PART E

Challenges in Statistics

CHAPTER 22

Multiple Comparisons Concepts

> If you torture your data long enough, they will tell you whatever you want to hear.
>
> MILLS (1993)

Coping 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 coping with multiple comparisons. Chapter 23 will show that the problem of multiple comparisons is pervasive, and Chapter 40 will explain special strategies for dealing with multiple comparisons after ANOVA.

THE PROBLEM OF MULTIPLE COMPARISONS

The problem of multiple comparisons is easy to understand

If you make two independent comparisons, what is the chance 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 not be statistically significant? The answer is the chance that the first comparison will not be statistically significant (0.95) times the chance that the second one will not be statistically significant (also 0.95), which equals 0.9025, or about 90%. 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 none will be statistically significant is 0.95^K, so the chance that one or more comparisons will be statistically significant is $1.0 - 0.95^K$. Figure 22.1 plots this probability for various numbers of independent comparisons.

Consider 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 will thus lead to a conclusion that the effect is statistically significant. In other words, with 13 independent comparisons, there is about a 50:50 chance of obtaining at least one false positive finding of statistical significance.

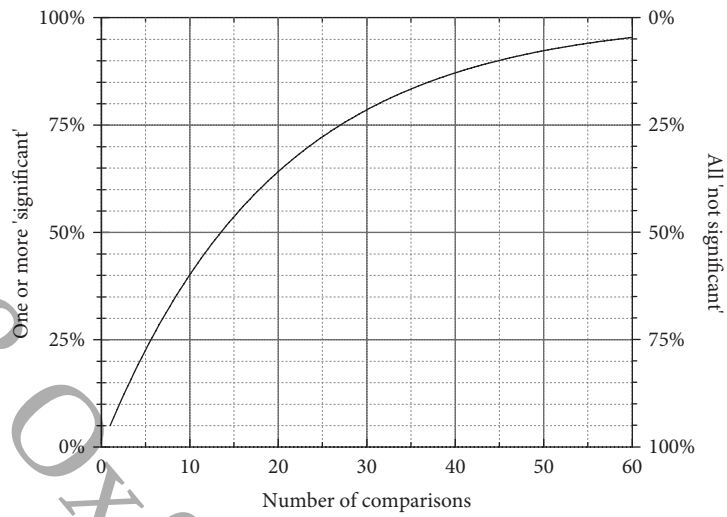

Figure 22.1. Probability of obtaining statistically significant results by chance.
The X-axis shows various numbers of statistical comparisons, each assumed to be independent of the others. The left Y-axis shows the chance of obtaining one or more statistically significant results ($P < 0.05$) by chance. The value on the left Y axis is computed as $1.0 - 0.95X$, where X is the number of comparisons.

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

A dramatic demonstration of the problem with multiple comparisons

Bennett and colleagues (2011) dramatically demonstrated the problem of multiple comparisons. They used functional magnetic resonance imaging (fMRI) to map blood flow in thousands of areas of a brain. Their experimental subject was shown a photograph of a person and asked to silently identify which emotion the person in the photograph was experiencing. The investigators measured blood flow to thousands of areas of the brain before and after presenting the photograph to their experimental subject. In two particular areas of the brain, blood flow increased substantially, and those differences were statistically significant ($P < 0.001$).

Sounds compelling. The investigators have identified regions of the brain involved with recognizing emotions. Right? Nope! The investigators could prove beyond any doubt that both findings were false positives resulting from noise in the instruments and were not caused by changes in blood flow. How could they be sure? Their experimental subject was a dead salmon! When they properly corrected for multiple comparisons to account for the thousands of brain regions at which they looked, neither finding was statistically significant.

CORRECTING FOR MULTIPLE COMPARISONS IS NOT ALWAYS NEEDED

Before considering two general approaches to correcting for multiple comparisons, let's pause to consider three scenarios in which corrections for multiple comparisons are not needed.

