Statistics for Data Science – Part IV: Hypothesis Testing

This article is the fourth and final part of a series and I will cover hypotheses testing. In the previous article, statistical inference was defined as the second major branch of statistics and also very important for the data scientist. The target was defined as making more meaningful estimates by specifying an interval of values on a number line, together with a statement of how confident you are that your interval contains the population parameter.

In this article, instead of making an estimate about a population parameter, I will stress on how to test a claim about a parameter.

You may find the first article of this series here and the second part is here.

Definitions

A hypothesis is an **educated guess** about something in the world around you. It should be testable, either by experiment or observation.

It is customary to write a statement before proposing a hypothesis. The statement will have a structure like this:

"If I…(do this to an independent variable)….then (this will happen to the dependent variable)."

As an example:

If I (decrease the amount of water given to herbs) then (the herbs will increase in size).

Wikipedia definition of a statistical hypothesis is: "A **statistical hypothesis** is a hypothesis that is testable on the basis of observed data modelled as the realized values taken by a collection of random variables."

A **hypothesis test** is a process that uses sample statistics to test a claim about the value of a population parameter. Researchers in fields such as medicine, medicine, psychology, and business rely on hypothesis testing to make informed decisions about new medicines, treatments, and marketing strategies.

Hypothesis Testing Example

As an example, let us assume that steel bolts are manufactured in a plant, and a certain hardness value is expected. The engineer hypothesized that the mean Rockwell C hardness values of all the parts are greater than 50 R_C. It will be very impractical to test all the parts if the daily production is in the order of a million parts, but it is still possible to make a reasonable decision about the mean hardness by taking a random sample from the population of bolts and testing the hardness of each. If the sample mean differs too much from the expected mean, it can be deduced that there is a problem with the production line or steel selection (or maybe measurement process).

To test that the mean hardness of all the parts is μ = 50 R_C, a random sample of n = 30 bolts could be picked up and tested. Suppose that the result is a sample mean of $\bar{x} = 47$ R_C and a sample standard deviation of s = 5.5 R_C.

Sampling Distribution of $\bar{x}$
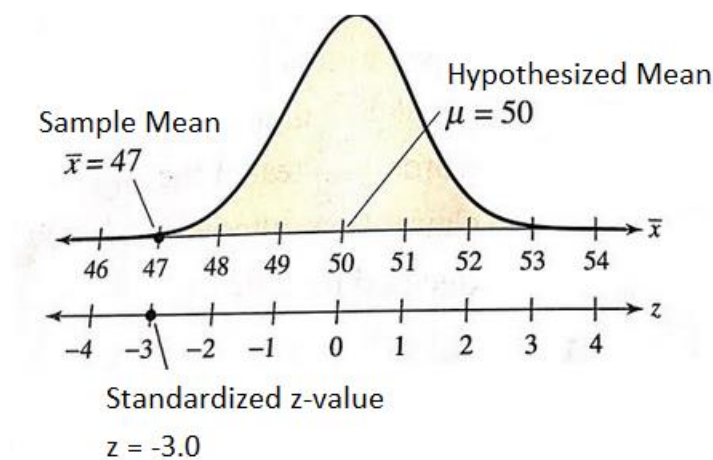
According to the graph, a sample mean of $\bar{x} = 47$ R$_C$ is highly unlikely - it is 3 z-values from the hypothesized mean. Probability of obtaining a sample mean of 47 or less is around 0.0013, which is highly unusual. Assuming that the testing equipment is working probably and well-calibrated, this was a very unusual sample, or there was a huge mistake in the production line.

Stating a Hypothesis

A statement about a population parameter is called a statistical hypothesis, which was defined above. To test a population parameter, a pair of hypotheses should be carefully stated – one that represents the claim and the other, its complement. When one of these hypotheses is false, the other must be true.

A **null hypothesis H_0** is a statistical hypothesis that contains a statement of equality, such as $\leq$, $=$ or $\geq$.

The **alternative hypothesis H_a** is the complement of the null hypothesis. It is a statement that must be true if H_0 is false and it contains a statement of strict inequality such as $>$, $\neq$ or $<$.

