ABSTRACT
There’s a strong belief in the empirical research community that there is a publication bias against null experimental results and non-significant experimental results. There have been many studies concerning this hypothesis in relatively small scale. In this paper, we study significance values in a large scale, validating around 25,000 published experiments significance values in the medical sciences. We present our findings regarding (i) publication bias and (ii) alteration in experiments and/or use of inappropriate statistical procedures in medical experiments. We show that both facts above are present in the considered dataset with high statistical confidence.

Keywords
Significance, p-values, analysis, ANOVA, t-test, publication bias, alteration in medical experiments.

1. INTRODUCTION
There’s a belief that a publication bias exists against publishing unsuccessful clinical trial attempts in experimental sciences. Easterbrook et al [1] study the publication bias in clinical trials against null results and non-significant trials. They “confirm a systematic selection bias in the publication process according to study results. Studies with a statistically significant result for the main outcome of interest were more likely to be submitted for publication and more likely to be published than studies with null results.” Even more research has been conducted with that emphasis. A survey of literature concerning these studies include Smith [2], Coursol and Wagner [3], Dickersins et al [4] and others.

Because of that well studied belief, researchers may 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.

Another effect on published significant experiments is the use of inappropriate procedures to test hypothesis. Godfrey [5] and Pocock et al [6] argue that multiple comparison problems are often not analyzed by appropriate procedures in medical publications. According to Pocock [7], the effect of multiplicity and selective publications is: “perhaps the majority of trial reports claiming a treatment difference are false-positives.”

Our conjuncture is that the reported significance p-values in the medical sciences are sometimes results of altered experiments or use of inappropriate procedures and therefore some of the clinical trials reported as successful are to be considered false-positives or non-significant. In this paper we test this conjuncture in large scale. About 35,000 publications including around 25,000 significance values for medical experiments are analyzed to verify this claim.

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

2.2 Data Collection
To collect data, we crawl medical publications from journals including Arthritis & Rheumatism, Cancer, British Journal of Pharmacology and many others. We collect the full text of a total of 35,000 publications.

We then data mine these publications to extract their reported significance values. We collect around 25,000 p-values from these publications with meta-data including the type of test resulting in the published p-value, degree(s) of freedom for the test, authors, journal and other underlying data.

3. DATA ANALYSIS METHODOLOGY
3.1 Introduction
Our goal is to assess, in large scale, (i) the level of bias in publications against null results or non-significant medical experiments and (ii) the level of alteration or use of inappropriate procedures in medical experiments.

For that, we construct a hypothesis to be tested in this study. Our hypothesis states that in medical experiments (i) there is publication bias against experimental results with significance values larger than 0.05 (which is the normative significance level in many cases), and (ii) there is use of inappropriate procedures and/or alterations in some of these experiments that shift significance values to lower values than they actually are.

3.2 Analysis Type
To test our hypothesis, we construct a null-hypothesis using a theoretical distribution that published p-values should follow under the assumptions of no publication bias and, no alteration and/or use of inappropriate procedures in experiments.

In more details, our null-hypothesis is that the \( \chi^2 \) distance between the observed distribution of p-values and the theoretical distribution described above is within an acceptable noise distance according to the Pearson’s \( \chi^2 \) test. Before holding the experiments or gathering the data, we set our significance level for the Pearson’s \( \chi^2 \) test to be 5%. That is, we will reject our null hypothesis only if the Pearson’s \( \chi^2 \) test produces a p-value less than 0.05.
3.3 Expected Distribution

As mentioned earlier, we perform the Pearson’s $\chi^2$ test to validate the observed distribution of p-values from the medical publications in our data set. For that matter, we need to construct an expected distribution of p-values and compare it to the observed distribution.

In this subsection, we discuss the construction of the p-values expected distribution. We start by giving a theoretical background of the student’s t-test for hypotheses testing in subsection 3.3.1 and lay down the basics for our methods of constructing an expected p-values distribution.

We then discuss, in subsection 3.3.2, a construction of the distribution by simulation. That is, we simulating p-values from t-distributions in case of t-test or F-distributions in case of ANOVA test in order to construct an expected distribution of p-values. In subsection 3.3.3 we discuss the construction of such a distribution by statistical theory.

Finally, in subsection 3.3.4, we discuss the validation of equivalence between constructing an expected distribution by simulation and a theoretical construction of an expected distribution. We perform the validation by running the Pearson’s $\chi^2$ test to compare the results of the two constructions.

