Two-sample independent t-test

In this example we are interested in comparing the amount of time elapsed between a client request and stream availability for the company AAVAIL's streaming servers. Specifically we want to compare our locally hosted servers to a cloud service in terms of speed. The data are arrival times (in seconds) for a stream, meaning the time it takes from submission to receive a link with the modified version of the stream.

Remember to formalize your hypothesis.

1. Pose your **question** - *Is it faster, on average, to process streams for viewing on a cloud service compared to our locally hosted servers?*
2. Find the relevant **population** - *The population consists of all possible streams*
3. Specify a **null hypothesis** - *There is no difference, on average, between local and hosted services for stream processing times location after I submit my ride request.*
4. Select the test and the significance level, two-sample independent t-test with **α=0.05**

Then you will collect your data, calculate the test statistic and evaluate for significance.

```
local_arrivals = np.array([3.99, 4.15, 6.88, 4.53, 5.65, 6.75, 7.13, 2.79, 6.20,
                           3.72, 7.28, 5.23, 4.72, 1.04, 4.25, 4.71, 2.16, 3.46,
                           3.41, 7.98, 0.75, 3.64, 6.25, 6.86, 4.71])
hosted_arrivals = np.array([5.82, 4.83, 7.19, 6.98, 5.82, 5.25, 5.71, 5.59,
                            7.93, 7.09, 6.37, 6.31, 6.28, 3.12, 6.02, 4.84,
                            4.16, 6.72, 7.44, 6.28, 7.37, 4.27, 6.15, 4.88,
                            7.78])
```

The test statistic will be calculated as part of the following code block.

```
test_statistic, pvalue = stats.ttest_ind(local_arrivals, hosted_arrivals)
print("p-value: {}".format(round(pvalue,5)))
```

```
p-value: 0.0069
```

In this case we would reject the **null hypothesis** in favor of the alternative that the average times are not the same.

Unequal variances t-test

The use of a Student's t-distribution, accounts for a specific bias that the Gaussian distribution does not, because it has heavier tails.

The t-distribution always has mean 0 and variance 1, and has one parameter, the **degrees of freedom**. Smaller degrees of freedom have heavier tails, with the distribution becoming more and more normal as the degrees of freedom gets larger.

The default version of a Student's t-test assumes that the sample sizes and variances of your two samples are equal. In the case of our arrival times above we cannot state that the variances of the two samples are suppose to be the same. The unequal variances t-test, also called Welch's t-test is a more appropriate variant of the t-test for this example.

There are many variants of the t-test and depending on the field of study some have different names for the same variant. The unequal variances t-test in Python can be accessed with the equal_var keyword argument.

```
1  test_statistic, pvalue = stats.ttest_ind(local_arrivals, hosted_arrivals,
     equal_var = False)
2  print("p-value: {}".format(round(pvalue,5)))
```