Refiltering hypothesis tests to control sign error

Art B. Owen
Stanford University

August 2019

Abstract

A common, though not recommended statistical practice is to report confidence intervals if and only if they exclude a null value of 0. The resulting filtered confidence intervals generally do not have their nominal confidence level. More worryingly, in low power settings they will frequently lie on the wrong side of zero. We assume that the confidence intervals were constructed using an asymptotically Gaussian parameter estimate accompanied by a weakly consistent estimate of its variance. In these cases, we can subject the given confidence interval(s) to a second filtering step such that the probability of a sign error is controled. This refiltering step retains only those confidence intervals that are sufficiently well separated from the origin.

1 Introduction

A serious problem arises for hypothesis tests of low statistical power, as pointed out by Gelman (2014). He considers an extreme case of testing a null hypothesis that $\theta = 0$ at significance level $\alpha = 0.05$ with power only 0.06. The type II error is then $\beta = 0.94$. In his formulation, reproduced below, any rejected hypotheses come with a very exaggerated estimate of the magnitude of θ. They frequently have the wrong sign too.

These points were made earlier by Gelman and Tuerlinckx (2000). Getting a statistically significant result with the incorrect sign is sometimes called a directional error. It is also sometimes called a type III error, though that term is also used to include tests that answer what is in some sense the 'wrong' question. Sign errors have potentially more serious consequences than the type I error that statisticians focus on. Sign errors have been discussed in the statistical and related literatures for a long time. See for instance, Shaffer (2002) and Jones and Tukey (2000) and references therein going back to the 1950s and 1960s.

In this article, we mitigate the sign error problem by inspecting the confidence interval for θ. We suppose that somebody has run one or more hypothesis tests and reports their confidence intervals, but only those confidence intervals that do not include the origin. We will see that when a confidence is well separated from the origin, then we can be confident that the sign of its center point is correct. In a refiltering step, we can declare that $\text{sign}(\theta) = \text{sign}(\hat{\theta})$ only in those well separated intervals.

There are better ways to handle multiple confidence intervals than simply reporting the ones that do not contain the origin. We would prefer the user to have used the proposal in Benjamini and Yekutieli (2005) which is a confidence interval version of the well-known procedure of Benjamini and Hochberg (1995), but we suppose that they have not done this. For notions of the false sign proportion in a set of decisions and the corresponding false sign rate, see Weinstein and Ramdas (2019) and articles cited there. Stephens (2016) presents a local false sign rate to parallel the local false discovery rate of Efron (2005).

Given a confidence interval, we do not know what power the corresponding test had. As mentioned above, we can reason that for a confidence interval that is well separated from the origin we have more cause to believe that the sign is correct, than we do for a confidence interval that just barely excludes the origin. When the tests and confidence intervals are based on an asymptotically normal test statistic $\hat{\theta}$ with a suitable estimate of $\text{var}(\hat{\theta})$ then we can say how far the confidence interval must be from the origin, in units of its own length, for us to have some confidence that $\text{sign}(\theta) = \text{sign}(\hat{\theta})$. That in turn leads us to a way to get a p-value for the sign of θ that takes into account the fact that we were only presented with confidence intervals that excluded 0.

The remainder of this paper is as follows. Section 2 introduces our model formulation. We consider power for a test of a scalar θ for which an asymptotically normal estimate $\hat{\theta}$ is available along with a consistent estimator of $\text{var}(\hat{\theta})$. When the confidence interval for θ is well separated from the origin, as described in Section 3, then we can be more confident that $\text{sign}(\theta) = \text{sign}(\hat{\theta})$. Equivalently, that confidence can come from repeating the test at a more stringent level. The result is a criterion for deciding when to declare the sign of θ subject to an upper bound on the probability of making an erroneous declaration. Section 4 computes and displays some operating characteristics of this testing procedure. Section 5 presents some 'sign p-values'. These are the smallest upper bounds on the probability of a sign error given the selected confidence interval. Section 6 briefly considers one-sided tests. Section 7 presents some conclusions. There is an appendix with some R code to handle some of the numerical tasks in this paper.

