Introduction to hypothesis testing

Ellie Trace

Summary

A hypothesis test is a statistical process used to determine whether there is enough evidence in a sample to support or reject a claim about a population. It involves comparing a null hypothesis, which represents the status quo, against an alternative hypothesis using sample data, a chosen significance level, and either critical values or p-values. This guide will focus on the general structure of a hypothesis test, critical values, how to choose which type of test to use, as well as when to reject, or not reject a hypothesis.

Before reading this guide, it is recommended that you read Guide: Introduction to probability, Guide: Expected value, variance, standard deviation, and Guide: PMFs, PDFs, CDFs.

What is a hypothesis test?

Statistics is the science of collecting, organizing and interpreting data. Making statements about characteristics of entire populations is important in statistics, as it gives a complete picture of the data. However, in real life, taking data from the entire population is incredibly difficult and could be impossible; for instance, could you imagine trying to co-ordinate the entire population of the world to measure their height?

Instead, statisticians take **samples** of data and use tools to potentially extend an observation to the entire population. This leads to the idea of a **hypothesis test**, one of the most important and most useful tools in statistics. Hypothesis tests help you to use data from a sample to test whether or not it is reasonable to believe a certain statistical characteristic is true for a whole population.

For example, the flagship branch of Cantor's Confectionery have 50 customers on average one week. Can they say that their daily average customers has now probably increased compared to their previous average of 45 per day? This is the sort of scenario that a hypothesis test can assess; trying to infer a statistical characteristic from a sample size. Hypothesis testing is therefore used extensively in any field that requires statistical analysis, such as biology and ecology, psychology and neuroscience, geography, sociology, and even in medicine.

This guide will introduce the idea of a hypothesis test. First, the null and alternative hypotheses are introduced. Then, a list of tests is introduced to assess certain statistical qualities such as mean and variance. The level of confidence required to make a conclusion from the test is introduced via statistical significance, and the equivalent ideas of critical and p-values are demonstrated and a fully worked example is given.

It's important to say that this guide is intended to be an overview as to how hypothesis testing works, rather than exploring the particular details of each hypothesis test. For more on this, see our associated guides.

Setting up a hypothesis test

ℹ Hypothesis, hypothesis test

A **statistical hypothesis** (shortened to **hypothesis** in this guide) is a statement showing an idea, or something that could possibly be true, about the whole of a particular data set.
A **hypothesis test** allows you to say, with a defined level of certainty, whether or not a hypothesis can be **rejected**. This determines whether there is enough statistical evidence to show that the original hypothesis is unlikely to be true.

Generally speaking, when you define hypotheses in statistics, they follow a very specific format, and refer to the *entire population* being studied.

- **Null hypothesis (H_0):** This hypothesis represents the 'status quo' or no effect. It is always a statement of equality.
- **Alternative hypothesis (H_1):** This is the hypothesis that you are trying to test. It is always a statement of **inequality**.

  Depending on the question, the alternative hypothesis H_1 can either be **one-tailed** or **two-tailed**:

  - One-tailed means you are you are testing whether the characteristic has increased or decreased.
  - Two-tailed means you are testing whether the tested characteristic is equal to the comparative characteristic or not.

Here's some examples of hypotheses that require testing. In these examples, the Greek letter μ (mu) is used for the **population mean**; see [Guide: Means, variance, and standard deviation] for more.

Example 1

Cantor's Confectionery wants to determine whether their waiting times in their corporate headquarters are longer than their target of 15 minutes per person.
You can set up a hypothesis test to evaluate this claim.
Null hypothesis (H_0): The average waiting time is equal to the target waiting time of 15 minutes. (Remember that the null hypothesis is always statement of equality!)
Alternative hypothesis (H_1): The average waiting time is longer than the target waiting time of 15 minutes.
So the null hypothesis is $H_0 : \mu = 15$ and the alternative hypothesis is $H_1 : \mu > 15$. This is a one-tailed alternative hypothesis.

Example 2

Context: Cantor's Confectionery wants to determine if a new sweet production method produces different average approval scores compared to their traditional method.
Null hypothesis (H_0): The mean approval score for the new method n is equal to the mean approval score for the traditional method t. So the null hypothesis is $H_0 : \mu_n = \mu_t$.
Alternative hypothesis (H_1): The mean approval score for the new method n is not equal to the mean approval score for the traditional method, indicating a difference in performance. So the alternative hypothesis here is $H_1 : \mu_n \neq \mu_t$

Example 3

Context: Cantor's Confectionery wants to test if the proportion P of defective products produced is lower than last year's average of 2%.
Null hypothesis (H_0): Here, the proportion would be equal to 2%; which means your null hypothesis is $H_0 : P = 0.02$.
Alternative hypothesis (H_1): The hypothesis that you want to test is that the proportion of defective products is less than 2%; so your alternative hypothesis should be $H_1 : P < 0.02$.

