**Hypothesis Testing with Python: Step by step hands-on tutorial with practical examples**

## [Ece Işık Polat](https://medium.com/@eceisikpolat)

Link : https://towardsdatascience.com/hypothesis-testing-with-python-step-by-step-hands-on-tutorial-with-practical-examples-e805975ea96e

Hypotheses are claims, and we can use statistics to prove or disprove them. At this point, *hypothesis testing* structures the problems so that we can use statistical evidence to test these claims. So we can check whether or not the claim is valid.

In this article, I want to show hypothesis testing with Python on several questions step-by-step. But before, let me explain the hypothesis testing process briefly. If you wish, you can move to the questions directly.

**1. Defining Hypotheses**

First of all, we should understand which scientific question we are looking for an answer to, and it should be formulated in the form of the *Null Hypothesis (H₀)* and the *Alternative Hypothesis (H₁ or Hₐ).*Please remember that H₀ and H₁ must be mutually exclusive, and H*₁* shouldn’t contain equality:

* **H₀:** μ=x, **H₁:** μ≠x
* **H₀:** μ≤x, **H₁:** μ>x
* **H₀:** μ≥x, **H₁:** μ<x

**2. Assumption Check**

To decide whether to use the **parametric**or **nonparametric**version of the test, we should check the specific requirements listed below:

* Observations in each sample are independent and identically distributed (IID).
* Observations in each sample are normally distributed.
* Observations in each sample have the same variance.

**3. Selecting the Proper Test**

Then we select the appropriate test to be used. When choosing the proper test, it is essential to analyze how many groups are being compared and whether the data are paired or not. To determine whether the data is matched, it is necessary to consider whether the data was collected from the same individuals. Accordingly, you can decide on the appropriate test using the chart below.

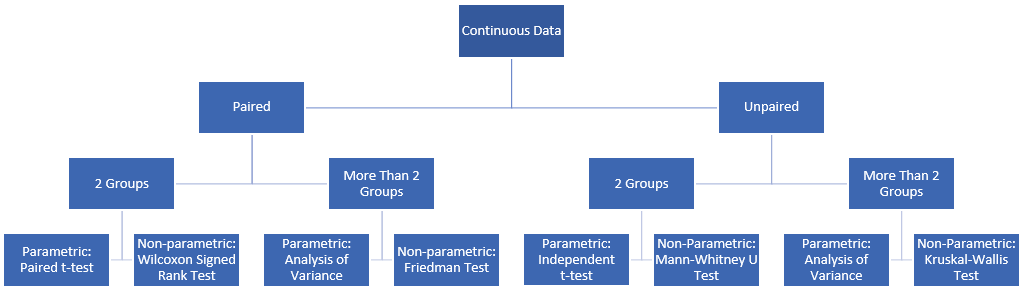


Image by the author.

**4. Decision and Conclusion**

After performing the hypothesis testing, we obtain a related ***p*-value** that shows the significance of the test.

If the *p*-value is smaller than the alpha (the significance level), in other words, there is enough evidence to prove H₀ is not valid; you can reject H₀. Otherwise, you fail to reject H₀. Please remember that rejecting H₀ validates H₁. However, failing to reject H₀ does not mean H₀ is valid, nor does it mean H₁ is wrong.

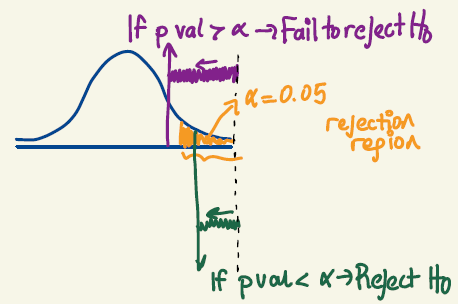


Image by the author.

Now we are ready to start the code part.

You can visit <https://github.com/eceisik/eip/blob/main/hypothesis_testing_examples.ipynb> to see the full implementation.

**Q1. t-test independent**

A university professor gave online lectures instead of face-to-face classes due to Covid-19. Later, he uploaded recorded lectures to the cloud for students who followed the course asynchronously (those who did not attend the lesson but later watched the records). However, he believes that the students who attend class at the class time and participate in the process are more successful. Therefore, he recorded the average grades of the students at the end of the semester. The data is below.

