11-2017

Using Simple Alternative Hypothesis to Increase Statistical Power in Sparse Categorical Data

Louis Mutter

Steven B. Kim

International Journal of Statistics and Probability; Vol. 6, No. 6; November 2017
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Using Simple Alternative Hypothesis to Increase Statistical Power in Sparse Categorical Data

Louis Mutter[1] & Steven B. Kim[1]

[1] Department of Mathematics and Statistics, California State University, Monterey Bay, Seaside, CA, USA

Correspondence: Steven B. Kim, Department of Mathematics and Statistics, California State University, Monterey Bay, Seaside, CA 93955, USA. Tel: 1-831-582-3954. E-mail: stkim@csumb.edu

Received: September 8, 2017 Accepted: September 22, 2017 Online Published: October 20, 2017

doi:10.5539/ijsp.v6n6p158 URL: https://doi.org/10.5539/ijsp.v6n6p158

Abstract

There are numerous statistical hypothesis tests for categorical data including Pearson's Chi-Square goodness-of-fit test and other discrete versions of goodness-of-fit tests. For these hypothesis tests, the null hypothesis is simple, and the alternative hypothesis is composite which negates the simple null hypothesis. For power calculation, a researcher specifies a significance level, a sample size, a simple null hypothesis, and a simple alternative hypothesis. In practice, there are cases when an experienced researcher has deep and broad scientific knowledge, but the researcher may suffer from a lack of statistical power due to a small sample size being available. In such a case, we may formulate hypothesis testing based on a simple alternative hypothesis instead of the composite alternative hypothesis. In this article, we investigate how much statistical power can be gained via a correctly specified simple alternative hypothesis and how much statistical power can be lost under a misspecified alternative hypothesis, particularly when an available sample size is small.

Keywords: chi-square goodness-of-fit test, simple alternative hypothesis, statistical power, likelihood ratio test, Neyman-Pearson Lemma, distractor analysis

1. Introduction

A researcher formulates a hypothesis based on scientific knowledge and then presents data to support the hypothesis. In this article, we focus on categorical data with three or more levels. In an expensive experiment or an observational study with a small sample size, despite researcher's deep and broad knowledge, the researcher may fail to provide empirical evidence due to a lack of statistical power. In such a case, many researchers may wish to increase statistical power without increasing sample size.

There are numerous methods of hypothesis testing for categorical data. Some tests are based on statistics with null sampling distributions following Chi-Square distributions (Pearson, 1900; Wilks, 1935; Neyman, 1949; Kullback, 1959), and some tests are based on discrete versions of goodness-of-fit statistics (Cramer, 1928; Kolmogorov, 1933; Smirnov, 1939; Anderson & Darling, 1952). These methods are equipped in statistical computing tools and widely used in practice.

We can perform power analysis by specifying a sample size, a significance level, a simple null hypothesis (denoted by H_0), and a simple alternative hypothesis (denoted by H_1). The power analysis can be done analytically (often based on asymptotic theory) or numerically (simulation). The general operating characteristic is that statistical power increases for a larger sample size, a larger significance level, and a larger degree of discrepancy between simple H_0 and simple H_1 (Cohen, 1988). Ampadu (2008) and Steel *et al.* (2009) compared various hypothesis tests for categorical data, and their results showed that the best test (in terms of statistical power) depends on H_1 when H_1 is true. For example, the Pearson's goodness-of-fit (GOF) test is outperformed by other tests when H_1 follows a monotonic trend, but it is competitive to the other tests when H_1 follows a triangular shape (Ampadu, 2008; Steel *et al.*, 2009).

Suppose a researcher can afford a small sample size. When there are multiple hypothesis tests under consideration, it is reasonable to choose the most powerful test under a specified H_1. The objective of our study is to compare statistical power when H_1 is simple and when H_1 is composite. For large sample sizes, we provide examples of power calculation and sample size calculation based on asymptotic theory. For small sample sizes ($n \leq 50$) we use simulation to study how much statistical power can be gained via a correctly specified H_1 and how much statistical power can be lost under a misspecified H_1 relative to the tests based on composite H_1.

2. Method

We define the following notation. Let K denote the number of levels in the categorical data. Let π_j denote the probability of observing the j-th level, where $\sum_{j=1}^{K} \pi_j = 1$. Let H_0 denote the null hypothesis, H_0: $\pi_j = p_{0j}$ for $j = 1, \ldots, K$. Let H_1 denote a simple alternative hypothesis, H_1: $\pi_j = p_{1j}$ for $j = 1, \ldots, K$. The sample size is denoted by n, and the

significance level is denoted by α. Let O_j denote the random variable which counts the number of cases in the j-th level for $j = 1, \ldots, K$, where $\sum_{j=1}^{K} O_j = n$.