If the claim value is k and the population parameter is μ, then some possible pairs of null and alternative hypotheses are:

$$\begin{cases} H_0: \mu \leq k \\ H_a: \mu > k \end{cases} \qquad \begin{cases} H_0: \mu \geq k \\ H_a: \mu < k \end{cases} \qquad \begin{cases} H_0: \mu = k \\ H_a: \mu \neq k \end{cases}$$

Regardless of which of the three pairs hypotheses is used, $\mu = k$ is always assumed, and the sampling distribution is examined based on this assumption.

Types of Errors and Level of Significance

No matter which hypothesis represents the claim, the first step in beginning a hypothesis test is by assuming that the equality condition in the null hypothesis is true. So, when you perform a hypothesis test, you make one of two decisions:

1. Reject the null hypothesis or
2. Fail to reject the null hypothesis.

Because the decision is based on a sample rather than the entire population, there is always the possibility that the decision will be wrong. The only way to be absolutely certain of whether H_0 is true or false is to test the entire population. Because the decision is based on a sample, it may be incorrect.

A **type I error** occurs if the null hypothesis is rejected when it is true.

A **type II error** occurs if the null hypothesis is not rejected when it is false.

The table shows the four possible outcomes of a hypothesis test.

	H_o True	H_o False
Reject H_o	Type I Error	Correct
"Accept" H_o	Correct	Type II Error

In a hypothesis test, the **level of significance (α)** is the maximum allowable probability of making a Type I Error.

The probability of a Type II Error is denoted by **β**.

Setting the level of significance at a small value means, the target is to get the probability of rejecting a true null hypothesis to be small. Three commonly used levels of significance are α = 0.10, α = 0.05 or α = 0.01.
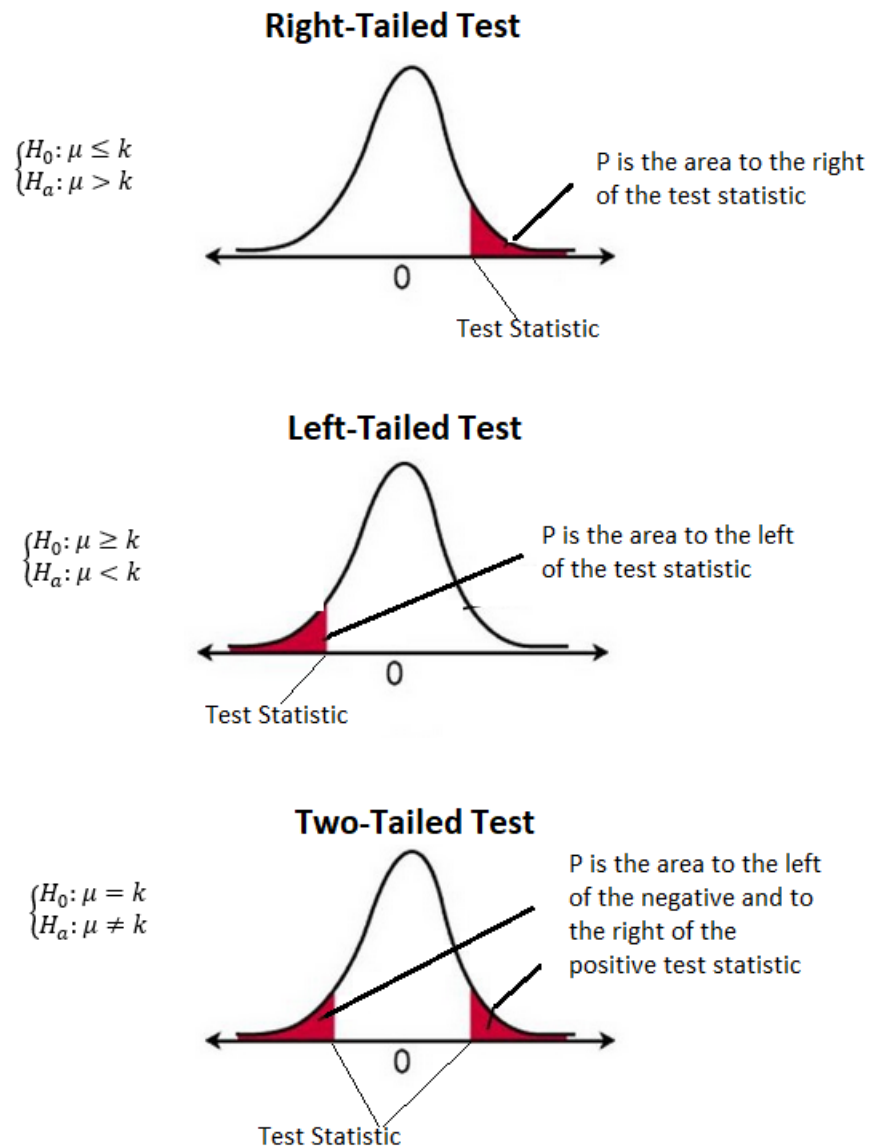