**synchronous** = [94. , 84.9, 82.6, 69.5, 80.1, 79.6, 81.4, 77.8, 81.7, 78.8, 73.2, 87.9, 87.9, 93.5, 82.3, 79.3, 78.3, 71.6, 88.6, 74.6, 74.1, 80.6]  
**asynchronous** = [77.1, 71.7, 91. , 72.2, 74.8, 85.1, 67.6, 69.9, 75.3, 71.7, 65.7, 72.6, 71.5, 78.2]

Conduct the hypothesis testing to check whether the professor’s belief is statistically significant by using a 0.05 significance level to evaluate the null and alternative hypotheses. Before doing hypothesis testing, check the related assumptions. Comment on the results.

**1. Defining Hypothesis**

Since the grades are obtained from the different individuals, the data is unpaired.

**H₀:** μₛ≤μₐ  
**H₁**: μₛ>μₐ

**2. Assumption Check**

**H₀:** The data is normally distributed.  
**H₁:** The data is not normally distributed.  
Assume that α=0.05. If the *p*-value is >0.05, it can be said that data is normally distributed.

For checking normality, I used Shapiro-Wilk’s W test which is generally preferred for smaller samples however there are other options like Kolmogorov-Smirnov and D’Agostino and Pearson’s test. Please visit <https://docs.scipy.org/doc/scipy/reference/stats.html> for more information.

p value:0.6556  
Fail to reject null hypothesis >> The data is normally distributed  
p value:0.0803  
Fail to reject null hypothesis >> The data is normally distributed

**H₀:** The variances of the samples are the same.  
**H₁:** The variances of the samples are different.

It tests the null hypothesis that the population variances are equal (called homogeneity of variance or homoscedasticity). Suppose the resulting *p*-value of Levene’s test is less than the significance level (typically 0.05). In that case, the obtained differences in sample variances are unlikely to have occurred based on random sampling from a population with equal variances.

For checking variance homogeneity, I preferred Levene’s test but you can also check Bartlett’s test from here: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bartlett.html#scipy.stats.bartlett>

p value:0.8149  
Fail to reject null hypothesis >> The variances of the samples the are same.

**3. Selecting the Proper Test**

Since assumptions are satisfied, we can perform the parametric version of the test for 2 groups and unpaired data.

p value:0.00753598  
since the hypothesis is one sided >> use p\_value/2 >> p\_value\_one\_sided:0.0038  
Reject null hypothesis

**4. Decision and Conclusion**

***At this significance level, there is enough evidence to conclude that the average grade of the students who follow the course synchronously is higher than the students who follow the course asynchronously.***

**Q2. ANOVA**

A pediatrician wants to see the effect of formula consumption on the average monthly weight gain (in gr) of babies. For this reason, she collected data from three different groups. The first group is exclusively breastfed children (receives only breast milk), the second group is children who are fed with only formula and the last group is both formula and breastfed children. These data are as below.

**only\_breast**=[794.1, 716.9, 993. , 724.7, 760.9, 908.2, 659.3 , 690.8, 768.7, 717.3 , 630.7, 729.5, 714.1, 810.3, 583.5, 679.9, 865.1]

**only\_formula**=[ 898.8, 881.2, 940.2, 966.2, 957.5, 1061.7, 1046.2, 980.4, 895.6, 919.7, 1074.1, 952.5, 796.3, 859.6, 871.1 , 1047.5, 919.1 , 1160.5, 996.9]

**both**=[976.4, 656.4, 861.2, 706.8, 718.5, 717.1, 759.8, 894.6, 867.6, 805.6, 765.4, 800.3, 789.9, 875.3, 740. , 799.4, 790.3, 795.2 , 823.6, 818.7, 926.8, 791.7, 948.3]

According to this information, conduct the hypothesis testing to check whether there is a difference between the average monthly gain of these three groups by using a 0.05 significance level. If there is a significant difference, perform further analysis to find what caused the difference. Before doing hypothesis testing, check the related assumptions.

**1. Defining Hypothesis**

**H₀:** μ₁=μ₂=μ₃ **or** The mean of the samples is the same.  
**H₁:** At least one of them is different.

**2. Assumption Check**

**H₀:** The data is normally distributed.  
**H₁:** The data is not normally distributed.

**H₀:** The variances of the samples are the same.  
**H₁:** The variances of the samples are different.

p value:0.4694  
Fail to reject null hypothesis >> The data is normally distributed  
p value:0.8879  
Fail to reject null hypothesis >> The data is normally distributed  
p value:0.7973  
Fail to reject null hypothesis >> The data is normally distributed  
p value:0.7673Fail to reject null hypothesis >> The variances of the samples are same.

