Author: Nico Herrig

Executive Summary

Evaluation of statistical power and size for parametric and non-parametric statistical tests using simulation techniques

Abstract

The size and power of a statistical test define their reliability under certain conditions. This paper examines how a Z-test and a χ^2-test perform in different scenarios through empirical simulation methods. Gender-grouped proportions of students admitted to a higher-education institute are compared, as the admission proportions of UC Berkeley from 1973 serve as a loose framework for data generation. The analysis shows that both tests are mainly affected by the number of applicants and the difference in proportions of acceptance. Given a distinct difference in proportions, significantly less applicants are needed for reliable results. In contrast, both tests show a lower level of reliability (expressed as power) if a large majority of applicants belong to one gender. For a given gender bias, the parametric test performs better for few applicants compared to the non-parametric test, while this difference decreases with increasing sample size. In a scenario where admission is not subject to a gender bias, test performance (expressed as size) shows no affection by the number of applicants and is generally in line with the underlying α-level used for testing.

Introduction

The scope of this paper is an empirical analysis of the statistical power and size of two appropriate tests to answer the research question whether UC Berkeley's student admission rate from 1973 was subject to a gender bias. Although the 1973 admission is a famous example for the Simpson paradox, this work will not answer the presence of a gender bias, but rather focus on the quality of statistical tests which could be used to answer such research question. As the statistical size and power of tests are a crucial factor in survey design , the conducted analysis gives valuable information about the sample size and the minimum effect size which are required for reliable results from both statistical tests.

Methodology

As size and power are analysed with empirical methods, artificial data is simulated under different scenarios and further used for analyzing the quality of a parametric testing method and a non-parametric counterpart. All testings are conducted at an α-level of 5%. The artificial data is loosely oriented on the real-world data from the *dslabs:admissions* data set[1], but altered to mimic different possible scenarios under which both tests are analysed. Data sets are simulated per gender, containing a binary variable which follows a given probability ($p_{male}; p_{female}$) of acceptance for the respective gender and containing the information of acceptance or rejection for each applicant. The expected gender bias (as proxy for the effect size d) is defined as $d = |p_{male} - p_{female}|$. The number of applications n can be seen as a proxy for a sample with n observations , resulting in n_{male} and n_{female} for the respective sample. For each simulated sample, an observed probability of acceptance, $\hat{p}_{male}$ and $\hat{p}_{female}$, is calculated and compared using a parametric and a non-parametric method. As the resulting $\hat{p}$ expresses a proportion of accepted students, it can be assumed that $\hat{p} \sim \mathcal{N}(\hat{p}, se(\hat{p})^2)$ if the interval $[\hat{p} - 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}};\ \hat{p} + 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$ lies completely within the interval $[0, 1]$ [2]. Under this assumption, a two-sided Z-test is used as a parametric testing procedure. Despite assuming that $\hat{p}$ follows a normal distribution, a χ^2-test is used to analyse a potential difference in statistical size and power between the parametric and non-parametric method. For each scenario, 1000 samples per gender are simulated and compared. The statistical power and size is then calculated from the 1000 results obtained, where power can be defined as the proportion of tests resulting in rejecting H_0 given H_1 is true ($Pr_{(H_1|H_1 true)}$). Given this logic, size is defined as $Pr_{(H_1|H_0 true)}$.

[1] (Irizarry & Gill, 2021)
[2] (Shafer & Zhang, 2012)

Scenarios

Scenario 1 - presence of gender bias (rejecting H_0): Under scenario 1, the general assumption is made that admission is subject to a gender bias and H_0 is treated as being rejected. Therefore, the following sub-scenarios 1.1 - 1.3 are computed to test the statistical power of both tests. Scenario 1.1.1 and 1.1.2 assumes a given gender bias, expressed as effect size, of 0.1 and 0.3, respectively, with a floating number of applicants from 10 to 1500, where $n_{male} = n_{female}$. Scenario 1.2 assumes a fixed number of applicants ($n_{male} = n_{female} = 100$) and analyses the test's power with increasing effect size from $d = 0$ to $d = 0.4$. Scenario 1.3 treats the effect size as well as the number of male applicants as fixed ($d = 0.1$, $n_{male} = 100$) and analyses the power for a change in n_{female} from 10 to 700.
Scenario 2 - no gender bias (not rejecting H_0): The underlying assumption for Scenario 2 is that admission is not subject to a gender bias ($p_{male} = p_{female}$, resulting in $d = 0$). As H_0 is treated as being true, the behavior of of both tests' statistical size with increasing number of applicants (n) are analyzed.
To test the statistical size and power of both tests, following parameters are assumed. Results are visually indicated by a red, dashed, vertical line.

