British Journal of Mathematical and Statistical Psychology (2019), 72, 1–17

The British Psychological Society

www.wileyonlinelibrary.com

A note on Type S/M errors in hypothesis testing

Jiannan Lu[1]*, Yixuan Qiu[2] and Alex Deng[1]

[1]Microsoft Corporation, Redmond, Washington, USA
[2]Purdue University, West Lafayette, Indiana, USA

Motivated by the recent replication and reproducibility crisis, Gelman and Carlin (2014, *Perspect. Psychol. Sci.*, *9*, 641) advocated focusing on controlling for Type S/M errors, instead of the classic Type I/II errors, when conducting hypothesis testing. In this paper, we aim to fill several theoretical gaps in the methodology proposed by Gelman and Carlin (2014, *Perspect. Psychol. Sci.*, *9*, 641). In particular, we derive the closed-form expression for the expected Type M error, and study the mathematical properties of the probability of Type S error as well as the expected Type M error, such as monotonicity. We demonstrate the advantages of our results through numerical and empirical examples.

1. Introduction

The recent replication and reproducibility crisis in psychological science (e.g., Anderson & Maxwell, 2017; Fiedler & Schwarz, 2016; Pashler & Wagenmakers, 2012) and within the broader scientific community (Baker, 2016; Ioannidis, 2005a, 2005b) has rekindled the debate on the highly controversial null hypothesis significance testing framework (NHST, Lehmann & Romano, 2006) that has lasted for over half a century (Krantz, 1999; Rozeboom, 1960). Among the critics, for example, Efron (2013) called out NHST for being 'opportunistic', because it only accumulates evidence against the null hypothesis. Bayarri, Benjamin, Berger, and Sellke (2016) pointed out that NHST has been 'overly relied on' by the scientific community, and Cumming (2014) stressed the 'need to shift from reliance on NHST to estimation and other preferred techniques'. Moreover, to replace the NHST framework, several researchers and practitioners (e.g., Berger, Boukai, & Wang, 1997; Deng, 2015; Deng, Lu, & Chen 2016; Johnson, 2013b; Kass & Raftery, 1995; Kruschke, 2013; Rouder, Speckman, Sun, Morey, & Iverson, 2009) have proposed several alternative frameworks, most of which are Bayesian in nature. Among the advocates, Gigerenzer and Swijtink (1990) praised NHST as the 'essential backbone of scientific reasoning'. Through several examples, Hagen (1997) illustrated the 'elegance and usefulness' of NHST. In a response to Cumming (2014), Morey, Rouder, Verhagen, and Wagenmakers (2014) argued that 'hypothesis tests are essential for psychological science'. For a comprehensive review of both the advantages and the pitfalls of NHST in psychological science, see Nickerson (2000).

As pointed out by several researchers (e.g., Senn, 2001), much of the criticism of the NHST framework is because of the scientific community's distorted obsession with

*Correspondence should be addressed to Jiannan Lu, One Microsoft Way, Redmond, WA 98004, USA (email: jiannl@microsoft.com).
The first two authors contributed equally to this work. This majority of the work was conducted when the second author was a research intern at Microsoft Corporation.

DOI:10.1111/bmsp.12132

p-values, which are essentially the cornerstone of NHST. To clarify the misconceptions about *p*-values, Wasserstein and Lazar (2016) issued an official statement on behalf of the American Statistical Association, which among other things emphasized that 'the widespread use of "statistical significance" (generally interpreted as $p \leq .05$) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process' and urged the scientific community to provide 'full reporting and transparency' in research communications. Indeed, *p*-values are often misinterpreted (Goodman, 2008; Huber, 2016; Peng, 2015), and, more importantly, are prone to (intentional or unintentional) human manipulations, often referred to as 'hacking' or 'cherry-picking' (Head, Holman, Lanfear, Kahn, & Jennions, 2015; Taylor & Tibshirani, 2015). Simmons, Nelson, and Simonsohn (2011) provided examples of such inappropriate practices in psychological science, which they acutely referred to as the 'researcher degrees of freedom'. Nevertheless, despite the controversies and criticisms, the NHST framework has remained the mainstream approach for scientific reporting. Exceptions include academic journals such as *Basic and Applied Social Psychology*, whose editorial team decided to ban the use of NHST, because 'the $p < .05$ bar is too easy to pass and sometimes serves as an excuse for lower quality research' (Trafimow & Marks, 2015).

To rebuild the credibility of the NHST framework and circumvent potential replication and reproducibility crises in the future, several researchers (e.g., Johnson, 2013a) have advocated establishing 'revised standards' (i.e., more stringent thresholds) for statistical significance. Others such as Gelman (2016) pointed out a more fundamental problem with typical NHST analyses: the lack of 'greater acceptance of uncertainty and embracing of variation'. Indeed, *p*-values and the associated power calculations only account for Type I/II errors, and may ignore other errors which are of more practical concern or have policy implications. To address this issue, in two illuminating papers Gelman and Tuerlinckx (2000) proposed two new statistical errors named Type S (sign) and Type M (magnitude) respectively, and Gelman and Carlin (2014) recommended calculating the probability of the Type S error and the expected Type M error, which are two guardrail metrics reflecting the trustworthiness of the NHST analysis conducted. Although conceptually sound, the 'applied' nature of Gelman and Carlin's (2014) paper resulted in a lack of in-depth theoretical discussions, which we aim to address in this paper. We believe that by filling the theoretical gaps, our work facilitates a better understanding of the proposed methodology by Gelman and Carlin (2014), which has been well received by the psychology community (Lishner, 2015).

The remainder of this paper is organized as follows. Section 2 introduces Type S/M errors in NHST, and reviews several measures to quantify them. Section 3 derives the closed-form expressions of the said measures, based on which some mathematical proprieties (e.g., bounds and monotonicity) are studied. Section 4 provides numerical and empirical examples to illustrate our theoretical results. Section 5 concludes and discusses future directions. We relegate all the proofs and other technical details to Appendices A and B.

