Statistical Testing with z-score and p-value

Introduction

In this lesson, you'll learn how the z-score is important when performing statistical tests. This lesson is meant to be a segue from standard normal distributions into testing, but you'll learn about testing in more detail later!

Objectives

You will be able to:

- Compare and contrast a population and a sample
- Explain what is meant by a "representative" sample
- Use the z-table and `scipy` methods to acquire the p value for a given z-score
- Define what null and alternative hypotheses mean and when they are used
- Define the significance threshold and its relation to p-value

Statistical significance

Statistical significance is one of those terms that is often used when someone claims that some data collection and analysis proves a **certain point** (or hypothesis). The terminology around statistical significance is often not well understood, however, it is a simple idea that can be understood fairly easily.

Statistical significance is based on a few concepts: samples and populations, hypothesis testing, the normal distribution, and p-values. In this lesson, we'll either repeat or introduce all of the foundational concepts associated with statistical significance.

First, let's look at how to differentiate between samples and populations.

Population vs sample

The first step of every statistical analysis you will perform is the population vs. sample check or to determine whether the data you are dealing with is either a **population** or a **sample**.

> A **population** is the collection of **all the items of interest in a study**. The numbers you obtain when using a population are called **parameters**.

> A **sample** is a **subset of the population**. The numbers you obtain when working with a sample are called **statistics**.

Example

Imagine we want to conduct a survey of the job prospects of the students studying at Flatiron School.

What is the population?

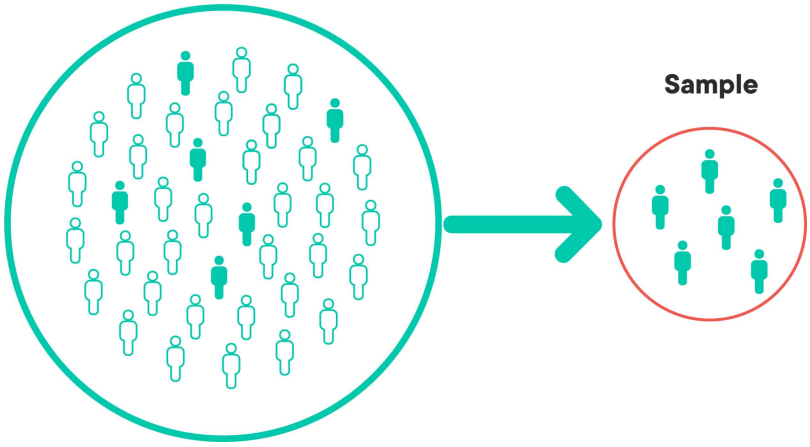
You can go ahead and contact all the students studying at Flatiron School campuses around the world in our physical campuses. But that would not be the whole population of Flatiron students. A lot of students take Flatiron's courses online. Including those students from all over the world would make a COMPLETE student population at Flatiron. Also, by the time you finish contacting all these students, you would realize that a lot of new students have enrolled in the courses since you last surveyed.

So you can see that populations are hard to describe and inspect in real life.

What is a sample?

A sample is much easier to describe and inspect. Conducting a survey on a sample is less time-consuming and less costly too. Time and resources are the main reasons we prefer drawing samples over working with entire populations.

As we first wanted to do, we can just go to the New York campus. We can visit during the lunch hour in the canteen because we know it will be full of people. We can then interview 50 of them. This would be called a student sample.



Is the sample "representative" ?

So what are the chances these 50 students can provide us answers that are a true representation of the whole student population of Flatiron School globally? You guessed it, the chances are pretty low. The sample is *neither* **random** nor **representative** of the whole population.

> A random sample is collected when each member of the sample is chosen from the population strictly by chance. In order for the sample to be random, each member should be equally likely to be chosen.

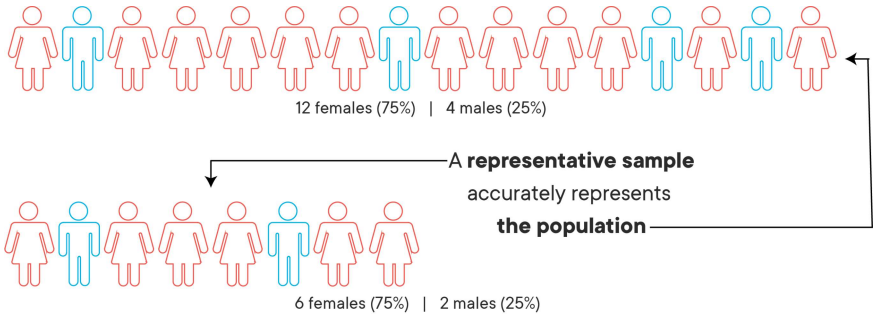
It is clear that these two requirements were violated by picking 50 students from the NYC campus.