2 Model formulation

In our setting there is a real-valued parameter θ and the null hypothesis is $H_0 : \theta = 0$. We wish to test this hypothesis. In many contexts, $\theta = 0$ means that some phenomenon is unimportant. For instance, θ might be the coefficient of x in a regression model for y and $\theta = 0$ then means that x does not affect the expected value of y.

While $|\theta|$ could well be small, it is a priori almost certain that θ is not exactly zero. Even if there is a scientific reason for $\theta = 0$ to be possible, we are unlikely to have perfect instruments, and so the θ governing our data will not be zero. Given that $\theta \neq 0$, it becomes interesting to justify doing the test. A potential explanation is that if we reject H_0 we have learned the sign of θ about which there may have been a reasonable doubt. Jones and Tukey (2000) formulate the problem as one of testing $\theta \geqslant 0$ versus $\theta \leqslant 0$ and show that the sign error probability is below $\alpha/2$. Unfortunately, things are more difficult if the interval we get has been selected for non-containment of the origin. As we will see, in the low power setting, rejecting H_0 can leave substantial doubt about $\mathrm{sign}(\theta)$. Another explanation is that while we may be confident that $\theta \neq 0$, we may still have to furnish evidence of that to other people. For the present purposes, we must work with $\theta \neq 0$ to measure the extent to which the significance filter gives large but finite values of $|\hat{\theta}/\theta|$.

To formulate a model for this selective process, we assume that there is an estimate $\hat{\theta} \sim \mathcal{N}(\theta, \tau_0^2)$ for some $\tau_0 > 0$. The user has an estimator s_0^2 of τ_0^2 with $s_0^2/\tau_0^2 \xrightarrow{\mathrm{d}} 1$. For instance, it might hold that $s_0^2 \sim \tau_0^2 \chi^2_{(\nu)}/\nu$ with $\nu \to \infty$. Then $\Pr(|s_0/\tau_0 - 1| > \epsilon) \to 0$ for all $\epsilon > 0$. To focus on essentials, we will work as if $s_0 = \tau_0$. The test of H_0 is based on $Z = \hat{\theta}/s_0$. It is a two-tailed test at level α that rejects H_0 when $\hat{\theta}^2/s_0^2 \geqslant \chi^{2,1-\alpha}_{(1)}$. It has the same asymptotic rejection probabilty as the test that rejects when $\hat{\theta}^2/\tau_0^2 \geqslant \chi^{2,1-\alpha}_{(1)}$.

Because $\theta \neq 0$, we may change units in our analysis to make $|\theta| = 1$. Then

$$\frac{\hat{\theta}}{\tau_0} = \frac{\hat{\theta}/|\theta|}{\tau} \sim \mathcal{N}\Bigl(\frac{\mathrm{sign}(\theta)}{\tau}, 1\Bigr)$$

where $\tau = \tau_0/|\theta|$. The test rejects H_0 when $(\hat{\theta}/\tau_0)^2 \geqslant \chi^{2,1-\alpha}_{(1)}$ and so it has power

$$\Pr(\chi'^{,2}(\tau^{-2}) \geqslant \chi^{2,1-\alpha}_{(1)}. \tag{1}$$

Given α and a level for power, we can solve (1) for a value of τ. Some R code to do this is in the appendix. Figure 1 shows the results for $\alpha = 0.05$ and power 0.06.

For $\alpha = 0.05$ and power 0.06, solving (1) leads to $\tau \doteq 3.39$. Then the test rejects H_0 when $|Z| \geqslant \Phi^{-1}(0.975) \times \tau \doteq 6.65$. Therefore any significant discovery must overestimate $|\theta|$ by at least 6.65-fold. When $\alpha = 0.05$ and the power is 0.06, the average value of $|\hat{\theta}/\theta|$ given that H_0 is rejected is 8.01 and the probability of a sign error is about 20%. These quantities were computed by numerical quadrature using R functions in the Appendix.

