What is Hypothesis Testing

Edris Safari

Bellevue University, Nebraska U.S.A.

Abstract

Hypothesis testing is an essential part of exploratory data analysis. The exercises that take place during this phase of the analysis can make or break the conclusions that are made. We describe this process in this paper and show the steps that are taken and some of the best practices that are used in testing any hypothesis made on any dataset.

What is Hypothesis Testing

Introduction

Hypothesis testing starts with a question and ends with an answer that can go many ways, but they all essentially say whether the hypothesis is correct or not. In statistics terms, the true answer is that after testing the hypothesis many times using sample data, the result applies to the larger population with a very high confidence. This is important because back in the example of building a bridge by first building a model (or sample) of a bridge, we ran many tests and now we want to build the real bridge. We want to make sure that all the hypothesis that we made applies to the real bridge. Hypothesis such as the bridge will have a wind resistance of up to 250 miles per hour, or the bridge will handle up to 9 million tons of weight, etc. The data that is gathered during the testing the model (or sample) will contribute to the hypothesis testing. Hypothesis testing begins by making a null hypothesis from the original hypothesis. As the name implies, a null hypothesis is the mathematical opposite of the hypothesis or the alternative hypothesis. In statistics, instead of proving something is true, we prove that it’s opposite is false. For example if the hypothesis is that is coin is biased because after 250 coin tosses(Downey), more heads than tails were observed, the null hypothesis is that the coin is not biased and the probability that there will be more heads than tails in a larger number of tries is less than 1 percent.

To accept or reject the null hypothesis, we must perform a test statistic on the data that is available (i.e. 140 heads and 110 tails in 250 tries). The available data, the nature of the data, the hypothesis, and the null hypothesis all help decide what test statistic to use. In the case of the coin toss and the available data, the test statistic is the difference between number of heads vs tails (30 in this example). In other cases, we may want to use difference of means of two groups of distributions. As an example, consider the hypothesis that the pregnancy length is longer among the first-born babies. The null hypothesis in this case would be that the length of pregnancy of all babies vary in the same way regardless of order in which the babies are born. In other words, the two distributions-pregnancy length of first-borns and others is the same. In this case, we can model the null hypothesis with permutation (Downey). In this model, we combine the two groups of first-born babies and others into one distribution and shuffle it randomly. We then create two groups from the shuffled distribution. The length of these two groups are the same and are set to the length of the first-born distribution (the choice is arbitrary). We can then compute the absolute value of the difference in the mean of the two groups as shown below:

```
def TestStatistic(self, data):
    group1, group2 = data
    test_stat = abs(group1.mean() - group2.mean())
    return test_stat
```

This computed value is the test statistics. With the test statistics at hand and the model created, we can run several simulations of the same data set and compute the p-value. This computation is done in the following way:

```
def PValue(self, iters=1000):
    self.test_stats = [self.TestStatistic(self.RunModel())
                       for _ in range(iters)]

    count = sum(1 for x in self.test_stats if x >= self.actual)
    return count / iters
```

The model is created 1000 times, and the resulting data is used to compute the test statistics for the data set created (from the shuffled dataset). The p-value is then the sum of all the mean differences test_stat = abs(group1.mean() - group2.mean()) divided by number of iterations. This value shows the probability that the null hypothesis is true. If it is low, it means null hypothesis is false and our original hypothesis is true. By convention, statistical significance is strong when the p-value is below 1%, tolerable below 10%, and not acceptable above 10%. p-value can change depending on the hypothesis, the null hypothesis, test statistic, the choice of modeling of the null hypothesis. For example, if the hypothesis were that the first-borns are late, regardless of what the null hypothesis, the distribution is limited to only first-borns because we want to prove something in a single group. In this case, the test is called a one-sided test and the p-value is less than its two-sided version (first-borns arrive later than non-first-borns).

While p-value is not the only indication that the hypothesis is correct, various ways of calculating it gives us more choices depending on our use case. For example, using z-score, we can get an idea of how far the null hypothesis by comparing the test results to a normal distribution. The z-score for the sample data is calculated using the formula below:

$$Z = (x - \mu)/(\sigma/\sqrt{n})$$

Where x is the sample mean, μ is the population mean, σ is the sample standard deviation, and is the number of records in the sample.