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

Some statisticians recommend that investigators never correct for multiple comparisons but instead report uncorrected P values and CIs with a clear explanation that no mathematical correction was made for multiple comparisons (Rothman, 1990). The people reading the results must then informally account for multiple comparisons. If all the null hypotheses are true, you'd expect 5% of the comparisons to have uncorrected P values of less than 0.05. Compare this number with the actual number of small P values.

This approach requires that all comparisons (or at least the *number* of comparisons) be reported. If the investigators only show the small P values, it is impossible to interpret the results.

Corrections for multiple comparisons are not essential when you have clearly defined one outcome as primary and the others as secondary

Many clinical trials clearly define, as part of the study protocol, that one outcome is primary. This is the key prespecified outcome on which the conclusion of the study is based. The study may do other secondary comparisons, but those are clearly labeled as secondary. Correction for multiple comparisons is often not used with a set of secondary comparisons.

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 planned as part of the experimental design. It is cheating to look at the data and then decide which comparisons you wish to make. When you make only a few preplanned comparisons, many statisticians think it is OK to not correct for multiple comparisons.

Corrections for multiple comparisons are not essential when the results are complementary

Ridker and colleagues (2008) asked whether lowering LDL cholesterol would prevent heart disease in patients who did not have high LDL concentrations, 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 the placebo. The drug worked.

The investigators also analyzed each of the end points separately. Those people taking the drug (compared with those taking the 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, and so on. In each of 25 subgroups, patients receiving the drug experienced fewer primary end points than those taking the 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, because these were planned as secondary analyses. The reader does not need to try to informally correct for multiple comparisons, 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 the placebo. In contrast, correction for multiple comparisons would be essential if the results showed that the drug worked in a few subsets of patients but not in other subsets.

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. This is therefore known as *the per-comparison error rate*, and 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.

When you correct for multiple comparisons, the significance level is redefined to be the chance of obtaining *one or more* statistically significant conclusions if *all* of the null hypotheses in the family 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*.

What is a family of comparisons?

What exactly is a family of related comparisons? Usually, a family consists of all the comparisons in one experiment or all the comparisons in one major part of an experiment. That definition leaves lots of room for ambiguity. When reading about results corrected for multiple comparisons, ask about how the investigators defined the family of comparisons.

The Bonferroni correction

The simplest approach to achieving a familywise error rate is to divide the value of α (often 5%) by the number of comparisons. Then define a particular comparison as statistically significant only when its P value is less than that ratio. This is called a *Bonferroni correction*.

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 22.1 and Figure 22.1 show 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, or 0.0025. This ensures that if all the null hypotheses are true, there is about 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 results. The 5% significance level applies to the entire family of comparisons rather than to each of the 20 individual comparisons.

Note a potential point of confusion. The value of α (usually 0.05) applies to the entire family of comparisons, but a particular comparison is declared to be statistically significant only when its P value is less than α/K (where K is the number of comparisons).

Example of a Bonferroni correction

Hunter and colleagues (1993) investigated whether vitamin supplementation could reduce the risk of breast cancer. The investigators sent dietary questionnaires

PROBABILITY OF OBSERVING SPECIFIED NUMBER OF SIGNIFICANT COMPARISONS		
NUMBER OF SIGNIFICANT COMPARISONS	**NO CORRECTION**	**BONFERRONI**
Zero	35.8%	95.1%
One	37.7%	4.8%
Two or more	26.4%	0.1%

Table 22.1. How many significant results will you find in 20 comparisons?