The exaggeration values and sign error estimates in Gelman (2014) (as of August 2019) do not match these results. There, the wrong sign probability is given as 24% and the minimum exaggeration factor is 9. It appears that the figure in the blog is based on slightly different numbers, corresponding to power less than 6%. For $\alpha = 0.05$ and $1 - \beta = 0.06$ we find that $\tau = 3.394507$. The point of all those digits in τ is that we can plug this value in to the `retrodesign`

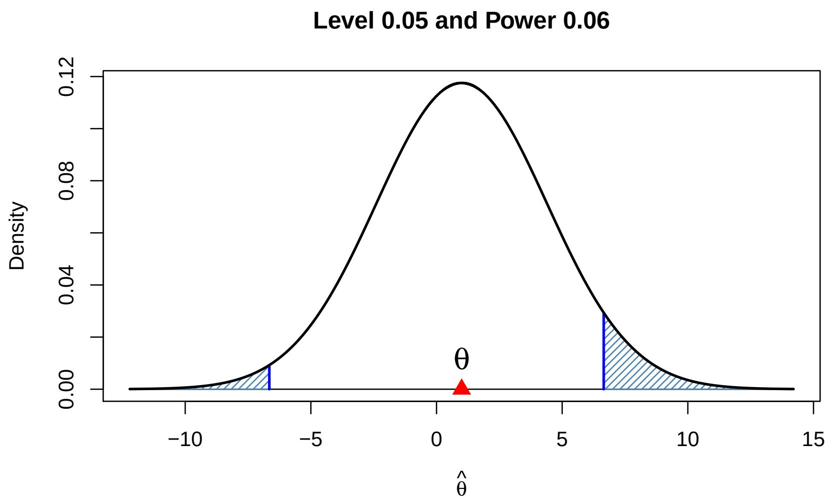

Figure 1: The curve shows the $\mathcal{N}(1, \tau^2)$ density of $\hat{\theta}$ when a test of $H_0 : \theta = 0$ at level 0.05 has power 0.06. A triangle marks the true θ. The shaded regions where $|\hat{\theta}| \geqslant 6.65$ comprise the rejection region for H_0. About 20% of rejections have the wrong sign.

R function from Gelman and Carlin (2014) and verify that it yields power 0.06 when $\alpha = 0.05$. See Figure 2. The numerical values from Gelman (2014) appear to describe power between 5.54% and 5.55% because within that range the minimum exaggeration factor is about 9.

3 Confidence interval separation

When evaluating a hypothesis test from the literature we cannot tell what the power was. There is often an accompanying confidence interval. Confidence intervals have the benefit of being in the same units as the effect θ and hence they facilitate the study of practical significance. Here they have the additional benefit that we can compare the width of the confidence interval to its distance from the origin to get an idea of how reliable $\text{sign}(\hat{\theta})$ is.

If the confidence interval is well separated from the origin, then its creators could have widened it to a higher level of confidence and still have excluded the origin. Equivalently, they could have rejected H_0 at an even smaller α than the one they used.

To illustrate, suppose that H_0 has been rejected at level $\alpha = 0.05$ because $|\hat{\theta}| \geqslant \Phi^{-1}(0.975)s \doteq 1.96s$. Call this event R. Suppose next, that we only declare the sign of θ to match that of $\hat{\theta}$ when the center $\hat{\theta}$ of the confidence interval is at least $2 \times 1.96s$ away from 0. Call this event S. That is, S happens

```
> retrodesign
function(A, s, alpha=.05, df=Inf, n.sims=10000){
z <- qt(1-alpha/2, df)
p.hi <- 1 - pt(z-A/s, df)
p.lo <- pt(-z-A/s, df)
power <- p.hi + p.lo
typeS <- p.lo/power
estimate <- A + s*rt(n.sims,df)
significant <- abs(estimate) > s*z
exaggeration <- mean(abs(estimate)[significant])/A
return(list(power=power, typeS=typeS, exaggeration=exaggeration))
}

> set.seed(1);retrodesign(1,3.394507)
$power
[1] 0.06

$typeS
[1] 0.2013426

$exaggeration
[1] 7.978919
```

Figure 2: R code from Gelman and Carlin (2014) to verify that standard error $\tau = 3.394507$ and effect size $\theta = 1$ yield power 0.06 at $\alpha = 0.05$. The random seed is set to make the results more reproducible.

when the separation between the confidence interval and 0 is at least half of the confidence interval's width. A sign error corresponds to $\hat\theta < -2 \times 1.96s$. We call this event T.

