Technical Note

A *p*-Value Paradox in Proportion Tests and Its Resolution

Hening Huang [*]

Posted Date: 13 February 2025

doi: 10.20944/preprints202502.0999.v1

Keywords: *p*-value paradox; *p*-values; proportion test; *z*-test

Technical Note

A *p*-Value Paradox in Proportion Tests and Its Resolution

Hening Huang

Teledyne RD Instruments (retired), San Diego, CA 92127, USA; heninghuang1@gmail.com

Abstract: This technical note investigates a *p*-value paradox that emerges in the conventional proportion test. The paradox is defined as the phenomenon where "decisions made on the same effect size from data of different sample sizes may be inconsistent." It is illustrated with two examples from clinical trial research. We argue that this *p*-value paradox stems from the use (or misuse) of *p*-values to compare two proportions and make decisions. We propose replacing the conventional proportion test and its *p*-value with estimation statistics that include both the observed effect size and a reliability measure known as the signal content index (SCI).

Keywords: *p*-value paradox; *p*-values; proportion test; *z*-test

1. Introduction

P-values generated from null hypothesis significance testing (NHST) are commonly used to assess the significance of scientific findings in many fields, including medicine. Within the NHST framework, a *p*-value below 0.05 is typically seen as evidence of "statistical significance" and hence, "proves" that, for example, an intervention has an effect on the outcome of interest (Dunkler et al., 2020). However, many authors have acknowledged that *p*-values are often misunderstood, misinterpreted, and misused, giving rise to the so-called "*p*-value fallacy" (e.g. Goodman, 1999; Dixon, 2003; Nuzzo, 2015) and the "*p*-value paradox" (e.g. Chén et al., 2023).

In this technical note, we examine a *p*-value paradox that arises in the conventional proportion test, as defined by Chén et al. (2023): "decisions made on the same effect size from data of different sample sizes may be inconsistent." Chén et al. (2023) illustrated this paradox using a one-sample proportion test in clinical trial research. Bonovas and Piovani (2023) demonstrated a similar phenomenon in a two-sample proportion test. By exploring the true meaning of the *p*-value produced by the conventional proportion test, we will show that the conventional proportion test and its *p*-value are actually misused for comparing two proportions and making decisions.

In the following sections, Section 2 details the two examples of the *p*-value paradox mentioned above. Section 3 presents a resolution to the *p*-value paradox. Section 4 proposes an alternative to the conventional proportion test and its *p*-value. Section 5 provides conclusion and recommendation.

2. Two Examples of the *p*-Value Paradox

Chén et al. (2023) provided the following one-sample proportion test example to illustrate the p-value paradox:

> Suppose a clinician wanted to test whether the prevalence of a disease was 10%. To do so, the clinician selected a sample of 10 individuals, found that two of the 10 had the disease, and used evidence from the sample (20% sample incident rate) to make inferences about the population prevalence. With *p*=0.26, the hypothesis was not rejected.
>
> …, suppose we increased the sample size from 10 to 50, of which 10 had the disorder (the sample incident rate remained at 20%). This yielded a *p* value of 0.02. Although the new sample had the same (20%) incident rate, the null hypothesis was rejected under a

significance level of 0.05. This test, however, would still fail to reject the null under a significance level of 0.005. Now consider an even larger sample of 100, of which 20 had the disease (the sample incident rate remained 20%), but the *p* value was 0.002. The hypothesis was rejected under 0.005.

In this example, the observed effect size is 0.1 (the difference between the sample incidence rate 0.2 and the population prevalence 0.1), which remains constant across sample sizes of 10, 50, and 100. However, the corresponding *p*-values differ significantly: 0.26, 0.02, and 0.002, respectively, leading to inconsistent conclusions about statistical significance.

Bonovas and Piovani (2023) presented an example that underscores the frequent misuse and misinterpretation of *p*-values and "statistical significance" in the biomedical field. They analyzed two hypothetical placebo-controlled trials designed to determine whether a new drug reduces mortality in patients hospitalized with severe COVID-19 pneumonia, who were otherwise receiving standard-of-care. The first trial involved 200 participants, and the second included 850 participants in both the active drug and placebo groups. Both trials yielded the same risk ratio of 70%. However, the *p*-values given by the conventional two-sample proportion tests differed markedly: 0.11 in the trial with 200 participants versus 0.001 in the trial with 850 participants. These differing *p*-values led to inconsistent conclusions about the effectiveness of the new drug.

