Table of Contents

[A/B Testing 1](#_Toc56068087)

[Type 1 and Type 2 Errors 1](#_Toc56068088)

[Preventing Type 1 and Type 2 Errors with Hypothesis Testing 1](#_Toc56068089)

[Step 1. Choose the statistic that is to be measured. 2](#_Toc56068090)

[Step 2. Define the hypothesis test. 2](#_Toc56068091)

[Step 2. Calculate the Effect Size 2](#_Toc56068092)

[Step 3. Calculate the Number of Samples Needed 3](#_Toc56068093)

[Step 4. Perform the Test 3](#_Toc56068094)

[Generate Random Data 4](#_Toc56068095)

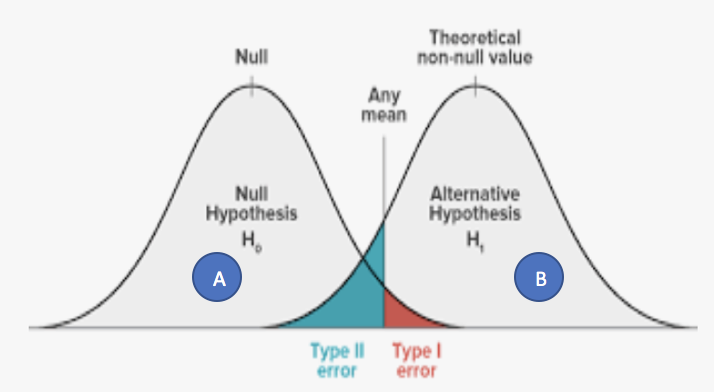
## A/B Testing

A/B testing involves testing whether or not one scenario performs better than another. For example, you could test whether visitors to a website purchase more if the page has a red background or a pink background. Or you might test if potential blood donors respond better to blood donor request message 1 or blood donor request message 2.

### Type 1 and Type 2 Errors

If we are not careful we may incorrectly interpret the results from our testing. A type 1 error occurs when we incorrectly conclude that samples are from distribution B when they are actually from distribution A. Type 2 errors occur when we incorrectly conclude that a sample is from distribution A when it is actually from distribution B (see Figure 1).

Figure : Type I and Type II Errors



Preventing Type 1 and Type 2 Errors with Hypothesis Testing

At this point you might be thinking “Wow this is potentially a serious problem! How can I avoid type A and type B errors?!!” Thankfully we can create a hypothesis test to help prevent these mistakes. To do this, we create a T-test for the means of *two independent* sample sets. The general design of the test is:

Change makes no difference. New sample is from same distribution. (p-value 0.05)

: Change makes a difference. Sample is from different distribution. (p-value < 0.05)

**Tips:**

When designing your hypothesis tests:

* Define a clear hypothesis.
* Design a test for one metric only.

Example : Selling Strawberry Cake

This example is based on the article at:

<https://www.mikulskibartosz.name/how-to-perform-an-ab-test-correctly-in-python/>

|  |  |
| --- | --- |
|  | A restaurant wants to sell more strawberry cakes to groups of customers. The owners wonder if their new menu card entices more groups to buy strawberry cake. The current mean number of dessert-eating customers is 123 groups per day with a standard deviation = 20. The restaurant wants the new mean to be 140 per day with a standard deviation of =20. |

This section will explain how to determine if the observations from A/B testing validate the hypothesis that more customers are buying desserts because of the new table card.

### Step 1. Choose the statistic that is to be measured.

First, identify the one statistic that we want to measure. For this case, the total number of customer groups that are purchasing desserts per day is the statistic to be measured.

### Step 2. Define the hypothesis test.

The next step involves defining the hypothesis. For our case, the tests are;

=The new table card makes no difference to the number of groups ordering dessert each day.

=The new table card increases the number of groups ordering dessert each day from an average of 123 to 140.

### Step 2. Calculate the Effect Size

Next, we need to calculate the effect size. The effect size helps us to identify an adequate minimum number of samples for the hypothesis test. For our case, we want the effect to help us evaluate whether or not the average strawberry cake desserts ordered increases to 140 from 123 per day. The effect size is calculated with Cohen’s d formula:

d =

For our example;

d = = = = = = -0.85

We can now use Cohen’s d value as a parameter in the next step.