We will find $\Pr(S, T \mid R)$ in case $\theta > 0$. The probability is the same for $\theta < 0$. The conditional probability that we wrongly call the sign is

$$\Pr(S, T \mid R) = \Pr(S, T, R)/\Pr(R) = \Pr(S, T)/\Pr(R)$$

because R occurs whenever S does. Next

$$\begin{aligned}
\Pr(S, T \mid R) &= \frac{\Pr(\hat\theta - \theta < -2 \times 1.96s - \theta)}{\Pr(R)} \\
&\leqslant \frac{\Pr(\hat\theta - \theta < -2 \times 1.96s)}{\Pr(R)} \\
&= \frac{\Phi(-3.92)}{2\Phi(-1.96)} \\
&\doteq 0.0009.
\end{aligned}$$

As before we have used $s=\tau$, which is reasonable when $\Pr(|s/\tau-1|>\epsilon)\to 0$.

More generally, suppose that R is the event that a $100(1-\alpha)\%$ confidence interval for θ excludes the origin, let S be the event that $|\hat\theta-\theta|\geqslant\lambda\Phi^{-1}(1-\alpha/2)$ for some $\lambda\geqslant 1$ and, assuming that $\theta\neq 0$, let T be the event that $\mathrm{sign}(\hat\theta)\neq\mathrm{sign}(\theta)$. Now suppose that we want $\Pr(S,T\mid R)\leqslant\alpha_S$ for a sign error quantity α_S. As in the example above,

$$\Pr(S,T\mid R)\leqslant\frac{\Phi(-\lambda\Phi^{-1}(1-\alpha/2))}{\alpha}. \tag{2}$$

We can keep the right side of (2) below α_S by taking

$$\lambda\geqslant\frac{\Phi^{-1}(\alpha\alpha_S)}{\Phi^{-1}(\alpha/2)}=\frac{\Phi^{-1}(1-\alpha\alpha_S)}{\Phi^{-1}(1-\alpha/2)}.$$

The second expression above is a ratio of positive quantities. Given an interval selected because $|\hat\theta|>\Phi^{-1}(1-\alpha/2)s$, if we declare that $\mathrm{sign}(\theta)=\mathrm{sign}(\hat\theta)$ when $|\hat\theta|>\lambda\Phi^{-1}(1-\alpha/2)s$, then the fraction of such intervals where we declare the wrong sign will be at most α_S.

We can describe this two step test in another way. The first step rejects H_0 if and only if $|\hat\theta|>\Phi^{-1}(1-\alpha_1/2)s$ for $\alpha_1\in(0,1)$. The second step rejects H_0 if and only if $|\hat\theta|>\Phi^{-1}(1-\alpha_2/2)s$ for $\alpha_2\in(0,\alpha_1]$. If both reject then we have conditional probability at most

$$\frac{\alpha_2/2}{\alpha_1}=\frac{\alpha_2}{2\alpha_1}$$

of making a sign error. The 2 in the denominator arises because at most half of the α_2 rejections have the wrong sign. The refiltering described here is an extremely basic form of inference after model selection, as described by Fithian et al. (2014) who also give a comprehensive bibliography. After confidence intervals are selected based on excluding the origin, we use a second test to draw an inference on $\mathrm{sign}(\theta)$.

This approach leaves an awkward intermediate possibility where we have rejected $H_0:\theta=0$ without deciding whether θ is positive or negative. This is reasonable. If we are only presented with a confidence interval for θ just barely misses the origin, then we cannot be confident of $\mathrm{sign}(\theta)$. Figure 3 displays 100 confidence intervals selected to omit the origin, in a setting where the test had 6% power at the 5% level. These intervals barely exclude the origin and many of them have center points with the wrong sign. The procedure in Benjamini and Yekutieli (2005) has a similar property. After deciding on the hypotheses to reject it produces confidence intervals at varying nominal levels in order to control the coverage events conditional on selection. That process lengthens the naive, one-at-a-time confidence intervals one would otherwise have made, and some of the resulting intervals for rejected hypotheses may contain the origin.