3. Resolution to the *p*-Value Paradox

3.1. What Does the p-Value Produced by the Conventional Proportion Test Really Mean?

To resolve the *p*-value paradox observed in the conventional proportion test, it is crucial to understand what the *p*-value truly represents when comparing two proportions.

Let $\hat{p}_1$ and $\hat{p}_2$ denote the observed proportions of events in Sample 1 and Sample 2, respectively, drawn from two populations of proportions p_1 and p_2, respectively. The sample sizes n_1 and n_2 are sufficiently large, so it is reasonable to assume that the sample proportions $\hat{P}_1$ and $\hat{P}_2$ are normally distributed: $\hat{P}_1 \sim N(\hat{p}_1, \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}})$ and $\hat{P}_2 \sim N(\hat{p}_2, \sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}})$, respectively. Let Δ denote the difference between the two sample proportions (two random variables), i.e. $\Delta = (\hat{P}_1 - \hat{P}_2)$, which is also normally distributed: $\Delta \sim N(\hat{p}_1 - \hat{p}_2, \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}})$, where $(\hat{p}_1 - \hat{p}_2)$ is the observed effect size and $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ is the standard deviation of the random variable $\Delta = (\hat{P}_1 - \hat{P}_2)$.

In the conventional two-sample proportion test, the test statistic Z represents the standardized effect size and is assumed to follow the standard normal distribution (i.e., $Z \sim N(0,1)$). Essentially, Z is obtained by taking the difference between the two sample proportions, $\Delta = (\hat{P}_1 - \hat{P}_2)$, subtracting the observed effect size, and then scaling by its standard deviation. That is, the process involves shifting Δ by the observed effect size and normalizing by its variability

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}. \tag{1}$$

On the other hand, the z-score is written as

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}. \tag{2}$$

Therefore, the proportion test becomes a z-test, and the two-tailed p-value for the two-sample z-test can be calculated as follows

$$p_{two-sample} = 1 - \Pr(-z \le Z \le z) = 1 - [\Pr(Z > -z) - \Pr(Z < -z)]. \tag{3}$$

To reveal the meaning of the p-value, we substitute the formulas for Z and z into Eq. (3). The term $\Pr(Z > -z)$ becomes

$$\Pr(Z > -z) = \Pr\left(\frac{(\hat{P}_1-\hat{P}_2)-(\hat{p}_1-\hat{p}_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} > -\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}\right) = \Pr(\hat{P}_1 > \hat{P}_2).$$

The term $\Pr(Z < -z)$ becomes

$$\Pr(Z < -z) = \Pr\left(\frac{(\hat{P}_1-\hat{P}_2)-(\hat{p}_1-\hat{p}_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} < -\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}\right) = \Pr(\hat{P}_1 < \hat{P}_2).$$

Thus, the p-value for the two-sample z-test can be rewritten as

$$p_{two-sample} = 1 - \left[\Pr(\hat{P}_1 > \hat{P}_2) - \Pr(\hat{P}_1 < \hat{P}_2)\right] = \psi_z, \tag{4}$$

where ψ_z is called the compatible probability (Huang, 2023). Therefore, $p_{two-sample}$ quantifies the degree of compatibility between the sampling distributions of the sample proportions $\hat{P}_1$ and $\hat{P}_2$. If $\Pr(\hat{P}_1 > \hat{P}_2) = \Pr(\hat{P}_1 < \hat{P}_2) = 0.5$, $p_{two-sample} = 1$, indicating that the two sampling distributions are completely compatible (for example, the two distributions completely overlap). If $\Pr(\hat{P}_1 > \hat{P}_2) = 1$, $\Pr(\hat{P}_1 < \hat{P}_2) = 0$, $p_{two-sample} = 0$, indicating that the two sampling distributions are completely incompatible (for example, the two distributions do not overlap). It is worth noting that the analysis and findings of this two-sample proportion test are similar to those of the usual two-sample z-test discussed by Huang (2024).