Test selection and statistic calculation

Choosing the appropriate test statistic is really important because they all have different inputs depending on the data you have available to you. The following checklist should be able to help you decide which test statistic to use.

In what follows, σ is the population standard deviation and n is the sample size.

- To **compare means**
  - with **one sample**
    - if you know σ, *or* you don't know σ *and* $n > 30$, use a Z-**test**
    - if you don't know σ *and* $n \leq 30$, use a t-**test**
  - with **two samples**
    - where the samples are **independent**
      - if you know σ for both samples, *or* you don't know σ *and* $n > 30$, use a **two-sample** z-**test**
      - if you don't know σ for both samples *and* $n \leq 30$, use a **two-sample** t-**test**
    - where the samples are **paired**
      - use a **paired** t-**test**
- To **test for variance** σ^2
  - use a F-**test**
- To **test for goodness of fit**
  - use a **chi-squared goodness of fit test**
- To **test for independence**
  - use a **chi-squared test for independence**

Once you have found the test you want to use, you can refer to their relevant guide to make sure you are doing the correct calculations. Once you have the test statistic, you can proceed to the next step to decide on what level of statistical significance you would like to test against.

Significance levels

Note

A **significance level** α (alpha) is the level of certainty you want to test your hypothesis with, or the line at which you would reject the null hypothesis H_0.
This significance level is the percentage of risk that you are willing to accept rejecting a true null hypothesis. The smaller the α, the smaller the risk of a false conclusion.

Most commonly, α is set to 0.05 (5%), 0.01 (1%) or 0.10 (10%). Usually, the most common choice is $\alpha = 0.05$ (so a significance level of 5%). **This means you are willing to accept a 5% risk of rejecting a true null hypothesis.**

Tip

Rejecting a true null hypothesis is known as a **Type I error**; better known as a false positive. There is also a **Type II error** (a false negative), which is where you fail to reject the null hypothesis (resulting in a negative test) but in fact the alternative hypothesis is true. For more information on these, see [Guide: Errors in hypothesis testing].

Critical values and p-values

You can use two different but equivalent methods to decide whether or not to reject your hypothesis to the stated significance level. These methods depend on **the probability distribution being used**; see Guide: PMFs, PDFs, CDFs and [Overview: Probability distributions] for more.

Critical value

A **critical value** is a boundary that defines the **rejection region** (or **critical region**) based on the significance level α.

For a **one-tailed test**, this critical region is at one end of the probability distribution with PDF $f(x)$. The critical value is then the unique z such that

- $\displaystyle\int_{-\infty}^{z} f(x)\,\mathrm{d}x = \alpha$ for a lower tail test
- $\displaystyle\int_{z}^{\infty} f(x)\,\mathrm{d}x = \alpha$ for an upper tail test.

For a **two-tailed test**, this critical region occupies **both ends** of the probability distribution with PDF $f(x)$. The **critical values** z_1, z_2 are the two boundaries such that

$$\int_{-\infty}^{z_1} f(x)\,\mathrm{d}x = \frac{\alpha}{2} \quad \text{and} \quad \int_{z_2}^{\infty} f(x)\,\mathrm{d}x = \frac{\alpha}{2}$$

You can notice that you want $\alpha/2$ in a two-tailed test; this is because the critical region is split in half!

> Tip
>
> If your probability distribution is **symmetric** (like the normal distribution), then $z_1 = -z_2$ in a two-tailed test. If it is not symmetric (like the χ^2 distribution), then this doesn't happen.

Finding critical values for different significance levels is hard for **humans** to do, let alone any reader of this guide! You can use statistical tables or a computer to find the critical value for the chosen test and then compare your test statistic to the critical value you find.

It can be helpful to sketch or visualize the graph whilst doing this to make sure you are not missing any negatives and you are using the right critical region depending on if you are doing a one-tailed or a two-tailed test.

Once the critical value(s) and critical region has been located, you can then perform your test. This is usually expressed in the language of p-**values:**

> p-value
>
> The p-**value** is the probability of obtaining a test statistic as extreme as the observed statistic, and is defined to be to the area under the distribution curve corresponding to the test statistic.

The p-value will be different depending on the type of test used, whether it was a one-tailed or a two-tailed test.