2.1. Pearson's Chi-Square GOF Test

The Pearson's Chi-Square GOF test is based on the test statistic

$$\sum_{j=1}^{K} \frac{(O_j - E_j)^2}{E_j},$$

where $E_j = np_{0j}$ for $j = 1, \ldots, K$. The asymptotic null distribution is the Chi-Square distribution with $K - 1$ degrees of freedom. Given simple H_0: $\pi_j = p_{0j}$ and simple H_1: $\pi_j = p_{1j}$ for $j = 1, \ldots, K$, Cohen (1988) defined the effect size as

$$w = \sqrt{\sum_{i=j}^{K} \frac{(p_{0j} - p_{1j})^2}{p_{0j}}}.$$

If H_1 is true, the asymptotic distribution of the test statistic is the non-central Chi-Square distribution with $K - 1$ degrees of freedom and the non-centrality parameter $\lambda = nw^2$ (Ferguson, 1996).

Example 1. Suppose we have a sample of $n = 100$ randomly selected car accidents involving deaths. Let π_1, π_2, π_3, π_4, π_5, π_6, and π_7 denote the proportions of death-involved accidents on Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday, respectively. Let H_0: $\pi_j = 1/7$ for $j = 1, \ldots, 7$ and H_1: $\pi_1 = .15$, $\pi_2 = \pi_3 = \pi_4 = .1$, $\pi_5 = .15$, $\pi_6 = \pi_7 = .2$. Then the effect size for the Chi-Square GOF test is

$$w = \sqrt{\frac{2(1/7 - .1)^2}{1/7} + \frac{3(1/7 - .15)^2}{1/7} + \frac{2(1/7 - .2)^2}{1/7}} = \sqrt{.085}.$$

The non-centrality parameter is $\lambda = nw^2 = 8.5$, and the degrees of freedom is $K - 1 = 6$. At $\alpha = .05$, the rejection of H_0 occurs when the Chi-Square GOF test statistic exceeds 12.59, the 95-th percentile of $\chi^2(6)$. We can approximate statistical power .56, for example `1 - pchisq( 12.59, df=6, ncp=8.5 )` using R (R Core Team, 2016). If a researcher desires statistical power of .95, the required sample size is $n = 246$ which can be calculated by numerical search using `pchisq` or by using `pwr` package in R (Champely, 2017).

2.2. Log-Likelihood Ratio Test (Simple H_0 vs. Simple H_1)

For simple H_0: $\pi_j = p_{0j}$ and simple H_1: $\pi_j = p_{1j}$ for $j = 1, \ldots, K$, the log-likelihood ratio test statistic is given by

$$\Lambda = 2 \sum_{j=1}^{K} O_j \cdot \ln\left(\frac{p_{1j}}{p_{0j}}\right). \tag{1}$$

If we can find a constant such that Λ exceeds (or equal to) the constant with probability α, we can formulate the most powerful test at significance level α (Neyman & Pearson, 1933). The K-variate random vector $\vec{O} = (O_1, \ldots, O_K)^T$ follows the multinomial distribution with parameter $\vec{\pi} = (\pi_1, \ldots, \pi_K)^T$. When the sample size n is large, $\vec{O} \dot{\sim} N_K(n\vec{\pi}, n\Sigma)$, where $\Sigma = \text{diag}(\pi_1, \ldots, \pi_K) - \vec{\pi}\vec{\pi}^T$ is the covariance matrix. The test statistic Λ can be written as a linear combination $\Lambda = \sum_{j=1}^{K} c_j O_j$, where

$$c_j = 2 \cdot \ln\left(\frac{p_{1j}}{p_{0j}}\right).$$

Therefore, when n is large, $\Lambda \dot{\sim} N(n\vec{c}^T\vec{\pi}, n\vec{c}^T\Sigma\vec{c})$, where $\vec{c} = (c_1, \ldots, c_K)^T$. This means that we can find λ which satisfies $P(\Lambda \geq \lambda) \approx \alpha$ for large n. Let $\vec{p}_0 = (p_{01}, \ldots, p_{0K})$ and $\vec{p}_1 = (p_{11}, \ldots, p_{1K})$. Let $\Sigma_m = \text{diag}(p_{m1}, \ldots, p_{mK}) - \vec{p}_m\vec{p}_m^T$ be the covariance matrix under simple H_m for $m = 0, 1$. Under H_0, we can standardize the log-likelihood test statistic Λ as

$$Z_{0,\Lambda} = \frac{\Lambda - n\vec{c}^T\vec{p}_0}{\sqrt{n\vec{c}^T\Sigma_0\vec{c}}} \dot{\sim} N(0, 1). \tag{2}$$

