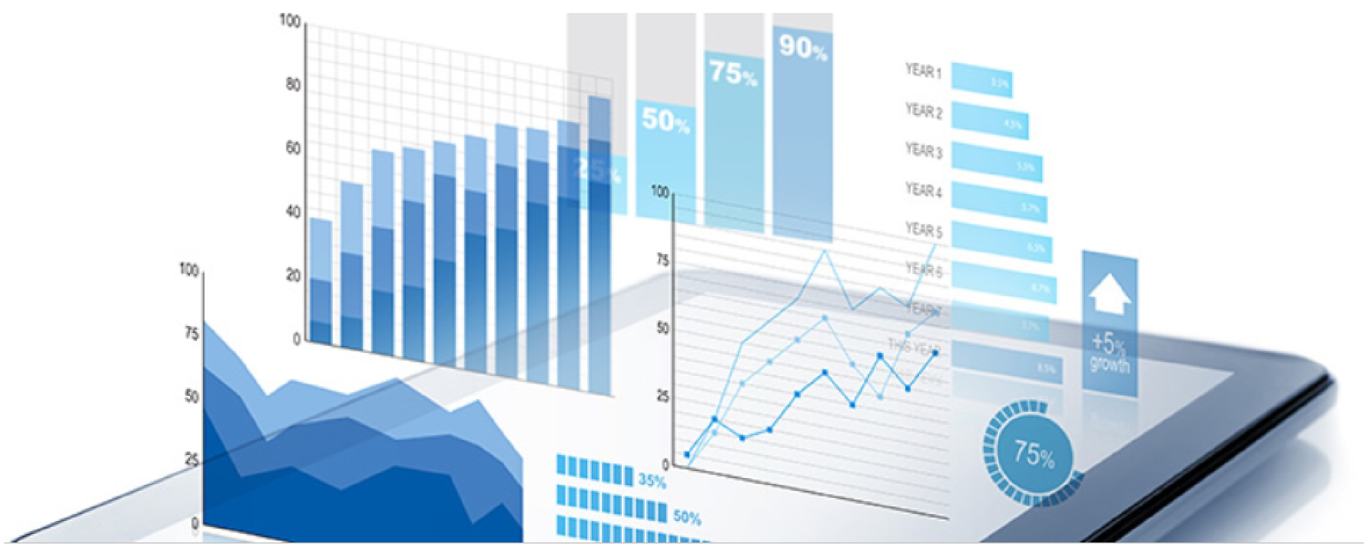
Demystifying hypothesis testing with simple Python examples

Hypothesis testing is the bread and butter of inferential statistics and a critical skill in the repertoire of a data scientist. We demonstrate the concept with very simple Python scripts.

Tirthajyoti Sarkar Follow

Sep 6 · 11 min read ★



Introduction

Hypothesis testing is a critical tool in inferential statistics, for determining what the value of a ***population parameter*** could be. We often draw this conclusion based on a ***sample data analysis***.

Not sure what exactly 'population' and 'sample' means in statistics? See this article.

What is a Population in Statistics?

Statistics Definitions > What is a Population in Statistics? What is a Population in Statistics? In stats, a sample is...

www.statisticshowto.datasciencecentral.com

With the advent of data-driven decision making in business, science, technology, social, and political undertakings, the concept of hypothesis testing has become critically important to understand and apply in the right context.

There are a plethora of tests, used in statistical analysis, for this purpose. It can get confusing. See this excellent article for a comprehensive overview of which test to use in what situation.

Statistical Tests — When to use Which ?

For a person being from a non-statistical background the most confusing aspect of statistics, are always the...

towardsdatascience.com

The basis of hypothesis testing has two attributes: (a) **Null Hypothesis** and (b) **Alternative Hypothesis.**

The null hypothesis is, in general, the **boring stuff** i.e. it assumes that nothing interesting happens/happened.