**3. Selecting the Proper Test**

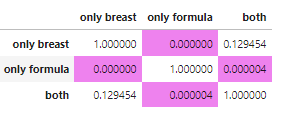
Since assumptions are satisfied, we can perform the parametric version of the test for more than 2 groups and unpaired data.

p value:0.000000  
Reject null hypothesis

**4. Decision and Conclusion**

***At this significance level, it can be concluded that at least one of the groups has a different average monthly weight gain.*** To find which group or groups cause the difference, we need to perform a posthoc test/pairwise comparison as below.

**Note:** To avoid family-wise *p*-value inflation, I used Bonferroni adjustment. You can see your other alternative from here: <https://scikit-posthocs.readthedocs.io/en/latest/generated/scikit_posthocs.posthoc_ttest/>



At this significance level, it can be concluded that:

**“only breast”** is different than **“only formula”**  
**“only formula”** is different than both **“only breast”** and **“both”**  
**“both”** is different than **“only formula”**

**Q3. Mann Whitney U**

A human resource specialist working in a technology company is interested in the overwork time of different teams. To investigate whether there is a difference between overtime of the software development team and the test team, she selected 17 employees randomly in each of the two teams and recorded their weekly average overwork time in terms of an hour. The data is below.

**test\_team**=[6.2, 7.1, 1.5, 2,3 , 2, 1.5, 6.1, 2.4, 2.3, 12.4, 1.8, 5.3, 3.1, 9.4, 2.3, 4.1]  
**developer\_team**=[2.3, 2.1, 1.4, 2.0, 8.7, 2.2, 3.1, 4.2, 3.6, 2.5, 3.1, 6.2, 12.1, 3.9, 2.2, 1.2 ,3.4]

According to this information, conduct the hypothesis testing to check whether there is a difference between the overwork time of two teams by using a 0.05 significance level. Before doing hypothesis testing, check the related assumptions.

**1. Defining Hypothesis**

**H₀:** μ₁≤μ₂  
**H₁**: μ₁>μ₂

**2. Assumption Check**

**H₀:** The data is normally distributed.  
**H₁:** The data is not normally distributed.

**H₀:** The variances of the samples are the same.  
**H₁:** The variances of the samples are different.

p value:0.0046  
Reject null hypothesis >> The data is not normally distributed  
p value:0.0005  
Reject null hypothesis >> The data is not normally distributed  
p value:0.5410  
Fail to reject null hypothesis >> The variances of the samples are same.

**3. Selecting the Proper Test**

There are two groups, and data is collected from different individuals, so it is not paired. However, the normality assumption is not satisfied; therefore, we need to use the nonparametric version of 2 group comparison for unpaired data: the Mann-Whitney U Test.

**4. Decision and Conclusion**

p-value:0.8226  
Fail to recejt null hypothesis

***At this significance level,******it can be said that there is no statistically significant difference between the average overwork time of the two teams.***

**Q4. Kruskal-Wallis**

An e-commerce company regularly advertises on YouTube, Instagram, and Facebook for its campaigns. However, the new manager was curious about if there was any difference between the number of customers attracted by these platforms. Therefore, she started to use Adjust, an application that allows you to find out where your users come from. The daily numbers reported from Adjust for each platform are as below.

**Youtube**=[1913, 1879, 1939, 2146, 2040, 2127, 2122, 2156, 2036, 1974, 1956, 2146, 2151, 1943, 2125]

**Instagram**= [2305., 2355., 2203., 2231., 2185., 2420., 2386., 2410., 2340., 2349., 2241., 2396., 2244., 2267., 2281.]

**Facebook**= [2133., 2522., 2124., 2551., 2293., 2367., 2460., 2311., 2178., 2113., 2048., 2443., 2265., 2095., 2528.]

According to this information, conduct the hypothesis testing to check whether there is a difference between the average customer acquisition of these three platforms using a 0.05 significance level. If there is a significant difference, perform further analysis to find that caused the difference. Before doing hypothesis testing, check the related assumptions.

