The Multiple Testing Problem

5

Contents

What You Will Learn in This Chapter

In Part I, we focused on the most basic statistical comparison, namely, the comparison of two means. We described the t-test, which has high power and, hence is a good test and should be used whenever possible. However, sometimes more than two means need to be compared; e.g., if we want to compare the population of trees in three regions of the world. In this case, a multiple testing problem arises that increases the risk of making a Type I error.

In this chapter, we will introduce the multiple testing problem and present Bonferroni corrections as one (rather suboptimal) way to cope with it.

5.1 Independent Tests

To understand the multiple testing problem, consider the following situation. If we compute one t-test, we know that if the null hypothesis is actually true then there is a Type I error rate of $\alpha = 0.05$. Another way of saying this is that we do not produce a Type I error (False Alarm) in $1 - 0.05$ of the cases when the null is true. If we compute two *independent* t-tests, the chance of not making any False Alarm is $0.95^2 = 0.9$. For 12 comparisons: $0.95^{12} = 0.54$. So the risk of making at least one False Alarm with 12 comparisons is $1 - 0.54 = 0.46$. Hence, it becomes more and more likely to produce False

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_5

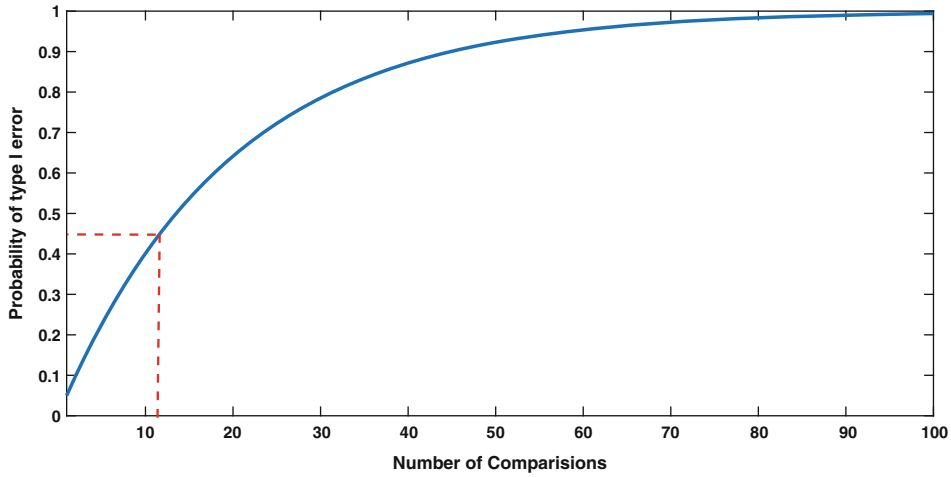

Fig. 5.1 The Type I error rate (False Alarm rate) strongly depends on the number of comparisons. For example with 12 comparisons (red dashed lines), the probability of making at least one False Alarm is 0.46, i.e., much increased compared to 0.05 with only 1 comparison

Alarms when the number of comparisons increases (Fig. 5.1). In general, the probability of making at least one Type I error for m independent tests is:

$$1 - (1 - \alpha)^m \tag{5.1}$$

or $1 - (1 - 0.05)^m$ for $\alpha = 0.05$.

Bonferroni Corrections One classic way to account for the increase in Type I error is to reduce the required significance level. If we want to set the Type I error rate for m independent tests to be equal to 0.05, we set Eq. 5.1 to be equal to 0.05 and solve for α:

$$\alpha = 1 - (0.95)^{\frac{1}{m}} \approx \frac{0.05}{m} \tag{5.2}$$

To have a False Alarm rate of 0.05 across all m tests, you need

$$p < \frac{0.05}{m} \tag{5.3}$$

Hence, with multiple tests we need a smaller p-value for any given test to reach significance. Signal Detection Theory tells us that a more conservative criterion always involves a trade off in Hits and False Alarms. Indeed, power (Hits) strongly decreases when using Bonferroni correction.

Statisticians are not in general agreement about whether, or when, Bonferroni (or other similar) corrections are appropriate. Obviously, you should not treat m as the total number of hypothesis tests you will perform over a scientific career. Indeed, if you run hypothesis tests on very different topics, then it seems appropriate to have a separate Type I error rate for each topic and no correction would be necessary.

A variation of the multiple testing situation is the following. You collected a sample with a certain hypothesis, which turned out to *not* be significant. You decide to test further hypotheses. For example, maybe you find no difference in memory for men and women on some task. You then decide to test whether younger women perform differently than older women and whether younger men perform differently than older men. For each of these hypotheses there is a risk of a False Alarm and you need to correct for it. Hence, asking too many questions can be problematic. Although these tests are not independent, a Bonferroni correction might effectively control the Type I error rate.

5.2 Dependent Tests

Equation 5.1 holds when all tests are independent. When you use one data set to try to answer many questions, the tests may not be independent because the data is being used multiple times. Although a Bonferroni correction might work to restrict Type I error, it may be overly conservative; but the impact depends on the nature of the dependencies.

For example, suppose we sample from gold fish in a pond and are interested whether larger tails predict larger hearts. By accident we sampled fish that suggest there is such a relationship whereas in fact the population does not have such a relationship: we picked a sample that produced a False Alarm. Now, we test a second hypothesis from the same sample, namely, that larger tails predict larger lungs. Suppose there is a perfect correlation between heart and lung size; our second analysis will produce another False Alarm.

Assume you are asking 10 questions about the fish in the pond. If you are unlucky you got the wrong sample and 10 wrong answers to your questions. In general, whether or not data are correlated is usually unknown, which is one more reason to abstain from asking more than one question about a sample.

5.3 How Many Scientific Results Are Wrong?

As mentioned, the Type I error is usually set to 5%. One might expect that hence 5% of all scientific results, where classic statistics is used, are wrong. However, this is not true. The statement would be true if the effect size is 0 ($\delta = 0$) for all experiments conducted. However, scientists usually aim for real effects, so for many experiments it is likely that there is actually a true effect and, thus, no chance of making a Type I error. Assume scientists conduct only experiments where there is a real effect. In this case there are no Type I errors, since the null hypothesis is wrong for all experiments. Hence, the number

of wrong scientific results depends on the incidence rate (see Chap. 1) of no effect. This number is largely unknown and, thus, we do not know how many results are False Alarms (or misses).

Take Home Messages

1. You can only ask one question for a set of data. Otherwise you need to account for multiple comparisons.
2. Keep your designs as simple as possible.
3. If you cannot keep your design simple and have more than one group comparison, read the next chapter.