2. Hypothesis testing, Type I/II/S/M errors, and their measures

2.1. Hypothesis testing and Type I/II/S/M errors

To better illustrate the methodology proposed by Gelman and Carlin (2014), we consider testing whether a true effect size μ is non-zero, that is,

$$H_0 : \mu = 0 \quad \text{versus} \quad H_1 : \mu \neq 0.$$

The above two-sided setting is common in psychological research, and we follow previous discussions (e.g., Gelman & Carlin, 2014; Gelman & Tuerlinckx, 2000) and adopt it throughout this paper, for the purpose of illustration. It is worth mentioning that our results can be trivially extended to other settings such as one-sided (which we discuss at the end of this section).

Assume that the test statistic follows a normal distribution:

$$Z \mid \mu, \sigma \sim N(\mu, \sigma^2) \quad (\sigma > 0).$$

As pointed out by Gelman and Tuerlinckx (2000), such setting is made plausible by the central limit theorem, and it covers a wide range of scenarios in psychological science, including averages, differences (e.g., between treatment and control groups in randomized experiments), and linear regression coefficients. The traditional hypothesis testing framework focuses on the following two statistical errors:

1. Type I, rejecting the null hypothesis H_0 when it is true;
2. Type II, failing to reject the null hypothesis H_0 when it is not true.

The key idea behind the traditional hypothesis testing is the significance level α, which controls the Type I error rate. To be more specific, let $z_\alpha = \Phi^{-1}(1 - \alpha/2)$ be the two-sided threshold value, and we reject the null hypothesis H_0 when $|Z| > z_\alpha$ so that

$$\Pr(|Z| > z_\alpha \mid H_0) = \alpha.$$

Although Type I/II errors are undoubtedly the cornerstone of the NHST framework, Gelman and Carlin (2014) argued that controlling for these two errors is insufficient to fully capture the risks of NHST analyses. Realizing this potential pitfall, Gelman and Carlin (2014) proposed to focus on two new errors:

1. Type S (sign), the test statistic Z is in the opposite direction to the effect size μ, given that it is statistically significant;
2. Type M (magnitude), the test statistic Z in magnitude exaggerates the effect size μ, given that it is statistically significant.

These two errors are of more practical concern and have policy implications, compared to the classic Type I/II errors. First, we may want to avoid the scenario where an actually harmless (beneficial) treatment is declared significantly harmful. Second, even if we correctly identify the sign of the treatment, we may prefer not to overstate its actual treatment effect.

To illustrate the definitions of Type S/M errors, we consider the case where $Z \sim N(0.5, 1)$. Figure 1 contains 5,000 repeated samplings of Z. First, the grey round points correspond to statistical non-significance. Second, the black triangular points correspond to occurrences of the Type S error. Third, the black (triangular and squared) points correspond to occurrences of the Type M error, because in this case all of them inevitably overestimate the true effect size $\mu(= 0.5)$ in magnitude.

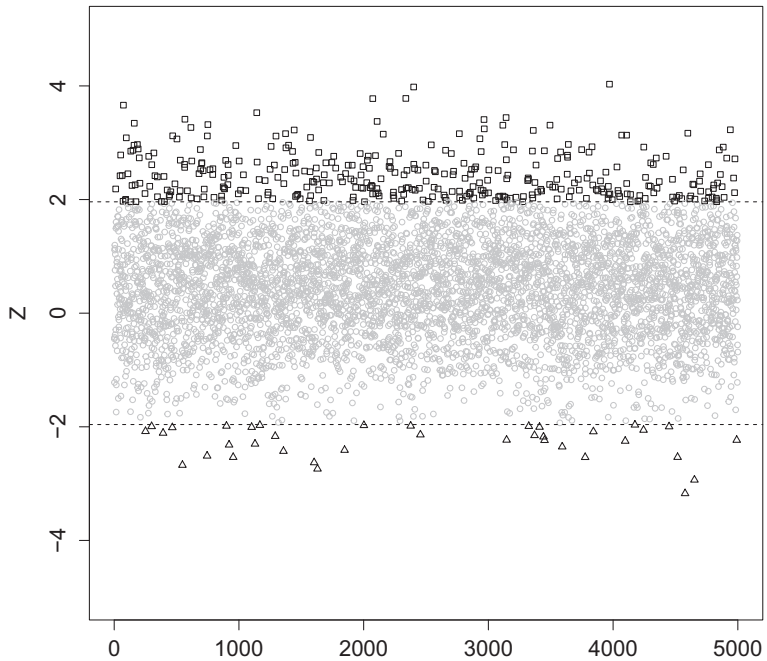


Figure 1. Five thousand repeated samplings of Z from $N(\mu = 0.5,\ \sigma = 1)$. The grey round points correspond to statistically non-significance. The black triangular points correspond to occurrences of the Type S error (statistically significant but in the opposite direction to μ). The black (triangular and squared) points together correspond to occurrences of the Type M error (statistically significant but overestimating the magnitude of μ).

2.2. Measures of Type II/S/M errors

Under the classic NHST framework, to access the credibility of the hypothesis test, we calculate the power function, which is a function of μ and formally defined as the probability of rejecting the null hypothesis H_0 under a non-zero μ (i.e., H_0 is indeed false):

$$p = \Pr(|Z| > \sigma z_\alpha \mid \mu). \tag{1}$$

By definition, the power function is one minus the probability of the Type II error. Therefore, the smaller the power is, the more likely it is that the Type II error occurs.

Having proposed new Type S/M errors, Gelman and Carlin (2014) advocated calculating two additional guardrail metrics for hypothesis testing, in conjunction with the traditional power calculation, which measure the severities of the Type S/M errors, respectively. Again, the two metrics are both functions of μ (without loss of generality, we assume that $\mu > 0$):

1. the probability of the Type S error,

.

$$s = \Pr(Z < 0 \mid \mu, |Z| > \sigma z_\alpha); \tag{2}$$

2. the expected Type M error ('exaggeration ratio'),

.

$$m = E(|Z|\,|\,\mu, |Z| > \sigma z_\alpha)/\mu. \tag{3}$$

Gelman and Carlin (2014) advocated taking (1)–(3) into account simultaneously when conducting NHST analyses, and through numerical and empirical examples demonstrated the benefits of adopting the two guardrail metrics (2) and (3).

