UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

MASTER THESIS

Development of a platform to do hypothesis testing

Jordi Bosch Bosch

Master:

Master en Estadística e Investigación Operativa UPC-UB

Director:

Dra. Maria José Serna Iglesias (UPC)

June 2022

Abstract

Development of a platform to do hypothesis testing

Jordi Bosch Bosch

Contents

Acknowledgements

1 The project

Hypothesis testing is a wildly used tool in statistics to verify assumptions from the data. The concept was first written down by Laplace, but modern significance testing was formalized by Karl Pearson and Ronald Fisher.

Statistics deal with analyzing data and taking out conclusions. Hypothesis testing is one of the techniques available to verify and reject assumptions from the data.

The goal of the project was to build a web platform that could, given some data and the hypothesis you want to verify, answer you back giving insights of the test (be it graphics, p-values and some hidden information about the data).

I decided to pursue this project in order to combine both my education in computer science and statistics. Building a web platform from scratch has many details and challenges that I have always wanted to tackle, and, given that I was doing a master in statistics I thought this would be a great mix.

2 Hypothesis testing fundamentals

We will formalize an hypothesis over our data. We will call it the null hypothesis H0. We want to know if we want to reject the null hypothesis H0 in favour of the alternate hypothesis H1.

If we reject the null hypothesis, we do not prove that the alternative hypothesis is true. If we do not reject the null hypothesis, we do not prove that the null hypothesis is true. We merely state that there is enough evidence to behave one way or the other. This is always true in statistics! Because of this, whatever the decision, there is always a chance that we made an error.

Definition: p-value. In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. The p-value serves as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.
As threshold, we usually choose $\alpha = 0.05$.
If p-value ≤ 0.05, we will reject H0.
If p-value is > 0.05 we won't reject the null hypothesis. We have to be really careful at this step. Not rejecting the null hypothesis does not mean that the null hypothesis is true. It just means there is not enough evidences to reject it in favour of the alternate hypotheses.

2.1 Types of Errors

Of course, since we are dealing with statistics and probability we will commit mistakes.

	H0 true	H1 true
Not reject H0	Correct choice	Type II error
Reject H0	Type I error	Correct choice

Two kind of errors may happen:

- Type I error: We reject the null hypothesis, but turns out it was true (false positive).
- Type II error: We accept the null hypothesis, but turns out it was false (false negative).

As we see, since we cannot have infinite samples, we will sometimes fall into this errors. Our goal should be to minimize them. The level of significance α is the probability that we do an Type I error. β is the probability that we do an Type II error.

2.2 Different kinds of hypothesis tests

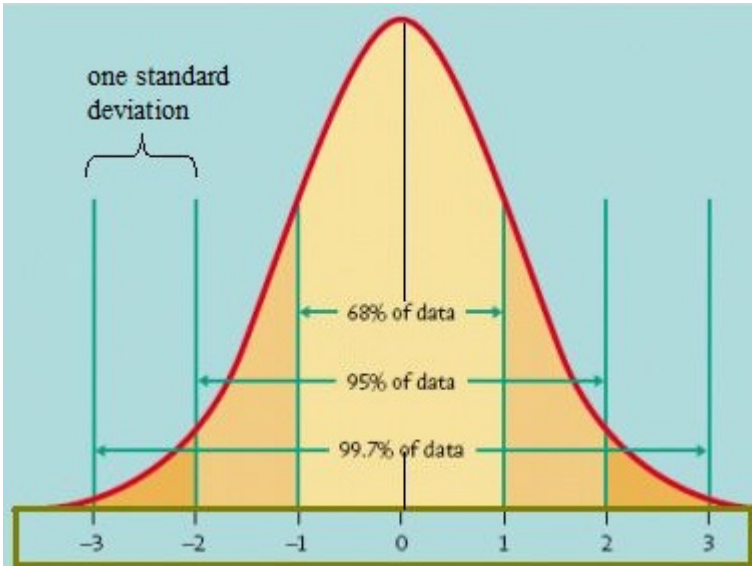
Before diving deep into each test we should state some properties about basic statistic distributions.

Normal distribution: Also know as Gaussian distribution. It's probability density function is expressed like this:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2}$$

We can quickly verify (using density function properties) that μ is the Expectation and σ is the standard deviation. The normal distribution plays a key role in statistics, probability and in nature.

