Type S Error Control in Hypothesis Testing

Andrew Neath[*]

Abstract

Hypothesis testing is commonly introduced with the problem of testing for a difference between treatment groups. A valid criticism of hypothesis testing problems in this setting is that all null hypotheses are wrong on the scientific grounds that treatments always have some difference in their effects. But there is a very simple answer to this criticism. If we instead consider the problem of testing for the direction of the treatment difference, a hypothesis test can be developed by controlling for the probability of a type S (or sign) error. A type S error occurs when the test claims the treatment difference is in the positive direction when the true direction is negative; or when the test claims the treatment difference is in the negative direction when the true direction is positive. Presenting a hypothesis testing problem in this fashion allows for a more meaningful introduction to the best uses of statistics in advancing scientific knowledge.

Key Words: statistics education, statistical inference

1. Introduction

We will use an example from Peck, Olsen, Devore (2016) to motivate our discussion on the teaching of hypothesis testing. Their textbook describes a study designed to compare the effect of human growth hormone (HGH) to the effect of steroid use (estradiol) on change in body fat mass. Hypothesis testing is almost universally introduced to students through control of the type I error probability. By letting μ_1, μ_2 denote the respective population means, the competing hypotheses are presented as $H_O : \mu_1 = \mu_2$, $H_A : \mu_1 \neq \mu_2$. Let t_o denote the appropriate t-test statistic. Type I error is controlled by using a decision rule for which $P\,(\text{reject}\, H_O \mid H_O\, \text{true}) = \alpha$. A decision rule which achieves type I error control can be stated as

$$\text{If } |t_o| > t_{\alpha/2}, \text{ then reject } H_o$$
$$\text{If } |t_o| < t_{\alpha/2}, \text{ then fail to reject } H_o.$$

The "reject H_O" and "fail to reject H_O" decisions represent common terminology where the latter is treated as a sort of "no decision" in that one is not able to decide conclusively in favor of the null hypothesis. It is here that we may begin to see a flaw in the traditional approach to hypothesis testing. An explanation for why we are unable to decide in favor of the null is that it is difficult to argue that a hypothesis involving a precise equality could actually be true. Let's think about our motivating example a bit further. HGH and estradiol are hormones involved in the regulation of the female reproductive cycle, and whose therapeutic use can help alleviate menopausal symptoms. However, HGH and estradiol serve different biological functions, so it is unrealistic to believe that HGH and estradiol could have precisely the same effect on body fat change (Chen, 2016). Cases where a precise null does not serve as a viable hypothesis are common (Gelman, Carlin, 2014). The situation is summarized nicely with the following sentiment: "I've never in my life made a type I error. In the applications I've worked on, I've never come across a null hypothesis that could actually be true" (Gelman, 2004).

[*] Southern Illinois University Edwardsville, College of Arts and Sciences, Edwardsville, IL, aneath@siue.edu

Type I error control is justified mathematically. But in cases where a precise null hypothesis is not scientifically meaningful, type I error control is not a valid concern. When we consider the underlying science instead of simply the underlying math, we can present a more thoughtful approach to hypothesis testing.

2. Type S error control

Instead of testing $H_O : \mu_1 = \mu_2$, $H_A : \mu_1 \neq \mu_2$, consider testing $H_1 : \mu_1 - \mu_2 > 0$, $H_2 : \mu_1 - \mu_2 < 0$. Once the precise null hypothesis is removed from consideration, the hypothesis testing problem reduces to a test on the *difference* between means. The competing hypotheses are defined in terms of the *sign* on this difference. A type S error is said to occur when a claim on the difference is in the wrong direction. That is, a type S error is one where we decide in favor of the hypothesis giving the wrong sign. Write $P\left(\text{type S error}\right) = P\left(\text{decide } H_k \mid H_l \text{ true, } l \neq k\right)$, for $l, k = 1, 2$.

Referring back to the motivating example from Section 1, having removed the hypothesis of HGH and estradiol having the exact same effect, we are left with the more scientifically relevant problem of testing for which of the competing therapies leads to the greater change in body fat mass. Instead of looking to reject a hypothesis we don't believe anyway, the testing problem under the type S error framework has the potential to provide a deeper scientific insight.