The one-sample proportion test is essentially a special case of the two-sample proportion test. In the one-sample test, we compare the observed proportion $\hat{p}_1$ against the known proportion p_2 (i.e. p_2 is treated as a fixed reference). Let n_2 be infinity in the above formulas. The difference between $\hat{P}_1$ and p_2 is also normally distributed: $(\hat{P}_1 - p_2) \sim N(\hat{p}_1 - p_2, \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}})$, where $(\hat{p}_1 - p_2)$ is the observed effect size and $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}$ is the standard deviation of the random variable $\Delta = (\hat{P}_1 - p_2)$. Then, the Z statistic becomes

$$Z = \frac{(\hat{P}_1 - p_2) - (\hat{p}_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}}. \tag{5}$$

The z-score becomes

$$z = \frac{\hat{p}_1 - p_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}}. \tag{6}$$

The p-value (two-tailed) for the one-sample z-test can be calculated as

$$p_{one-sample} = 1 - \left[\Pr(\hat{P}_1 > p_2) - \Pr(\hat{P}_1 < p_2)\right]. \tag{7}$$

Thus, $p_{one-sample}$ quantifies the degree of compatibility between the sampling distribution of the sample proportion $\hat{P}_1$ and the known proportion p_2. If the distribution of $\hat{P}_1$ is symmetric about p_2, then $\Pr(\hat{P}_1 > p_2) = \Pr(\hat{P}_1 < p_2) = 0.5$, and $p_{one-sample} = 1$. If the sampling distribution of $\hat{P}_1$ does not have any overlap with p_2, then $\Pr(\hat{P}_1 > p_2) = 1$, $\Pr(\hat{P}_1 < p_2) = 0$, and $p_{one-sample} = 0$.

It should be noted that the z-score shown in Eq. (6) differs slightly from the usual z-score used in a one-sample proportion test, which is typically expressed as $\sqrt{\frac{p_0(1-p_0)}{n}}$, where $p_0 = p_2$ and $n = n_1$. This formulation implies that, for Z to follow the standard normal distribution, the sampling distribution of $\hat{P}_1$ must be assumed to be $\hat{P}_1 \sim N(\hat{p}_2, \sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_1}})$. Because of this assumption, the usual one-sample proportion test is actually incompatible with the two-sample proportion test, which employs a different formulation for standardizing the effect size.

doi:10.20944/preprints202502.0999.v1

3.2. Resolution to the p-Value Paradox

The analysis above indicates that the *p*-value from the conventional proportion test quantifies the compatibility between either two sampling distributions or between a single sampling distribution and a known proportion. In other words, the *p*-value provides information about the behavior of the sampling distributions of the sample proportions, not about the underlying populations themselves.

According to the fundamental principle of scientific inductive reasoning, scientific claims must be based on statistical inference and domain knowledge about the *population* properties (or population information) of the quantity under consideration (e.g. effect size) (Huang 2024). Statistical inference relies on using sample statistics as estimates of the corresponding population parameters. However, in the conventional proportion test, the *p*-value is a sample statistic, for which no corresponding population parameter exists. In other words, the *p*-value does not provide inferential information about the underlying populations. Therefore, using the *p*-value to make decisions violates the fundamental principle of scientific inductive reasoning. The *p*-value paradox is actually caused by the use (or misuse) of the *p*-value to compare two proportions and make decisions. To avoid the *p*-value paradox, when comparing two proportions, the conventional proportion test and its *p*-value *should not be used* in the first place.

4. Alternative to the Conventional Proportional Test and Its *p*-Value

When comparing two proportions, we propose the following alternative approach to the conventional proportion test and its *p*-value: (1) evaluate the observed effect size using domain-specific knowledge to determine whether it is of practical importance, and (2) assess the reliability of the observed effect size using a descriptive statistic to ensure that it is a credible estimate of the true (population) effect size. This approach complies with the fundamental principle of scientific inductive reasoning.