How can you make it a representative sample?

> A representative sample is a subset of the population that accurately reflects the members of the entire population

Our sample represented the NYC campus students who have lunch at the canteen. If we ran a survey about job prospects of Flatiron students who eat in the NYC campus canteen, we would have done well.

By now, you must be wondering how to draw a sample that is both random and representative. Well, the safest way would be to get access to the student database with **all students around the world** and contact individuals in a random manner.



We said populations are hard to define and observe. Then, we saw that sampling is difficult. But samples have two big advantages:

1. Once well-understood, it is not that hard to recognize if a sample is a representative one
2. Statistical tests are designed to work with incomplete data, so making a small sampling error is not always a problem

Now that we understand using samples vs. populations, we can move on with statistical testing.

Statistical Testing

There are a huge number of statistical tests and you will always try to select the one that best fits your research design. [This link (https://stats.idre.ucla.edu/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/)](https://stats.idre.ucla.edu/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/) provides a quick introduction to a number of testing criteria. We'll cover some of these in-depth later on.

Right now, let's talk about one of the most basic types of statistical testing techniques based on the z-score, the one-sample z-test.

The One-sample z-test

The one-sample z-test is used when you want to know if your sample comes from a particular population.

For instance, when collecting data from successive cohorts of students taking the Data Science Bootcamp, you may want to know if this particular sample of students is similar to or different from Flatiron students in general.

> **The one-sample z-test is used only for tests related to the sample mean.**

When running a one-sample z-test, you test whether the average of the sample suggests that the students come from a certain population with a known mean or whether it may come from a different population.

You already know what a z-score is and how to standardize a dataset into a z-distribution. By itself, the z-score doesn't provide a lot of information to conclude a question significantly (except for saying how many standard deviations some observation is from a mean). The real value from a z-test comes from comparing it against a **z-table**.

The z-table

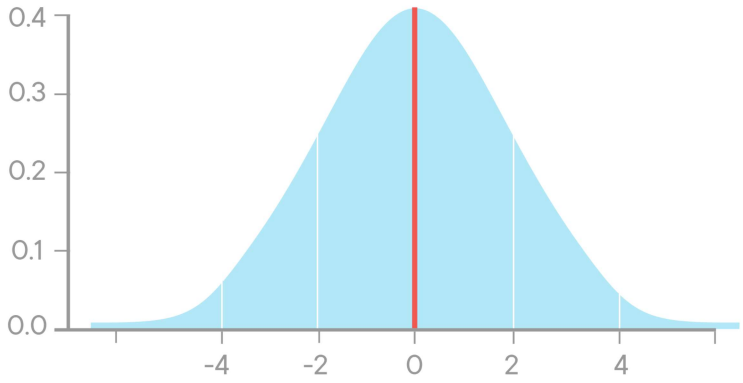
A z-table contains cumulative probabilities of a standard normal distribution up until a given z-score value. By transforming normal distributions with various means and standard deviations, you can use this z-table for any value that follows a normal distribution. The z-table is short for the "Standard Normal z-table".

> **The area under the whole of a normal distribution curve is 1, or 100 percent. The z-table helps by telling us what percentage is under the curve up to any particular point.**

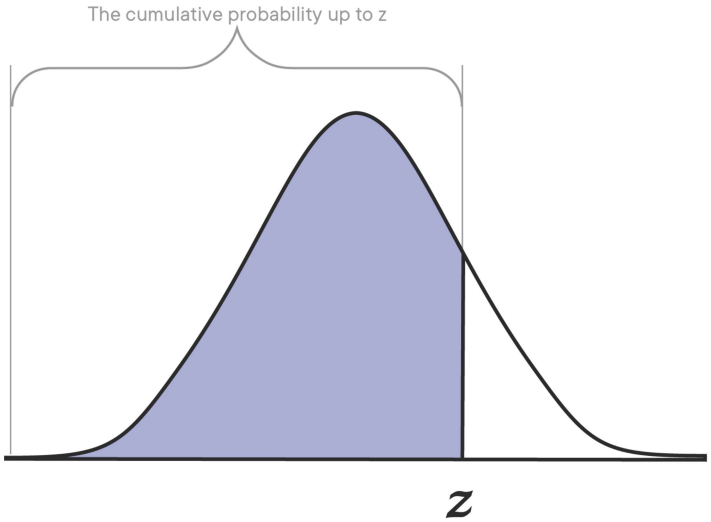
Here is a link to an online version of z-table (http://www.z-table.com/). This lesson's GitHub repository also contains a pdf version of this table, `z-table.pdf` .