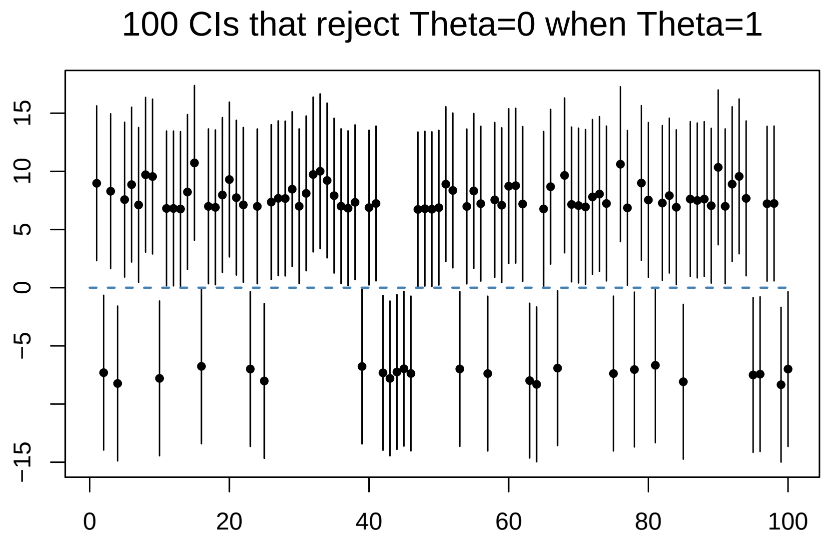

Figure 3: The figure shows 100 randomly selected confidence intervals out of 10^5 that were generated to have 6% power at level $\alpha = 0.05$. The ones shown are the first 100 to exclude 0.

4 Some operating characteristics

Here we make some numerical illustrations of how the tests described here behave. We illustrate with $\alpha = 0.05$, because that is a well known level. Whether it is a good choice in practice depends on context. This has been the subject of much discussion (Wasserstein and Lazar, 2016).

Figure 4 shows the conditional probability of a sign error given that H_0 is rejected at the 5% level. The power of the underlying test ranges from 6% to 80%. The probability of a sign error becomes negligible at power greater than about 30%. Gelman and Carlin (2014) also plot this quantity.

Next we consider magnitude errors in $\hat{\theta}$. The solid curve in Figure 5 shows the minimal value of $|\hat{\theta}/\theta|$ for which H_0 can be rejected at level 0.05. Whenever the power is below 0.5 this minimal ratio is above 1. When the power exceeds 50% then H_0 can be rejected with either an exaggerated magnitude or an underestimated magnitude. The dashed curve in Figure 5 shows the expected value of $|\hat{\theta}/\theta|$ given that H_0 has been rejected, versus the power of the test. Even at 50% power the average exaggeration can be as high as 1.4 fold. Gelman and Carlin (2014) also plot the average exaggeration.

Suppose for example that we want to test H_0 at level $\alpha = 0.05$ but we are very averse to sign errors and want $\alpha_S = 0.001$. That is at most one in one thousand rejections of H_0 should come with a sign wrongly declared by the test at level $\alpha_2 = 0.0001$. Figure 6 shows both the power to reject H_0

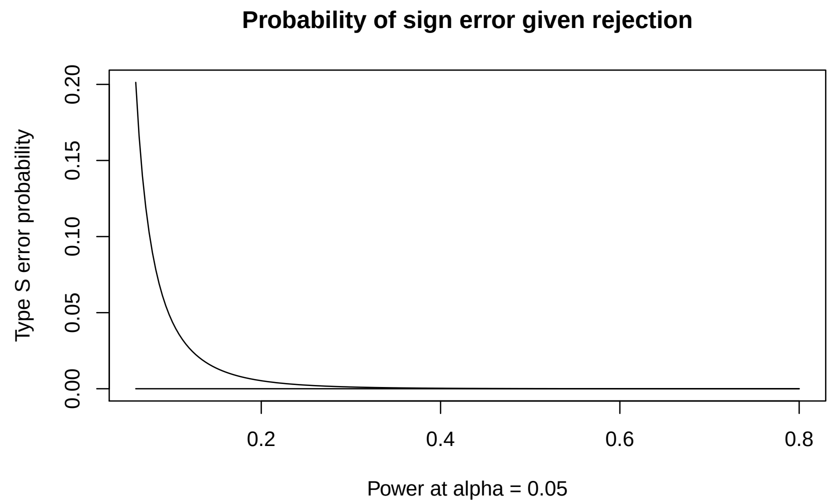

Figure 4: The figure shows how the probability of a sign error, conditional on rejecting H_0, depends on power.