Table 1: Scenario Parameters

	p male	p female	effect size	n male	n female
1.1.1	0.4	0.5	0.1	300	300
1.1.2	0.3	0.6	0.3	300	300
1.2	0.1	0.2	0.2	100	100
1.3	0.4	0.5	0.1	200	800
2.1	0.2	0.2	0.0	300	300

Results

Scenario 1.1 (1.1.1 & 1.1.2)

Computing the power of both tests as a function of n (the number of applicants) with a given effect size of 0.1 (scenario 1.1.1) and 0.3 (scenario 1.1.2), the following behavior presented in figure 1 can be observed. Both tests tend be behave in similar ways.

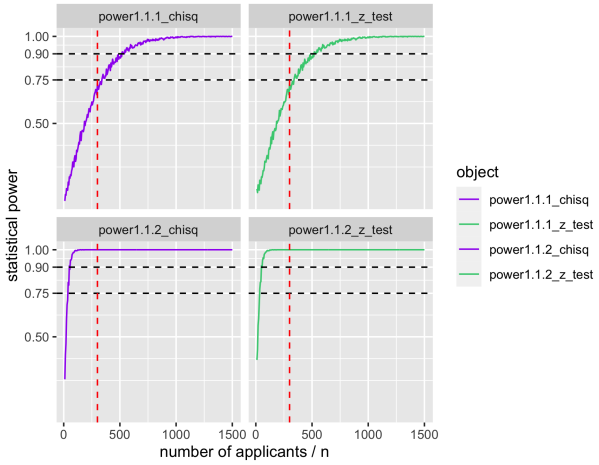

Figure 1: Scenario 1.1.1/1.1.2 - stat. power

Scenario 1.1.1 ($d = 0.1$) shows that at least 520 applicants per group are needed to reach a statistical power of 90%. In contrast, scenario 1.1.2 indicates that with a larger effect size ($d = 0.3$), only 50 observations per group are needed for the same power. Figure 1 also shows that the the parametric Z-test performs better for small sample sizes. Under the given parameters from table 1, the power of both tests is 0.729 for a gender bias of 0.1 and 1 for a gender bias of 0.3. Figure 2 shows the difference in total numbers between the Z-test and the χ^2-test. The findings underline the assumption that the parametric test offers a larger statistical power for small sample sizes. However, this difference decreases with increasing n.

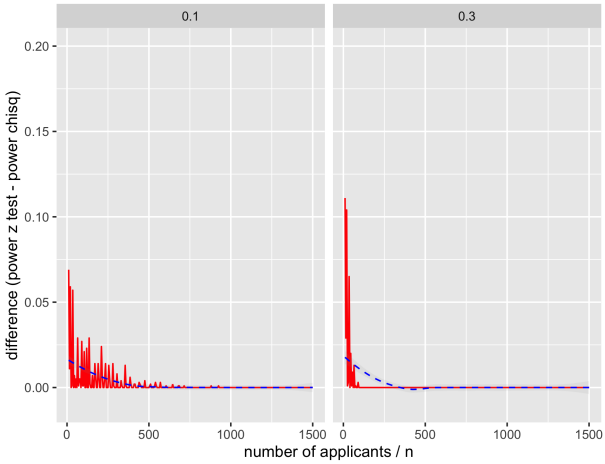

Figure 2: Scenario 1.1.1/1.1.2 - difference between tests

Scenario 1.2

Scenario 1.2 observes the statistical power as a function of the effect size d with $n_{male} = n_{female} = 100$. Figure 3. shows the behavior as d increases.