To illustrate the proposed measures in (1)–(3), we revisit the simulation results in Figure 1, under the setting where $\mu = 0.5$ and $\sigma = 1$. First, by (1) the power can be approximated by the proportion of red/blue points, which is .0796 in this case. Intuitively, such lower power study typically will suffer from Type S/M errors, as we will show later. Second, by (2) the probability of Type S error can be approximated by the proportion of red points among the red/blue, which is .093 in this case. Third, by (3) the exaggeration ratio can be approximated by the average absolute value of the red and blue points divided by 0.5, which is 4.82 in this case.

The above numerical evaluations can help us determine the values of the three measures in (1)–(3) under different settings. However, as mentioned before, we believe that a more thorough theoretical study is imperative to ensure better understanding of the proposed methodology regarding Type S/M errors. In particular, only simulation code was provided by Gelman and Carlin (2014) to compute (3), and we believe that a closed-form expression for would be beneficial for both practitioners and methodology researchers. On the one hand, closed-form expressions enable faster and more accurate computations. On the other hand, we can more conveniently study mathematical properties and gain insights from them.

To end this section, we briefly discuss the one-sided analogues of the probability of the Type S error and expected Type M error in (2) and (3), respectively. In this case, without loss of generality, the null and alternative hypotheses are H_0: $\mu = 0$ and H_1: $\mu > 0$, respectively. Consequently, the corresponding probabilities of the Type S error and expected Type M error are

$$s_{\text{one-sided}} = \Pr(Z < 0 \mid \mu, Z > \sigma z_{2\alpha}), \quad m_{\text{one-sided}} = E(Z \mid \mu, Z > \sigma z_{2\alpha})/\mu,$$

respectively, where $z_{2\alpha} = \Phi^{-1}(1 - \alpha)$.

3. Theory behind Type S/M errors

3.1. Closed-form expressions

We provide the closed-form expressions for the power function, the probability of Type S error, and the exaggeration ratio, defined in (1)–(3), respectively. It is worth mentioning that, although the closed-form expression for (1) is well known among the statistical and psychological science communities, and that for (2) is rather straightforward to obtain and has already been tersely sketched by Gelman and Carlin (2014), we choose to explicitly derive both in the Appendix B, to ensure that this paper is self-contained.

Theorem 1. *Let $\Phi(\cdot)$ and $\phi(\cdot)$ denote the cumulative distribution function and the probability density function of the standard normal distribution, respectively, and $\lambda = \mu/\sigma$ be the 'signal' to 'noise' ratio. The closed-form expressions for (1) and (2) are*

$$p = \Phi(-z_\alpha - \lambda) + 1 - \Phi(z_\alpha - \lambda) \tag{4}$$

and

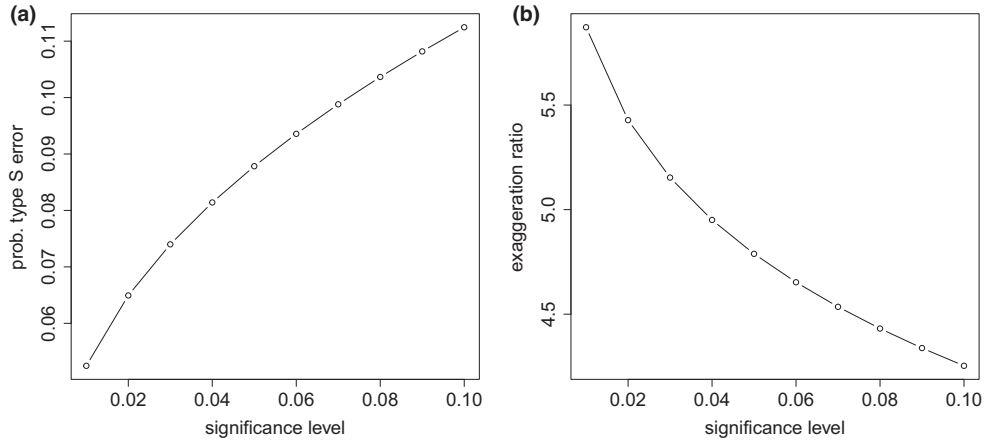
$$s = \frac{\Phi(-z_\alpha - \lambda)}{\Phi(-z_\alpha - \lambda) + 1 - \Phi(z_\alpha - \lambda)}, \tag{5}$$

respectively. The closed-form expression for (3) is

$$m = \frac{\phi(\lambda + z_\alpha) + \phi(\lambda - z_\alpha) + \lambda\{\Phi(\lambda + z_\alpha) + \Phi(\lambda - z_\alpha) - 1\}}{\lambda\{1 - \Phi(\lambda + z_\alpha) + \Phi(\lambda - z_\alpha)\}}. \tag{6}$$

To facilitate better understanding, we consider Theorem 1 from two perspectives. First, we fix the significance level α, and Theorem 1 indicates that the measures in (4)–(6) are functions of $\lambda = \mu/\sigma$, where μ and σ are the mean and standard deviation of the test statistic Z, respectively. In practice, however, λ is sometimes proportional to the square root of the sample size n. For example, in the one-sample t-test, given observations $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma_0^2)$, the corresponding $\lambda = \sqrt{n}\mu/\sigma_0$. From a super-population point of view, λ is closely related to Cohen's d (cf. Cohen, 1969; Orwin, 1983), which plays a crucial role in quantitative psychological research and beyond. It is well known that the power function can be solely determined by λ, and Gelman and Carlin (2014) implicitly stated that the probability of Type S error and the exaggeration ratio possess the same characteristic, which we rigorously prove in the next section.

Second, we fix λ, and Theorem 1 indicates that the measures in (4)–(6) are functions of the significance level α. For illustration, we again let $\mu = 0.5$ and $\sigma = 1$, and allow α to vary in $\{.01, .02, \ldots, .1\}$. In Figure 2 we plot the probability of Type S error s and the expected Type M error m, for different values of α. We also consider the special case $\alpha = 1$, which implies that $z_\alpha = 0$. Consequently, in the following corollary we obtain the expectation of



Probability of Type S error. Expected Type M error (exaggeration ratio).

