Litterature review - Hypothesis testing

Département de mathématiques et de statistique - Université Laval

Stéphane Caron

12 December, 2019

Contents

1 Introduction

This document provides a quick review of hypothesis testing theory. We first cover the basic of simple hypothesis testing. Then, we'll cover briefly some concepts about multiple hypotheses testing and also why it is relevant in the context of *reliable* anomaly detection. Finally, we'll describe 2 tests that might be used in order to conclude if an observation comes from a certain distribution (inlier) or not (outlier) in our specific context.

2 Simple hypothesis testing

A hypothesis test is a method of statistical inference. It aims to test a **null hypothesis** (H_0) versus an **alternative hypothesis** (H_1). In thoery, the best way to test an hypothesis would be to gather data about the entire population and test our hypothesis. However, in real life we often do not have access to data of the entire population. In fact, we usually have access to a sample only, on which we'll use the framework of hypothesis testing to confirm or infirm certain assumptions.

As an example, let's say we want to test if the average height of the canadian male adults is greater or equal than 1.75 m or less. Indeed, we probably don't have access to the informations of all canadians, we then need to test on a given sample of canadians male adults. We can formulate our test like that :

$$\begin{cases} H_0 : \mu >= 1.75 \\ H_1 : \mu < 1.75 \end{cases}$$

where μ is the average height of canadian male adults.

Once we have defined our hypothesis, we need a test method that will includes statistical assumptions about our sample. In the above example, we can use the one-sample t-test, that determines whether the hypothesized mean (μ) differs from the observed sample mean ($\bar{X}$). That will allows us to compute a certain statistic and then compare it with a known distribution. The one-sample t-test (denoted by t) is calculated using the following formula:

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

where $s_{\bar{x}} = \frac{s}{\sqrt{n}}$ and s is the sample standard deviation. That test allows us to compute a test statistic and compare it with a known distribution (Student's t-distribution in that case).

This example is considered a simple hypothesis test since we do one single test. We'll cover some concepts of testing more than one hypothesis at the time in further sections.

2.1 *p*-value

Once we have a computed test statistic coming from a known distribution, we can compute what we call a *p*-value. A *p*-value is defined as the probability, under null hypothesis, to observe the test statistic from the known distribution. When we have a small the *p*-value, it means that the hypothesis may not adequatly explain the observation. The null hypothesis is then rejected when the value is less than a certain threshold α, which is referred to as the **level of significance**.

When the null hypothesis is true and the underlying random variable is continuous, then the probability distribution of the *p*-value is uniform on the interval [0, 1].

2.2 Type of errors

When a decision is made from comparing a *p*-value with a level of significance, two kinds of errors can occur:

- Type I error (false positive): occurs when the null hypothesis is rejected when it is in fact true. The probability of making that error is called the **significance level** (also called alpha).
- Type II error (false negative): occurs when failing to reject null hypothesis that is false. The probability of commiting that kind of error is often called **Beta**. The probability of not commiting a type II error is called the **power** of the test.

3 Multiple comparison procedures

Multiple comparison procedures refers to the testing of more than one hypothesis at a time (Shaffer 1995). This framework aims to counteract the problem of multiple comparisons, where the more hypotheses we check, the higher the probability of Type I error (false positive) arises. The multiple comparison procedures often adjust the rejection criteria for each individual hypothesis, so that we do not have a single level of significance. In anomaly detection, that concept becomes important because we do a statistical test for each observation in our dataset, which should prevent false positive errors, or the detection of outliers that are in fact inliers.

3.1 Family-wise error rate

In statistics, family-wise error rate (FWER) is the probability of making one or more Type I errors when performing multiple hypotheses tests. In fact, it is simply defined by:

$$FWER = Pr(V \geq 1)$$

where V is the number of false positives (Type I error).

There are multiple definitions of "family" in the litterature, but the idea is that a family regroups a collection of inferences for which it is relevant to take into account a combined measure of error. When taking into consideration that family notion, we prevent from "data fishing" or "p-hacking" situations.

There are different procedures designed to control the Type I errors, some are called weak and some are called strong. We are more interested in strong procedures as they guarantee to control the α level for any distributions of true versus false null hypotheses. At the opposite, weak procedures provide guarantees only when all null hypotheses are true (which is not really interesting in our case since we want to find outliers).