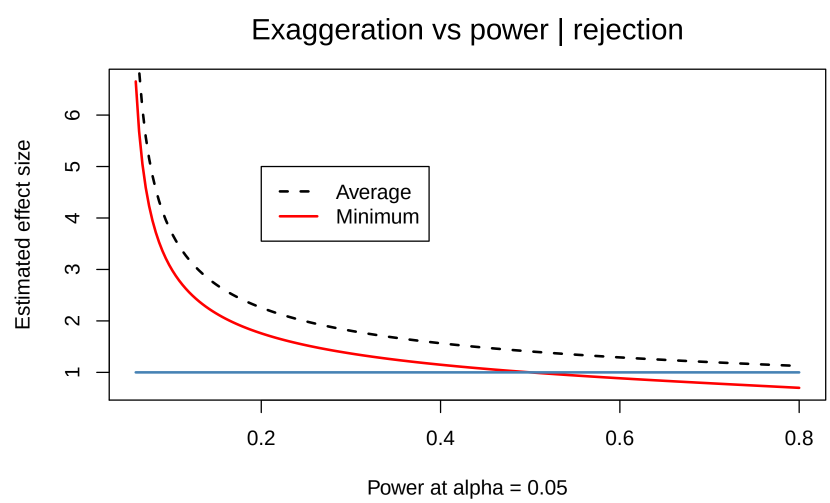

Figure 5: The horizontal axis is the power of a test at level $\alpha = 0.05$. The solid curve is the expected value of $|\hat{\theta}/\theta|$ given that H_0 has been rejected. The lighter curve is the minimum value of $|\hat{\theta}/\theta|$ for which H_0 could be rejected.

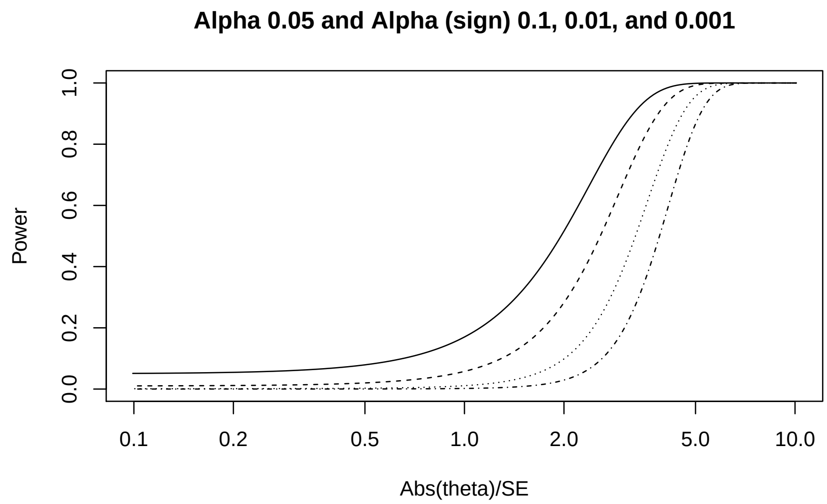

Figure 6: The horizontal axis is the effect size $|\theta|/\tau$. The solid curve is the power of a test of $H_0 : \theta = 0$ at level $\alpha = 0.05$. The next curves from top to bottom are the probabilities of also making a definitive sign declaration at $\alpha_S \in \{0.1, 0.01, 0.001\}$.

and the power to identify $\mathrm{sign}(\theta)$ as a function of $|\theta|/\tau$. It also depicts more lenient sign conditions given by $\alpha_S = 0.01$ and $\alpha_S = 0.1$. Ordinarily $\tau \doteq \sigma/\sqrt{n}$ where σ is an asymptotic standard deviation. There is a range of effect sizes in which rejections of H_0 without decisions on $\mathrm{sign}(\theta)$ will be common. But once $|\theta|\sqrt{n} \geqslant 5\sigma$ or so, there is very high power that H_0 will be rejected and a sign will be identified. The hypotheses in limbo with H_0 rejected but no sign determination are the ones for which obtaining additional data may be most valuable.