Figure 2. Illustration of Theorem 1 for $\mu = 0.5$ and $\sigma = 1$. In each sub-figure, the horizontal axis denotes the significance level. In (a), the vertical axis denotes the probability of Type S error. In (b), the vertical axis denotes the expected Type M error.

the folded-normal random variable Leone, Nelson, and Nottingham (1961), which is useful in engineering statistics.

Corollary 1. *The expectation of the folded-normal is*

$$E(|Z|) = \sigma(\pi/2)^{-1/2}e^{-\mu^2/2\sigma^2} + \mu\{1 - 2\Phi(-\mu/\sigma)\}.$$

3.2. Monotonicity properties

We now discuss the monotonicity properties of the power function, probability of the Type S error, and exaggeration ratio, for a fixed significance level α. Although it is well known among the statistical and psychological science communities that the power function monotonically increases as λ increases, we choose to include it in this paper for the purpose of completeness. Furthermore, intuitively speaking, the probability of Type S error and the exaggeration ratio should be monotonically decreasing, as indeed demonstrated through numerical examples (and in some sense implied) by Gelman and Carlin (2014). However, as we will show later, rigorously proving such a statement turns out to be a non-trivial task, and to the best of our knowledge this paper presents the first rigorous proof.

To help us with the proofs, we first present two lemmas. The first lemma contains a simple yet fundamental result regarding the tail probabilities of normal distributions. The lemma not only plays an important role in proving the main theorem, but also is of independent interest.

Lemma 1. *For all $x \in R$, we have*

$$x\{1 - \Phi(x)\} < \phi(x).$$

Lemma 2. *Define functions*

$$s_1 = \phi(\lambda + z_\alpha) + \phi(\lambda - z_\alpha), \quad s_2 = \phi(\lambda + z_\alpha) - \phi(\lambda - z_\alpha),$$

and

$$t_1 = \Phi(\lambda + z_\alpha) + \Phi(\lambda - z_\alpha), \quad t_2 = \Phi(\lambda + z_\alpha) - \Phi(\lambda - z_\alpha).$$

Then the function

$$u = -z_\alpha + z_\alpha t_2 + s_1 + \lambda t_1 - \lambda$$

is positive for all $\lambda > 0$.

We provide the proofs of Lemmas 1 and 2 in Appendix A. With the help of the lemmas, we now present the main theorem of this paper.

Theorem 2. *As λ increases:*

1. the power function (4) monotonically increases;
2. the probability of Type S error (5) monotonically decreases;
3. the exaggeration ratio (6) monotonically decreases.

We believe that Theorem 2 has a twofold use. On the one hand, it can serve as a theoretical complement to the heuristic discussion by Gelman and Carlin (2014). On the other hand, from a more practical perspective, only with the above monotonicity guarantees can a policy-maker asks questions like 'in order to control the probability of Type S error below .01 and the extent of exaggeration below 10%, what power do I need?' We will continue to discuss this matter later.

To end this section, we briefly discuss the sharp bounds of the power function, the probability of Type S error, and the exaggeration ratio, defined in (1)–(3) respectively. We emphasize that although the bounds seem straightforward and intuitive, only with the monotonicity guarantees can we rigorously derive them.

Corollary 2. *The sharp bounds of the power function, the probability of Type S error, and the exaggeration ratio are [α, 1], [0, .5] and [1, ∞], respectively.*

The above bounds are indeed intuitive, as previously pointed out by Gelman and Carlin (2014) and Gelman and Tuerlinckx (2000). For illustration we consider two extreme cases:

1. When the signal-to-noise ratio λ approaches zero,
   (a). the power function (4) approaches the significance level α by definition;
   (b). the probability of Type S error (3) approaches .5, because the sign of any realization of the test statistic Z (statistically significant or not) is essentially determined by a fair coin flip;
   (c). the exaggeration ratio (6) approaches infinity.
2. When the signal-to-noise ratio λ approaches infinity,
   (a). the power approaches 1, because we are almost sure to detect the non-zero treatment effect;
   (b). the probability of Type S error approaches zero, because there will be no realization of Z in the opposite direction;
   (c). the exaggeration ratio approaches 1, because overestimation no longer exists.

In the next section, we will provide some examples to illustrate the above arguments.