The alternative hypothesis is, **where the action is** i.e. some observation/phenomenon is real (i.e. not a fluke) and statistical analysis will give us more insights on that.

Check this document for a quick and comprehensive guide on the topic.

What’s the process?

Statisticians take a pessimistic sort of view and start with the Null hypothesis, and compute some sort of test-statistic in the sample data. It is given by,

Here, the ‘*best estimate*’ comes from the sample e.g. sample mean or proportion of some data in the sample. Standard error represents the variability in this estimate and often depends on the variance and sample size.

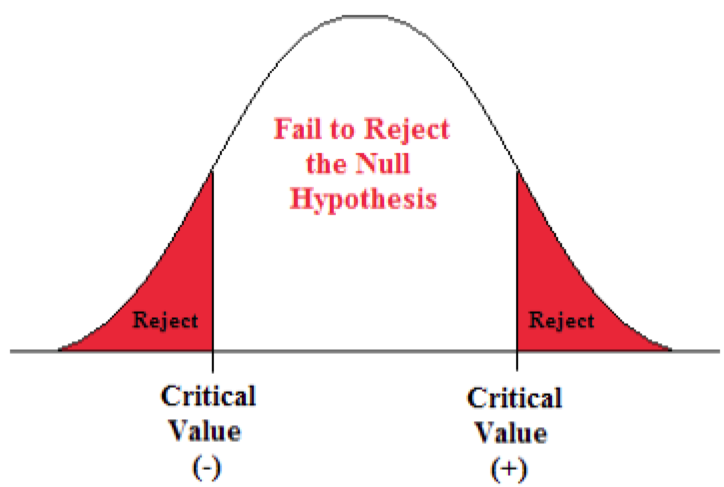
Then they ask this simple question,

> “What is the chance of observing the test-statistic, this extreme, for this sample (considering its size and a suitable probabilistic dynamics governing it),

purely randomly aka if the Null hypothesis were true?”

This chance — probability value of observing the test-statistic — is the so-called **p-value**.

Therefore, the p-value is the probability of observing the test-statistic, as is, given the Null hypothesis is true. And this probability is calculated under the assumption of a certain probability distribution (that the test statistic is generated from). Here is the idea,



If this particular value is very small (less than a pre-determined **Critical Value**), we can reject the Null hypothesis.

Note, in some situations, we have to use both sides of the probability distribution (as shown in red above). That is called ‘two-sided’ or ‘two-tailed’ test. In these situations, the alternative hypothesis is generally expressed in the form “***x is not equal to y***”.

In other situations, the hypothesis deals with questions in the form of “***x is greater than y***” or “***x is lesser than y***”. In those cases, only one side of the probability distributions have to be checked for and we call them ‘one-sided’ or ‘one-tailed’ test.

Source: https://www.fromthegenesis.com/why-hypothesis-testing/

In effect, ***we have injected sufficient doubt in the mind of the observer (ourselves) about the validity of our base assumption*** — that Null hypothesis is true.

Here is a good article summarizing p-value and its uses.

Everything you Should Know about p-value from Scratch for Data Science

What is p-value? Where is it used in data science? And how can we calculate it? We answer all these questions in this…