- If you have an upper-tailed test
  $$H_1 : \mu > \mu_0$$
  (where μ_0 is the value you are testing your sample data against), then the p-value is the area to the **right** of the test statistic under the probability distribution curve.

- If you have a lower-tailed test
  $$H_1 : \mu < \mu_0$$
  the p-value is the area to the left of the test statistic under the probability distribution curve.

- In the case of a two-tailed test

$$H_1 : \mu \neq \mu_0$$

  , then the p-value corresponds to the areas beyond the critical values in both cases; you would need to **halve** p in this case.

You would then compare the p-value to α. If the p-value is less than α, **you reject** H_0. If the p-value is greater than or equal to α, **you fail to reject** H_0.

This has a different impact based on whether you are performing a one-tailed or a two-tailed test. As shown in Figure 1 and Figure 2, for a one-tailed test, that whole 5% rejection region will be at one end of the results. This is because you are only testing whether the characteristic is greater than or equal to our comparative characteristic. In the case of a two-tailed test, that 5% will be split across either end of the distribution to reject the 2.5% at both far left and far right of the curve; this is shown in Figure 3. This is because you are testing whether the test characteristic is both greater than or less than the comparative characteristic.

If your test statistic falls in the critical regions as defined by your test, that is when **you reject the null hypothesis** H_0.

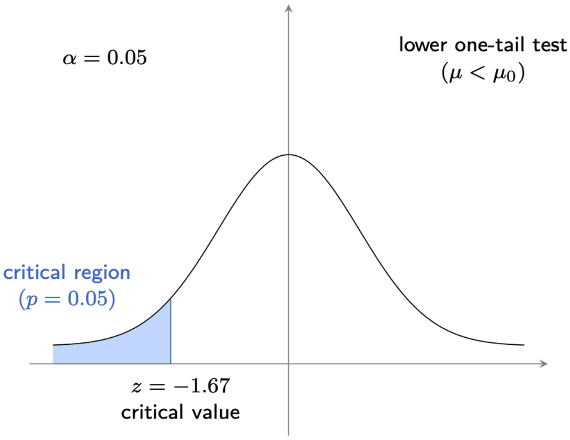

Figure 1: A **lower one-tailed probability distribution** where $\alpha = 5\%$, the critical value is -1.67 and there is a highlighted area on the left tail of the curve representing a probability area of $p = 0.05$. This highlighted area represents the critical region or area where you would reject the null hypothesis H_0

Forming a conclusion

To form a conclusion you then use the test to decide whether or not to reject your null hypothesis H_0.

If the test statistic falls in the critical region, or if the p-value is less than α, then you would **reject** H_0. This suggests there is enough evidence to support the alternative hypothesis.

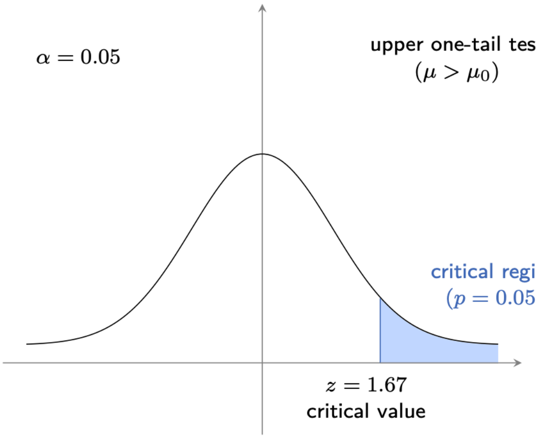

Figure 2: An **upper one-tail probability distribution** where $\alpha = 5\%$, the critical value is 1.67 and there is a highlighted area on the right tail of the curve representing a probability area of $p = 0.05$. This highlighted area represents the critical region or the area where you would reject the null hypothesis H_0.

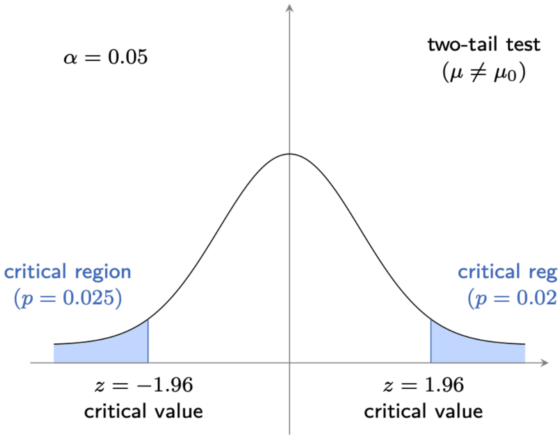

Figure 3: A **two-tailed test on the normal distribution** where $\alpha = 5\%$, the critical value is 1.96 and there is a highlighted area on both the left and right tails of the curve representing a probability area of $p = 0.025$ each. These highlighted areas represent the critical regions or the areas where you would reject the null hypothesis H_0.