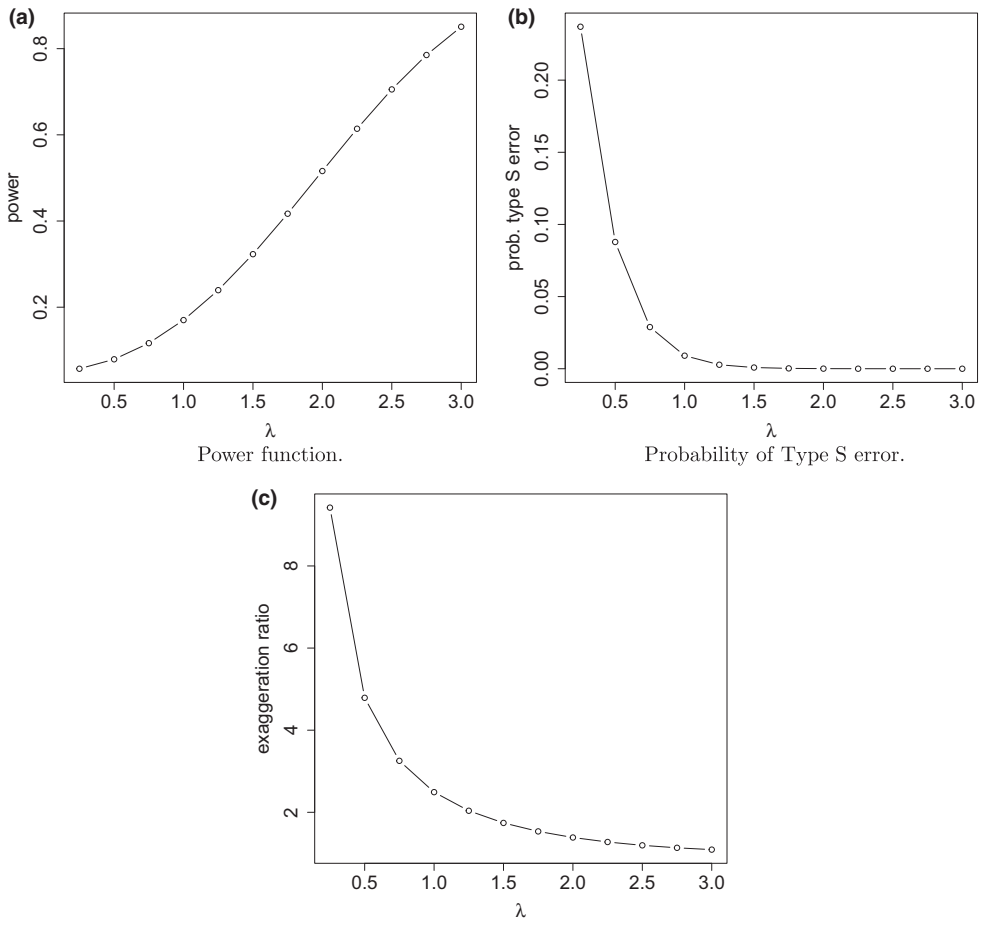
4. Numerical and empirical examples

4.1. Numerical examples

We first present some numerical examples to illustrate our theoretical results. We allow λ to vary over (0.25, 0.5, . . . , 2.75, 3), representing a wide range of typical psychological experiment settings. For each value of λ, we use (4)–(6) to calculate the corresponding power, the probability of Type S error and the exaggeration ratio, respectively. Results are in Figure 3, from which we can draw several conclusions, some of which echo the discussions by Gelman and Carlin (2014):

1. The numeric results corroborate our findings regarding the monotonicity properties of the measures in Theorem 2.
2. The probability of Type S error decreases fast. As a matter of fact, to ensure that $s \leq .1$ and $s \leq .01$, we only need $\lambda \geq 0.5$ ($p = .08$) and $\lambda \geq 1$ ($p = .17$), respectively. In other words, even moderately low-power studies are sufficient to control the Type S errors.
3. The exaggeration ratio decreases relatively slowly. To ensure that $m \leq 1.5$ and $m \leq 1.1$, we need $\lambda \geq 2$ ($p = .52$) and $\lambda \geq 3$ ($p = .85$), respectively. In other words, only high-power studies are sufficient to control the extent of Type M errors.

To summarize, it appears while that Type S errors are rare in practice as long as the analysis is conducted in a principled way, Type M errors are rather common. This conclusion corroborates several discussions in the existing literature. For example, as mentioned by Gelman and Carlin (2014), Button *et al.* (2013) emphasized that 'using



Power function.

Probability of Type S error.

Expected Type M error (exaggeration ratio).

Figure 3. Numerical results for $\alpha = .05$. In each sub-figure, the horizontal axis denotes the signal to noise ratio λ. In (a), the vertical axis represents power. In (b), the vertical axis represents the probability of Type S error. In (c), the vertical axis represents the expected Type M error.

statistical significance as a screener can lead researchers to drastically overestimate the magnitude of an effect'. As a matter of fact, Figure 3 suggests that even for studies with moderately high power (e.g., between .6 and .7), on average we will overestimate the true effect size μ by 20–30%.

We conclude this section by answering the motivating question raised earlier – to simultaneously achieve the two goals of controlling the probability of Type S error below .01 and the extent of exaggeration below 10%, we need the power to be at least 85%.

4.2. Empirical example

By focusing on Type S/M errors in additional to the classic Type I/II errors, Gelman and Carlin (2014) proposed a new retrospective analysis tool that complements the traditional NHST framework. In particular, it can 'provide useful insight, beyond what was revealed by the estimate, confidence interval, and *p*-value that came from the original data summary'. In practice, for an existing data analysis, the following steps were taken:

1. Assume that the sampling standard deviation equals (or at least accurately approximates) the true standard deviation σ.
2. Make an educated guess at the true effect size μ, through, for example, extensive external literature reviews.
3. Use (5) and (6) to calculate the minimal required standard error σ to meet the thresholds, and compare it to the standard deviation in the existing study.

Gelman and Carlin (2014) used this tool to reanalyse three empirical studies in psychological and political sciences (all of which are in published papers), and revealed that existing results might have grossly overestimated the true effect sizes. What is worse, there was a decent chance that the point estimates were in the wrong directions.

In this section, we illustrate how to obtain additional insights from a different perspective, by proposing some slight modifications to Gelman and Carlin (2014) analytic tool, without any fundamental changes:

1. Depending on the context of the study, propose 'toleration' levels of Type S/M errors – thresholds for the probability of Type S error and exaggeration ratio.
2. Make an educated guess at the true effect size μ.
3. Use (5) and (6) to calculate the minimal required standard error σ to meet the thresholds, and compare it to the standard deviation in the existing study.

We believe that our approach has some merits. For example, suppose that we find that for an existing study, the standard deviation is twice the desired minimal standard error that meets the thresholds for controlling Type S/M errors. If circumstances (time, budget) permit, it is possible to repeat the study with four times the sample size. Therefore, instead of declaring the existing study untrustworthy, we advocate 'salvage' by accruing more evidence.

We revisit the controversial 'beauty and sex ratio' study by Kanazawa (2007). For illustration, we require that $s \leq .01$ and $m \leq 1.1$, which is equivalent to $\lambda \geq 3$. As mentioned by Gelman and Carlin (2014), a review of the literature suggested that the true effect size can be at most 1%, which implies that in order to meet the thresholds the standard deviation should be at most 0.33%. The existing study has a sample size of 2,972 and a standard error of 3.3%. Therefore, by assuming that the sample is representative of the target population, we recommend reconducting the study with a sample size of at least 300,000. This recommendation more or less echoes Gelman and Carlin (2014)

recommendation of 500,000. It is worth mentioning that, if such a large-scale study were indeed conducted, it is very likely that we would have obtained a point estimate much less than the 8% reported by Kanazawa (2007). According to the existing literature, such a large number is highly unlikely to represent the truth.

5. Concluding remarks

In this paper, we have studied the proposed methodology for Type S/M errors in null hypothesis significance testing in Gelman and Carlin (2014), and filled several theoretical gaps. In particular, we derived the closed-form expression of the exaggeration ratio, and proved its monotonicity property. Through several numerical and empirical examples, we demonstrated that our results can complement of the heuristic discussion by Gelman and Carlin (2014), and, moreover, are of both theoretical and practical interest. We also discussed how to apply our retrospective analysis tool to real-life data sets.