**1. Defining Hypothesis**

**H₀:** μ₁=μ₂=μ₃ **or** The mean of the samples is the same.  
**H₁:** At least one of them is different.

**2. Assumption Check**

**H₀:** The data is normally distributed.  
**H₁:** The data is not normally distributed.

**H₀:** The variances of the samples are the same.  
**H₁:** The variances of the samples are different.

p value:0.0285  
Reject null hypothesis >> The data is not normally distributed  
p value:0.4156  
Fail to reject null hypothesis >> The data is normally distributed  
p value:0.1716  
Fail to reject null hypothesis >> The data is normally distributed  
p value:0.0012  
Reject null hypothesis >> The variances of the samples are different.

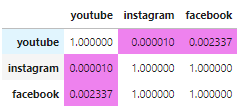
**3. Selecting the Proper Test**

The normality and variance homogeneity assumptions are not satisfied, therefore we need to use the nonparametric version of ANOVA for unpaired data (the data is collected from different sources).

**4. Decision and Conclusion**

p value:0.000015  
Reject null hypothesis

***At this significance level, at least one of the average customer acquisition number is different.***  
**Note:** Since the data is not normal, the nonparametric version of posthoc test is used.



The average number of customers coming from YouTube is different than the other (actually smaller than the others).

**Q5. t-test dependent**

The University Health Center diagnosed eighteen students with high cholesterol in the previous semester. Healthcare personnel told these patients about the dangers of high cholesterol and prescribed a diet program. One month later, the patients came for control, and their cholesterol level was reexamined. Test whether there is a difference in the cholesterol levels of the patients.

According to this information, conduct the hypothesis testing to check whether there is a decrease in the cholesterol levels of the patients after the diet by using a 0.05 significance level. Before doing hypothesis testing, check the related assumptions. Comment on the results

**test\_results\_before\_diet**=[224, 235, 223, 253, 253, 224, 244, 225, 259, 220, 242, 240, 239, 229, 276, 254, 237, 227]  
**test\_results\_after\_diet**=[198, 195, 213, 190, 246, 206, 225, 199, 214, 210, 188, 205, 200, 220, 190, 199, 191, 218]

**1. Defining Hypothesis**

**H₀:**μd>=0 **or** The true mean difference is equal to or bigger than zero.  
**H₁:**μd<0 **or** The true mean difference is smaller than zero.

**2. Assumption Check**

• The dependent variable must be continuous (interval/ratio)  
• The observations are independent of one another.  
• The dependent variable should be approximately normally distributed.

**H₀:** The data is normally distributed.  
**H₁:** The data is not normally distributed.

p value:0.1635  
Fail to reject null hypothesis >> The data is normally distributed  
p value:0.1003  
Fail to reject null hypothesis >> The data is normally distributed

**3. Selecting the Proper Test**

The data is paired since data is collected from the same individuals and assumptions are satisfied, then we can use the dependent t-test.

p value:0.000008 one tailed p value:0.000004  
Reject null hypothesis

**4. Decision and Conclusion**

***At this significance level, there is enough evidence to conclude mean cholesterol level of patients has decreased after the diet.***

**Q6. Wilcoxon signed-rank test**

GIF from giphy.com

A venture capitalist wanted to invest in a startup that provides data compression without any loss in quality, but there are two competitors: PiedPiper and EndFrame. Initially, she believed the performance of the EndFrame could be better but still wanted to test it before the investment. Then, she gave the same files to each company to compress and recorded their performance scores. The data is below.

**piedpiper**=[4.57, 4.55, 5.47, 4.67, 5.41, 5.55, 5.53, 5.63, 3.86, 3.97, 5.44, 3.93, 5.31, 5.17, 4.39, 4.28, 5.25]  
**endframe** = [4.27, 3.93, 4.01, 4.07, 3.87, 4. , 4. , 3.72, 4.16, 4.1 , 3.9 , 3.97, 4.08, 3.96, 3.96, 3.77, 4.09]

According to this information, conduct the related hypothesis testing by using a 0.05 significance level. Before doing hypothesis testing, check the related assumptions. Comment on the results.

**1. Defining Hypothesis**

Since the performance scores are obtained from the same files, the data is paired.

**H₀:**μd>=0 **or** The true mean difference is equal to or bigger than zero.  
**H₁:**μd<0 **or** The true mean difference is smaller than zero.

**2. Assumption Check**

• The dependent variable must be continuous (interval/ratio)  
• The observations are independent of one another.  
• The dependent variable should be approximately normally distributed.