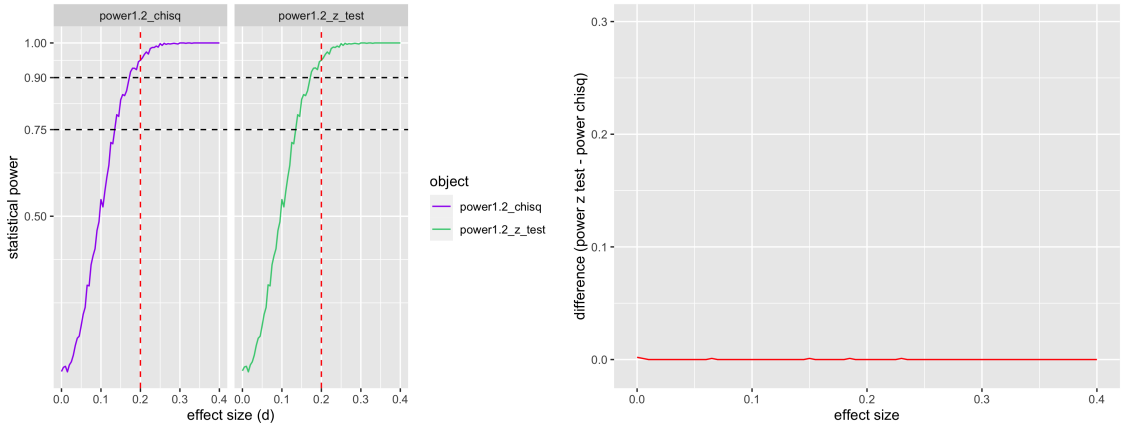

Figure 3: Scenario 1.2 - stat. power and difference

It can be inferred that, given $n = 100$, both tests require an effect size of 0.135 and 0.175, respectively, for a statistical power of 75% and 90%. Figure 3 also shows that the effect size does not affect the difference in

power between the parametric and non-parametric test. For an expected gender bias of $d = 0.2$ in admission proportions, a power of 0.95 can be observed for the parametric test and its counterpart.

Scenario 1.3

Scenario 1.3 underlies a fixed difference between p_{male} and p_{female} of 0.1 as well as a fixed rate of male applicants. It is of interest how the power of both tests behave when $n_{male} \neq n_{female}$.

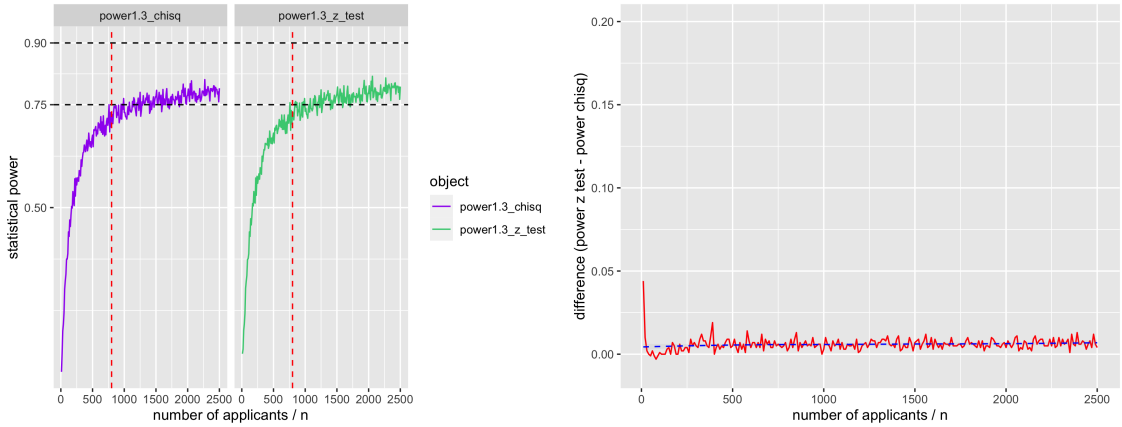

Figure 4: Scenario 1.3 - stat. power and difference

Comparing the outcome of scenario 1.3 ($n_{male} \neq n_{female}$) and 1.1 ($n_{male} = n_{female}$), it can be inferred that, given a comparably low sample size of group 1 (here: $n_{male} = 200$), the increase in power will stagnate at a certain level despite increasing the sample size of group 2 and shows a power of 0.727 (Z-test) and 0.725 (χ^2-test) for scenario 1.3 ($n_{female} = 800$). The difference between the power of both tests shows no distinct pattern and tends to stay the same for all n_{female}.

Scenario 2.1

Scenario 2.1 analyzes the statistical size of both tests. H_0 is treated as being true and $d = 0$. The size of both tests is therefore modeled as a function of n, where $n_{male} = n_{female}$.