Let z_k denote the k-th percentile of $N(0,1)$. Under H_1, the approximate statistical power is

$$\begin{aligned} 1-\beta &= P\left(\frac{\Lambda - n\vec{c}^T\vec{p}_0}{\sqrt{n\vec{c}^T\Sigma_0\vec{c}}} \geq z_{1-\alpha}\right) \\ &= P\left(\Lambda \geq z_{1-\alpha}\sqrt{n\vec{c}^T\Sigma_0\vec{c}} + n\vec{c}^T\vec{p}_0\right) \\ &= P\left(\frac{\Lambda - n\vec{c}^T\vec{p}_1}{\sqrt{n\vec{c}^T\Sigma_1\vec{c}}} \geq \frac{z_{1-\alpha}\sqrt{n\vec{c}^T\Sigma_0\vec{c}} + n\vec{c}^T(\vec{p}_0-\vec{p}_1)}{\sqrt{n\vec{c}^T\Sigma_1\vec{c}}}\right) \\ &= 1-\Phi\left(\frac{z_{1-\alpha}\sqrt{\vec{c}^T\Sigma_0\vec{c}} + \sqrt{n}\vec{c}^T(\vec{p}_0-\vec{p}_1)}{\sqrt{\vec{c}^T\Sigma_1\vec{c}}}\right), \end{aligned} \tag{3}$$

where Φ denotes the cumulative distribution function (CDF) for $N(0,1)$. Given α, β, H_0 and H_1, the required sample size is approximately

$$n = \left(\frac{z_\beta\sqrt{\vec{c}^T\Sigma_1\vec{c}} - z_{1-\alpha}\sqrt{\vec{c}^T\Sigma_0\vec{c}}}{\vec{c}^T(\vec{p}_0-\vec{p}_1)}\right)^2. \tag{4}$$

Example 2. It is continued from Example 1. For given $\vec{p}_0 = (1/7,\ldots,1/7)^T$ and $\vec{p}_1 = (.15,.1,.1,.1,.15,.2,.2)^T$, we can obtain the constant vector $\vec{c} = (.09758, -0.71335, -0.71335, -0.71335, .09758, .67294, .67294)^T$. Then, we can calculate statistical power $1-\beta = 1-\Phi(-1.2839) = \Phi(1.2839) = .90$ from Equation (3). If a researcher desires statistical power $1-\beta = .95$ at significance level $\alpha = .05$, we can calculate $n = 126$ from Equation (4), and it is nearly one half of $n = 246$ calculated in Example 1 for the Chi-Square GOF test.

2.3. Numeric Transformation

In this section, we discuss an alternative perspective of the standardized test statistic $Z_{0,\Lambda}$ of Equation (2). The standardized statistic

$$Z_{0,\Lambda} = \frac{\Lambda - n\vec{c}^T\vec{p}_0}{\sqrt{n\vec{c}^T\Sigma_0\vec{c}}} = \frac{\frac{1}{n}\Lambda - \vec{c}^T\vec{p}_0}{\sqrt{\frac{\vec{c}^T\Sigma_0\vec{c}}{n}}}$$

can be viewed as a test statistic for H_0: $\mu = \mu_0$, where $\mu_0 = \sum_{j=1}^K c_j p_{0j}$, by transforming the j-th categorical value to the numeric value $c_j = 2\cdot\ln\left(\frac{p_{1j}}{p_{0j}}\right)$. From this perspective, under H_0, we have

$$Z_{0,\Lambda} = \frac{\bar{X}-\mu_0}{\sigma_0/\sqrt{n}} \dot{\sim} N(0,1),$$

where $\bar{X} = \frac{1}{n}\Lambda = \frac{1}{n}\sum_{j=1}^K c_j O_j$ and $\sigma_0^2 = \sum_{j=1}^K (c_j-\mu_0)^2 p_{0j}$. Under H_1, when n is large, we have $\bar{X} \dot{\sim} N(\mu_1,\sigma_1)$, where $\mu_1 = \sum_{j=1}^K c_j p_{1j}$ and $\sigma_1^2 = \sum_{j=1}^K (c_j-\mu_1)^2 p_{1j}$.

Proposition. *Let Φ denote the CDF of $N(0,1)$, and let z_t be the percentile such that $\Phi(z_t) = t$. Assume $\mu = \mu_1$ is true. When n is large, in either case $\mu_1 > \mu_0$ or $\mu_1 < \mu_0$, the statistical power is approximately*

$$1-\beta = \Phi\left(z_\alpha\frac{\sigma_0}{\sigma_1} + \frac{|\mu_0-\mu_1|}{\sigma_1/\sqrt{n}}\right).$$

The proof of the proposition is provided in the appendix (Section 6). The proposition has three implications. First, the statistical power depends on the distance between the null value μ_0 and the alternative value μ_1 when $\mu = \mu_1$. (In Section 4, using simulation, we show that the statistical power is maintained closely even when the true value of μ is not exactly equal to μ_1.) Second, the statistical power also depends on the standard deviation σ_1, and we shall prefer smaller σ_1. Third, assuming $\vec{\pi} = \vec{p}_1$ is true, consider replacing c_j by another real number x_j for $j = 1,\ldots,K$. Then, for given $\vec{p}_0$ and $\vec{p}_1$, the means (μ_0 and μ_1) and the standard deviations (σ_0 and σ_1) depend on the choice of $(x_1,\ldots,x_K)$, so we write

$$h(x_1,\ldots,x_K) = z_\alpha\frac{\sigma_0}{\sigma_1} + \frac{|\mu_0-\mu_1|}{\sigma_1/\sqrt{n}}. \tag{5}$$