The reason is because many data coming from nature and psychology follows a normal distribution.

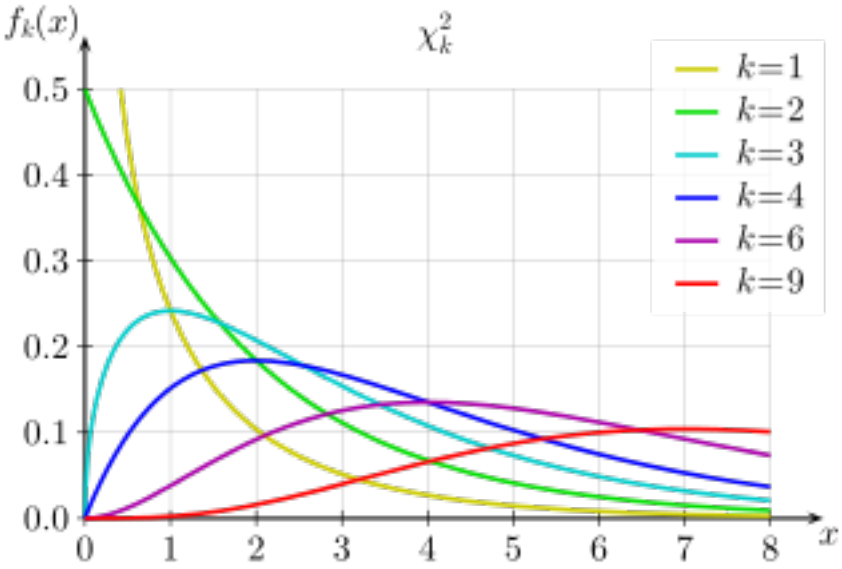


The central limit theorem proofs us why and how, many data follows a Gaussian distribution.

Chi square distribution of degree k:

$$\sum_{i=1}^{i=k} Z_i^2$$

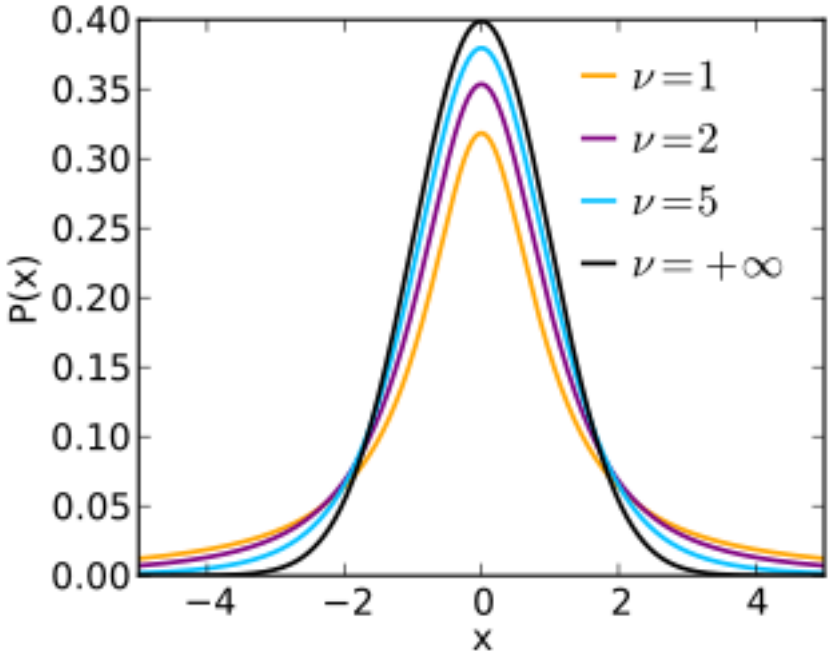
where Z_i are n independent normal distributions $N(0,1)$.



T Student distribution: Let $X_1, \ldots, X_n$ be independently and identically drawn from the distribution $\mathcal{N}(\mu, \sigma^2)$, i.e. this is a sample of size n from a normally distributed population with expected mean value μ and variance σ^2. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \bar{X}$ be the sample mean and let

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ be the (Bessel-corrected) sample variance.

Then the random variable $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ i.e where S has been substituted for σ, has a Student's t-distribution with $n-1$ degrees of freedom.