### Step 3. Calculate the Number of Samples Needed

Once you have the effect size, you can use it as a parameter to calculate the appropriate number of samples needed to detect such a change between both groups. Running the following code outputs a result that indicates that 23 separate days are needed for testing with the new cards since we are testing the average numbers of groups that order dessert per day.

|  |
| --- |
| from statsmodels.stats.power import TTestIndPower  effect = -0.85 # Obtained from previous step.  alpha = 0.05 # Enable 95% confidence for two tail test.  power = 0.8 # One minus the probability of a type II error.  # Limits possibility of type II error to 20%.  analysis = TTestIndPower()  numSamplesNeeded = analysis.solve\_power(effect, power=power, alpha=alpha)  print(numSamplesNeeded) |

### Step 4. Perform the Test

The next step involves performing the test. This can be done using 23 observations from days when the old table card is used and with 23 observations from days when the new card is used. It happens that the variance for both distributions is the same but we will treat the variances as if they are independent so *equal\_var* is set to False.

testResult = stats.ttest\_ind(new\_menu\_sales, old\_menu\_sales, equal\_var=False)

Here is the full program.

|  |
| --- |
| from scipy import stats  old\_menu\_sales = [101, 110, 115, 136, 140, 108,  80, 89, 131, 98, 121, 117, 106,  141, 119, 153, 184, 127, 103,  139, 130, 146, 130]  new\_menu\_sales = [158, 145, 134, 130, 113, 135,  163, 128, 166, 154, 143, 147, 132,  132, 136, 99, 163, 106, 143, 168, 136,  123, 159]  testResult = stats.ttest\_ind(new\_menu\_sales,  old\_menu\_sales, equal\_var=False)  import numpy as np  print("Hypothesis test p-value: " + str(testResult))  print("New sales mean: " + str( np.mean(new\_menu\_sales)))  print("New sales std: " + str(np.std(new\_menu\_sales))) |

The output shows a p value of 0.009. Therefore, we can reject the null hypothesis which suggests the new card makes no difference. In other words, we can conclude that the new card likely does help to obtain more orders of strawberry cake dessert. The higher number of dessert purchases after using the new card appears to be the result of having the new card:

|  |
| --- |
| Hypothesis test p-value:  Ttest\_indResult(statistic=2.7250996310490008, pvalue=0.009307319110185647)  New sales mean: 139.69565217391303  New sales std: 18.425961794515096 |

Exercise (8 marks)

The government of Canada recently shut-down gym class in high school due to COVID-19 cases. Prior to the shut-down it appeared that an average of 10 new cases per day were linked to high school across the country. After the shut-down it appears that the cases linked to high school had dropped to 5 new cases per day. The standard deviation in both cases appears to be around 4.5 before and after the shut-down. Treat the standard deviations as if they are from separate distributions. The government is facing requests to lift these restrictions on gym class so it wants to verify if the shut-down has been effective statistically.

You are to create the following test:

=The shut-down makes no difference to the number of new COVID cases linked to high school per day.

=The shut-down has helped to reduce the number of new COVID cases linked to high school per day.

Calculate Cohen’s d value for this case. Show your calculations here:

|  |
| --- |
| d =  d =  =1.11111 |

Using the effect size that is obtained in the last step, calculate an adequate number of sample days that are needed to compare COVID case counts before and after the shut-down. The government only wants to allow for a 5% probability of a type 2 error. Show your code here:

|  |
| --- |
| 22.060362579722707 so 23 days |

Show the optimal number of sample days here:

|  |
| --- |
| from statsmodels.stats.power import TTestIndPower effect = 1.11 # Obtained from previous step. alpha = 0.05 # Enable 95% confidence for two tail test. power = 0.95 # One minus the probability of a type II error.  # Limits possibility of type II error to 5%. analysis = TTestIndPower() numSamplesNeeded = analysis.solve\_power(effect, power=power, alpha=alpha) print(numSamplesNeeded) |

Exercise (4 marks)

We happen to have 30 samples of new case counts from before and after the gym shut-down.

Case counts linked to high school before gym class shut down:

[10, 11, 6, 18, 11, 9, 13, 9, 3, 12, 3, 13, 14, 4, 12, 8, 18, 17, 15, 18, 6, 1, 13, 9, 11, 15, 11, 7, 12, 14]

Case counts linked to high school after gym class shut down:

[ 5, 7, 3, 12, 0, 7, 0, 8, 9, 5, 5, 2, 2, 2, 4, 6, 6, 7, 1, 6, 10, 1, 0, 2, 2, 4, 5, 1, 4, 3]

Using only the number of samples that are recommended in Exercise 1, determine if the cases

Show your p-value and state whether the closing of gym class appears to reduce the COVID cases or not.

|  |
| --- |
| P value is pvalue=3.7411156335575167e-05 so closeing gym is helpful to reduce the covid cases |

## Monte Carlo Simulations

Monte Carlo simulations allow us to perform risk analysis by using randomly selected variable values. Monte Carlo simulations assume the variables fall within a specific distribution.

Exercise (12 marks)

This video explains how to implement a Monte Carlo simulation with Excel.

<https://www.youtube.com/watch?v=HwVBi--mE4M>

For this exercise, complete the simulation and upload your spreadsheet with the word document to the drop box for this lesson.

During the lesson we will examine how to implement Monte Carlo simulations for different distributions with Python.