Let $x_1^*,\ldots,x_K^*$ be the values which maximize h, and let $\mu_0^* = \sum_{j=1}^K x_j^* p_{0j}$ and $\sigma_0^{*2} = \sum_{j=1}^K (x_j^*-\mu_0)^2 p_{0j}$. Then, the test statistic

$$Z_{0,\Lambda^*} = \frac{\bar{X}^*-\mu_0^*}{\sigma_0^*/\sqrt{n}} \dot{\sim} N(0,1), \tag{6}$$

must be more powerful than $Z_{0,\Lambda}$ asymptotically (at least not less powerful) at significance level α.

Example 3. Continuing from Example 2, under the same $\vec{p}_0 = (1/7, \ldots, 1/7)^T$, $\vec{p}_1 = (.15, .1, .1, .1, .15, .2, .2)^T$, $\alpha = .05$ and $n = 100$, our goal is to find a set of maximizers for the objective function $h(x_1, \ldots, x_K)$ in Equation (5). There is no closed-form solution, and we can use a numerical method (e.g., `optim` function in R). We find $x_1^* = 4.416653$, $x_2^* = x_3^* = x_4^* = 7.474607$, and $x_6^* = x_7^* = 2.164017$ are a set of maximizers. The transformation from categorical data to numeric data (e.g., Monday $\rightarrow$ 4.416653) leads to $\mu_0 = 5.083594$, $\sigma_0 = 2.238887$, $\mu_1 = 4.432985$, and $\sigma_1 = 2.198819$. Therefore, we can formulate the hypothesis testing as H_0: $\mu = 5.083594$ versus H_1: $\mu < 5.083594$ according to the numeric transformation. Note that

$$h(x_1^*, \ldots, x_K^*) = -1.644854 \left(\frac{2.238887}{2.198819} \right) + \frac{|5.083594 - 4.432985|}{2.198819/\sqrt{100}} = 1.2841 .$$

If $\vec{\pi} = \vec{p}_1$ is the truth (i.e., $\mu = 4.432985$), the approximate statistical power is $\Phi(1.2841) = .90$ which is substantially greater than .56 from the Chi-Square GOF test (Example 1) and fairly similar to $\Phi(1.2839) = .90$ from the log-likelihood ratio test (Example 2).

Figure 1 presents the statistical power of the Chi-Square GOF, the log-likelihood ratio and the numeric transformation methods using the asymptotic calculations. There is nearly no difference between the log-likelihood ratio and the numeric transformation, while both methods yield significantly greater power than the Chi-Square GOF when the specified H_1: $\vec{\pi} = \vec{p}_1$ is true.

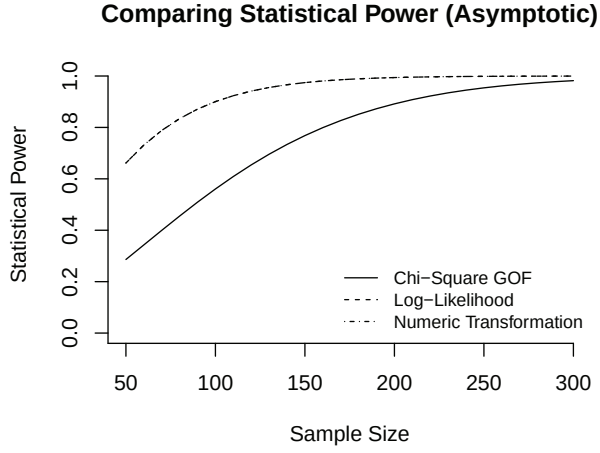

Figure 1. Comparing statistical power among Chi-Square GOF, log-likelihood ratio and numeric transformation

From practical perspective, our interest should be when the specified H_1 is not true and when n is small. It will be investigated numerically in the following section.