2.2.1 Student t test

2.2.2 Anova

ANOVA (One-way analysis of variance) is widely used to test when you have a categorical independent variable normally distributed and you want to know if there exists significantly differences in the means depending on the factor. ANOVA is widley used in medicine (to test the effects of vaccines for example) among many other fields.

Hypothesis needed

- Independence between different samples. No hidden relationships between observations should exist.
- Homocedasticity. Variance among the different factors should be the same. In case this hypothesis is not satisfied, ANOVA looses rebustness.
- Normality in each group. ANOVA requires normality in the samples of each factor of the variable. If this assumption is not verified, ANOVA looses robustness. This requirement becomes less needed when the size of the samples increases.

This hypothesis can be checked using QQplot and Saphiro-Wilk test (for the normality assumption) and Levene's test for the homocedasticity assumption. In both cases, we will choose a significance factor α and we would be provided with a $p-$value in order to check.

How does ANOVA work?

Lemma 2.1. *Suppose we are dealing with random variable with k factors, x_{ij} referes to the j-th sample of the i-th group. The total variance can be decomposed as:*

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{\bar{x}})^2}_{SST} = \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x_j})^2}_{SSE} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x_i}-\bar{\bar{x}})^2}_{SSB} \tag{1}$$

Proof.

$$\begin{aligned} SST &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{\bar{x}})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x_j}+\bar{x_j}-\bar{\bar{x}})^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}[(x_{ij}-\bar{x_i})^2+(\bar{x_j}-\bar{\bar{x}})^2+2(x_{ij}-\bar{x_j})(\bar{x_j}-\bar{\bar{x}})] \end{aligned} \tag{2}$$

This last term can be refactored as:

$$\begin{aligned} \sum_{i=1}^{k}\sum_{j=1}^{n_i}2(x_{ij}-\bar{x_j})(\bar{x_j}-\bar{\bar{x}}) &= 2\sum_{i=1}^{k}(\bar{x_i}-\bar{\bar{x}})\sum_{j=1}^{n_i}(x_{ij}-\bar{x_i}) \\ &= 2\sum_{i=1}^{k}(\bar{x_i}-\bar{\bar{x}})(\sum_{j=1}^{n_i}(x_{ij}-\sum_{j=1}^{n_i}\bar{x_i})) \\ &= 2\sum_{i=1}^{k}(\bar{x_i}-\bar{\bar{x}})(\sum_{j=1}^{n_i}x_{ij}-n_i\bar{x_i}) \\ &= 2\sum_{i=1}^{k}(\bar{x_i}-\bar{\bar{x}})(\sum_{j=1}^{n_i}x_{ij}-n_i\bar{x_i}) \\ &= 2\sum_{i=1}^{k}(\bar{x_i}-\bar{\bar{x}})(\sum_{j=1}^{n_i}\bar{x_i}-\sum_{j=1}^{n_i}\bar{x_i}) \\ &= 0 \end{aligned} \tag{3}$$

Thus proving that:

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{\bar{x}})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x_j})^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x_i}-\bar{\bar{x}})^2}_{\text{SSB}} \tag{4}$$

□

If we look carrefully at the terms we can infer that the total variance decomposes in two very distinguished factors.

The first one is the between variance (sum of the variances of each group), whereas the second term is the between-groups variance (which is, nonetheless the variance of the means). Intuitively we should see some things:

- If the division of $\frac{SSB}{SSE}$ is < 1. This probably means both groups have the same mean, since the variability on the means is really low compared to the one

between groups. Whereas if $\frac{SSB}{SSE}$ is $>> 1$, we could probably reject the null hypothesis $H0$.

In order to quantify things properly we will define the magnitude $\frac{\frac{SSB}{k-1}}{\frac{SSE}{n-k}}$ This magnitude (under the hypothesis of homocedasticity and normality) should follow the F-distribution since SSB follows a Chi-square and so does SSE. By definition, the division of two Chi-squareds by its degrees of freedom follows the F-distribution.

Now that we have obtained a statistic that follows a distribution we are well suited to use $p-$values.

2.2.3 Test U de Mann-Whitney-Wilcoxon

2.2.4 Test Kolmogorov–Smirnov

3 Building of the platform

Building a platform from scratch can be a tedious job. Although many web creation editors exist (Wix, Wordpress, WebFlow), none of them offers a solution when we want to build complex things, with interactions and graphic generation.

