EXPERIMENTS SIGNIFICANCE ANALYSIS IN EMPIRICAL SCIENCES

Background

In this work we will study the distribution of significance p-values among experiments in the empirical sciences.

In statistics, the significance p-value is the probability of obtaining a test statistic (i.e. hypothesis test) that is at least as *extreme* as the one that was observed under the assumption that the null hypothesis is true. That is $P(T \geq O \mid H_0)$ where T is the probable outcome of a test statistic, O is the observed test and H_0 is the null-hypothesis. This value is correlated with the probability of a false positive (i.e. obtaining these results by chance). An example of such a significance test is the *t-test* invented by Gosset [1] (Published under the pseudo-name "BY STUDENT") in the early 1900's.

Moreover, the significance level (or the critical p-value) is the largest acceptable p-value for an experiment to be accepted as unlikely to have arisen by chance. These significance levels are set before holding the experiment and are usually set by the publishing party. Typical significance levels are 5% and 1%.

If all hypotheses that were reported to be correct are actually correct, then the distribution of their reported significance p-values should be "smooth", meaning that there should not be any discrimination towards any specific p-value.

Introduction

Since publishers set the significance levels, and because unsuccessful hypotheses tests do not get published as much as the successful attempts, authors want to stay in the region where their experiments' significance p-values are lower than the significance levels. This may sometimes lead to alterations in the experiments themselves in order to achieve publication.

As an example, consider that a control group of 200 subjects in a certain experiment achieved significance p-value that is larger than the required significance level but a subgroup of 150 subjects yielded a significance p-value that is "publishable", the authors may report the experiment using the subgroup results only and by that get published.

Godfrey [2] and Pocock *et al* [3] argue that multiple comparison problems are often not analyzed by appropriate procedures in medical publications. According to Pocock [4], the effect of multiplicity and selective publications is: 'perhaps the majority of trial reports claiming a treatment difference are false-positives.'

Our hypothesis is that the reported significance p-values in the empirical sciences are sometimes results of altered experiments and therefore some of the clinical trials reported as successful are to be considered false-positives. In this project we will test this hypothesis in the context of medical, pharmaceutical and healthcare experiments.

Resources

To the best of our knowledge, there is no ready (off the shelf) data set that would serve the needs of this project. Therefore, we will build our own data set by collecting papers from journals that report empirical hypothesis testing in the medical fields.

Approach

We will data mine the crawled publications, which will generally be in PDF format. For that, we will convert them to XML format. After converting the data to XML, the data mining process will become tractable. Our goal is to extract as much information as possible from the publication including, but not limited to:

1) Significance p-value
2) Type of significance test taken (t-test, Z-test, etc.)
3) Publisher
4) Author
5) Experiment keywords (HIV, ACHD, etc.)

We aim to analyze and check our hypothesis under different categorizations. For example, to analyze the significance p-values for a certain publisher, for a certain significance test type, for a specific keyword or a combination of these categories. We would also want to give evaluations on the estimated amount of alteration in experiments (if any).

Roadmap

1) Crawl papers and publications of clinical trials and medical experiments
2) For each paper
    a. Convert to XML
    b. Extract significance p-values, publisher's name, authors' names, type of significance test (t-test, Z-test, etc.), etc.
    c. Add the extracted information to the data-set
3) Analyze data

References

[1] B. Student and W. S. Gosset, "The Probable Error of a Mean," vol. 6, no. 1, p. 1, Mar. 1908.

[2] K. Godfrey, "Comparing the means of several groups," *New England Journal of Medicine*, 1985.

[3] S. Pocock and M. Hughes, "Statistical problems in the reporting of clinical trials," *New England Journal of*, 1987.

[4] S. Pocock, *Clinical Trials - A Practical Approach*. John Wiley & Sons, Inc., 1984.