There are multiple possible future directions based on our work. In particular, although Gelman and Carlin (2014) current framework aims to 'use prior information without necessarily using Bayesian inference', it is possible to embed the proposed measures of Type S/M errors in a fully Bayesian setting. We leave this to future research.

Acknowledgements

The authors thank Dr. David Afshartous at Amazon for early conversations during the 2017 IMS/ASA Spring Research Conference held at Rutgers University that motivated this work, and Professor Eric-Jan Wagenmakers at University of Amsterdam for helpful suggestions on an early draft. We are grateful to members of the Analysis and Experimentation Team at Microsoft, especially Brian Frasca, Greg Linden and Ronny Kohavi, for continuous encouragement and support. Thoughtful comments from an Associate Editor and two anonymous reviewers have substantially improved the quality and presentation of the paper.

References

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the 'replication crisis': Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, *52*, 305–324. https://doi.org/10.1080/00273171.2017.1289361

Baker, M. (2016). 1500 scientists lift the lid on reproducibility. *Nature*, *533*, 452–454. https://doi.org/10.1038/533452a

Bayarri, M. J., Benjamin, D. J., Berger, J. O., & Sellke, T. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, *72*, 90–103. https://doi.org/10.1016/j.jmp.2015.12.007

Berger, J. O., Boukai, B., & Wang, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis (with discussion). *Statistical Science*, *12*, 133–160. https://doi.org/10.1214/ss/1030037904

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376. https://doi.org/10.1038/nrn3475

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. https://doi.org/10.1177/0956797613504966

Deng, A. (2015). Objective Bayesian two sample hypothesis testing for online controlled experiments. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 923–928). Republic and Canton of Geneva: International World Wide Web Conferences Steering Committee.

Deng, A., Lu, J., & Chen, S. (2016). Continuous monitoring of A/B tests without pain: Optional stopping in Bayesian testing. In Proceedings of 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). https://doi.org/10.1109/DSAA.2016.33

Efron, B. (2013). A 250-year argument: Belief, behavior, and the bootstrap. *Bulletin of the American Mathematical Society*, *50*, 129–146. https://doi.org/10.1090/S0273-0979-2012-01374-5

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*, 45–52. https://doi.org/10.1177/1948550615612150

Gelman, A. (2016). The problems with *p*-values are not just with *p*-values. *American Statistician*, *70*. (Online discussion of the ASA statement on p-values.)

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*, 641–651. https://doi.org/10.1177/1745691614551642

Gelman, A., & Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, *15*, 373–390. https://doi.org/10.1007/s001800000040

Gigerenzer, G., & Swijtink, Z. (1990). *The empire of chance: How probability changed science and everyday life*. Cambridge, UK: Cambridge University Press.

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, *45*, 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, *52*, 15–24. https://doi.org/10.1037/0003-066X.52.1.15

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of *p*-hacking in science. *PLoS Biology*, *13*, e1002106. https://doi.org/10.1371/journal.pbio.1002106

Huber, W. (2016). A clash of cultures in discussions of the *P* value. *Nature Methods*, *13*, 607. https://doi.org/10.1038/nmeth.3934

Ioannidis, J. P. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, *294*, 218–228. https://doi.org/10.1001/jama.294.2.218

Ioannidis, J. P. (2005b). Why most published research findings are false. *PloS Medicine*, *2*, e124. https://doi.org/10.1371/journal.pmed.0020124

Johnson, V. E. (2013a). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, *110*, 19313–19317. https://doi.org/10.1073/pnas.1313476110

Johnson, V. E. (2013b). Uniformly most powerful Bayesian tests. *Annals of Statistics*, *41*, 1716–1741. https://doi.org/10.1214/13-AOS1123

Kanazawa, S. (2007). Beautiful parents have more daughters: A further implication of the generalized Trivers-Willard hypothesis. *Journal of Theoretical Biology*, *244*, 133–140. https://doi.org/10.1016/j.jtbi.2006.07.017

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. https://doi.org/10.1080/01621459.1995.10476572

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, *94*, 1372–1381. https://doi.org/10.1080/01621459.1999.10473888

Kruschke, J. K. (2013). Bayesian estimation supersedes the t-test. *Journal of Experimental Psychology: General*, *142*, 573–603. https://doi.org/10.1037/a0029146

Lehmann, E. L., & Romano, J. P. (2006). *Testing statistical hypotheses* (3rd ed.). New York, NY: Springer.

Leone, F. C., Nelson, L. S., & Nottingham, R. B. (1961). The folded normal distribution. *Technometrics*, *3*, 543–550. https://doi.org/10.1080/00401706.1961.10489974

Lishner, D. A. (2015). A concise set of core recommendations to improve the dependability of psychological research. *Review of General Psychology*, *19*, 52. https://doi.org/10.1037/gpr0000028

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E. J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming (2014). *Psychological Science*, *25*, 1289–1290. https://doi.org/10.1177/0956797614525969

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241. https://doi.org/10.1037/1082-989X.5.2.241

Orwin, R. G. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics*, *8*, 157–159. https://doi.org/10.2307/1164923

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530. https://doi.org/10.1177/1745691612465253

Peng, R. D. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance*, *12*(3), 30–32. https://doi.org/10.1111/j.1740-9713.2015.00827.x

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237. https://doi.org/10.3758/PBR.16.2.225

Rozeboom, W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, *57*, 416–428. https://doi.org/10.1037/h0042040

Senn, S. (2001). Two cheers for p-values? *Journal of Epidemiology and Biostatistics*, *6*, 193–204. https://doi.org/10.1080/135952201753172953

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. https://doi.org/10.1177/0956797611417632

Taylor, J., & Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, *112*, 7629–7634. https://doi.org/10.1073/pnas.1507583112

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*, 1–2. https://doi.org/10.1080/01973533.2015.1012991

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *American Statistician*, *70*, 129–133. https://doi.org/10.1080/00031305.2016.1154108