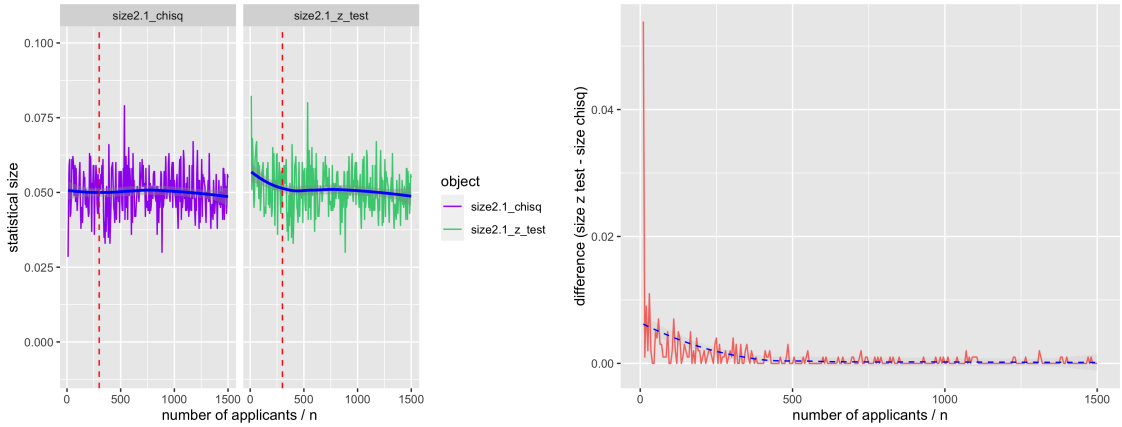

Figure 5: Scenario 2.1 - stat. size and difference

It can be inferred that the sample size n does not affect the statistical size of both the Z-test as well as the χ^2-test, as figure 5 does not show a distinct pattern when n increases. The statistical size tends to be at a constant level around 0.05, which is in line with the underlying alpha-level of 5 %. As for power in scenario 1, the Z-test shows a slightly higher size than its counterpart for smaller sample sizes. Given the parameters under this scenario, both tests show a statistical size of 0.054 and 0.053, respectively.

Summary & Conclusion

Table 2: Results - Overview

	1.1.1	1.1.2	1.2	1.3	2.1
Z-Test	0.729	1	0.95	0.727	0.054
ChiSq-Test	0.729	1	0.95	0.725	0.053

Answering the question on whether a university admission is subject to a gender bias, important metrics are the number of applicants as well as the observed difference in proportions of admitted students. Given a difference in proportions of 0.1, approximately 500 applicants *per gender* would be necessary for a 90% probability of detecting such difference and correctly rejecting H_0, while only $\frac{1}{10}$ of applications would be required given a bias of 0.3. A large difference between the number of female and male applicants, i.e., a small number of female applicants compared to male applicants, would cause an issue in getting a reliable test result. Such scenario is conceivable for subjects which tend to be historically dominated by a certain gender[3]. With regards to the UC Berkeley admission proportions of 1973, subject B from the dslabs::admissions data set shows such pattern, with 25 female applicants compared to 560 male applicants. In contrast, both tests produce reliable results around their underlying α-level if admission rates are not subject to a gender bias. Comparing both tests, the parametric Z-test tends to obtain a slightly larger power for small sample sizes. However, this difference decreases with increasing sample sizes.
Given the UC Berkeley's total number of applicants per gender (not grouped by subject) in 1973[4], both tests are expected to generate reliable results and answer the question of an underlying gender bias.

[3] i.e., Delaney & Devereux (2019) show a 22 % gender gap for STEM subjects
[4] $p_{male} = 0.445; p_{female} = 0.304; n_{male} = 2691; n_{female} = 1835; d = 0.141$

References:

1. Delaney, J. M., & Devereux, P. J. (2019). Understanding gender differences in STEM: Evidence from college applications. Economics of Education Review, 72, 219-238. doi:10.1016/j.econedurev.2019.06.002/
2. Irizarry, R., & Gill, A. (2021). Retrieved November 09, 2022, from https://CRAN.R-project.org/package=dslabs./
3. Shafer, D., & Zhang, Z. (2012). Introductory statistics. Minneapolis: Saylor Foundation Washington, DC.