Type I Error Rates

Neil J. Hatfield

Mar 25, 2024

The purpose of this guide is to introduce and discuss the five kinds of Type I Error rates.

Type I Errors and the MC/SI Problem

Recall that a Type I error occurs whenever we claim a statistical discovery that turns out false. That is to say, we reject a null hypothesis that turns out to be the better model. We might also think about a Type I error as a false positive test result.

While we often talk about Type I errors at the level of a single hypothesis test, we need to keep in mind that *each* hypothesis test we do carries a non-zero risk of making a Type I error. Thus, as we conduct more and more hypothesis tests, our overall risk of making at least one Type I error will increase. This leads us to the Multiple Comparison/Simultaneous Inference (MC/SI) Problem: how do we control our overall Type I error rate as we use the same data to conduct multiple hypothesis tests and/or perform simultaneous inference?

Type I Error Rates

When we are talking about a single inference act, we often talk about our Type I Error Risk: the probability that for that particular act we will make a Type I error. When we deal with a collection of inference acts, we shift to talking about our *Type I Error Rate*. We can think about this as the proportion of Type I Errors we might make relative to the total number of inference acts. We use the rate term here to help signal that we're keeping in mind more than one statistical act.

There are five Type I Error Rates that we can choose to control. Each of these rates lie along a continuum that we can imagine go from + Least to Most Conservative (low evidentiary to high evidentiary standard) + Most to Least Powerful (guarding against Type II error) + Least to Most Control over the Type I Error

In this regard we can think about the Type I Error Rates ordered in the following way

$$CER \preceq EER \preceq FDR \preceq MEER \preceq SCI$$

CER: Comparisonwise Error Rate

The first Type I Error Rate is the Comparisonwise Error Rate (CER). If you recall any statistical test you did in Introductory Statistics, this is the error rate you inherently controlled.

CER is just the probability of making a Type I error on any single inference act. That is, whatever comparison we're making between two things (e.g., two values of the location parameter, one empirical value and one theoretical value, etc.) becomes our entire world and no other comparisons exist. Essentially, CER resolves the MC/SI Problem by pretending that every hypothesis test is completely isolated in its own testing family of one.

EER: Experimentwise Error Rate

The second Type I Error Rate is the Experimentwise Error Rate (EER). As the name suggests, we are no longer thinking about a single comparison but rather an entire *experiment* (or more generally, an entire study). EER is the probability that we will make at least one Type I error when *all* of the null models in the testing family reflect reality. That is, if we set EER to 0.1, then we expect that in 10% of the infinite repetitions of the experiment, we would mistakenly decide against at least one null model when all of the null models do the better job.

Notice that 1) we have to imagine a family consisting of all inference acts for the experiment, and 2) we are thinking about all of the null hypotheses (models) doing the better job.

EER places an upper bound on CER: $ EER CER $. Some methods that control EER include both most omnibus tests such as the parametric ANOVA F test, the nonparametric Kruskal-Wallis H test, and Protected Least Significant Difference (LSD).

FDR: False Discovery Rate

The third Type I Error Rate is the False Discovery Rate (FDR). This centers on the idea that in order to make a Type I error, we first have to make a statistical discovery (i.e., reject a null hypothesis). If we make no statistical discoveries, then our Type I Error rate should be 0. This idea led people to think about quantifying the Expected False Discovery Fraction: the proportion of False Discoveries to the total number of discoveries made within a testing family.

If we set FDR to 0.1, then we allow ourselves to make up to 1 false discovery for every 10 discoveries. Thus, we could have

$$FDR = 0.1 \Rightarrow \frac{1 \text{ False}}{10 \text{ Discoveries}} \equiv \frac{10 \text{ False}}{100 \text{ Discoveries}}$$

This allows us to now have some null hypotheses be rejected whereas EER worked under the principle that all null hypotheses did the better job.

FDR places an upper bound on EER: $FDR \geq EER \geq CER$. Some methods that control FDR Benjamini-Hocberg, and [Student-] Newman-Keuls. Most of these methods will require that the tests be independent of each other.

MEER: Maximum Experimentwise Error Rate

The fourth Type I Error Rate has two names: the Maximum Experimentwise Error Rate (MEER) and the Strong Familywise Error Rate. The MEER serves as an adjustment to the FDR. Notice that we can allow ourselves to make more False Discoveries by simply making more discoveries. The MEER functions as the probability of making *any* false discoveries within a testing family and thereby limits allowance of extra false discoveries.

MEER serves as an upper bound for the FDR: $MEER \geq FDR \geq EER \geq CER$. Some methods work for MEER include Hochberg, Holm, REGWR, Gabriel.

SCI: Simultaneous Confidence Intervals

The last Type I Error Rate is Simultaneous Confidence Intervals (SCI). This gets its name from the fact that these methods first came about in relation to constructing multiple confidence intervals for different parameters using the same data. However, this error rate (and subsequent methods) can be used for any act of inference–testing or confidence intervals.

The SCI is the strongest control we have for Type I Errors in a testing family. If we set SCI to 0.1, then we are setting the probability making at least one Type I error *in the entire testing family* is no more than 10%. Alternative, we can express SCI in terms of confidence. With an SCI of 0.1, then the success rate of *all* of our

interval construction methods *simultaneously* capturing are target parameters is 90%. I need to point out that the success rate of any one of the interval construction methods is **not** 90% in this setting. The 90% success rate is on the *entire* suite of construction methods working in concert.

SCI is the upper bound for MEER: $SCI \geq MEER \geq FDR \geq EER \geq CER$. Some examples of methods controlling SCI include Bonferroni, Šidák, Tukey HSD, and Scheffé.

Some Final Thoughts to Think About

I want to leave you with some additional thoughts for you to think about in relation to the MC/SI Problem and Type I Error Rates. I don't have solid answers to these questions (I have opinions on them though); treat them as curiosities for you to think on over the coming months.

While generally associated with the Frequentist approach to inference, a good question to ask ourselves is whether we also have this problem with Bayesian methods? This is a matter of some debate where some individuals say that Bayesian methods don't need adjustments and other say that the we do.

Another place to think about the MC/SI Problem is the concept of testing families. When Neyman first came up with his theory of interval estimation, he had in mind *all future acts* of estimation a person might undertake. This raises a question about whether testing families should be constrained to one particular study, all of the studies conducted by a particular lab group, or all of the studies a researcher have/are/will ever conduct? Suppose that we partner with another researcher to collect data on two different topics at the same time from the same sample of participants. Do we need to include the other researcher's inference acts in our testing families and vice versa? After all, we are using the same data to answer our research questions.