Data Science
Soo Min Kwon

Independent Study

Hypothesis Testing in Practice

Preamble

This guide is a brief review on hypothesis testing in practice. Some of the examples have been obtained from open sources online as well as some of the material. This is a good source to follow for a more in-depth discussion on hypothesis testing:

`https://stattrek.com/hypothesis-test`

1 Prelimiaries

1.1 Student's t-Distribution

The t-distribution, or Student's t-distribution, is a probability distribution used to estimate the parameters of a population when either the sample size is too small or the variance is unknown. When either of these problems arise, we rely on the distribution of the t-statistic, defined as

$$t = \frac{x - \mu}{\sigma/\sqrt{n}}, \tag{1}$$

where x is the sample mean, μ is the population mean, σ is the standard deviation of the sample, and n is the sample size.

Degrees of Freedom. As with other probability distributions, there are different "forms" of t-distributions, where these "forms" are determined by the degrees of freedom. The degrees of freedom for the t-distribution is determined by the sample size minus one.

Properties. Some properties of the t-distribution are the following:

- The mean of the distribution is equal to zero.
- The variance is equal to

$$\sigma^2 = \frac{v}{v-2}, \tag{2}$$

  where v is the degrees of freedom.
- When the number of samples is sufficiently large, then σ^2 approaches 1, and the t-distribution converges to the standard Normal distribution.

We demonstrate the use of the t-distribution using an example.

Example:

Acme Corporation manufactures light bulbs. The CEO claims that an average Acme light bulb lasts 300 days. A researcher randomly selects 15 bulbs for testing. The sampled bulbs last an average of 290 days, with a standard deviation of 50 days. If the CEO's claim were true, what is the probability that 15 randomly selected bulbs would have an average life of no more than 290 days?

Solution. We can compute the t-statistic as

$$t = \frac{x - \mu}{\sigma/\sqrt{n}} \tag{3}$$

$$= \frac{290 - 300}{50/\sqrt{15}} \tag{4}$$

$$= -0.775. \tag{5}$$

Using the t-distribution table, we can see that the cumulative probability is 0.226.

1.2 Significance Level & p-Values

The significance level in hypothesis testing, generally denoted by α, is mathematically defined as

$$\alpha = 1 - \text{Confidence Level}, \tag{6}$$

where the confidence level measures the uncertainty associated with a type of sampling method. For example, if we have a confidence level of 95%, this would mean that if we sample from this distribution multiple times again using the same sampling technique, we can assume that 95% of the samples would come from the distribution defined by its confidence interval. Given a confidence level of 95%, this would mean that our significance level is $\alpha = 0.05$. With the definition of the confidence level, the definition of the significance level comes out somewhat naturally – this is the probability in which we reject the null hypothesis when it is true (or the probability of false alarm).

When performing hypothesis tests, we say that we reject the null hypothesis when the p-value is less than or equal to the significance level α. To understand what this means, we need to define what the p-value actually is. By definition, the p-value is just the probability in which we observe a sample that is "in the extreme" under the null hypothesis. Hence, if the p-value is less than the significance level (also often called the critical level), it means it is likely that the samples that we observed comes from the alternate distribution and can reject the null hypothesis. We go into more detail in the next section with concrete examples on hypothesis testing.

2 Hypothesis Testing

There are actually many ways in which we can perform hypothesis testing, and the variation in methods comes from the choice in the parameter. For example, if we have a discrete parameter (e.g. click through rate), then we need to use the p-value method to perform hypothesis testing, where we can compute the p-value using the **Pearson's Chi-squared test** or the **Fisher's exact test**. If we have a continuous parameter (e.g. the mean), then we can perform the Student's t-test or the z-test, along with many others. We will first go over some hypothesis testing examples using the t-test.

2.1 Example: One-sided Test

Example:

Suppose an engineer measured the Brinell hardness of 25 pieces of ductile iron that were sub-critically annealed. The mean of the Brinell hardness was stated to be 170, and the sample mean of the 25 pieces was 172.52 with a standard deviation of 10.31. The engineer hypothesizes that the average hardness score is greater than 170. At a 95% confidence level, is there enough evidence to support this hypothesis?

Solution. The null and alternative hypotheses for this problem are the following:

$$\mathbf{H}_0 : \mu = 170 \tag{7}$$

$$\mathbf{H}_1 : \mu > 170. \tag{8}$$

This is clearly a one-sided test. Using the t-distribution table, the value corresponding to a significance level of $\alpha = 0.05$ is 1.71. If the t-statistic or value is greater than 1.71, then we can reject the null hypothesis. We can compute the t-statistic as

$$t = \frac{172.52 - 170}{10.31/\sqrt{25}} \tag{9}$$

$$= 1.222. \tag{10}$$

Since the t-statistic is less than 1.71, we accept the null hypothesis. Hence, at the significance level $\alpha = 0.05$, there is insufficient evidence to believe that the mean is greater than 170.

Similarly, we can use the p-value approach to solve this problem. To do this, we compute the area under the curve to the right of the t-statistic $t = 1.22$, which is 0.117. Since this p-value is greater than our significance level, we again fail the reject the null hypothesis.

Note that this example was a **one-sided**, one-tailed test since we are comparing the mean to one value. A **two-sided** test would be if we compared the means from two different populations. We demonstrate this in the next example. The next example is actually an interview question, which we will "hypothetically" try to answer.

Example:

Given a list of users who only share their height (in cm) and gender on their profile, how would you test the hypothesis that men on average are taller than women?

Solution. Since we do not have concrete data, we need to speculate how we would answer this problem. The null and alternate hypotheses for this problem would be

$$\mathbf{H}_0 : \mu_M = \mu_F \tag{11}$$

$$\mathbf{H}_1 : \mu_M > \mu_F. \tag{12}$$

As we are comparing if the average height of the male population is greater than that of the female population, this is a one-tailed test. Note that since we do not know the standard deviation of the average heights of the population, we need to use the t-test. To compute the t-statistic, we need to find the average height and standard deviation of both the male and female population. Then, we need to compute the t-statistic using the following formula:

$$t = \frac{X_M - X_F}{\sqrt{\sigma_M^2/N_M + \sigma_F^2/N_F}}, \tag{13}$$

where X_M is the sample mean for the male population, σ_M^2 is the variance for the male population, and N_M is the sample size for the male population, and the F subscript is the female population. We then need to pick a significance level. Upon picking the significance level, we just need to compute the t-value (critical value) corresponding to the area of the significance level, and then compare the t-statistic to the corresponding to that critical value. If the t-statistic is greater than the critical value, we reject the null hypothesis.

2.2 Hypothesis Testing with Non-Normal Metrics

In practice, there are often times where we need to perform hypothesis testing, but the samples from a distribution that is not Normal. For example, **zero-inflated distributions** are not normally distributed, and they occur when a user for example does not buy anything at all. There are also **multi-modal distributions**, where a market splits up expensive and cheap purchases. For both of these distributions, we can often still use z-test and the Student's t-test if we are given enough samples due to the Central Limit Theorem. However, if the sample size is too small to assume normality, we can use a non-parametric approach such as

the **Mann-Whitney U Test**. This test is non-parametric, as it makes no assumption on the nature of the sample distribution. We can use this test to obtain the p-value and perform hypothesis testing.