- **The rows** of the table contain z-values in the form $x.x$ along the left margins of the table, specifying the ones and tenths.
- **The columns** fine-tune these values to hundredths, allowing us to look up the probability of being below any standardized value z of the form $x.xx$.

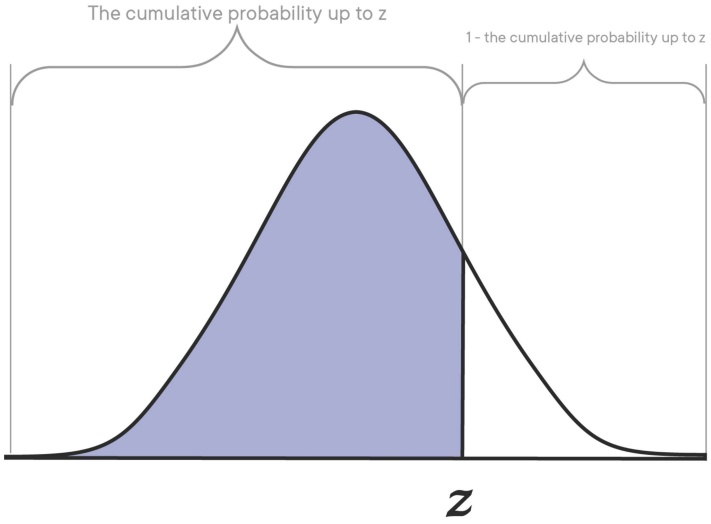
You know that a cumulative probability is the sum of the probabilities of all values up until a given point. An easy example is the mean. The mean is the exact middle of the normal distribution, so we know that the sum of all probabilities of getting values from the left side up until the mean is 0.5. Also, the sum of probabilities from the mean to right tail would also sum up to 0.5. The mean is denoted by the red line in the image below.



When using the idea of cumulative probability in the context of the standard normal distribution, we look at the cumulative distribution until a point z.



As the sum of all probabilities is equal to 1 or 100%, you can use the z-table to calculate probabilities on both sides of the z-score under the standard normal distribution.



Using z-scores, you can answer questions like "how far is a value from the mean" and "how likely is a value this far from the mean to be from the same group of observations?"

Example

> What is the probability of a z-score being less than or equal to 1.5?

To find the answer using the z-table, you have to go and look where the row for 1.5 intersects with the column for 0.00; this value is 0.9332. The z-table shows only “less than” probabilities so it gives you exactly what you need for this question. The probability of a z-score being less than or equal to 1.5 is 0.9332.

> What is the probability of a z-score being greater or equal to 1.34?

Use the z-table to find where the row for 1.3 intersects with the column for 0.04, which is 0.9099. Because the z-table gives you only "less than" probabilities, subtract P(Z < 1.34) from 1 (remember that the total probability is 1.00 or 100%). So, **1 - 0.9099 = 0.0901**.

Here is a short video on how to use a z-table (https://www.youtube.com/watch?v=lgwT6tDniko)

The z-table alternative in python

When programming in Python, SciPy provides a handful of features so you won't have to go out consulting z-tables for automated analysis. For normal distributions, probabilities **up to the z-score** can be calculated with `.cdf` -method as shown below:

In [2]:
```
# Z-table in Python
import scipy.stats as stats

# Probabilities up to z-score of 1.5
print(stats.norm.cdf(1.5))

# Probabilities greater than z-score of 1.34
print (1-stats.norm.cdf(1.34))
```

```
0.9331927987311419
0.09012267246445238
```

What Are Hypotheses ?

A very important aspect of statistical testing is setting up a hypothesis to be tested through data analysis.

The hypothesis is a data scientist's initial understanding of an observation prior to the testing. This hypothesis is known as the **Alternative Hypothesis** (written as H_a).

The opposite to the Alternative Hypothesis is known as a **Null Hypothesis** (written as H_0).

The table below shows three sets of Null and Alternative Hypotheses. Each makes a statement about how the mean μ is related to some hypothesized value M.

Set	H_0	H_a	Tails
1	$\mu = M$	$\mu \neq M$	2
2	$\mu \geq M$	$\mu < M$	1
3	$\mu \leq M$	$\mu > M$	1

Here the tails represent if we are testing both sides of the distribution or only one side.

As an example, an analyst may want to check how effective a new drug is by setting an Alternative Hypothesis that the new drug reduces blood pressure by 10%.

In this case, the Null Hypothesis states that the new drug has no effect on the patients. When performing hypothesis testing, you generally will try to reject the null hypothesis to obtain what we call "Statistically Significant Results".

