11.4 Multiple Comparison

Zongyi Liu

2023-06-24

11.4 Multiple Comparison

In most experiments there are multiple conditions or treatments.

When comparing multiple conditions, the constraint in NHST is to keep the overall false alarm rate down to the desired level, e.g. 5%. Abiding by this constraint depends on the number of comparisons that are to be made, which in turn depends on the intentions of the experimenter. In a Bayesian analysis, however, there is just one posterior distribution over the parameters that describe the conditions. That posterior distribution is unaffected by the intentions of the experimenter, and the posterior distribution can be examined from multiple perspectives however is suggested by insight and curiosity. The next two sections expand on NHST and Bayesian approaches to multiple comparisons

11.4.1 NHST Correction for Experimentwise Error

When there are multiple groups, it often makes sense to compare each group to every other group. With four groups, for example, there are six different pairwise comparisons we can make; e.g., groups 1 vs 2, 2 vs 3, 1 vs 3, etc.

In NHST, we have to take into account which comparisons we intend to run for the whole experiment. The problem is that each comparison involves a decision with the potential for false alarm. Suppose we set a criterion for rejecting the null such that each decision has a "per-comparison" (PC) false alarm rate of α_{PC}, e.g., 5%. Our goal is to determine the overall false alarm rate when we conduct several comparisons.

We do not get a false alarm on the complementary proportion $1 - \alpha_{PC}$ of replications. If we run c independent comparison tests, then the probability of not getting a false alarm on any of the tests is $(1 - \alpha_{PC})^c$. Consequently, the probability of getting at least one false alarm is $1 - (1 - \alpha_{PC})^c$. We call that probability of getting at least one false alarm, across all the comparisons in the experiment, the "experimentwise" false alarm rate, denoted α_{EW}. Here's the rub: α_{EW} is greater than α_{PC} For example,if $\alpha_{PC} = 0.05$ and $c = 6$,then $\alpha_{EW} = 1 - (1 - \alpha_{PC})^c = 0.26$. Thus,evenwhenthe null hypothesis is true, and there are really no differences between groups, if we conduct six independent comparisons, we have a 26% chance of rejecting the null hypothesis for at least one of the comparisons.

One way to keep the experimentwise false alarm rate down to 5% is by reducing the permitted false alarm rate for the individual comparisons, i.e., setting a more stringent criterion for rejecting the null hypothesis in individual comparisons. One often-used re-setting is the **Bonferonni correction**, which sets $\alpha_{PC} = \alpha_{EW}^{desired}/c$. For example, if the desired experimentwise false alarm rate is .05, and there are 6 comparisons planned, then we set each individual comparison's false alarm rate to .05/6.

There are many different corrections available to the discerning NHST aficionado . Not only do the correction factors depend on the structural relationships of the comparisons, but the correction factors also depend on whether the analyst intended to conduct the comparison before seeing the data, or was provoked into conducting the comparison only after seeing the data. If the comparison was intended in advance, it is called

a **planned comparison**. If the comparison was thought of only after seeing a trend in the data, it is called a **post-hoc comparison**.

The **point** is not correction to use for our propurse. The point is that the NHST analyst must make some correction, and the correction depends on the number and type of comparisons that the analyst intends to make. This creates a problem because two analysts can come to the same data but draw different conclusions because of the variety of comparisons that they find interesting enough to conduct, and what provoked their interest.

11.4.2 Just One Bayesian Posterior

The data from an experiment, or from an observational study, are carefully collected so to be totally insulated from the experimenter's intentions regarding subsequent tests. Indeed, the data should be uninfluenced by the presence or absence of any other condition or subject in the experiment! For example, it doesn't matter to an individual in a filtration group whether or not the experiment includes the other filtration group, or the condensation groups, or still yet other conditions, or how many subjects there are in the groups.

In a Bayesian analysis, the interpretation of the data is **uninfluenced** by the experimenter's intentions. A Bayesian analysis yields a posterior distribution over the parameters of the model.

11.4.3 How Bayesian Analysis Mitigates False Alarms

No analysis is immune to false alarms, because randomly sampled data will occasionally contain accidental coincidences of outlying values. Bayesian analysis eschews the use of p values as a criterion for decision making, however, because the probability of false alarm depends dramatically the experimenter's intentions. Bayesian analysis instead accepts the fact that the posterior is the best inference we can make, given the observed data and the prior beliefs.

Bayesian analysis can address the problem of false alarms by incorporating prior knowledge into the structure of the model. Specifically, if we know that different groups have some overarching commonality, even if their specific treatments are different, we can nevertheless model the different group parameters as having been drawn from an overarching distribution that expresses the commonality.