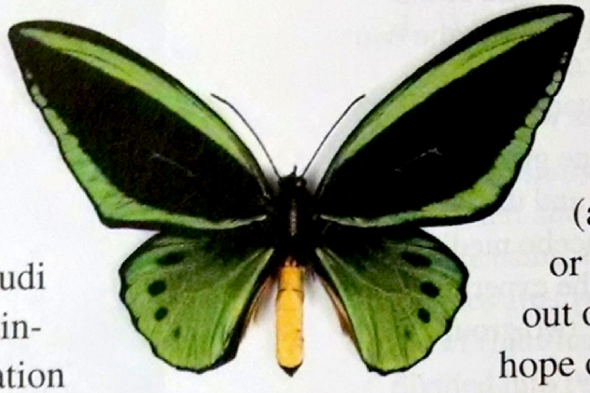
Data dredging



In the spoof journal *Annals of Improbable Research*, a satirical article reported on a study of the so-called butterfly effect (Inaudi et al. 1995). This effect, a mainstay of the popular representation of chaos theory, says that small initial causes, like the flapping of a butterfly's wings, can ultimately have large effects, like a hurricane, on the other side of the world. The fearless researchers set out to measure this effect by capturing several dozen butterflies and holding them in captivity in Switzerland. Each day, they checked the butterflies and recorded whether or not they flapped their wings. Then, using the lab's phone, they called their girlfriends in Paris each day to ask whether or not it was raining.[1] At the end of the study, the students tested each butterfly for an association between its daily flapping behavior and the daily weather in Paris. They found that the flapping days of one of the butterflies closely matched the rainfall days in Paris ($P < 0.05$). They exulted, "Not only have we proven that the butterfly effect exists, we have found the butterfly."

These guys were clearly joking, but statistically speaking, where did they go wrong? The answer is that they went "data dredging." They performed many statistical tests and eventually one of them was significant. Data dredging (also called "data snooping" or "data fishing") is the carrying out of many statistical tests in hope of finding at least one statistically significant result.

The problem with data dredging is that the probability of making *at least one* Type I error (i.e., of obtaining a false positive) is greater than the significance level α when many tests are conducted, if the null hypothesis is true (as it surely is in the butterfly example). Each hypothesis test has some chance of error, and these errors are compounded over multiple tests. There is a much larger probability of getting an error out of several tries than in any one try. By analogy, we might get away with playing Russian roulette once, but we would be unlikely to survive a month of playing once a night.

It's useful to do a few calculations to see how big the problem might be. The probability of making no Type I errors in N independent tests is $(1 - \alpha)^N$. Thus, the chance of making at least one Type I error from N independent tests is $1 - (1 - \alpha)^N$. This means that, if we use $\alpha = 0.05$ and carry out 20 independent tests of true null hypotheses, the probability that at least one of these tests will falsely reject the null hypothesis is about 65%. If we carry out 100

1. They continued the experiment "until the first phone bill reached our Office of Financial Services."

tests, then the chance of rejecting at least one of the null hypotheses becomes 99.4%, even if all the null hypotheses are true. With data dredging, a false positive result is almost inevitable.

Nevertheless, multiple testing is common in biology, and for good reasons. A dedicated experimentalist on human participants might measure many conceivable responses (e.g., blood pressure, body temperature, red blood cell count, white blood cell count, speed of recovery, appetite, and weight change) and perhaps even a few extra variables that might be long shots. The result is that the clinician might end up carrying out 10 or 20 tests of treatment effects, raising the probability of a false positive result. This level of multiple testing pales next to that seen in gene mapping. Locating a gene for a single trait, such as a genetic disease, typically involves thousands of statistical tests (one for each section of the genome). What should be done about the soaring Type I error rates resulting from so much testing?

The answer to this question depends on your goals. If your goal is simply to *explore* the data, to discover the possibilities but not to provide rigorous tests, then you need do nothing special about multiple testing except report the number of tests that you carried out and note which ones yielded a significant result. If you admit that you dredged the data, your results can still be useful. New hypotheses and unexpected discoveries can emerge from a thorough fishing expedition. However, the individual significant results that pop up from data dredging cannot yet be taken seriously, due to the high probability of one or more Type I errors. Some of the significant results might indeed be real, but it will be difficult to establish which ones. Rather, a new study must be carried out with new data to test any promising results that emerged from the exploratory approach. Another strategy sometimes used when exploring data is to divide the data randomly into two independent parts. One part is used for data dredging, and the other part is used to confirm any positive results suggested by the dredging.

If your goals from multiple testing are more rigorous (e.g., you want to determine which variable really did respond to treatment in a clinical trial, or which location in the genome really does contain a gene for a heritable disease), then steps must be taken to *correct* for the inflation of Type I error rates that occurs with multiple testing. The simplest way to accomplish this is to use a more stringent significance level—that is, one smaller than the usual $\alpha = 0.05$.

The most common correction for multiple comparisons is the **Bonferroni correction**. In the simplest version of this method, each test uses a significance level α^* rather than α, where

$$\alpha^* = \frac{\alpha}{\text{number of tests}}$$

For example, if we typically adopt the significance level $\alpha = 0.05$ when carrying out a single test, then to carry out 12 separate tests we should use the significance level $\alpha^* = 0.05/12 = 0.00417$ instead. In this case, we would reject H_0 in each test only if P were less than or equal to 0.00417. With the Bonferroni correction, the probability of getting at least one Type I error during the course of carrying out all 12 tests is approximately equal to the initial α-value (i.e., 0.05 in this case).

Keep in mind, though, that applying the Bonferroni correction greatly reduces the power of single tests. This is the price paid

for asking many questions of the data. More than ever, we should be mindful not to "accept the null hypothesis." It is okay to be skeptical when a null hypothesis is not rejected and power is so limited, but there is little to do about it except to repeat the study and look again.

Another, increasingly popular approach to correct for multiple comparisons is called the **false discovery rate** (**FDR**). To use this approach, carry out all of the multiple tests at a fixed significance level α (e.g., the usual 0.05). Gather all of the tests that yield a statistically significant result (i.e., all of the tests for which $P \leq \alpha$). We can call this subset of tests the "discoveries." The FDR estimates the proportion of discoveries that are false positives. In other words, the FDR is the proportion of tests for which the null hypothesis was rejected yet the null hypothesis was true. For example, Brem et al. (2005) carried out hundreds of statistical hypothesis tests of interactions between pairs of yeast genes. Of these tests, 225 yielded a statistically significant result (the "discoveries"). Using the false discovery rate method, they estimated that 12 of these 225 tests were false positives, leaving 213 "true" discoveries.

An extension of the FDR calculates a quantity called the q-value for each discovery. The q-value is analogous to a P-value, providing a measure of the strength of support from the data that the null hypothesis is false in a specific test. The smaller the q-value, the stronger is the evidence that H_0 is false and should be rejected. Unlike the P-value, the q-value takes into account other tests carried out at the same time. The idea is that, by choosing to reject H_0 only if the q-value is 0.05 or less, we reject the null hypothesis falsely in only 5% of tests. FDR and q-values are a more powerful approach to dealing with multiple comparisons, and we expect their use to increase in biological applications over the next decade. Consult Benjamini and Hochberg (1995) or Storey and Tibshirani (2003) for more details.