3.3.1 Theoretical Background

We provide a quick overview of the theoretical background behind hypotheses testing. In general, a hypothesis $H$ can be tested by ensuing the following recipe

- Define an experiment to test the hypothesis $H$ (modify $H$ if needed)
- Choose a population and a sampling method
- Construct a null hypothesis $H_0$
- Construct a test statistic
- Choose a significance level and a sample size
- Run the experiment, analyze the results

We will review one way of computing a test statistic and testing its significance. The method is known as the Student’s t-test introduced by Gosset [8], in 1908, under the pseudo-name By Student.

In a nutshell, assuming two sample populations X and Y of sizes $n_X$ and $n_Y$, respectively, the test statistic compatible with the t-test is of the following form, assuming samples are of equal variance

$$ t = \frac{\bar{X} - \bar{Y}}{S_{xy} \cdot \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \tag{3.1} $$

where

$$ S_{xy} = \sqrt{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2} \quad \frac{n_1 + n_2 - 2}{n_X + n_Y} \tag{3.2} $$

and the total number of degrees of freedom is given by $d.f. = n_X + n_Y - 2$

In case the samples are of unequal variances, then

$$ t = \frac{\bar{X} - \bar{Y}}{S_{\tau,T}} \tag{3.3} $$

where

$$ S_{\tau,T} = \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}} \tag{3.4} $$

And the total number of degrees of freedom, in this case, is given by $d.f. = \frac{(s_X^2/n_X + s_Y^2/n_Y)^2}{(s_X^2/n_X - 1) + (s_Y^2/n_Y - 1)}$.

Note that $S_{\tau}^2$ is an unbiased variance estimator of group $i \in \{X,Y\}$.

Gosset [8] established that under the assumption that the null hypothesis $H_0$ is true, the test statistic $t$ is distributed according to the t-distribution, which is given by the probability density function

$$ P(t;df) = \frac{\Gamma \left( \frac{df + 1}{2} \right)}{\sqrt{\pi} \cdot \Gamma \left( \frac{df}{2} \right)} \left( 1 + \frac{t^2}{df} \right)^{-\frac{df+1}{2}} \tag{3.5} $$

where $df$ is the total number of degrees of freedom, $\Gamma$ is the gamma function and $t$ is the test statistic described in equations (3.1) and (3.3).

Thus, one could report the value $P(x > t; df)$ as an estimate of the extremeness of the results $t$ (one tailed test). In other cases, the value $P(x > t; df)$ can be reported (two tailed test). This value is called, the p-value or the significance value.

In the case where the null hypothesis is false, we expect this p-value to be low (indicating the unlikelihood of the results and their resulting test statistic). The p-value threshold for rejecting the null hypothesis is called the significance level. If the experiment’s significance value is lower than the significance level, we say that the null hypothesis is rejected with high statistical confidence (and therefore, the original hypothesis is accepted).

3.3.2 Construction by Simulation

As mentioned earlier, in order to perform the Pearson’s $\chi^2$ test we need a reference distribution that constitutes our hypothesis. A straightforward, intuitive way to construct such a distribution is by simulating a large number p-values and considering the distribution of the simulated p-values as the expected distribution. In this subsection, we will describe the process of simulating p-values that were results of t-tests. The same technique is used to generate simulated p-values distribution from ANOVA tests, but using the F-distribution instead. We will let the reader work the similar details for constructing an expected ANOVA p-values distribution by simulation.

The simulation process is intuitive and can be described as follows. To simulate a single p-value that is a sample from the expected distribution, we first create two effective samples of data from two different normal distributions $\left[ X_i \right]_{i=1}^{n_X} \sim Norm(\mu_1, \sigma_1)$ and $\left[ Y_i \right]_{i=1}^{n_Y} \sim Norm(\mu_2, \sigma_2)$, where $n_X, n_Y$ are the sample sizes, $\mu_1 \neq \mu_2$ are the means of the normal distributions and $\sigma_1, \sigma_2$ are the standard deviations of the normal distributions, respectively.

Then, we perform a two samples student t-test to compare $\left[ X_i \right]_{i=1}^{n_X}$ and $\left[ Y_i \right]_{i=1}^{n_Y}$ and record the result p-value.
We perform the described process 10,000 times to create a distribution of p-values (that were created under the assumption of false null-hypothesis since $\mu_0 \neq \mu_i$ for each iteration). We consider this an expected p-values distribution. Figure 1 is an example of such output for t-tests.

