

CHAPTER 22



Multiple Comparisons Concepts

If you torture your data long enough, they will tell you whatever you want to hear.

JAMES L. MILLS (1993)

*C*oping with multiple comparisons is one of the biggest challenges in data analysis. If you calculate many P values, some are likely to be small just by random chance. Therefore, it is impossible to interpret small P values without knowing how many comparisons were made. This chapter explains three approaches to cope with multiple comparisons.

THE PROBLEM OF MULTIPLE COMPARISONS

If you make two independent comparisons, what is the chance of that one or both comparisons will result in a statistically significant conclusion just by chance? It is easier to answer the opposite question. Assuming both null hypotheses are true, what is the chance that both comparisons will be not statistically significant? The answer is the chance that the first comparisons will be not significant (0.95) times the chance that the second one will be not significant (also 0.95), or 0.9025. That leaves about a 10% chance of obtaining at least one statistically significant conclusion by chance.

It is easy to generalize that logic to more comparisons. With K independent comparisons (where K is some positive integer), the chance that all will be not significant is 0.95^K , so the chance that one or more comparison will be statistically significant is $1.0 - 0.95^K$. Figure 22.1 plots this probability for various numbers of independent comparisons.

Remember the unlucky number 13. If you perform 13 independent comparisons (with the null hypothesis true in all cases), the chance is about 50% that one or more of these P values will be less than 0.05, and thus lead to a conclusion of statistically significant.

With more than 13 comparisons, it is more likely than not that one or more conclusions will be significant just by chance. With 100 independent null hypotheses that are all true, the chance of obtaining at least one significant P value is 99%.

The multiple-comparisons problem is clear. If you make lots of comparisons (and make no special correction for the multiple comparisons), you are likely to find some statistically significant results just by chance.

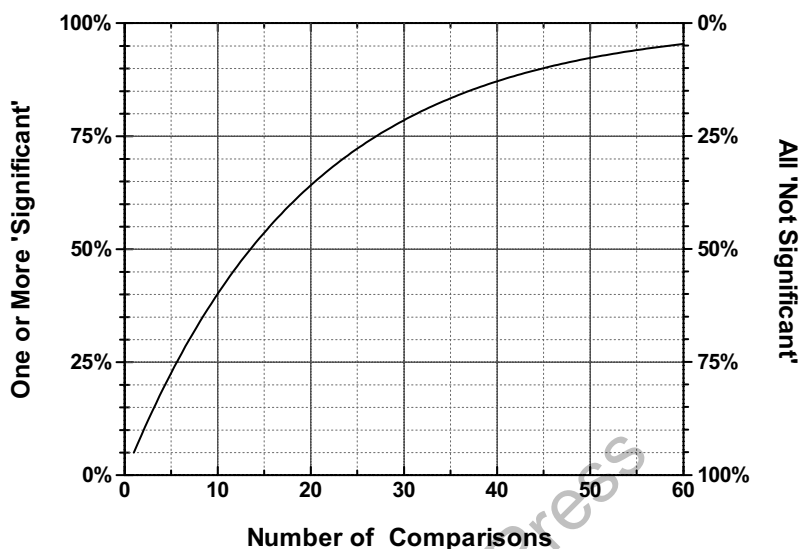


Figure 22.1. Chance of obtaining statistically significant results by chance.

The X-axis shows various numbers of statistical comparisons, each assumed to be independent of the others. The left Y-axis shows the chance of obtaining one or more statistically significant results ($P < 0.05$) by chance.

CORRECTING FOR MULTIPLE COMPARISONS IS NOT ALWAYS NEEDED

Corrections for multiple comparisons are not needed if the people reading the data take into account the number of comparisons

Some statisticians recommend never correcting for multiple comparisons (Rothman, 1990). The alternative is to report all of the individual P values and CIs and make it clear that no mathematical correction was made for multiple comparisons. This approach requires that all comparisons (or at least the *number* of comparisons) be reported. When you interpret these results, you must informally account for multiple comparisons. If all the null hypotheses are true, you'd expect 5% of the comparisons to have uncorrected P values less than 0.05. Compare this number with the actual number of small P values.

Corrections for multiple comparisons may not be needed if you make only a few planned comparisons

Even if a study collects lots of data, you may want to focus on only a few scientifically sensible comparisons, rather than every possible comparison. The term *planned comparison* is used to describe this situation. These comparisons must be designed into the experiment and cannot be decided upon after inspecting the

data. When you make only a few planned comparisons, many statisticians think it is OK to not correct for multiple comparisons.

Corrections for multiple comparisons are not needed when the comparisons are complementary

Ridker and colleagues (2008) asked whether lowering LDL cholesterol would prevent heart disease in patients who did not have high LDL concentrations and did not have a prior history of heart disease (but did have an abnormal blood test suggesting the presence of some inflammatory disease). The study included almost 18,000 people. Half received a statin drug to lower LDL cholesterol and half received a placebo.

The investigators' primary goal (planned as part of the protocol) was to compare the number of "end points" that occurred in the two groups, including deaths from a heart attack or stroke, nonfatal heart attacks or strokes, and hospitalization for chest pain. These events happened about half as often to people treated with the drug, compared with people taking placebo. The drug worked.

The investigators also analyzed each of the end points separately. Those taking the drug (compared with those taking placebo) had fewer deaths, fewer heart attacks, fewer strokes, and fewer hospitalizations for chest pain.

The data from various demographic groups were then analyzed separately. Separate analyses were done for men and women, old and young, smokers and nonsmokers, people with hypertension and those without, people with a family history of heart disease and those without, etc. In each of 25 subgroups, patients receiving the drug experienced fewer primary end points than those taking placebo, and all of these effects were statistically significant.