P-value

You will now learn about the **p-value** as a statistical summary of the compatibility between the observed data and what you would expect to see in a population assuming the statistical model is correct. The concepts of p-value and level of significance are vital components of hypothesis testing and methods like regression. However, they can be a little tricky to understand. We'll try to explain the concept in an easy, logical way.

> In hypothesis testing you set a null hypothesis, then draw a sample, and test your null hypothesis based on that sample.

For example, imagine your null hypothesis H_0 is that the population mean μ is $\mu = 10$. Upon drawing a sample, you get a mean of 12. **With the p-value, you are going to obtain a probability that, given a null hypothesis of $\mu = 10$, you would observe a sample mean of 12.**

> If your p-value is low, you will reject your null hypothesis. You will basically say that **based on current evidence and testing, the null hypothesis is not true.**

> If your p-value is high, you will fail to reject your null hypothesis. You will fail to reject the null hypothesis, that is, you will say that **based on current evidence and testing, the null hypothesis cannot be rejected.**

You'll see that the phrase "accepting a null hypothesis" is not used. This is because conclusions of hypothesis tests will state that "we reject H_0 in favor of H_a" or that "we cannot reject H_0 in favor of H_a", which is less definitive and leaves room for errors while testing. You reject or fail to reject a null hypothesis based on the evidence you have.

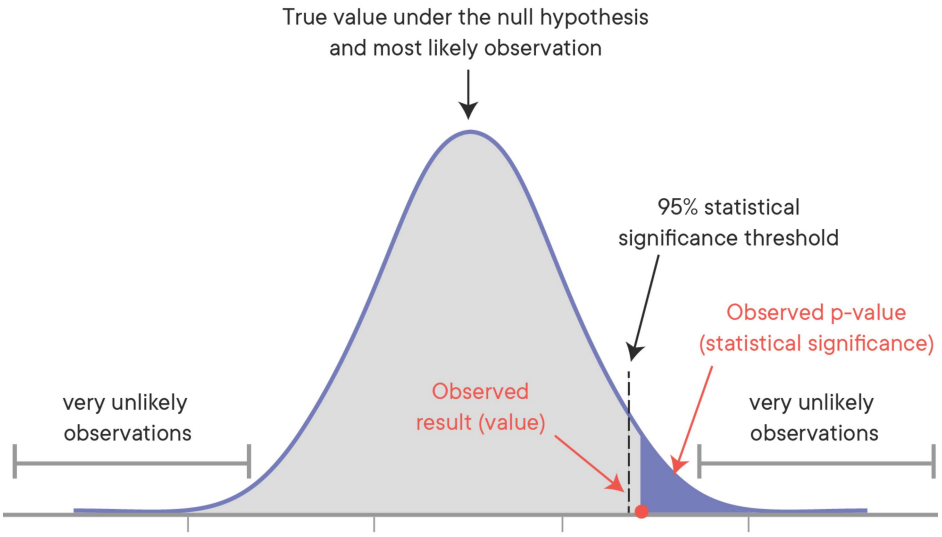
> It is important to understand what you have **assumed** and what you have **observed**

You assumed your population mean is 10, without actually observing that. You have observed a sample mean of 12 after testing your sample.

You then verify whether the sample mean you obtained is consistent with the population mean you have assumed. In other words, what are the chances of getting the result (sample mean) if the assumption is actually true (population mean). What is the probability that the sample mean is 12, assuming that the population mean is 10?

This chance or probability is called a p-value.

If your p-value is low, you say that that the result is **significant**, in the sense that you conclude that the sample mean is **significantly different** from the population mean.



What is the Significance Threshold (alpha, α)?

You noticed that we talked about "high" and "low" p-values, but that is pretty vague. What number is high and what number is low?

This is where the significance level, also denoted as alpha or α comes in. α is the threshold value that defines whether a p-value is low or high. You can define your alpha level yourself, but you'll see that an alpha level of $\alpha = 0.05$ is most commonly used. You'll see $\alpha = 0.1$ and $\alpha = 0.01$ appear frequently as well.

What level of alpha to use depends on your situation. Choosing a lower alpha leads to a test that is more strict, so you will be less likely to be able to reject your null-hypothesis (which is generally what you want). Choosing a higher alpha or significance level leads to a higher probability of rejecting the null-hypothesis. The downside of using a higher alpha level, however, is that you run a higher risk of falsely concluding that there is a difference between your null-hypothesis and your observed results when there actually isn't any.

This may all seem a little vague for now. You'll get a better understanding when we dig deeper later on.

Summary

In this lesson, you learned about the basics of hypothesis testing and went through the concepts and terminologies that are used while performing statistical tests. Now, let's move on to perform a one-sample z-test to validate a hypothesis that the sample drawn belongs to the same population.