This table assumes you are making 20 comparisons, that all 20 null hypotheses are true, and that α is set to its conventional value of 0.05. 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%.

to over 100,000 nurses in 1980. From the questionnaires, they determined the participants' intake of vitamins A, C, and E and divided the women into quintiles for each vitamin (i.e., the first quintile contains the 20% of the women who consumed the smallest amount). They then followed these women for eight years to determine the incidence rate of breast cancer. Using a test called the chi-square test for trend, the investigators calculated a P value to test the null hypothesis that there is no linear trend between vitamin-intake quintile and the incidence of breast cancer. There would be a linear trend if increasing vitamin intake was associated with increasing (or decreasing) incidence of breast cancer. There would not be a linear trend if (for example) the lowest and highest quintiles had a low incidence of breast cancer compared with the three middle quintiles. The authors determined a different P value for each vitamin. For Vitamin C, P = 0.60; for Vitamin E, P = 0.07; and for Vitamin A, P = 0.001.

Interpreting each P value is easy: if the null hypothesis is true, the P value is the chance that random selection of subjects will result in as large (or larger) a linear trend as was observed in this study. If the null hypothesis is true, there is a 5% chance of randomly selecting subjects such that the trend is statistically significant.

If no correction is made for multiple comparisons, there is a 14% chance of observing one or more significant P values, even if all three null hypotheses are true. The Bonferroni method sets a stricter significance threshold by dividing the significance level (0.05) by the number of comparisons (three), so a difference is declared statistically significant only when its P value is less than 0.050/3, or 0.017. According to this criterion, the relationship between Vitamin A intake and the incidence of breast cancer is statistically significant, but the intakes of Vitamins C and E are not significantly related to the incidence of breast cancer.

The terminology can be confusing. The significance level is still 5%, so α still equals 0.05. But now the significance level applies to the family of comparisons. The lower threshold (0.017) is used to decide whether each particular comparison is statistically significant, but α (now the familywise error rate) remains 0.05.

Example of Bonferroni corrections: Genome-wide association studies

Genome-wide association studies (GWAS) use the Bonferroni correction to cope with a huge number of comparisons. These studies look for associations between diseases (or traits) and genetic variations (usually single nucleotide polymorphisms). They compare the prevalence of many (up to 1 million) genetic variations between many (often tens of thousands) cases and controls, essentially combining about 1 million case-control studies (explained in Chapter 28). To correct for multiple comparisons, the threshold is set using the Bonferroni correction. Divide α (0.05) by the number of comparisons (1,000,000) to calculate the threshold, which is 0.00000005 (5×10^{-8}).

A tiny P value (less than is 0.00000005) in a GWAS is evidence that the prevalence of a particular genetic variation differs in the two groups you are studying. One of the groups is composed of people with a disease and the other group is composed of controls, so one explanation is that the disease is associated with that genetic variant. But there is another possible explanation. If the patients and

controls are not well matched and have different ancestry (say the patients are largely of Italian ancestry and the controls are largely Jewish), you'd expect genetic differences between the two groups that have nothing to do with the disease being studied (see "The Challenge of Case-Control Studies" in Chapter 28).

CORRECTING FOR MULTIPLE COMPARISONS WITH THE FALSE DISCOVERY RATE

The false discovery rate (FDR) approach is an alternative approach to multiple comparisons that is especially useful when the number of simultaneous comparisons is large (Benjamini & Hochberg, 1995). To learn more about this approach, read the super clear nonmathematical review by Glickman, Rao and Schultz (2014).

Lingo: FDR

This approach does not use the term *statistically significant* but instead uses the term *discovery*. A finding is deemed to be a discovery when its P value is lower than a certain threshold. A discovery is false when the null hypothesis is actually true for that comparison. The FDR is the answer to these two equivalent questions (this definition actually defines the positive FDR, or pFDR, but the distinction between the pFDR and the FDR is subtle and won't be explained in this book):

- If a comparison is classified as a discovery, what is the chance that the null hypothesis is true?
- Of all discoveries, what fraction is expected to be false?

Controlling the FDR

When analyzing a set of P values, you can set the FDR to a desired value, abbreviated Q. If you set Q to 10%, then your goal is for at least 90% of the discoveries to be true and no more than 10% to be false discoveries (for which the null hypothesis is actually true). Of course, you can't know which are which.