The investigators made no correction for multiple comparisons for all these separate analyses of outcomes and subgroups. No corrections were needed, because the results are so consistent. The multiple comparisons each ask the same basic question, and all the comparisons lead to the same conclusion—people taking the drug had fewer cardiovascular events than those taking placebo.

IF YOU DON'T ACCOUNT FOR MULTIPLE COMPARISONS

Table 22.1 shows the results of many comparisons. You can't create this table with actual data, because the entries in the rows assume that you are "mother nature" and therefore know whether each null hypothesis is actually true. In fact, you never know that, so this table is conceptual.

The top row represents the results of comparisons where the null hypothesis is in fact true—the treatment really doesn't work. The second row shows the results of comparisons where there truly is a difference. The first column tabulates comparisons where the P value was low enough to be deemed statistically significant (or a "discovery" in the lingo of the FDR method discussed below).

	DECISION: "STATISTICALLY SIGNIFICANT" OR "DISCOVERY"	DECISION: "NOT STATISTICALLY SIGNIFICANT" OR "NOT A DISCOVERY"	TOTAL
Null hypothesis: True	A	B	A + B
Null hypothesis: False	C	D	C + D
Total	A + C	B + D	A + B + C + D

Table 22.1. This table (identical to Table 18.1) shows the results of many statistical analyses, each analyzed to reach a decision to reject or not reject the null hypothesis.

The top row tabulates results for experiments where the null hypothesis is really true. The second row tabulates experiments where the null hypothesis is not true. When you analyze data, you don't know whether the null hypothesis is true, so you could never create this table from an actual series of experiments. A, B, C, and D are integers (not proportions) that count the number of analyses.

The second column tabulates comparisons where the P value was high enough to

APPROACH	WHAT YOU CONTROL	FROM TABLE 22.1
Significance level (α) with no correction for multiple comparisons	α = If all null hypotheses are true, the fraction of all experiments where the conclusion is statistically significant	$\alpha = A/(A + B)$
Familywise significance level	α = The chance of obtaining one or more statistically significant conclusions if all null hypotheses are true	$\alpha = \text{probability}(A > 0)$
False Discovery Rate (FDR)	Q = The fraction of all the discoveries where the null hypothesis really true	$Q = A/(A + C)$

Table 22.2. Three approaches to handling multiple comparisons.

be deemed not statistically significant (or not a discovery).

It would be nice if all comparisons ended up in cells B or C, leaving A and D empty. This is rarely the case. Even if the null hypothesis is true, random sampling will ensure that some comparisons will mistakenly yield a statistically significant conclusion and contribute to box A. And even if the null hypothesis is false, random sampling will ensure that some results will be not statistically significant and will contribute to box D.

A, B, C, and D each represent a number of comparisons, so the sum of $A + B + C + D$ equals the total number of comparisons you are making.

What happens if you make no correction for multiple comparisons and set α to its conventional value of 5%? Of all experiments done when the null hypothesis is true, you expect 5% to be statistically significant just by chance. You expect the ratio $A/(A + B)$ to equal 5%. This 5% value applies to each comparison separately, so is called a *per-comparison error rate*. In any particular set of comparisons, that ratio might be greater than 5% or less than 5%. But on average, if you make many comparisons, that is the value you'd expect. Table 22.2 summarizes three methods of dealing with multiple comparisons.

NO. OF "SIGNIFICANT" FINDINGS	NO CORRECTION (%)	BONFERRONI (%)
Zero	35.8	95.1
One	37.7	4.8
Two or more	26.4	0.1

Table 22.3. How many significant results will you find in 20 comparisons?

This table assumes you are making 20 comparisons, and all 20 null hypotheses are true. If there is no correction for multiple comparisons, there is only a 36% chance of observing no statistically significant findings. With the Bonferroni correction, this probability goes up to 95%.

THE TRADITIONAL APPROACH TO CORRECTING FOR MULTIPLE COMPARISONS

The Familywise Error Rate

When each comparison is made individually without any correction for multiple comparisons, the traditional 5% significance level applies to each individual comparison so it is known as *the per-comparison error rate*. It is the chance that random sampling would lead *this particular comparison* to an incorrect conclusion that the difference is statistically significant when this particular null hypothesis is true.

With multiple comparisons, the significance level is redefined to be the chance of obtaining *one or more* statistically significant conclusions if the *all* of the null hypotheses are actually true. The idea is to make a stricter threshold for defining significance. If α is set to the usual value of 5% and all the null hypotheses are true, then the goal is to have a 95% chance of obtaining zero statistically significant results and a 5% chance of obtaining one or more statistically significant results. That 5% chance applies to the entire family of comparisons performed in the experiment, so it is called a *familywise error rate* or the *per-experiment error rate*.

The Bonferroni correction

The simplest approach to achieve a family wise error rate is to divide the value of α (often 5%) by the number of comparisons. Then define any of the comparisons to be statistically significant only when its P value is less than that ratio. This is called the *Bonferroni method*.

Imagine that an experiment makes 20 comparisons. If all 20 null hypotheses are true and there are no corrections for multiple comparisons, about 5% of these comparisons are expected to be statistically significant (using the usual definition of α). Table 22.3 shows that there is about a 65% chance of obtaining one (or more) statistically significant result.

If the Bonferroni correction is used, a result is only declared to be statistically significant when its P value is less than $0.05/20 = 0.0025$. This ensures there is a 95% chance of seeing no statistically significant results among all 20 comparisons and only a 5% chance of seeing one (or more) statistically significant result.