Usually this web editors are suitable when you want to build an informative, static website. Or maybe an e-commerce (Shopify).
In my case, I need to communicate with a server (the backend, where I will run all algorithms) and for those complex websites you need to build everything from scratch.

3.1 Frontend

A website is created using the following technologies:

- HTML (Hypertext Markup Language): For building the schema. Every website is rendered using HTML. Browsers understand HTML and have the ability to render the tags.
- CSS (Cascading Style Sheets): For styling the HTML. This gives us the option to order HTML elements, put them as we desire, decide the colors, adapt the position of the elements if we have see a website using the phone or the computer, etc.
- JS (Javascript). JavaScript is a Turing Complete language. That is, we can perform any computation we desire. In web development it was introduced in order to simulate interaction between elements. For example, when a user logs in, we may want to display a popup that tells the user to put the email and password.

This are the basic web technologies. Over the years, and with the increased popularity of web, new complexities appeared. Building large web applications was getting more popular and new frameworks appeared.

A web framework provides the ability to automate and deal with the overhead of building a large web application. It lets you reuse components all over the website and helps you dealing with Database and server connections.
In this project case, we have decided to use Angular Framework. Angular is a web framework designed by Google which will help us reuse components, deal with URL routing, and backend connections. At the end, it helps us not write too much code, which would make this project extremely long and tedious.

3.2 Backend

We need to generate graphics, interact with the data, return p-values and this is not suitable to be done with Javascript. In our master, we have been using R and Python and this are the two most common languages nowadays into which perform data analysis.
Since Python is more widely used and has more documentation about how to build a server with it, I decided to go with this second option.

3.3 How to structure a backend

We know that we will have code running in Python somewhere in a server. But, how do we communicate with it? Traditional way (using SSH or TCP) don't apply in web develpment. We cannot require that much in a user. Our website should answer fast, without problems and without any more action needed from the user.
This is where an API comes into hand. An API (Application programming interface) is a software interface that let computers talk to each other. This middleman agent will allow the web in a computer client to talk with our server in an easy way. In this case we will set up an API that listens for requests, parses the request and sends it to our Python server. We have two options here:

- Flask framework. Flask is widely used with Python servers. It lets you listen to requests and send them and has been there for a long time. I

started using this solution but changed to FastAPI because the documentation was not that well written

- FastAPI. Extremely light web framework that does a great job validating inputs. It runs over Guvicorn and simplifies a lot the connection. The documentation is really understandable and easy to deploy.

3.4 Cloud

We need our backend server to be physically somewhere. Traditional hosting of servers has many drawbacks compared with the opportunities that cloud services offer. Advantages of using cloud vendors instead of hosting services:

- Pay for what you use. Instead of having one server for you all the time, you'll just need to pay for the amount of time your servers are being used. This is a great difference with respect to the traditional server hosting services and can reduce your monthly bill by a lot.
- Cloud vendors offer many more features than traditional web hosting. To list some of them:
  - CDN (Content Delivery Network): Cloud vendors can serve static content as close to the user as possible. Data is replicated to multiple places and a connection from South Korea - Spain that would take a lot (due to many DNS resolutions) can be easily be resolved thanks to the CDN that has duplicated your data and serves South Korea users from a DataCenter really close to them.
  - Security improvement. DDOS attacks are one of the most common attacks received. They consist in sending a huge spike of traffic to a server, up to the point that the maximum bandwidth of the server has been achieved and no more connections from trusted services can be answered. Cloud vendors offer you protection from these attacks. If they feel the traffic is malicious, they won't let it arrive to your servers. They also offer you services for secret managing, code analysers and many others.
  - Backups and data recovery are really easy to implement.

3.5 DNS Server

The web application was originally deployed in Firebase hosting using a mockup URL https://statistics-test-74f2e.web.app/. Firebase belongs to Google Cloud and offers us this option. The problem is that we may not want this domain name.

Domain name Services offers us a mapping between names (google.com) and the IP where the website currently is. This way we don't need to remember a 32-bit, but just a common name. In order to select another website name, we need to purchase it and then tell a DNS service where our website lives. For that, I've used Cloudfare as my DNS manager.

https://online.stat.psu.edu/statprogram/reviews/statistical-concepts/terminology
Turing Complete: https://www.youtube.com/watch?v=RPQD7-AOjMI