For the problem under consideration, the true effect size is the difference between the two population proportions, $p_1 - p_2$. For the comparison of two sample proportions $\hat{p}_1$ and $\hat{p}_2$, the observed effect size is $\hat{p}_1 - \hat{p}_2$. Although we can evaluate the practical importance of this observed effect size directly using domain-specific knowledge, it is often useful to consider a normalized measure, i.e. the relative effect size (RES), defined as

$$\mathrm{RES} = \frac{\hat{p}_1 - \hat{p}_2}{\frac{1}{2}(\hat{p}_1 + \hat{p}_2)}. \tag{8}$$

For the comparison of a sample proportion $\hat{p}_1$ with the known population proportion p_2, the observed effect size is $|\hat{p}_1 - p_2|$ and the RES is defined as

$$\mathrm{RES} = \frac{\hat{p}_1 - p_2}{\frac{1}{2}(\hat{p}_1 + p_2)}. \tag{9}$$

In practice, the practical importance of the observed effect size or RES should be determined using domain-specific knowledge, while the reliability of the observed effect size (i.e. how precisely it estimates the true effect size) should be assessed using descriptive statistics that quantify its uncertainty.

There are several descriptive statistics available to measure the reliability of observed effect sizes, including the confidence interval and the standard error. Here, we recommend using the signal content index (SCI) as proposed by Huang (2019). In this context, the SCI is defined as

$$\mathrm{SCI} = \frac{\mathit{\Delta}_{ob}^2}{\mathit{\Delta}_{ob}^2 + u^2(\mathit{\Delta})}, \tag{10}$$

where $\mathit{\Delta}_{ob}$ is the observed effect size, $u^2(\mathit{\Delta})$ is the variance of $\mathit{\Delta}$, and $u(\mathit{\Delta})$ is called standard uncertainty (SU) in measurement science (e.g. JCGM, 2008).

For the comparison of two sample proportions $\hat{p}_1$ and $\hat{p}_2$, the variance of $\mathit{\Delta} = (\hat{P}_1 - \hat{P}_2)$ is given by

$$u^2(\mathit{\Delta}) = \mathrm{Var}(\hat{P}_1) + \mathrm{Var}(\hat{P}_2) = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}. \tag{11}$$

For the comparison of a sample proportion $\hat{p}_1$ with the known population proportion p_2, the variance of $\mathit{\Delta} = (\hat{P}_1 - p_2)$ is given by

$$u^2(\mathit{\Delta}) = \mathrm{Var}(\hat{P}_1) = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1}. \tag{12}$$

The signal content index (SCI) is defined based on the law of **conservation of energy;** it is the ratio of signal energy to the total energy (signal plus noise) (Huang 2019). In measurement science, the SCI for an estimate of the quantity of interest is a measure of the effectiveness of the measurement process or the "relative trueness" of the estimate (Huang 2019). SCI values range from 0 to 1. The interpretation of SCI is intuitive and meaningful. SCI values of 0.25, 0.5, and 0.75 correspond to small, moderate and high levels of effectiveness or trueness, respectively (Huang 2019).

For the problem under consideration, the observed effect size $(\mathit{\Delta}_{ob})$ represents the signal, while the variance of $\mathit{\Delta}$ represents the noise. Then, the SCI is an indicator of the reliability of the observed effect size. A high SCI value (e.g. close to 1) implies that the observed effect size is reliable, whereas a low SCI value (e.g. close to 0) suggests that the observed effect size is unreliable due to experiment noise. As a rule of thumb, if the SCI value is 0.75 or higher, the observed effect size is considered reliable and therefore credible. It is important to note that SCI depends on sample size; larger sample sizes typically yield higher SCI values, thus enhancing the reliability of the observed effect size. Conversely, if the SCI is too low (say, less than 0.5) the observed effect size is deemed unreliable, and additional samples or trials should be conducted to obtain a more precise estimate.

Table 1 shows the data and analysis results for the example of Chén et al. (2023). Table 2 shows the data and estimation results for the example of Bonovas and Piovani (2023).

Table 1. Data and analysis results for the example of Chén et al. (2023).

n	$\hat{p}_1$	p_2	Observed effect size	RES (%)	$u(\mathit{\Delta})$	SCI
10	0.2	0.1	0.1	66.7	0.126	0.38
50	0.2	0.1	0.1	66.7	0.057	0.76
100	0.2	0.1	0.1	66.7	0.040	0.86

Table 2. Data and analysis results for the example of Bonovas and Piovani (2023).

n	$\hat{p}_1$	$\hat{p}_2$	Observed effect size	RES (%)	$u(\Delta)$	SCI
200	0.14	0.2	0.06	35.3	0.037	0.72
850	0.14	0.2	0.06	35.3	0.018	0.92

As shown in Table 1, when the sample size is 10, the SCI value is only 0.38. This low SCI value indicates that the observed effect size of 0.1 (or a RES of 66.7%) is unreliable and should not be trusted. Consequently, the clinician should not make any inferences about the population prevalence based on this small sample. In contrast, for sample sizes of 50 and 100, the SCI values are 0.76 and 0.86, respectively. These higher SCI values suggest that the observed effect size of 0.1 (or RES of 66.7%) is reliable and credible when derived from the larger samples. Therefore, the clinician can confidently conclude that the population prevalence is 20% rather than 10%.