If the test statistic does not fall in the critical region, or if the p-value is greater than or equal to α, this means there is not enough evidence to support the alternative hypothesis and you **fail to reject** H_0.

Tip

To complete the problem, you then must formally state your conclusion. It's important that any conclusion is robust. You could use the following templates for help.

> "I reject H_0 as there is sufficient evidence to conclude that the mean is greater than 50."

or

> "I do not reject H_0 as there is not sufficient evidence to conclude that the mean is greater than 50."

Warning

Hypothesis tests are based on sample data, not on the entire population. This means **you can never accept either hypothesis**; it is always a rejection or a failure to reject. This is because you cannot definitively 'prove' the null hypothesis to be true.

Example 4

Your local mathematical sweet shop, Cantor's Confectionery, claims that the average weight of its popular Boole Bar is 5 grams. You suspect that the actual average weight is less than 5 grams, so you decide to perform a hypothesis test. You take a random sample of 30 Boole Bars and find that the sample mean weight of the sample is $\bar{x} = 4.5$ grams with a sample standard deviation of $s = 1.5$ grams. Because of the amazing production methods of Cantor's Confectionery, you can safely assume that the population standard deviation $\sigma = 1.5$ grams.
You want to test if the average weight of the Boole Bars is less than 5 grams at a significance level of 5%. This means that $\alpha = 0.05$.

State your hypotheses
Your null hypothesis is that the average weight of the bars is $\mu_0 = 5$ grams and your alternative hypothesis is the average weight of bars is less than $\mu_0 = 5$ grams. So $H_0 : \mu = 5$ and $H_1 : \mu < 5$.

Calculating the test statistic
In this case you have a sample of 30 bars and you know the population standard deviation $\sigma = 1.5$. You are also testing the alternative hypothesis that $\mu < 5$. This implies that you would want to use a **lower one-tailed Z-test**. This means calculating a test statistic according to a Z-test, which is outlined here.
First, you would need to calculate the standard error from the mean. This is

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{1.5}{\sqrt{30}}$$

Next, the test statistic (Z-score) is given by

$$Z = \frac{\bar{x} - \mu_0}{SE} = \frac{4.5 - 5}{\frac{1.5}{\sqrt{30}}} \approx -1.826$$

(See [Guide: Introduction to Z-testing] for more.)

Example 4 continued

Find the critical value or p-value

Since this is a lower one-tailed test, you need to find the critical value for $\alpha = 0.05$ from a Z-table. The critical Z-value for a significance level of 0.05 in a lower one-tailed test is -1.645.

Alternatively, you can find the p-value associated with the test statistic $Z = -1.826$. You can use Calculator: Z-testing to see that a Z-score of -1.826 corresponds to a p-value of approximately 0.033925.

Conclusion

If your test statistic is less than the critical value -1.645 or $p \leq \alpha$, you can reject the null hypothesis H_0. If neither condition is met, you fail to reject the null hypothesis. In this case, $Z = -1.826$, which is less than the critical value of -1.645. Similarly $p = 0.033925$, which is smaller than $\alpha = 0.05$.

So you can reject the null hypothesis H_0. There is enough evidence to conclude that the average weight of the Boole Bars is less than 5 grams. It looks like Cantor Confectionery need to get to work...

Quick check problems

1. What would your hypotheses be when testing whether the average battery life of a laptop is less than 10 hours?

(a) $H_0 : \mu > 10$ $H_1 : \mu < 10$

(b) $H_0 : \mu = 10$ $H_1 : \mu \neq 10$

(c) $H_0 : \mu = 10$ $H_1 : \mu < 10$

2. What α corresponds to a 10% level of significance?

3. What test would you use to compare means when you have one sample and you know the population standard deviation?

4. You are performing a two-tailed test with $\alpha = 0.05$ and critical values of 1.96 and -1.96. Your test statistic is -2.34. What conclusion do you draw?

Further reading

For more questions on the subject, please go to Questions: Hypothesis testing.

[For more information on how to perform a Z-test, please see Guide: Z-testing.]

[For more information on type I and type II errors please see Guide: Errors in hypothesis testing.]

Version history

v1.0: initial version created 12/24 by Ellie Trace as part of a University of St Andrews VIP project.