Statistical Tests and p-Values

After stating the null and alternative hypotheses and specifying the level of significance, the next step is to obtain a random sample from the population and calculate sample statistics, such as the mean and the standard deviation. The statistic that will be compared with the parameter in the null hypothesis is called the test statistic.

One way to decide whether to reject the null hypothesis is to determine whether the probability of obtaining the standardized test statistic is less than the level of significance.

If the null hypothesis is true, a **P-value** (probability value) of a hypothesis test is the probability of obtaining a sample statistic with a value as extreme or more extreme than the one determined from the sample data.

The p-value of a hypothesis test depends on the nature of the test. Basically, there are three types of hypothesis tests – a left-, right-, or two-tailed test.

Right-Tailed Test

$\begin{cases} H_0: \mu \leq k \\ H_a: \mu > k \end{cases}$

P is the area to the right of the test statistic

0

Test Statistic

Left-Tailed Test

$\begin{cases} H_0: \mu \geq k \\ H_a: \mu < k \end{cases}$

P is the area to the left of the test statistic

0

Test Statistic

Two-Tailed Test

$\begin{cases} H_0: \mu = k \\ H_a: \mu \neq k \end{cases}$

P is the area to the left of the negative and to the right of the positive test statistic

0

Test Statistic

The smaller the p-value of the test, the more evidence there is to reject the null hypothesis. A very small p-value indicates an unusual event. But, even a very low p-value does not give proof that the null hypothesis is false; only that it is probably false.
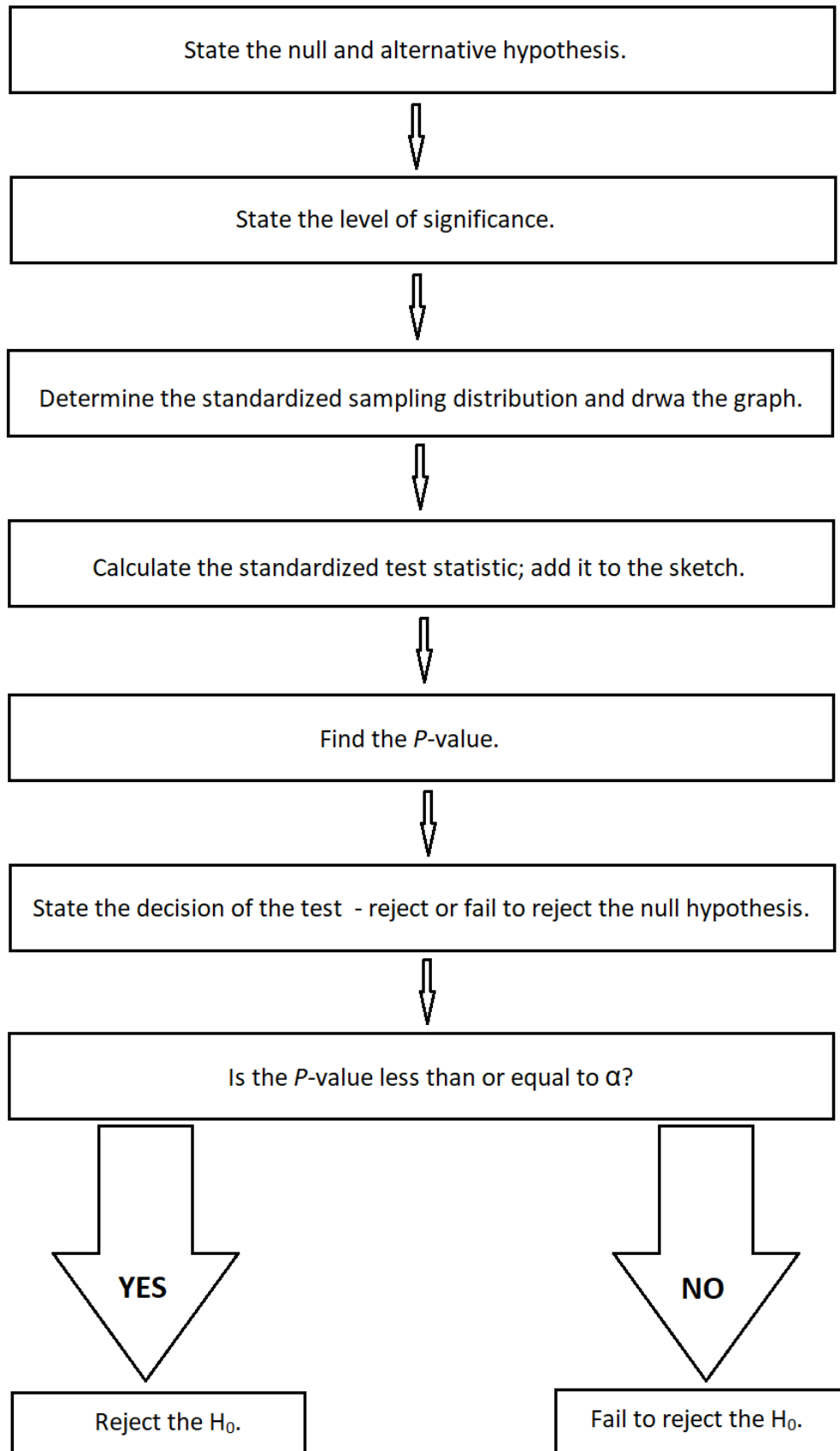
Interpreting the Decision

To make a conclusion in a hypothesis test, the P-value should be compared with the level of significance, α.

1. If $P \leq \alpha$, reject H_0.
2. If $P > \alpha$, fail to reject H_0.