Table 2 shows that the SCI values are 0.72 for a sample size of 200 and 0.92 for a sample size of 850. These SCI values indicate that the observed effect size of 0.06 (or a RES of 35.3% and a risk ratio of 70%) is fairly reliable in the first trial and very reliable in the second trial. Therefore, the clinician can confidently conclude that the new drug is effective.

5. Conclusion and Recommendation

The *p*-value produced by the conventional proportion test is a sample statistic that does not provide inferential information about the underlying populations. In other words, the *p*-value is not an inferential statistic for comparing two proportions. Using the *p*-value to make decisions violates the fundamental principle of scientific inductive reasoning. The *p*-value paradox actually stems from the use (or misuse) of the *p*-value to compare two proportions and make decisions. Therefore, to avoid the *p*-value paradox, when comparing two proportions, the conventional proportion test and its *p*-value *should not be used* in the first place.

We propose replacing the conventional proportion test and its *p*-value with the following approach: (1) evaluate the observed effect size using domain-specific knowledge to determine whether it is of practical importance, and (2) assess the reliability of the observed effect size using the signal content index (SCI) to ensure that it is a credible estimate of the true effect size. This approach complies with the fundamental principle of scientific inductive reasoning.

Conflict of Interest: The authors declare no conflicts of interest.

References

Bonovas, S. & Piovani, D. (2023). On *p*-Values and Statistical Significance. *J. Clin. Med. 12*, 900. https://doi.org/10.3390/ jcm12030900

Chén, O. Y., Bodelet, J. S., Saraiva, R. G., Phan, H., Di, J., Nagels, G., Schwantje, T., Cao, H., Gou, J., Reinen, J. M., Xiong, B., Zhi, B., Wang, X., & de Vos, M. (2023). The roles, challenges, and merits of the p value. *Patterns (New York, N.Y.)*, *4*(12), 100878. https://doi.org/10.1016/j.patter.2023.100878

Dixon P. (2003). The p-value fallacy and how to avoid it. *Canadian journal of experimental psychology = Revue canadienne de psychologie experimentale*, *57*(3), 189–202. https://doi.org/10.1037/h0087425

Dunkler D, Haller M, Oberbauer R, Heinze G. (2020). To test or to estimate? P-values versus effect sizes. *Transpl Int.* 33(1), 50-55. doi: 10.1111/tri.13535. Epub 2019 Oct 21. PMID: 31560143; PMCID: PMC6972498.

Goodman S. N. (1999). Toward evidence-based medical statistics. 1: The *p* value fallacy. *Annals of Internal Medicine*. 130(12), 995–1004.

Huang H. (2019). Signal content index (SCI): a measure of the effectiveness of measurements and an alternative to *p*-value for comparing two means. *Measurement Science and Technology*, 31, 045008 https://doi.org/10.1088/1361-6501/ab46fd

Huang H. (2023). Probability of net superiority for comparing two groups or group means. *Lobachevskii Journal of Mathematics*, 44(11), 42-54.

Huang, H. (2024). Comments on "The Roles, Challenges, and Merits of the *p* Value" by Chén et al. *Basic and Applied Social Psychology*, 1–7. https://doi.org/10.1080/01973533.2024.2442957

Joint Committee for Guides in Metrology (JCGM) (2008). *Evaluation of Measurement Data - Guide to the Expression of Uncertainty in Measurement* (GUM 1995 with minor corrections). Sevres, France.

Nuzzo R. L. (2015). The inverse fallacy and interpreting P values. *PM & R: the journal of injury, function, and rehabilitation*, *7*(3), 311–314. https://doi.org/10.1016/j.pmrj.2015.02.011

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.