A method developed by Benjamini and Hochberg (1995) sets the threshold values for deciding when a P value is low enough to be deemed a discovery. The method actually sets a different threshold value for each comparison. The threshold is tiny for the comparison with the smallest P value and much larger for the comparison with the largest P value. To understand why this makes sense, imagine that you computed 100 P values and all the null hypotheses were true. You'd expect the P values to be randomly distributed between 0.0 and 1.0. It would not be at all surprising for the smallest P value to equal 0.01. But it would be surprising (if all null hypotheses were true) for the median P value to be 0.01. You'd expect that value to be about 0.5. So it makes perfect sense to rank the P values from low to high and use that ranking when choosing the threshold that defines a discovery.

Here is a brief explanation of how those thresholds are determined. If all the null hypotheses were true, you'd expect the P values to be randomly scattered between zero and 1. Half would be less than 0.50, 10% would be less than 0.10, and so on. Let's imagine that you are making 100 comparisons and you have set

Q (the desired FDR) to 5%. If all the null hypotheses were true, you'd expect that the smallest P value would be about 1/100, or 1%. Multiply that value by Q. So you declare the smallest P value to be a discovery if its P value is less than 0.0005. You'd expect the second-smallest P value to be about 2/100, or 0.02. So you'd call that comparison a discovery if its P value is less than 0.0010. The threshold for the third-smallest P value is 0.0015, and so on. The comparison with the largest P value is called a discovery only if its value is less than 0.05. This description is a bit simplified but does provide the general idea behind the method. This is only one of several methods used to control the FDR.

If you set Q for the FDR method to equal alpha in the conventional method, note these similarities. For the smallest P value, the threshold used for the FDR method is α/k, which is the same threshold used by the Bonferroni method. For the largest P value, the threshold for the FDR method is α, which is equivalent to not correcting for multiple comparisons. P values between the smallest and largest are compared to a threshold that is a blend of the two methods.

Other ways to use the FDR

The previous section explained one approach: choose a desired FDR and use that to decide which results count as a discovery. In this approach, there is one FDR for the whole set of comparisons.

An alternative approach is to first decide on a threshold for defining discovery. For example, in a gene-chip assay, you might choose the 5% of genes whose expression changed the most. Given that definition, you would then compute the FDR. Again, there is one FDR for the entire set of comparisons.

A third alternative approach is to compute a distinct FDR for each comparison. For each comparison, define a threshold for discovery so that particular comparison just barely satisfies the definition. Using that definition, compute the overall FDR for all the comparisons. This value is called a q value (note the use of a lower-case letter). Repeat for each comparison. If you make 1,000 comparisons, you'll end up with 1,000 distinct q values.

COMPARING THE TWO METHODS OF CORRECTING FOR MULTIPLE COMPARISONS

Table 22.2 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 for which the null hypothesis is in fact true—the treatment really doesn't work. The second row shows the results of comparisons for which there truly is a difference. The first column tabulates comparisons for which the P value was low enough to be deemed statistically significant (or a discovery, in the lingo of the FDR method discussed earlier). The second column tabulates comparisons for which the P value was high enough to be deemed not statistically significant (or not a discovery).

	DECISION: STATISTICALLY SIGNIFICANT, OR DISCOVERY	DECISION: NOT STATISTICALLY SIGNIFICANT OR NOT A DISCOVERY	TOTAL
Null hypothesis: True	A (false positive)	B	A + B
Null hypothesis: False	C	D (false negative)	C + D
Total	A + C	B+D	A + B+ C + D

Table 22.2. This table (identical to Table 18.2) 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 for which the null hypothesis is really true. The second row tabulates experiments for which 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.

It would be nice if all comparisons ended up in Cells B or C, leaving A and D empty. This will rarely be 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 Cell 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 Cell 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.

If you make no correction for multiple comparisons

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 would expect 5% to be statistically significant just by chance. In other words, you would 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.