![Histogram of t-test p-values distribution constructed by simulation.](Image)

Figure 1: t-test p-values distribution constructed by simulation.

Note that, in each iteration, the simulation process can be performed with different sample sizes in order to create a better estimation of the expected p-values distribution. The reason is, the observed p-values we obtained come from different experiments with different numbers of subjects. We tried to estimate this effect by simulating p-values from different t-distributions taking into account the span of observed t-test degrees of freedom from the data set.

3.3.3 Theoretical Construction

Another way to construct an expected p-values distribution is by analyzing the t-distribution in case of t-test p-values or the F-distribution in case of ANOVA p-values.

We will discuss the construction of a theoretical expected p-values distribution from t-tests. The construction of a theoretical ANOVA p-values expected distribution can be done using the same technique and therefore will be left to the reader to fill in the details.

We’ll start off by considering the t-distribution probability density function given by equation (3.5). The density function’s graph, with 10 degrees of freedom, is demonstrated in Figure 2.

![t-distribution with 10 degrees of freedom](Image)

Figure 2: t-distribution with 10 degrees of freedom

This graph is the prior over the test statistics under the assumption that the null hypothesis is true. As defined earlier, the p-value (in case of a one tailed test) is $P(x > t; df)$. Since $cdf_{stdist}(t; df) = P(x < t; df)$, then the p-value can be defined as

$$p-value(t) = 1 - cdf_{stdist}(t; df)$$  \hspace{1cm} (3.6)

The graph of p-values vs. t-statistic with 10 degrees of freedom is shown in Figure 3.

![p-values vs. t-statistic with 10 degrees of freedom](Image)

Figure 3: p-values vs. t-statistic with 10 degrees of freedom

Assuming we want to build a histogram of p-values between $a$ and $b$ with bin size $s$, the percentage of t-statistics that yield p-values in the bin $[a_i, a_i + s]$ is equal to:

$$P(p \in [a, a + s] | p \in [a, b]) = \frac{t_{[a, a + s]}}{t_{[a, b]}}$$  \hspace{1cm} (3.7)

where $t_{[a, b]}$ is the length of interval of t-statistic values that yield p-values between $x$ and $y$. In a formal way

$$t_{[a, b]} = \sup(t_i - t_j)$$  \hspace{1cm} (3.8)

If we assume equal prior on experimental effects that may occur yielding each t-statistic, then the value in equation (3.7) is also equal to the percentage of possible experimental effects that yield p-values in the bin $[a_i, a_i + s]$.

This assumption isn’t completely accurate. Also, under this assumption the percentage of effects that will yield p-values in the bin $[0, s]$ under any full set $[0, b]: b \geq s$ isn’t well defined and by using bins $[t, t + s]: t \to 0$ the percentage will tend to 100%. Same problem, maybe less interestingly, will happen with any bin $[1 - s, 1]$ under any full set $[a, 1]$.

![Theoretical construction of t-test p-values distribution with 10 degrees of freedom](Image)

Figure 4: Theoretical construction of t-test p-values distribution with 10 degrees of freedom

Since we mostly care about analyzing the region around 0.05, our solution is to limit our full set to $[a, b]$ where $a > 0$ and $b < 1$. We will refer to these selections in section 4. Under this case, the
assumption we made is a good approximation. We will show that in the next subsection. Unless stated otherwise, when we mention the interval \([0,b]\) we refer to the interval \([10^{-3}, b]\). After running this scheme, the resulting p-values distribution (with 10 degrees of freedom) is shown in Figure 4.

3.3.4 Equivalence Validation

Figure 5: Equivalence test between the simulated distribution and the theoretically constructed distribution in the interval 0 to 0.09. With high confidence (p>0.99), the two histograms come from the same distribution.

In subsections 3.3.2 and 3.3.3 we provided two constructions of an expected distribution of p-values. We would like to measure the equivalence between the two constructions. For that matter, we perform the Pearson’s \(\chi^2\) test to compare the histograms obtained by the two constructions.

We compare the two histograms under different spans. As can be seen from Figure 5 and Figure 6, which summarizes the test results in a graphical manner, the two distributions can be considered equivalent with a high confidence, especially around the p-value 0.05 which is our region of interest.

Since both histograms are equivalent with a high confidence, the theoretical construction will be used, from now on, in order to evaluate the observed distribution of p-values. We pick the theoretical construction since it is an exact construction as opposed to the probabilistic nature of the construction by simulation.