3. Simulation

We observed a remarkable increase in statistical power by the use of simple alternative hypothesis, and the log-likelihood ratio and the numeric transformation yielded nearly same statistical power according to the asymptotic calculations. For practical purpose, we considered scenarios when we have discrepancy between alternative $\vec{p}_1$ and true $\vec{\pi}$, particularly in small samples. We simulated data and approximated statistical power to investigate the impact of wrongly assumed H_1 in the numeric transformation (NT) and the simple-versus-simple log-likelihood ratio test (S-LR) for $n \leq 50$. We also compared NT and S-LR to other hypothesis tests studied in Ampadu (2008) and Steel *et al.* (2009) including Chi-Square GOF (χ^2 GOF), discrete Kolmogorov-Smirnov (DKS), log-likelihood ratio (LR), Freeman-Tukey (FT), power divergence (PD), discrete Cramer-von Mises (DCM), and discrete Anderson-Darling (DAD). The test statistics are as follow:

- χ^2 GOF: $\sum_{j=1}^{K} \frac{(O_j - E_j)^2}{E_j}$, where $E_j = np_{0j}$ is the expected count under H_0.
- DKS: $\max_{j=1}^{K} |Z_j|$, where $Z_j = \sum_{i=1}^{j} (O_i - E_i)$.
- LR: $2 \sum_{j=1}^{K} O_j \cdot \ln\left(\frac{O_j}{E_j}\right)$

- FT: $4\sum_{j=1}^{K}\left(\sqrt{O_j}-\sqrt{E_j}\right)^2$
- PD: $\frac{2}{\lambda(1+\lambda)}\sum_{j=1}^{K}O_j\left[\left(\frac{O_j}{E_j}\right)^{\lambda}-1\right]$ with $\lambda = 2/3$ (Cressie and Read, 1984)
- DCM: $\sum_{j=1}^{K}O_jZ_j^2$
- DAD: $\frac{1}{n}\sum_{j=1}^{K}\frac{Z_j^2 p_{0j}}{H_i(1-H_i)}$, where $H_j=\sum_{i=1}^{j}E_i$.
- S-LR: $2\sum_{j=1}^{K}O_j\cdot\ln\left(\frac{p_{1j}}{p_{0j}}\right)$ as defined in Equation (1).
- NT: $\frac{\bar{X}^*-\mu_0^*}{\sigma_0^*/\sqrt{n}}$ as defined in Equation (6).

3.1. Simulation Design

For simulation designs, we fixed $K = 5$ levels with the uniform null hypothesis H_0: $\pi_j = 1/5$ for $j = 1,\ldots,5$. We considered $n = 10$, 20, 30 and 50 for sample sizes and $\alpha = .05$ for significance level, we generated data under four scenarios (Scenarios A to D). The true probabilities were set at $\vec{\pi}$ = (.45, .19, .14, .12, .10) in Scenario A (monotonically decreasing), $\vec{\pi}$ = (.15, .15, .20, .25, .25) in Scenario B (step), $\vec{\pi}$ = (.27, .18, .10, .18, .27) in Scenario C (triangular), and $\vec{\pi}$ = (.095, .27, .27, .27, .095) in Scenario D (platykurtic). These designs were used by Steel *et al.* (2009).

The statistical power for S-LR and NT depends on the choice of simple alternative hypothesis H_1. For each scenario, we considered four cases for H_1. In Case 1, $\vec{p}_1$ was between $\vec{p}_0$ and $\vec{\pi}$. In Case 2, $\vec{\pi}$ was equal to $\vec{p}_1$. In Case 3, $\vec{p}_1$ had a stronger trend than $\vec{\pi}$ in the scenario. In Case 4, $\vec{p}_1$ was completely misspecified with an opposite trend of $\vec{\pi}$. Table 1 presents the four cases in each scenario.

Table 1. Four simulation scenarios (A, B, C and D) and four cases (1, 2, 3 and 4) in each scenario

	Scenario A: Decreasing	Scenario B: Step	Scenario C: Triangular	Scenario D: Platykurtic
$\vec{\pi}$ (truth)	(.45, .19, .14, .12, .10)	(.15, .15, .20, .25, .25)	(.27, .18, .10, .18, .27)	(.095, .27, .27, .27, .095)
$\vec{p}_1$ (Case 1)	(.325, .195, .17, .16, .15)	(.175, .175, .20, .225, .225)	(.235, .19, .15, .19, .235)	(.1475, .235, .235, .235, .1475)
$\vec{p}_1$ (Case 2)	(.45, .19, .14, .12, .10)	(.15, .15, .20, .25, .25)	(.27, .18, .10, .18, .27)	(.095, .27, .27, .27, .095)
$\vec{p}_1$ (Case 3)	(.55, .17, .13, .10, .05)	(.10, .10, .20, .30, .30)	(.30, .16, .08, .16, .30)	(.08, .28, .28, .28, .08)
$\vec{p}_1$ (Case 4)	(.10, .12, .14, .19, .45)	(.25, .25, .20, .15, .15)	(.13, .22, .30, .22, .13)	(.305, .13, .13, .13, .305)

To account for the discreteness of test statistics due to small sample size, we generated the null sampling distribution and the alternative sampling distribution of each test statistic by $m = 100,000$ repetitions. By letting $q_{.95}$ be the simulated 95-th percentile of the null sampling distribution and s_i be the i-th simulated value for the alternative sampling distribution, the statistical power was approximated by $\frac{1}{m}\sum_{i=1}^{m}1_{s_i\geq q_{.95}}$. By doing so, the probability of Type I error is at or below $\alpha = .05$ for any test statistic.