In creating a decision rule which controls for the probability of a type S error, we are able to write one which closely resembles the traditional rule presented for type I error control. We propose the following, using the same t_o test statistic as before:

If $t_o > t_{\alpha/2}$, then decide H_1
If $t_o < -t_{\alpha/2}$, then decide H_2
If $|t_o| < t_{\alpha/2}$, then make no decision.

Not much is required to change the focus to type S error control. What was once called a rejection of the null is now interpreted as sufficient evidence favoring one sign hypothesis over the the other. There is still a no decision aspect to the rule, but the meaning is more clear. If the evidence in favor of one direction over the other is insufficient to make a claim, we are simply left with no decision on the sign of the difference in means.

To see why type S error control is achieved, we begin with an equivalent statement of the decision rule written in terms of confidence intervals. Let $t_o = \widehat{\delta}/SE_{\widehat{\delta}}$, where $\widehat{\delta}$ is a difference in sample means. A confidence interval for $\delta = \mu_1 - \mu_2$ is stated as $\widehat{\delta} \pm t_{\alpha/2} SE_{\widehat{\delta}}$. We can then rewrite the decision rule as

If CI for δ is everywhere greater than 0, then decide H_1
If CI for δ is everywhere less than 0, then decide H_2
If CI for δ includes 0, then make no decision.

To establish type S error control, we need to show for all $\mu_1 \neq \mu_2$, $P\left(\text{type S error}\right) < \alpha$. Since the case $\mu_1 = \mu_2$ is not considered to be viable, we have no need for error control under the (former) null hypothesis. Take μ_1, μ_2 to be fixed, but arbitrary. Without loss of generality, we can take $\mu_1 < \mu_2$. Call this difference $\delta^* < 0$. A type S error is then committed when the rule decides in favor of H_1. That is, a type S error is committed when the CI for δ is everywhere greater than 0. If the CI is correct, then the interval includes δ^*. Since $\delta^* < 0$, then the interval can not be everywhere greater than 0 and no type S error is committed. The only way for a type S error to occur is if the CI is incorrect. But the interval can be wrong in not including δ^*, but still include negative values for δ. No type S error is committed under this scenario. Since $P\,(\text{CI incorrect}) = \alpha$, it must be that $P\,(\text{type S error}) < \alpha$.

One could argue for a tighter bound on the probability of a type S error. Since a type S error can occur only if the CI is incorrect *and* in the wrong direction, we can say that for all $\mu_1 \neq \mu_2$, $P\,(\text{type S error}) < \alpha/2$. Performing statistical inference under the more realistic setting of type S error control does not change the actual computations, but does put the interpretations on a stronger scientific foundation. The similarities between frameworks could provide an entry for educators to think more carefully about the underlying science behind the statistical procedures we teach without having to make broad changes to their curriculum.

3. Concluding Remark

The traditional approach to teaching hypothesis testing is to focus on the control of a type I error probability. Recently, there has been much written and discussed in the statistical community about how this traditional approach is out of line with other valid approaches to scientific inquiry. For example, see Wasserstein and Lazar (2016), along with the corresponding discussion papers. It is important for those of us in the statistical education community to be aware of the ongoing debate regarding the different frameworks for hypothesis testing problems. As statistical educators, we should always be looking for ways to improve our communication with the future practitioners of statistical science.

REFERENCES

Chen, R. (2016), "Human Growth Hormone and Sex Steriod Supplementation," *HVMN Blog*, July 28.

Gelman, A. (2004), "Type 1, Type 2, Type S, and Type M Errors," *Statistical Modeling, Causal Inference, and Social Science Blog*, December 29.

Gelman, A. and Carlin, J. (2014), "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors," *Perspectives on Psychological Science*, 9, 641-651.

Peck, R., Olsen, C., and Devore, J. (2016), *Introduction to Statistics and Data Analysis*, Cengage Learning: Boston.

Wasserstein, R. and Lazar, N. (2016), "The ASA's Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129-133.