3.1.1 Simple Bonferroni procedure

The simple Bonferroni procedure takes the form of:

$$\text{Reject } H_i \text{ if } p_i \leq \alpha_i$$

where the sum of all α_i equals to α. Usually, all α_i are equal, that means $\alpha_i = \alpha/n$ for all tests.

3.1.2 Holm–Bonferroni procedure

That procedure involves first ordering p-values from lowest to highest $(P_{(1)}, ..., P_{(m)})$ and the associated hypotheses $(H_{(1)}, ..., H_{(m)})$. Then, it defines k as the minimal index such that :

$$P_{(k)} > \frac{\alpha}{m + 1 - k}$$

Finally, it rejects all the null hypotheses $H_{(1)}, ..., H_{(k-1)}$. It's basicaly an improved version of simple Bonferroni procedure.

3.2 False discovery rate

False discovery rate (FDR) is another method of defining the Type I errors in null hypotheses when conducting multiple hypotheses tests. The controlling procedures under that framework provides less stringent control on Type I errors compared to family-wise error rate procedure. Compared to these methods, the FDR procedures provide greater power to the tests, at the cost of increasing Type I errors. In brief, when missing an anomaly detection is critical, the FDR framework becomes more interesting.

In more formal terms, the FDR can be defined as:

$$FDR = E[Q]$$

where Q is the proportion of false discoveries among all discoveries (rejections of null hypothesis). From that definition, we can see how it is less stringent than the FWER for which the probability (or the definition) is calculated from having **at least** 1 false positive.

3.2.1 Benjamini–Hochberg procedure

In the same fashion as the Holm–Bonferroni procedure, we need to first order the *p*-values and their hypotheses. Afterward, we need to find the largest k such that:

$$P_{(k)} \leq \frac{k}{m} \times \alpha$$

Then, we reject the null hypotheses for all H_i for $i = 1, ..., k$.

4 Examples of anomaly tests

To do *reliable* anomaly testing, we basically want to test if a given observation comes from a known distribution or not. Multiple tests exist to achieve that, but we described more in details 2 examples.

4.1 Chi-square test

A well-known test is the Chi-square test. Under that test, we suppose the known distribution is a chi-squared distribution when the null hypothesis is true. The hypotheses can be simply defined as :

$$\begin{cases} H_0 : \text{sample distribution is } \chi^2 \\ H_1 : \text{sample distribution is not } \chi^2 \end{cases}$$

Once we have our hypotheses, we need to compute a certain statistic that we think should follow the null hypothesis assumptions. If we have a sample of observations, let's say n representations of p dimensions following a Normal distribution, we can compute the mahalanobis distances of each of those representations with the "true" parameters of the distribution. That will allow us to retrieve a distance for each observation that will follow a $N(0, 1)$ distribution. By taking the square of that distance, we retrieve a χ^2 with p degrees of freedom.

Once we have the statistic value for each observation, we can computed the *p*-value for each of those using the H_0 assumption. Finally, we can reject null hypotheses (detect outliers) based on multiple comparison procedures

4.2 Hoeffding's *universal* test

The *universal* hypothesis testing problem is the idea of testing whether a certain sample (or an empirical distribution) of i.i.d . observations come from a known null hypothesis distribution (p_0) or some other distribution. It could be formalized as:

$$\begin{cases} H_0 : X_i \sim p_0 \\ H_1 : X_i \sim p \neq p_0, \quad p \text{ is unknown} \end{cases}$$

In the nineteen sixties, Hoeffding's proposed a universal test statistic that can be written in terms of Kullbach-Leibler (K-L) divergence between 2 distributions. The Hoeffding test is optimal in error-exponent sense and can be defined as:

$$\hat{H} = \mathcal{I}\{D(p_n || p_0) > \tau\}$$

where the divergence term $D(p_n || p_0)$ is based on K-L. This test provides weak convergence under p_0:

$$2nD(p_n||p_0) \xrightarrow[n\to\infty]{} \chi^2_{p-1}$$

where p is the number of dimensions.

References

Shaffer, Juliet. 1995. “Multiple Hypothesis Testing.” *Annual Review of Psychology* 46 (January): 561–84. https://doi.org/10.1146/annurev.ps.46.020195.003021.