3.2. Simulation Result

Table 2 presents simulation results, and it addresses three key points. First, we could increase statistical power by the S-LR or the NT even when the specified H_1 was not exactly equal to the truth (Cases 1 to 3 under all Scenarios A to D). An increase in statistical power was sometimes more than double when compared to the other seven tests (χ^2GOF, DKS, LR, FT, PD, DCM and DAD). Second, when the specified H_1 was in an opposite trend of the truth, statistical power was close to zero (Case 4 under all Scenarios A to D). Third, NT and S-LR showed similar statistical power in many cases (with a difference less than .05), but they showed significantly different statistical power in some cases (e.g., Cases 1 to 3 under Scenario D with $n = 20$). The discreteness of test statistic in small samples could play a role in such a remarkable difference. We could not generalize the outperformance of NT over S-LR in Scenario D because we have not exhausted all simple alternative hypotheses.

Table 2. Simulation results (statistical power)

	Scenario A: Decreasing				Scenario B: Step				Scenario C: Triangular				Scenario D: Platykurtic			
n	10	20	30	50	10	20	30	50	10	20	30	50	10	20	30	50
χ^2 GOF	.294	.577	.754	.939	.060	.085	.128	.196	.083	.139	.231	.397	.115	.249	.436	.708
DKS	.370	.537	.767	.940	.083	.086	.151	.236	.097	.076	.120	.168	.032	.020	.035	.118
LR	.291	.538	.720	.927	.061	.097	.130	.210	.081	.168	.254	.409	.115	.287	.483	.738
FT	.215	.284	.660	.912	.066	.084	.127	.201	.091	.144	.248	.426	.142	.252	.463	.741
PD	.281	.573	.748	.939	.057	.100	.135	.206	.079	.163	.247	.412	.111	.291	.452	.725
DCM	.401	.692	.851	.973	.089	.155	.213	.329	.088	.107	.135	.198	.027	.054	.096	.257
DAD	.534	.771	.901	.985	.062	.081	.131	.217	.120	.154	.177	.245	.019	.138	.224	.418
S-LR (Case 1)	.587	.833	.939	.992	.151	.256	.327	.477	.248	.426	.569	.765	.407	.458	.805	.952
NT (Case 1)	.584	.828	.936	.992	.164	.258	.332	.477	.241	.389	.552	.749	.413	.666	.812	.954
S-LR (Case 2)	.589	.833	.940	.993	.170	.257	.330	.481	.248	.426	.556	.768	.407	.520	.803	.953
NT (Case 2)	.587	.833	.939	.993	.165	.259	.333	.478	.241	.423	.566	.768	.413	.666	.810	.954
S-LR (Case 3)	.582	.826	.934	.991	.152	.261	.331	.477	.248	.419	.563	.766	.407	.511	.809	.949
NT (Case 3)	.587	.826	.929	.988	.168	.261	.334	.478	.241	.423	.565	.767	.412	.665	.811	.953
S-LR (Case 4)	.001	.000	.000	.000	.006	.004	.002	.001	.002	.001	.000	.000	.000	.000	.000	.000
NT (Case 4)	.002	.000	.000	.000	.009	.003	.002	.001	.002	.001	.000	.000	.000	.000	.000	.000

4. Examples

4.1. Multiple-Choice Questions

An exam writer often designs multiple-choice questions to reduce the burden of grading. For a four-choice question (one correct answer and three distractors), let π_A, π_B, π_C and π_D denote the probability that each letter (A, B, C, and D) is a correct answer. Let H_0: $\pi_A = \pi_B = \pi_C = \pi_D = .25$ which is an ideal distribution. Students' common conception is that "C" is the most common answer for four-choice questions. Based on their common conception, let H_1: $\pi_A = .1$, $\pi_B = .25$, $\pi_C = .4$, $\pi_D = .25$. We analyzed a mathematics test written by a college professor which consists of $n = 40$ four-choice questions (one correct answer and three distractors). In the answer key, there were 5 A's, 11 B's, 13 C's, and 11 D's. The significance level of hypothesis testing was fixed at $\alpha = .05$, and we implemented each hypothesis test discussed in Section 3. As done in the simulation study, we generated the null sampling distribution of each test statistic and then calculated the p-value. The resulting p-values are given in Table 4. The two tests based on the simple alternative hypothesis (S-LR and NT) achieved the statistical significance, while the other tests could not.