**H₀:** The data is normally distributed.  
**H₁:** The data is not normally distributed.

p value:0.0304  
Reject null hypothesis >> The data is not normally distributed  
p value:0.9587  
Fail to reject null hypothesis >> The data is normally distributed

**3. Selecting the Proper Test**

The normality assumption is not satisfied; therefore, we need to use the nonparametric version of the paired test, namely the Wilcoxon Signed Rank test.

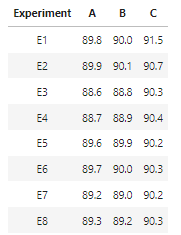
**4. Decision and Conclusion**

p-value:0.000214 >> one\_tailed\_pval:0.000107  
one sided pvalue:0.000107  
Reject null hypothesis

***At this significance level, there is enough evidence to conclude that the performance of the PiedPaper is better than the EndFrame.***

**Q7. Friedman Chi-Square**

A researcher was curious about whether there is a difference between the methodology she developed, C, and baseline methods A and B in terms of performance. Therefore, she decided to design different experiments and recorded the achieved accuracy by each method. The below table shows the achieved accuracy on test sets by each method. Please note that the same train and test sets were used for each method.



According to this information, conduct the hypothesis testing to check whether there is a difference between the performance of the methods by using a 0.05 significance level. If there is a significant difference, perform further analysis to find which one caused the difference. Before doing hypothesis testing, check the related assumptions. Comment on the results.

**1. Defining Hypothesis**

**H₀:** μ₁=μ₂=μ₃ **or** The mean of the samples is the same.  
**H₁:** At least one of them is different.

**2. Assumption Check**

**H₀:** The data is normally distributed.  
**H₁:** The data is not normally distributed.

**H₀:** The variances of the samples are the same.  
**H₁:** The variances of the samples are different.

p value:0.3076  
Fail to reject null hypothesis >> The data is normally distributed  
p value:0.0515  
Fail to reject null hypothesis >> The data is normally distributed  
p value:0.0016  
Reject null hypothesis >> The data is not normally distributed  
p value:0.1953  
Fail to reject null hypothesis >> The variances of the samples are same.

**3. Selecting the Proper Test**

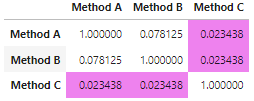
There are three groups, but the normality assumption is violated. So, we need to use the nonparametric version of ANOVA for paired data since the accuracy scores are obtained from the same test sets.

**4. Decision and Conclusion**

p value:0.0015  
Reject null hypothesis  
89.35 89.49 90.49

***At this significance level, at least one of the methods has a different performance.***

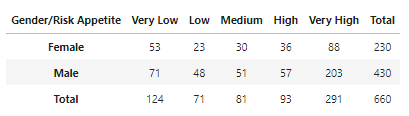
**Note:** Since the data is not normal, the nonparametric version of the posthoc test is used.



Method C outperformed others and achieved better accuracy scores than the others.

**Q8. The goodness of Fit (Bonus :)**

An analyst of a financial investment company is curious about the relationship between gender and risk appetite. A random sample was taken of 660 customers from the database. The customers in the sample were classified according to their gender and their risk appetite. The result is given in the following table.



Test the hypothesis that the risk appetite of the customers in this company is independent of their gender. Use **α = 0.01**.

**1. Defining Hypothesis**

**H₀:** Gender and risk appetite are independent.  
**H₁:** Gender and risk appetite are dependent.

**2. Selecting the Proper Test and Assumption Check**

chi2 test should be used for this question. This test is known as the goodness-of-fit test. It implies that if the observed data are very close to the expected data. The assumption of this test every Ei ≥ 5 (in at least 80% of the cells) is satisfied.

expected frequencies:  
 [[ 43.21 24.74 28.23 32.41 101.41]  
 [ 80.79 46.26 52.77 60.59 189.59]]  
degrees of freedom: 4  
test stat :7.0942  
p value:0.1310

**3. Decision and Conclusion**

critical stat:13.2767

Since the p-value is larger than α=0.01 ( or calculated statistic=7.14 is smaller than the critical statistic=13.28) → Fail to Reject H₀. At this significance level, it can be concluded that gender and risk appetite are independent.

You can visit <https://github.com/eceisik/eip/blob/main/hypothesis_testing_examples.ipynb> to see the full implementation.