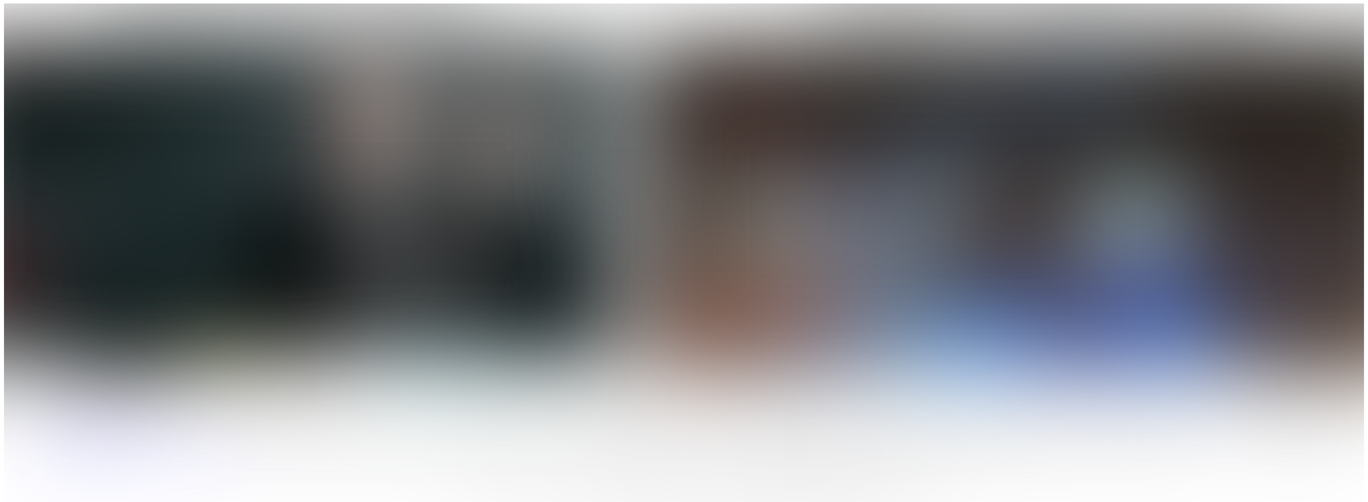
medium.com

And, here is the hypothesis testing flow, summarized, There is a lot of lively and not-so-flattering discussion about this p-value approach but this has worked for a long time reasonably well. So, we will follow this for this article. However, keep an open mind and look for other approaches as well.

One quick example of hypothesis testing?

Do you watch movies or TV series on Netflix? Then, you have surely been subjected to a form of hypothesis testing called A/B testing.

Netflix shows the same show, differently designed, to different user groups. Responses of the users (click/no-click/browse/no-browse/comes-back-to-watch-or-not) are recorded and analyzed using the good old hypothesis testing method.



Hypothesis testing with Python

One of the most important factors driving Python's popularity as a statistical modeling language is its widespread use as the language of choice in data science and machine learning.

Today, there's a huge demand for data science expertise as more and more businesses apply it within their operations. Python offers the right mix of power, versatility, and support from its community to lead the way.

While Python is most popular for data wrangling, visualization, general machine learning, deep learning and associated linear algebra (tensor and matrix operations), and web integration, **its statistical modeling abilities are far less advertised**.

A large percentage of data scientists still use other special statistical languages such as R, MATLAB, or SAS over Python for their modeling and analysis. While each of these alternatives offers their own unique blend of features and power for statistical analyses, it's useful for an up-and-coming data scientist to know more about various Python frameworks and methods that can be used for routine operations of descriptive and inferential statistics.

See this article for a comprehensive discussion about how to get started with statistical modeling with Python.

Statistical Modeling with Python: How-to & Top Libraries - Kite Blog

One of the most important factors driving Python's popularity as a statistical modeling language is its widespread use...

kite.com

In this article, we show four simple examples of hypothesis testing with Python (using functions from the **Statsmodels package**). These four situations appear in a large fraction of statistical analyses,

- One Population Proportion
- A difference in Population Proportions
- One Population Mean
- A difference in Population Means

Code examples of real-life situations

The whole **Jupyter notebook is here**. Individual cases are illustrated with real-life examples. Here they are,

One Population Proportion

Research question: In previous years, 52% of parents believed that electronics and social media was the cause of their teenager's lack of sleep. *Do more parents today believe that their teenager's lack of sleep is caused due to electronics and social media*?

Population: Parents with a teenager (age 13–18)
Parameter of Interest: p (proportion)

Null Hypothesis: $p = 0.52$
Alternative Hypothesis: $p > 0.52$ (note that this is a one-sided test)

Data: 1018 people were surveyed. 56% of those who were surveyed believe that their teenager's lack of sleep is caused due to electronics and social media.