Table 3. Resulting p-values

χ^2 GOF	DKS	LR	FT	PD	DCM	DAD	S-LR	NT
.324	.264	.278	.242	.293	.171	.074	.030	.028

Varying alternative hypothesis (H_1) after the calculation of p-value is not allowed in practice. For illustration purpose only, we considered another alternative hypotheses H_1: $\pi_A = .1$, $\pi_B = .2$, $\pi_C = .5$, $\pi_D = .2$. The resulting p-values were .045 for S-LR and .044 for NT. This example illustrates that S-LR and NT serve as an efficient test when we have a plausible alternative hypothesis based on accumulated experiences before observing data.

4.2. Distractor Analysis

A multiple choice question can be an effective method to assess students conceptual thinking (if well designed), and it reduces the burden of grading. The effectiveness of a multiple-choice question depends on its distractors, choices which serve as wrong answers (University of Wisconsin Oshkosh Testing Services, 2017). For example, in a four-choice question (one correct answer and three distractors), if two distractors are easily identified by students as wrong answers, the four-

choice question may seem to be a true-or-false question. An ideal (conditional) distribution of students choices on three distractors would be one-third for each distractor.

To assess students' understanding for the interpretation of a confidence interval, the following sentence was given in a quiz. "Based on a sample of size 132, a 95% confidence interval is calculated as (.48, .72) for the proportion of female students in the campus." Students were asked to select the correct interpretation among the following four choices: (A) 95% of 132 students in the sample were female, (B) Before collecting the sample, a 5% chance was allowed for missing the population proportion of female students, and an estimated proportion of female students is from .48 to .72 based on the collected sample, (C) There is a 95% chance that the true proportion of female students is between .48 and .72, and (D) If we take a sample from the population a large number of times, the true population proportion will fall between .48 and .72. If the three distractors (A), (C) and (D) are plausible, the null hypothesis H_0: $\pi_A = \pi_C = \pi_D = 1/3$ could be a reasonable assumption. Assuming (C) is the most common misconception and (A) is not as a strong distractor, the simple alternative hypothesis was specified as H_1: $\pi_A = .1$, $\pi_C = .6$, $\pi_D = .3$. Among the 68 students who took the test, $n = 26$ students selected one of the three distractors; 4 selected (A), 13 selected (C), and 9 selected (D), where the respective observed proportions are .154, .500, and .346.

The significance level of hypothesis testing was fixed at $\alpha = .05$, and we implemented each hypothesis test discussed in Section 3. As done in the simulation study, we generated the null sampling distribution of each test statistic and then calculated the p-value. The resulting p-values are given in Table 4. The two tests based on the simple alternative hypothesis (S-LR and NT) achieved the statistical significance, while the other tests could not.

Table 4. Resulting p-values

χ^2 GOF	DKS	LR	FT	PD	DCM	DAD	S-LR	NT
0.111	0.109	0.082	0.082	0.111	0.150	0.073	0.015	0.015

For illustration purpose only, we considered other alternative hypotheses. When H_1: $\pi_A = .25$, $\pi_C = .5$, $\pi_D = .25$, the resulting p-values were .047 for S-LR and .033 for NT. When H_1: $\pi_A = .3$, $\pi_C = .4$, $\pi_D = .3$, the resulting p-values were .056 for S-LR and .028 for NT. When H_1: $\pi_A = .4$, $\pi_C = .3$, $\pi_D = .4$, which is not supported by the observed data, the resulting p-values were .947 for S-LR and .946 for NT. This example illustrates the benefit of using S-LR and NT for experienced and knowledgeable researchers, but not for any researchers.

5. Discussion

It is difficult to reject H_0 with composite H_1 when n is small and particularly when K is large. When a researcher has specific scientific rationale and/or experience to argue a simple alternative hypothesis, statistical power can be significantly increased by the use of simple H_1 instead of composite H_1. The simulation results show that we can gain statistical power when a researcher specifies a correct trend such as decreasing, step, triangular, platykurtic or etc. For NT and S-LR, a simple H_1: $\vec{\pi} = \vec{p}_1$ does not have to be exactly the truth, and a loss of statistical power due to a small degree of discrepancy between simple alternative $\vec{p}_1$ and the truth $\vec{\pi}$ was negligible. In particular, a researcher can gain statistical power (relative to other tests based on composite H_1) when the direction of one-sided H_1 in terms of μ is consistent with the true value of μ. In other words, if we denote the null, alternative and true values of μ by μ_0, μ_1 and m_T, respectively, NT and S-LR have consistently shown higher statistical power than the other tests when $(\mu_1 - \mu_0)(m_T - \mu_0) > 0$. On the other hand, when $(\mu_1 - \mu_0)(m_T - \mu_0) < 0$, NT and S-LR have resulted in nearly zero power. The benefit of using a simple alternative hypothesis is (i) for those who know their scientific problems reasonably well and/or (ii) for those who have practically meaningful simple H_1 to be tested.

Acknowledgements

Louis Mutter was supported by the Undergraduate Opportunities Center (UROC) at California State University, Monterey Bay and the U.S. Department of Education (#P031C11021 and #P031C160221).