Failing to reject the null hypothesis does not mean that the null hypothesis is accepted as true. It means that there is not enough evidence to reject the null hypothesis.

Steps for hypothesis testing are summarized as:

State the null and alternative hypothesis.

↓

State the level of significance.

↓

Determine the standardized sampling distribution and drwa the graph.

↓

Calculate the standardized test statistic; add it to the sketch.

↓

Find the P-value.

↓

State the decision of the test - reject or fail to reject the null hypothesis.

↓

Is the P-value less than or equal to α?

YES   NO

Reject the H_0.   Fail to reject the H_0.

I advise you to check statistics.howto.com webpage – there are very nice examples that will help clarify the theoretical knowledge.

Conclusion

Hypothesis testing is useful in many different fields because it gives a scientific procedure for assessing the validity of a claim about a population.

The entire theory of hypothesis testing is based on the fact that the sample is randomly selected. If the sample is not random, it is not possible to use it to infer anything about a population parameter.

Also, it should be kept in mind that if the *P*-value for a hypothesis test is greater than the level of significance, there is still no proof that the null hypothesis is true – only that there is not enough evidence to reject it.

Finally, type I error means rejecting a null hypothesis that is true and a type II error is failing to reject a null hypothesis that is false.

- It is possible to decrease the probability of a type I error by lowering the level of significance.
- Generally, decreasing the probability of making a type I error will increase the probability of making a type II error.
- Increasing the sample size will decrease the chance of making both types of errors.

You can access to this article and similar ones here.

Photo by [Clay Banks](#) on [Unsplash](#)