Received 12 July 2017; revised version received 8 December 2017

Appendix A: Proofs of lemmas

Proof of Lemma 1. For $x \leq 0$, the lemma holds trivially. For $x > 0$, note that

$$\begin{aligned} x\{1-\Phi(x)\} &= (2\pi)^{-1/2}\int_x^\infty xe^{-s^2/2}\mathrm{d}s \\ &< (2\pi)^{-1/2}\int_x^\infty se^{-s^2/2}\mathrm{d}s \\ &= (2\pi)^{-1/2}e^{-x^2/2} \\ &= \phi(x), \end{aligned}$$

which completes the proof. □

Proof of Lemma 2. By the definitions of s_1, s_2, t_1 and t_2, we can rewrite u as follows:

$$\begin{aligned} u &= s_1 - z_\alpha - \lambda + z_\alpha\{\Phi(\lambda + z_\alpha) - \Phi(\lambda - z_\alpha)\} + \lambda\{\Phi(\lambda + z_\alpha) + \Phi(\lambda - z_\alpha)\} \\ &= s_1 - z_\alpha - \lambda + (\lambda + z_\alpha)\Phi(\lambda + z_\alpha) + (\lambda - z_\alpha)\Phi(\lambda - z_\alpha) \\ &= (\lambda + z_\alpha)\{\Phi(\lambda + z_\alpha) - 1\} + (\lambda - z_\alpha)\Phi(\lambda - z_\alpha) + \phi(\lambda + z_\alpha) + \phi(\lambda - z_\alpha). \end{aligned}$$

By Lemma 1,

$$(\lambda + z_\alpha)\{\Phi(\lambda + z_\alpha) - 1\} + \phi(\lambda + z_\alpha) > -\phi(\lambda + z_\alpha) + \phi(\lambda + z_\alpha) = 0$$

and

$$\begin{aligned} (\lambda - z_\alpha)\Phi(\lambda - z_\alpha) + \phi(\lambda - z_\alpha) &= -(z_\alpha - \lambda)\{1 - \Phi(z_\alpha - \lambda)\} + \phi(z_\alpha - \lambda) \\ &> -\phi(z_\alpha - \lambda) + \phi(z_\alpha - \lambda) = 0. \end{aligned}$$

The proof is complete. □

Appendix B: Proofs of theorems and corollaries

Proof of Theorem 1. By definition, $Z/\sigma \sim N(\lambda, 1)$. First, by (1),

$$p = \Pr(Z/\sigma < -z_\alpha) + \Pr(Z/\sigma > z_\alpha) = \Phi(-z_\alpha - \lambda) + 1 - \Phi(z_\alpha - \lambda).$$

Second, by (2),

$$\begin{aligned} s &= \frac{\Pr(Z < 0, |Z| > \sigma z_\alpha)}{\Pr(|Z| > \sigma z_\alpha)} \\ &= \frac{\Pr(Z/\sigma < -z_\alpha)}{\Pr(Z/\sigma > z_\alpha) + \Pr(Z/\sigma < -z_\alpha)} \\ &= \frac{\Phi(-z_\alpha - \lambda)}{\Phi(-z_\alpha - \lambda) + 1 - \Phi(z_\alpha - \lambda)}. \end{aligned}$$

Third, let $Y = |Z/\sigma|$ denote the corresponding folded-normal random variable, whose probability density function is (cf. Leone et al., 1961)

$$f(y) = (2\pi)^{-1/2}\left\{e^{-\frac{(y-\lambda)^2}{2}} + e^{-\frac{(y+\lambda)^2}{2}}\right\} \quad (y \geq 0).$$

By definition, we can rewrite (3) as

$$\begin{aligned} m(\lambda) &= \frac{E(|Z| \mid |Z/\sigma| > z_\alpha)/\sigma}{\mu/\sigma} \\ &= \frac{E(Y \mid Y > z_\alpha)}{\lambda}, \end{aligned} \tag{7}$$

which implies that

$$m(\lambda) = \frac{(2\pi)^{-1/2} \int_{z_\alpha}^{\infty} y \left\{ e^{-\frac{(y-\lambda)^2}{2}} + e^{-\frac{(y+\lambda)^2}{2}} \right\} \mathrm{d}y}{\lambda (2\pi)^{-1/2} \int_{z_\alpha}^{\infty} \left\{ e^{-\frac{(y-\lambda)^2}{2}} + e^{-\frac{(y+\lambda)^2}{2}} \right\} \mathrm{d}y}. \tag{8}$$

To further simplify (8), note that

$$(2\pi)^{-1/2} \int_{z_\alpha}^{\infty} e^{-\frac{(y-\lambda)^2}{2}} \mathrm{d}y = \Phi(\lambda - z_\alpha), \quad (2\pi)^{-1/2} \int_{z_\alpha}^{\infty} e^{-\frac{(y+\lambda)^2}{2}} \mathrm{d}y = 1 - \Phi(\lambda + z_\alpha),$$

that

$$\begin{aligned} (2\pi)^{-1/2} \int_{z_\alpha}^{\infty} y e^{-\frac{(y-\lambda)^2}{2}} \mathrm{d}y &= -(2\pi)^{-1/2} \int_{z_\alpha - \lambda}^{\infty} de^{-\frac{s^2}{2}} + \lambda (2\pi)^{-1/2} \int_{z_\alpha}^{\infty} e^{-\frac{(y-\lambda)^2}{2}} \mathrm{d}y \\ &= \phi(\lambda - z_\alpha) + \lambda \Phi(\lambda - z_\alpha), \end{aligned}$$

and that

$$\begin{aligned} (2\pi)^{-1/2} \int_{z_\alpha}^{\infty} y e^{-\frac{(y+\lambda)^2}{2}} \mathrm{d}y &= -(2\pi)^{-1/2} \int_{z_\alpha + \lambda}^{\infty} de^{-\frac{s^2}{2}} - \lambda (2\pi)^{-1/2} \int_{z_\alpha}^{\infty} e^{-\frac{(y+\lambda)^2}{2}} \mathrm{d}y \\ &= \phi(\lambda + z_\alpha) + \lambda \{ \Phi(\lambda + z_\alpha) - 1 \}. \end{aligned}$$