References

Ampadu, C. (2008). On the powers of some new chi-square type statistics. *Far East Journal of Theoretical Statistics, 26*, 59-72.

Anderson, T. W., & Darling, D. A. (1952). Asymptotic theory of certain goodness of fit criteria based on stochastic processes. *The Annals of Mathematical Statistics, 23*, 193-212. https://doi.org/10.1214/aoms/1177729437

Champely, S. (2017). pwr: Basic Functions for Power Analysis. R package version 1.2-1. https://CRAN.R-project.org/package=pwr

Cramer, H. (1928). On the composition of elementary errors. *Scandinavian Actuarial Journal, 1928*(1), 13-74. https://doi.org/10.1080/03461238.1928.10416862

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences, 2nd Edition*. Hillsdale, NJ: Lawrence Erlbaum.

Ferguson, T. S. (1996). *A course in large sample theory*. London: Chapman & Hall. https://doi.org/10.1007/978-1-4899-4549-5

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell Istituto degli Attuari, 4*, 83-91.

Kullback, S. (1959). *Information Theory and Statistics*. John Wiley, New York.

Neyman, J., & Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 231*, 289-337. https://doi.org/10.1098/rsta.1933.0009

Neyman, J. (1949). Contribution to the theory of the χ^2 test. *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability*, 239-273. https://projecteuclid.org/euclid.bsmsp/1166219208

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, 5*, 157-175. https://doi.org/10.1080/14786440009463897

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Smirnov, N. V. (1939). Estimate of deviation between empirical distribution functions in two independent samples. *Bull Moscow University, 2*, 3-16.

Steele, M. C., Smart, N. A., Hurst, C. P., & Chaseling J. (2009). Evaluating the statistical power of goodness-of-fit tests for health and medicine survey data. *The Modelling and Simulation Society of Australia and New Zealand Inc. (MODSIM) and the International Association for Mathematics and Computers in Simulation (IMACS)*, Cairns, Australia.

University of Wisconsin Oshkosh Testing Services (2017). Distractor & Effectiveness. Retrieved from: http://www.uwosh.edu/testing/faculty-information/test-scoring/score-report-interpretation/item-analysis-1/distractors-and-effectiveness.

Wilks, S. S. (1935). The likelihood test of independence in contingency tables. *The Annals of Mathematical Statistics, 6*, 190-196. https://doi.org/10.1214/aoms/1177732564

Appendix

The proof of the proposition in Section 2.3 is provided below.

Proof. When $\mu_1 > \mu_0$, the approximate statistical power for one-sided right tail test, H_0: $\mu = \mu_0$ and H_1: $\mu > \mu_0$, is

$$\begin{aligned} P(Z \geq z_{1-\alpha}) &= P\left(\frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} \geq z_{1-\alpha}\right) \\ &= P\left(\bar{X} \geq z_{1-\alpha}\frac{\sigma_0}{\sqrt{n}} + \mu_0\right) \\ &= P\left(\frac{\bar{X} - \mu_1}{\sigma_1/\sqrt{n}} \geq \frac{z_{1-\alpha}\frac{\sigma_0}{\sqrt{n}} + \mu_0 - \mu_1}{\sigma_1/\sqrt{n}}\right) \\ &= 1 - \Phi\left(z_{1-\alpha}\frac{\sigma_0}{\sigma_1} + \frac{\mu_0 - \mu_1}{\sigma_1/\sqrt{n}}\right) \\ &= \Phi\left(z_{\alpha}\frac{\sigma_0}{\sigma_1} + \frac{\mu_1 - \mu_0}{\sigma_1/\sqrt{n}}\right). \end{aligned} \tag{7}$$

Similarly, when $\mu_1 < \mu_0$, the approximate power for one-sided left tail test, H_0: $\mu = \mu_0$ and H_1: $\mu < \mu_0$, is

$$\begin{aligned} P(Z \leq z_{\alpha}) &= P\left(\frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} \leq z_{\alpha}\right) \\ &= P\left(\bar{X} \leq z_{\alpha}\frac{\sigma_0}{\sqrt{n}} + \mu_0\right) \\ &= P\left(\frac{\bar{X} - \mu_1}{\sigma_1/\sqrt{n}} \leq \frac{z_{\alpha}\frac{\sigma_0}{\sqrt{n}} + \mu_0 - \mu_1}{\sigma_1/\sqrt{n}}\right) \\ &= \Phi\left(z_{\alpha}\frac{\sigma_0}{\sigma_1} + \frac{\mu_0 - \mu_1}{\sigma_1/\sqrt{n}}\right). \end{aligned} \tag{8}$$

To this end, we can express both Equations (7) and (8) as

$$1 - \beta = \Phi\left(z_{\alpha}\frac{\sigma_0}{\sigma_1} + \frac{|\mu_0 - \mu_1|}{\sigma_1/\sqrt{n}}\right). \tag{9}$$