If you use the Bonferroni method to correct for multiple comparisons

If you use the Bonferroni method and set α to its conventional value of 5%, then there is a 5% chance of designating one or more comparisons (for which the null hypothesis is really true) as statistically significant. In other words, there is a 5% chance that the value A in Table 22.2 is greater than zero, and a 95% chance that A equals zero.

If you use the false discovery method to correct for multiple comparisons

In Table 22.2, the total number of discoveries equals A + C. If you set Q (the desired FDR) to 5%, then you would expect the ratio A/(A + C) to be no higher than 5% and the ratio C/(A + C) to exceed 95%.

Table 22.3 compares the three methods for dealing with multiple comparisons.

APPROACH	WHAT YOU CONTROL	FROM TABLE 22.1
No correction for multiple comparisons	α = if all null hypotheses are true, the fraction of all experiments for which the conclusion is statistically significant	$\alpha = A/(A + B)$
Bonferroni	α = if all null hypotheses are true, the chance of obtaining one or more statistically significant conclusions	α = probability that $A > 0$
False discovery rate	Q = the fraction of all discoveries for which the null hypothesis really is true	$Q = A/(A + C)$

Table 22.3. Three approaches to handling multiple comparisons.

Q & A

If you make 10 independent comparisons and all null hypotheses are true, what is the chance that none will be statistically significant?

If you use the usual 5% significance level, the probability that each test will be not statistically significant is 0.95. The chance that all 10 will be not significant is 0.95^{10}, or 59.9%.

Is the definition of a family of comparisons always clear?

No, it is a somewhat arbitrary definition, and different people seeing the same data could apply that definition differently.

If you make only a few planned comparisons, why don't you have to correct for those comparisons?

It makes sense to correct for all comparisons that were made, whether planned or not. But some texts say that if you only plan a few comparisons, you should get rewarded by not having to correct for multiple comparisons. This recommendation doesn't really make sense to me.

When using the Bonferroni correction, does the variable α refer to the overall familywide significance level (usually 5%) or the threshold used to decide when a particular comparison is statistically significant (usually 0.05/K, where K is the number of comparisons)?

The significance level α refers to the overall familywide significance level (usually 5%).

Will you know when you make a false discovery?

No. You never know for sure whether the null hypothesis is really true, so you won't know when a discovery is false. Similarly, you won't know when you make Type I or Type II errors with statistical hypothesis testing.

Are there other ways to deal with multiple comparisons?

Yes. One way is to fit a multilevel hierarchical model, but this requires special expertise and is not commonly used, at least in biology (Gelman, 2012).

Physicists refer to the "look elsewhere effect." How does that relate to multiple comparisons?

It is another term for the same concept. When physicists search a tracing for a signal, they won't get too excited when they find a small one, knowing that they also looked elsewhere for a signal. They need to account for all the places they looked, which is the same as accounting for multiple comparisons.

CHAPTER SUMMARY

- 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.
- Coping with multiple comparisons is one of the biggest challenges in data analysis.
- One way to deal with multiple comparisons is to analyze the data as usual but to fully report the number of comparisons that were made and then let the reader account for the number of comparisons.
- If you make 13 independent comparisons with all null hypotheses true, there is about a 50:50 chance that one or more P values will be less than 0.05.
- The most common approach to multiple comparisons is to define the significance level to apply to an entire family of comparisons, rather than to each individual comparison.
- A newer approach to multiple comparisons is to control or define the FDR.

TERMS INTRODUCED IN THIS CHAPTER

- Bonferroni correction (p. 207)
- Discovery (p. 209)
- Family of comparisons (p. 206)
- Familywise error rate (p. 206)
- Genome-wide association studies (GWAS) (p. 208)
- Multiple comparisons (p. 203)
- Per-comparison error rate (p. 206)
- Per-experiment error rate (p. 206)
- Planned comparisons (p. 205)

au: these terms do not appear in italics like the other terms on the list. Should we change?