We complete the proof by combining the above. □

Proof of Corollary 1. By (7) and (8) in the proof of Theorem 1,

$$\begin{aligned} E(Y \mid Y > z_\alpha) &= \lambda m(\lambda) \\ &= \frac{\phi(\lambda + z_\alpha) + \phi(\lambda - z_\alpha) + \lambda \{ \Phi(\lambda + z_\alpha) + \Phi(\lambda - z_\alpha) - 1 \}}{1 - \Phi(\lambda + z_\alpha) + \Phi(\lambda - z_\alpha)}. \end{aligned}$$

We let $\alpha = 1$ and $z_\alpha = 0$. Consequently, $E(Y) = 2\phi(\lambda) + \lambda\{1 - 2\Phi(-\lambda)\}$, which implies that

$$\begin{aligned} E(|Z|) &= 2\sigma\phi(\lambda) + \lambda\sigma\{1 - 2\Phi(-\lambda)\} \\ &= \sigma\sqrt{2/\pi} e^{-\mu^2/2\sigma^2} + \mu\{1 - 2\Phi(-\mu/\sigma)\}, \end{aligned}$$

and the proof is complete. □

Proof of Proposition 2. First, to prove that (4) monotonically increases, note that its derivative

$$\frac{\partial p}{\partial \lambda} = \phi(z_\alpha - \lambda) - \phi(z_\alpha + \lambda) > 0$$

for all $\alpha \in (0,\ 1)$ and $\lambda > 0$. The last step holds because $(z_\alpha - \lambda)^2 < (z_\alpha + \lambda)^2$.

Second, proving that (5) monotonically decreases is equivalent to proving that $h = \Phi(\lambda - z_\alpha)/\Phi(-\lambda - z_\alpha)$ monotonically increases, which holds because its derivative

$$\frac{\partial h}{\partial \lambda} = \frac{\phi(\lambda - z_\alpha)\Phi(-\lambda - z_\alpha) + \Phi(\lambda - z_\alpha)\phi(-\lambda - z_\alpha)}{\Phi^2(-\lambda - z_\alpha)} > 0$$

for all $\lambda > 0$.

Third, to prove that (6) monotonically decreases, first note that by the definitions of s_1, s_2, t_1 and t_2, we can express $m(\lambda)$ as m_1/m_2, where

$$m_1 = s_1 + \lambda(t_1 - 1), \quad m_2 = \lambda(1 - t_2).$$

Furthermore, the derivative of m_1 is

$$\begin{aligned} m_1' &= -(\lambda + z_\alpha)\phi(\lambda + z_\alpha) - (\lambda - z_\alpha)\phi(\lambda - z_\alpha) + t_1 - 1 + \lambda s_1 \\ &= -z_\alpha s_2 + t_1 - 1, \end{aligned}$$

and that of m_2 is $m_2' = 1 - t_2 - \lambda s_2$. Therefore, we have

$$\begin{aligned} m_1' m_2 &= \lambda(1 - t_2)(-z_\alpha s_2 + t_1 - 1) \\ &= -\lambda s_2 z_\alpha (1 - t_2) + \lambda(1 - t_2)(t_1 - 1) \end{aligned} \tag{9}$$

and

$$\begin{aligned} m_2' m_1 &= (1 - t_2 - \lambda s_2)(s_1 + \lambda t_1 - \lambda) \\ &= s_1(1 - t_2) - \lambda s_1 s_2 + \lambda(1 - t_2)(t_1 - 1) - \lambda^2 s_2(t_1 - 1). \end{aligned} \tag{10}$$

Combining (9) and (10), we have

$$\begin{aligned} m_1' m_2 - m_2' m_1 &= -\lambda s_2 z_\alpha (1 - t_2) - s_1(1 - t_2) + \lambda s_1 s_2 + \lambda^2 s_2(t_1 - 1) \\ &= \lambda s_2(-z_\alpha + z_\alpha t_2 + s_1 + \lambda t_1 - \lambda) - s_1(1 - t_2) \\ &= \lambda s_2 u - s_1(1 - t_2). \end{aligned}$$

By the definitions of s_1, s_2 and t_2, and by Lemma 2,

$$s_1 > 0, \quad s_2 \leq 0, \quad 1 - t_2 > 0, \quad u > 0,$$

for all $\lambda > 0$. The proof is complete. □

Proof of Corollary 2. By the monotonicity properties in Theorem 2:

1. $p(0) \leq p \leq p(\infty)$, where

$$p(0) = \alpha, \quad p(\infty) = 1;$$

2. $s(\infty) \leq s \leq s(0)$, where

$$s(\infty) = 0, \quad p(0) = 0.5;$$

3. $m(\infty) \leq m \leq m(0)$, where

$$\begin{aligned} m(\infty) &= \lim_{\lambda\to\infty} \frac{\phi(\lambda + z_\alpha) + \phi(\lambda - z_\alpha)}{\lambda\{1 - \Phi(\lambda + z_\alpha) + \Phi(\lambda - z_\alpha)\}} + \lim_{\lambda\to\infty} \frac{\Phi(\lambda + z_\alpha) + \Phi(\lambda - z_\alpha) - 1}{1 - \Phi(\lambda + z_\alpha) + \Phi(\lambda - z_\alpha)} \\ &= \lim_{\lambda\to\infty} \frac{0}{\lambda} + \lim_{\lambda\to\infty} \frac{1}{1} \\ &= 1 \end{aligned}$$

and

$$\begin{aligned} m(0) &= \lim_{\lambda\to 0} \frac{\phi(\lambda + z_\alpha) + \phi(\lambda - z_\alpha)}{\lambda\{1 - \Phi(\lambda + z_\alpha) + \Phi(\lambda - z_\alpha)\}} + \lim_{\lambda\to 0} \frac{\Phi(\lambda + z_\alpha) + \Phi(\lambda - z_\alpha) - 1}{1 - \Phi(\lambda + z_\alpha) + \Phi(\lambda - z_\alpha)} \\ &= \lim_{\lambda\to 0} \frac{2\phi(z_\alpha)}{\lambda\alpha} + \lim_{\lambda\to\infty} \frac{\alpha}{\alpha} \\ &= \infty. \end{aligned}$$

The proof is complete. □