Approach: Single group proportion uses **z-statistic test**. We use the `proportions_ztest()` function from the Statsmodels package. Note the argument `alternative="larger"` indicating a one-sided test. The function returns two values - the z-statistic and the corresponding p-value.

```
import statsmodels.api as sm
import numpy as np
import matplotlib.pyplot as plt

n = 1018
pnull = .52
phat = .56

sm.stats.proportions_ztest(phat * n, n, pnull, alternative='larger')

>> (2.571067795759113, 0.005069273865860533)
```

Conclusion: Since the calculated p-value (~0.005) of the z-test is pretty small, we **can reject** the Null hypothesis that *the percentage of parents, who believe that their teenager's lack of sleep is caused due to electronics and social media, is as same as previous years' estimate i.e. 52%*. Although we do not accept the alternative hypothesis, this informally means that there is a good chance of this proportion being more than 52%.

A difference in Population Proportions

Source: http://www.ymcamidtn.org/swim-lessons

Research question: *Is there a significant difference between the population proportions of parents of black children and parents of Hispanic children who report that their child has had some swimming lessons*?

Populations: All parents of black children age 6–18 and all parents of Hispanic children age 6–18
Parameter of Interest: p1 — p2, where p1 = black and p2 = Hispanic

Null Hypothesis: p1 — p2 = 0
Alternative Hypothesis: p1 — p2 ≠ 0

Data: 247 Parents of Black Children. 36.8% of parents report that their child has had some swimming lessons. 308 Parents of Hispanic Children. 38.9% of parents report that their child has had some swimming lessons.

Approach: The difference in population proportion needs **t-test**. Also, the population follows a binomial distribution here. We can just pass on the two population quantities with the appropriate binomial distribution parameters to the t-test function.

We can use the `ttest_ind()` function from Statsmodels. The function returns three values: (a) test statistic, (b) p-value of the t-test, and (c) degrees of freedom used in the t-test.

```
n1 = 247
p1 = .37
```

```
n2 = 308
p2 = .39

population1 = np.random.binomial(1, p1, n1)
population2 = np.random.binomial(1, p2, n2)


sm.stats.ttest_ind(population1, population2)


>> (0.9309838177540464, 0.3522681761633615, 553.0)
```

Conclusion: Since the p-value (~0.352) is quite high, we **cannot reject** the Null hypothesis *that the difference in the population proportions is zero*.

But what happens if we could sample 5000 parents instead of only a few hundreds? Check the Notebook to see how the conclusion changes.

One Population Mean

Research Question: Let’s say a cartwheeling competition was organized for some adults. The data looks like following,

(80.57, 98.96, 85.28, 83.83, 69.94, 89.59, 91.09, 66.25, 91.21, 82.7 , 73.54, 81.99, 54.01, 82.89, 75.88, 98.32, 107.2 , 85.53, 79.08, 84.3 , 89.32, 86.35, 78.98, 92.26, 87.01)

Is the average cartwheel distance (in inches) for adults more than 80 inches?

Population: All adults
Parameter of Interest: μ, the population mean cartwheel distance.

Null Hypothesis: $\mu = 80$
Alternative Hypothesis: $\mu > 80$

Data:
25 adult participants.
μ=83.84
σ=10.72

Approach: We use the z-test from Statsmodels package with `alternate='larger'` as an argument to denote the one-tailed test.

```
sm.stats.ztest(cwdata, value = 80, alternative = "larger")

>> (1.756973189172546, 0.039461189601168366)
```

Conclusion: Since the p-value (0.0394) is lower than the standard confidence level 0.05, we **can reject** the Null hypothesis *that the mean cartwheel distance for adults (a population quantity) is equal to 80 inches*. There is strong evidence in support for the alternative hypothesis that the mean cartwheel distance is, in fact, higher than 80 inches.