5 Sign p-values

In high throughput settings, one might make a very large number of primary hypothesis tests at level α_1. Suppose that they have been selected in a naive way, simply providing the confidence intervals that omit zero and withholding the others. For each one that is rejected, there is then a smallest $\alpha_2 \leqslant \alpha_1$ at which H_0 is also rejected. Call this value p_2. Then the smallest α_S for which we would have found the sign significant is $p_S = p_2/(2\alpha_1)$. Note that p_S depends on the chosen value α_1, which we assume has been fixed in advance and not modified to reduce or increase p_S. Though we call it a p-value because it governs a probability of misinterpretation, note that it cannot be larger than $1/2$.

$\hat{\theta}$	$1.96s$	$\lambda = \lvert\hat{\theta}\rvert/1.96s$	$p_S = \Phi(-1.96\lambda)/0.05$
$-.431$	0.379	1.137	0.258
$-.236$	0.226	1.044	0.407
$-.755$	0.755	1.000	0.500

Table 1: This table shows some examples of p_S, the refiltered p-value for $\mathrm{sign}(\theta)$.

Benjamini and Yekutieli (2005) consider Giovannucci et al. (1995) who tested relationships between diet variables and cancers. Their abstract reports three 95% confidence intervals for relative risk. The most promising one has a relative risk estimate of 0.65 with a 95% confidence interval from 0.44 to 0.95. The face value interpretation is of a protective effect versus prostate cancer from consumption of tomatoes, tomato sauce, tomato juice and pizza. It is natural to measure relative risk on a logarithmic scale. Their confidence interval is then -0.431 ± 0.379, which is approximate due to rounding. The 95% confidence interval for lycopene intake (which overlaps tomatoes et cetera) was, on a log scale, -0.236 ± 0.226. The third interval was approximately -0.755 ± 0.755.

We use these numerical examples below. We do not make any claim about the relationship between diet and cancer. That would require revisiting their data analyses, model selections and considering the related literature on diet and cancer. Here we just show the values of p_S corresponding to those intervals, in Table 1. None of those selected intervals would lead us to be confident about $\mathrm{sign}(\theta)$.

6 One-sided tests

Here we investigate the possibility of sign errors in one-tailed tests. Those tests are sometimes justified by an assumption that the direction of the effect is certain a priori. Then sign errors are not possible, unless our certain opinion is wrong. The a priori uncertainty about the sign of θ must however be small compared to the critical level α in use, and that is a very strong assumption to work under.

A second justification is that sometimes only one direction is consequential. An opinion about one direction being inconsequential might also be mistaken, but for sake of argument we work with it. In this situation we might well make a sign error by rejecting H_0. Suppose that $\hat{\theta} \sim \mathcal{N}(1, \tau^2)$, so that $\hat{\theta}/\tau \sim \mathcal{N}(1/\tau, 1)$. The true effect is positive but we might be testing for a (consequential) negative effect.

A one-tailed test at level α in the negative direction would reject H_0 if $\hat{\theta}/\tau \leqslant \Phi^{-1}(\alpha)$. This happens with probability $\Phi(\Phi^{-1}(\alpha) - 1/\tau)$.

Now we revisit the case of 6% power for a two-tailed test at $\alpha = 0.05$. Then $\tau \doteq 3.39$ and the sign error is 0.20. The wrong direction one-tailed test at level 0.05 will reject H_0 with probability $\Phi(\Phi^{-1}(0.05) - 1/\tau) \doteq 0.026$.

The chance of a wrong sign rejection is actually larger than $\alpha/2$ here. That is not to say the conditional probability of a rejection being wrong is over 50%. Indeed in this setting any rejection at all is a sign error. What it does mean is that in this low power setting, the probability of a sign error is not small compared to α. As a result, using one-tailed tests does not correct the problem of sign errors in low power settings.

7 Conclusions

The practice of reporting only the hypothesis tests that are significant at some level α is not recommended because it can bring large sign and magnitude errors. When one sees such results based on an asymptotically normal test statistic, it is possible to filter them in such a way that the tests passing the filter have a small probability of a sign error. This article presented one such method assuming that $(\hat{\theta}-\theta)/s$ had an asymptotic Gaussian distribution. It could be modified for student t distributions in the event that one also knew the appropriate degrees of freedom to use.