4. RESULTS ANALYSIS

All pieces are in place for the Pearson’s \(\chi^2\) test, we constructed an expected p-values distribution for both ANOVA and t-test hypotheses tests. In the following subsections, we will discuss the analysis in details and present the results.

4.1 t-test Results Analysis

In this subsection, we discuss the analysis of the p-values distribution that resulted from t-tests and the process of comparing the p-values distribution that resulted from t-tests in the observed data to the expected p-values distribution we constructed in section 3.3.

The histogram of p-values in the range \([0, 0.1]\) is reported in Figure 7. Two abnormalities can be observed in Figure 7. The first is that after the p-value 0.05, there’s a suspicious sudden drop in the number of publications reporting these p-values. The second suspicious observation is the bump around the p-value 0.05. It is noteworthy to mention that the expected distribution of p-values is smooth in that region with relatively small derivative.

The more interesting set of questions is, how likely is this bump to occur by luck/noise? How unexpected is the sudden drop in the distribution? As mentioned earlier, a standard way to compare histograms is by performing the Pearson’s \(\chi^2\) test. We first compare the observed histogram and the expected histogram in the range \([0, 0.1]\) which is unable to reject the hypothesis that the two histograms come from the same distribution with \(p = 0.86\).

Figure 7: observed t-test p-values distribution in the range [0, 0.1]

However, a more careful test around \(p = 0.05\) (in the interval [0.03, 0.07]) rejects the hypothesis that the two histograms come
from the same distribution with \( p = 0.039 \). The results are graphically reported in Figure 8 and Figure 9.

4.2 ANOVA Results Analysis

We perform the same tests on the ANOVA p-values distribution. We would like to mention that the number of p-values that are results of ANOVA tests in our dataset is smaller than the number of p-values from t-tests (by a factor of 2), but still enough to perform analysis. In Figure 10, the observed p-values histogram from ANOVA tests is plotted.

4.3 Results Summary

The results summary can be found in Table 1.

<table>
<thead>
<tr>
<th>RANGE</th>
<th>TEST TYPE</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0, 0.1]</td>
<td>t-test</td>
<td>0.86</td>
</tr>
<tr>
<td>ANOVA</td>
<td></td>
<td>0.63</td>
</tr>
<tr>
<td>[0.03, 0.07]</td>
<td>0.039</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS

We showed that around \( p = 0.05 \), both observed p-values distributions, for t-test and for ANOVA don’t match the expected p-values distributions for t-test and for ANOVA, respectively. The significance of these conjunctures in terms of the Pearson’s \( \chi^2 \) test significance values were \( p = 0.039 \) for the t-test p-values distribution and \( p < 0.001 \) for the ANOVA p-values distribution. Both tests’ significance values are less than the significance level we set for the experiment, namely \( p = 0.05 \).

Our interpretation of these results is as follows:

1) The sudden drop in observed number of publications after \( p = 0.05 \) is an indication of publication bias against non-significant experiments and experiments with null results.

2) The bump (unexpected increase) in the number of publications right before \( p = 0.05 \) is an indication of experiments alterations for some experiments and/or use of inappropriate procedures in analyzing them. These experiments were reported significantly successful but should be either considered non-significant or false-positives.
6. FUTURE WORK
In this study, we performed analysis to validate significance values of experiments that were performed in the medical sciences. We divided our analysis by test type that resulted in the significance value reported. There are multiple ways to proceed from this point.

An interesting insight is that the FDA regulates experiments in the medical sciences. Moreover, medical researchers register experiments beforehand and therefore are more careful in reporting experiments results even for unsuccessful experiments. With that being said, we still found publication bias against unsuccessful experiments and possible experiments fixing in the medical sciences. We expect to see more of that trend in other, less regulated, sciences like psychology (non-drug related trials), social sciences, biology, etc. It is interesting to see if this effect is indeed present and possibly even more dominant in these fields.

Another interesting way to go is to analyze the distributions of reported p-values per experiment category. For instance, it would be insightful to understand the nature of experiments and analyze the level of bias and/or the use of inappropriate procedures in conducting these experiments in fields like HIV, cancer, etc.

7. ACKNOWLEDGMENTS
We would like to thank John Canny for his thoughtful insights and fruitful discussions. We also would like to thank Keng-hao Chang for his support and advice.

8. REFERENCES