Visual test of normality: For single population mean test with z-statistic, the fundamental assumption is the Normality of the data. We should check that always.

The difference in Population Means

Source: https://www.shdoonline.org/blog-5/

Research Question: *Considering adults in the NHANES data, do males have a significantly higher mean Body Mass Index than females*?

Perspective: The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations.

Population: Adults in the NHANES data.
Parameter of Interest: $\mu1-\mu2\mu1-\mu2$, Body Mass Index.

Null Hypothesis: $\mu1=\mu2$
Alternative Hypothesis: $\mu1\neq\mu2$

Data:

2976 Females, $\mu1=29.94$, $\sigma1=7.75$
2759 Male Adults, $\mu2=28.78$, $\sigma2=6.25$

$\mu1-\mu2=1.16$

Approach: We can again use the z-statistic for this hypothesis testing. But here the test has to be "two-sided" as an inequality appears in the alternative hypothesis i.e. the BMI can be either higher or lower for males than females. Both side probabilities have to be added for p-value calculation.

```
url =
"https://raw.githubusercontent.com/kshedden/statswpy/master/NHANES/me
rged/nhanes_2015_2016.csv"
da = pd.read_csv(url)

females = da[da["RIAGENDR"] == 2]
male = da[da["RIAGENDR"] == 1]

sm.stats.ztest(females["BMXBMI"].dropna(),
male["BMXBMI"].dropna(),alternative='two-sided')

>> (6.1755933531383205, 6.591544431126401e-10)
```

Conclusion: Since the p-value (6.59e-10) is extremely small, we **can reject** the Null hypothesis *that the mean BMI of males is the same as that of females*.

Takeaways

There are some simple takeaways from these examples,

- It may not be too difficult to **cast your statistical inference problem into one of the set categories**. If you are doing a clinical drug trial or an A/B testing on a website design choice, you may be comparing two population means. If you are trying to judge the effect of some manufacturing process change, you may be dealing with a question of proportion.
- Once you have cast your problem, **follow a tried-and-tested procedure**. Identify your population, parameter of interest, construct a null and alternative hypothesis. Pre-determine a p-value threshold. Often it is 0.05 but suit it as per your application.
- Thereafter, **choose a test** which suits the type of inference problem you have — t-test, z-test, chi-square test, or something similar. Simple and intuitive functions exist in Python packages such as Scipy or Statsmodels. Run the test with the chosen function and **examine the p-value** and **reject/fail to reject the null hypothesis** accordingly.
- Understand the true meaning of p-value. **We never accept the alternative hypothesis**, we just show sufficient/not-sufficient evidence in favor of rejecting the null hypothesis.
- You can experiment with various sample sizes, different types of probability distributions or generative models, to examine the impact of

those on your conclusion. But the basic process remains the same.

- You may wonder what happens when you reject a null hypothesis when it was true, or vice versa. These questions relate to the **Type-I and Type-II errors**, which you may know by other names in data science parlance — **false positive** and **false negative**. Understand how they arise and what to do about it by doing experimentation with your data and samples.

Summary

In this article, we showed simple but intuitive real-life examples with inferential statistics research questions that can be answered using compact Python code.

These examples illustrate the application of the concept of hypothesis testing, which is an extremely valuable skill to master for data scientists and analytics professionals working with data to gather insights and making business/scientific decisions.

. . .

If you have any questions or ideas to share, please contact the author at **tirthajyoti[AT]gmail.com**. Also, you can check the author's **GitHub repositories** for other fun code snippets in Python, R, and machine learning resources. If you are, like me, passionate about machine learning/data science, please feel free to add me on LinkedIn or follow me on Twitter.

Tirthajyoti Sarkar - Sr. Principal Engineer - Semiconductor, AI, Machine Learning - ON...
Making data science/ML concepts easy to understand through writing: https://medium.com/@tirthajyoti Open-source and fun...
www.linkedin.com

Thanks to Amber Teng.

Data Science Technology Python Programming Analytics