For example, we have the following problem statement:

The average weight of residents in a town is 168 lbs. A recent survey of 36 individuals showed the average weight is 169.5 lbs. with a standard deviation of 3.9 lbs. At 95% confidence level, can we conclude that the average weight is 168 lbs and not 169.5 lbs. to calculate Z:

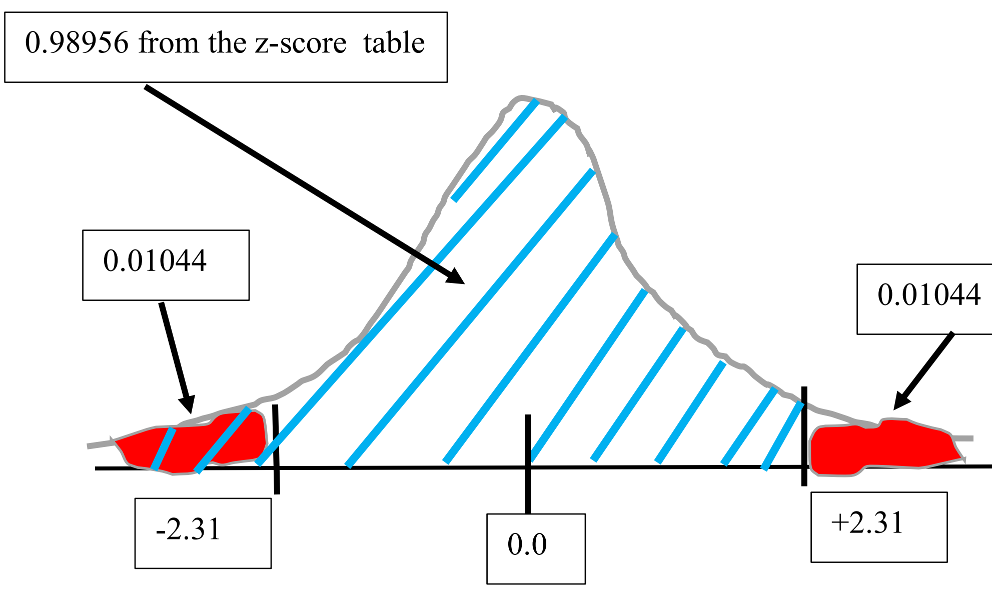
In this case, the null hypothesis is that the average weight is 168 and the alternative hypothesis is the average weight is 169.5. So, the numbers for the equation are as follows:

$$\mu = 168, n = 36, x = 169.5, \sigma = 3.9, n = 36$$

And with 95% confidence requirement, the significance level α=1-.95 = 0.05. The p-value must be greater than the significance level in order to accept the null hypothesis. Plugging in the numbers to calculate the Z value, we get

$$Z = (169.5 - 168)/\left(3.9/\sqrt{36}\right) \;=\; 2.31$$

The z-score on the normal distribution looks like below. We are interested in the area on the right side of 2.31. To get this value we must subtract the entire area to the left of 2.31 from the total area under the curve (which is 1). The area on the left is obtained from the z table shown below and in the blue-shaded area in the graph. This makes the area we seek to 1-.98956 = 0.01044 as shown in the graph below. The p-value is the sum of the areas on the right and left side of the distribution. So, the p-value = 0.01044 * 2 = 0.2088. With α=0.05, we can conclude that the null hypothesis is false and can be rejected.



Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
+0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
+0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55966	0.56360	0.56749	0.57142	0.57535
+0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
+0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
+0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
+0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
+0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
+0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
+0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
+0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
+1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
+1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
+1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
+1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
+1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
+1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
+1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
+1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
+1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
+1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
+2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
+2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
+2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
+2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
+2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
+2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
+2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
+2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
+2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
+2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
+3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
+3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
+3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
+3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
+3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
+3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
+3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
+3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
+3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995

Conclusion

Hypothesis testing is an essential part of data exploration. There are many statistical methods and BKMs that can only be learned by doing more research. More specifically, the use cases govern which methods to use, but at the core of it, hypothesis testing involves creating a null hypothesis, then create an appropriate(use-case specific) statistical test, then evaluate the result to decide whether to accept or reject the null hypothesis. Easier said that done!

References

Downey, Allen B.. Think Stats: Exploratory Data Analysis . O'Reilly Media. Kindle Edition.

Intro to Hypothesis Testing in Statistics - Hypothesis Testing Statistics Problems & Examples- https://www.youtube.com/watch?v=VK-rnA3-41c&list=TLPQMzAwMTIwMjB6n0F8Yhb1gw&index=1

Calculate the P-Value in Statistics - Formula to Find the P-Value in Hypothesis Testing https://www.youtube.com/watch?v=KLnGOL_AUgA&list=TLPQMzAwMTIwMjB6n0F8Yhb1gw&index=3

P-Value Method For Hypothesis Testing- https://www.youtube.com/watch?v=8Aw45HN5lnA&t=224s

Z-Score: Definition, Formula and Calculation - https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/