviation(s) = 1.9173
* Sample size(n) = 24 (number of sampled values of blue cars in postal code 75016)

Now that we have the sample mean, sample standard deviation, and sample size, we can begin our calculations.

Since our *sample size is less than 30*, we will be using the **T-test statistic** instead of the Z-test statistic to establish a generally normal distribution.

First, we need to calculate our critical value or **t-value**. In order to be able to do that, we need to identify our degree of freedom and alpha level.

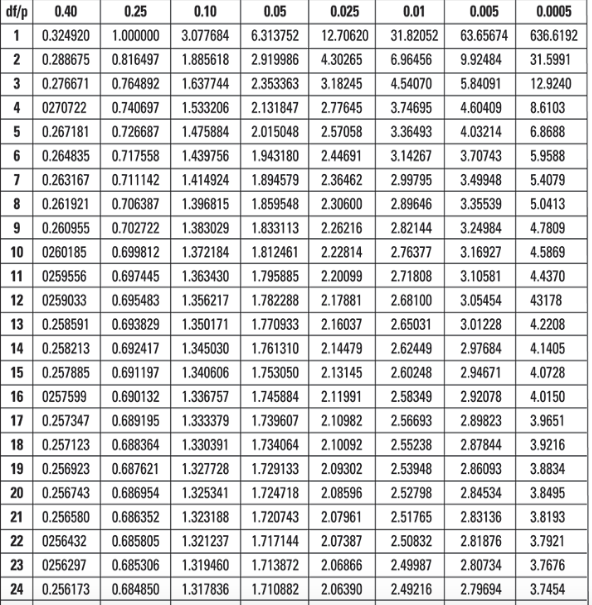
To get our **degree of freedom(df)**, we just take n(sample size) - 1 ; In this case it will be 24 - 1 = **23**

Since we are going to use a **95% confidence level** for our analysis, our **alpha/significance level** will be determined by subtracting the confidence level, 95% or 0.95, from 1; which gives us **0.05**

Looking at our Ha, the sign we used is *not equal to*, meaning that the mean number of cars could either be less or greater than 2 - this indicates that this is a **two-tailed test**. Hence, our alpha value will be divided by 2; resulting in **0.025**

To get our **critical t-value**, we need to refer to the T-distribution table where one will check the value that aligns with the df(degree of freedom) 23 and the alpha level 0.025; you get your t-value as **2.0687**.

The T-distribution table is down below:



**Hypothesis Testing Results.**

To get our results for the test, we need to incorporate the sample mean, the t-value (t) we found, sample standard deviation(s) and the sample size(n) into this formula:

**Mean ± t(s / √n)**

t(s / √n) --> **EBM**, margin of error

To substitute our values into the formula:

**Result/confidence interval** -> 2 ± 2.0687 (1.9173 / √24) = **(1.1904 , 2.8096)**

Our result is the range of values above; hence, indicating that ***we have failed to reject the null hypothesis***. (Or in other words, our null hypothesis is supported or there is not enough evidence to support our alternative hypothesis)

On to calculating our *P-value* or otherwise, the probability of our null hypothesis being correct or adequately supported by evidence...

To get the P-value, we will refer to the T-distribution table once more and use our df(degree of freedom) and our critical value or t-value to get an estimated value.

From the table, we get the value as 0.025. Since it is a two-tailed test, we have to multiply this value by 2 so as to get our **P-value** as **0.05**; which is equal to our established significance level above.

From getting our P-value equal to our predetermined significance level, we can support our null hypothesis.

**Summary.**

The hypothesis test involved data collected over weekends in the month of April concerning the number of blue cars used in two specific postal codes in Paris (75014 and 75016).

After randomly sampling our data using the strata “Postal code”, we began our calculations. First, we calculated the mean number of blue cars used in the postal code 75014 as 2; this would be in order to compare the mean number of blue cars in the postal code 75016. Hence our hypotheses are:

* **Null** hypothesis (H0): ***μ = 2*** (here, we are assuming that the mean number of the cars in postal code 75016 is equal to that of postal code 75014 which is 2)
* **Alternative** hypothesis (Ha): ***μ ≠ 2***  (here, we assume that the mean number of the cars in postal code 75016 is not equal to that of postal code 75014 hence opposing the null hypothesis)

Also, we generated the sample mean, sample standard deviation, and sample size under the postal code 75016 in order to begin our calculations. Since our sample size was < 30, we used the T-test statistic. We used our confidence level as 95% and our alpha/significance level as 5%.

Using the formula: **Mean ± t(s / √n),** we were able to calculate the confidence interval as **(1.1904, 2.8096)** where our intended value of 2 lies. Hence, *we failed to reject the null hypothesis.*

We also calculated our **P-value** from our t-value or critical value and our degree of freedom as 0.05 which is equal to our predetermined significance level hence supporting our null hypothesis.