Acknowledgments

I thank Jelena Markovic and Andrew Gelman for discussions on an earlier version of this article. Thanks to Daniel Yekutieli, Aaditya Ramdas, and Edgar Dobriban for further discussions at WHOA-PSI 4. This work was supported by the U.S. National Science Foundation under grants DMS-1407397, DMS-1521145 and IIS-1837931.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81.

Efron, B. (2005). Local false discovery rates. Available at `http://statweb.stanford.edu/~ckirby/brad/papers/2005LocalFDR.pdf`.

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. Technical report, arXiv:1410.2597.

Gelman, A. (2014). *This is what "power = .06" looks like. Get used to it.* `http://andrewgelman.com/2014/11/17/power-06-looks-like-get-used/`.

Gelman, A. and Carlin, J. (2014). Beyond power calculations assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651.

Gelman, A. and Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, 15(3):373–390.

Giovannucci, E., Ascherio, A., Rimm, E. B., Stampfer, M. J., Colditz, G. A., and Willett, W. C. (1995). Intake of carotenoids and retino in relation to risk of prostate cancer. *Journal of the National Cancer Institute*, 87(23):1767–1776.

Jones, L. V. and Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological methods*, 5(4):411.

Shaffer, J. P. (2002). Multiplicity, directional (type III) errors, and the null hypothesis. *Psychological methods*, 7(3):356.

Stephens, M. (2016). False discovery rates: a new deal. *Biostatistics*, 18(2):275–294.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133.

Weinstein, A. and Ramdas, A. (2019). Online control of the false coverage rate and false sign rate. Technical report, Carnegie-Mellon University.

Appendix: some R code

This function solves equation (1) for τ given the level and power of a test.

```
tau = function(alpha,powr){
# A Gaussian test at level alpha
# has power powr for Y ~ N( 1, tau^2 ).
# Solve for tau using noncentral chisquare.

aux = function(tau){
  1-powr - pchisq(qchisq(1-alpha,1),1,ncp=1/tau^2)
}

ur = uniroot( aux, lower=10^-9,upper=10^6,
    tol = .Machine$double.eps^0.9)
ur$root
}
```

The function `exag` below computes the expected value of $\hat\theta$ conditionally on H_0 being rejected and on $\mathrm{sign}(\hat\theta)$. It uses `gauscondmean` which computes $\mathbb{E}(Z \mid A \leqslant Z \leqslant B)$ for $Z \sim \mathcal{N}(\mu, \sigma^2)$.

```
exag = function(alpha=.05,powr=.06){
# Compute exaggeration factors

critse  = tau(alpha,powr)  # Critical Standard Error
critz   = critse*qnorm(1-alpha/2)
posmean = gauscondmean(mu=1,sigma=critse,A=critz, B=Inf, n=10^5 )
negmean = gauscondmean(mu=1,sigma=critse,B=-critz, n=10^5 )
typeS   = pnorm(-critz,1,critse)/powr

exager = posmean*(1-typeS)+abs(negmean)*typeS

list(critse=critse,critz=critz,exager=exager,posmean=posmean,negmean=negmean,typeS=typeS)
}
```

The function `gauscondmean` below computes $\mathbb{E}(Z \mid A \leqslant Z \leqslant B)$ for $Z \sim \mathcal{N}(\mu, \sigma^2)$ using `stdgauscondmean` which handles the case $\mu = 0$ and $\sigma = 1$.

```
gauscondmean = function(mu=0,sigma=1,A=-Inf, B=Inf, n=10^5 ){
# Average of N( mu, sigma^2 ) over interval (A,B)

mu + abs(sigma)*stdgauscondmean((A-mu)/abs(sigma),(B-mu)/abs(sigma),n)
}
```

The function `stdgauscondmean` below computes $\mathbb{E}(Z \mid A \leqslant Z \leqslant B)$ for $Z \sim \mathcal{N}(0, 1)$. It uses a midpoint rule.

```
stdgauscondmean = function(A=-Inf, B=Inf, n=10^5, plot=FALSE ){
# Average of N( 0, 1 ) over interval (A,B)

u = ((1:n)-0.5)/n
u = pnorm(A) + (pnorm(B)-pnorm(A)) * u
z = qnorm(u)
if( plot )hist(z,50) # for testing/debug
mean(z)
}
```