Chapter 1

Introduction

1.1 The Multiplicity Problem

Practically every day, you find in the newspaper or other popular press, some claim of association between a stimulus and outcome, with consequences for health or general welfare of the population at large. Many of these associations are suspect at best, and often do not hold up under scrutiny. Examples taken from recent periodicals include the following claimed associations: cellular phones with brain tumors, power lines with leukemia (more recently overturned by the scientific community), vitamins with IQ, season of the year with mental performance (but only in men!), genetics with homosexuality (the "gay gene"), abortions with breast cancer (but not spontaneous abortions), remarriage with cancer, electric razors with cancer, and on and on. Many of these claims have shaky foundations *a priori*, and some have been found not to replicate in further studies. With so much conflicting information in the popular press, the general public has learned to mistrust the results of statistical studies, and to shy away from the use of statistics in general.

How do such incorrect conclusions become part of the scientific and popular landscape? While scientists typically fault such things as improper study design and poor data, there is another explanation that is the focus of this book. Data analysts can easily make such incorrect claims when they analyze data from large studies, reporting any test that is "statistically significant" (usually defined as "$p < 0.05$," where "p" denotes "p-value") as a "real" effect. On the surface, this practice seems innocuous. After all, isn't that the rule we learn in statistics classes—to report results where we find "$p < 0.05$" as "real"?

The problem, briefly stated, is that when multiple tests are performed, "$p < 0.05$" outcomes can often occur even when there are no real effects. Historically, the rule was devised for a single test, with the following logic: if the $p < 0.05$ outcome was observed, then the analyst has two options. He or she can believe that there is no real effect, and that the data are so anomalous that they are within the range of values that would be observed only 1 in 20 samples; or, he or she may choose to believe that the observed effect is real. Because the 1 in 20 chance is relatively small, the common decision is to "reject" the hypothesis of no real effect, and "accept" the conclusion that the effect is real.

This logic breaks down when you consider multiple tests or comparisons in a single study. If you consider 20 or more tests, then you *expect* at least one "1 in 20" significant outcome, even when none of the effects are real. Thus, there is little protection offered by the "1 in 20" rule, and incorrect claims can result. While problems of faulty study design, bad data, etc., can and do cause faulty conclusions, you should be aware that multiplicity is also a likely cause, especially in large studies where many tests or comparisons are made. Such studies are common, as the following examples indicate.

1.2 Examples of Multiplicity in Practice

Multiple comparisons and multiple tests occur in all areas of data analysis. The following sections contain descriptions of situations where the problem occurs, and discuss its practical consequences.

1.2.1 Multiple Comparisons in a Marketing Experiment

Suppose a market researcher shows five different advertisements (labeled, say, as A, B, C, D, and E) to focus groups of 20 males and 20 females. Advertisement E is the current one in circulation, and since there's a cost to pulling an old ad and starting up a new one, the market researcher would like to replace the current one with one that is assuredly better. Each person is shown all five ads via videotape, in random order, and each is allowed to return to previously viewed ads. At the end of the viewing each subject rates the ads on a standard set of attributes. Questions of interest include

- Is one of the new advertisements better than the old one?
- Are the males' ratings generally different than the females' ratings?

To answer these questions, researchers must perform many comparisons of advertisements, both within and between sexes. Without considering the multiple testing aspect, the analyst might be led to conclude, incorrectly, that advertisement "C" is better than "E," when in fact none of the new ads is really significantly better. In such a case, he or she might suggest a nationwide campaign for "C," potentially costing millions of dollars for no gain in revenue. With multiple comparisons methods, the conclusion of the data analysis is more likely to be that "E" and "C" are not significantly different.

On the other hand, if the conclusion that "C" is best is made after proper adjustment for multiple comparisons, then the analyst can proceed more confidently with the "C" recommendation.

An additional wrinkle to this problem is the analysis of the multiple questions on the questionnaire. The previous discussion presumes that there is a primary question of interest, such as "Overall, how much did you like this ad?" As such, the methodology is an example of multiple comparisons, although it is somewhat more complicated than usual with the different sources of variation (within and between subjects) and gender comparisons. However, in reality, there will be multiple questions pertaining to various aspects of "like" and "dislike." When all such questions are analyzed, the data analysis contains multiple tests as well as multiple comparisons.

Even in this simple example, there might be dozens or even hundreds of implicit multiple tests or comparisons. Thus, the opportunity for incorrect conclusions to arise by chance alone is great, unless the data are analyzed thoughtfully with this possibility in mind.

1.2.2 Subgroup Analysis in a Clinical Trial

As a part of the pharmaceutical development process, new therapies usually are evaluated using randomized clinical trials. In such studies, a cohort of patients is identified, and randomly assigned to either active or placebo therapy. After the conclusion of the study, the active and placebo groups are compared to see which is better, using a single predefined outcome of interest (e.g., whether the patient was cured). At this stage, there is no multiplicity problem, as there is only one test.

However, there are often good reasons to evaluate patient subgroups. The therapy might work better for men than for women, better for older patients, better for patients with mild conditions as opposed to severe, etc. While it is well and good to ask such questions, such data must be analyzed carefully, and with the multiplicity problem in mind. If the data are thus subdivided into many subgroups, it can easily happen that a patient subgroup shows "statistical significance" by chance alone, leading analysts to (incorrectly) recommend it for that subgroup, or worse yet, to recommend it for all groups based on the evidence from the single subgroup.

While such practice seems so obviously wrong, we recount two examples where it actually has happened. The first is reported in Fleming (1992), regarding a preoperative radiation therapy for colorectal cancer patients. The study was stopped early due to lack of significance; however, follow-up analysis revealed "significant" improvement in a particular subgroup. The trial's conclusions were then revised to recommend "universal use" of the therapy. A follow-up study involving the same therapy and a larger sample size revealed no statistical significance, so it seems likely that the original finding of a therapeutic effect was an incorrect claim, likely caused by the multiplicity effect.

Another case, reported in the *Wall Street Journal* (King, 1995) concerned the development of "Blue Goo," a salve meant to heal foot wounds of diabetic patients, by the Biotechnology firm ProCyte Corp. The firm decided to proceed with an expensive, large-scale clinical trial to assess efficacy of the salve, based on statistically significant efficacy results found in a subgroup of patients in a preliminary clinical trial. The larger study found no significant effect of the "Blue Goo" therapy, and as reported by King, "Within minutes [of the announcement of no therapeutic effect], ProCyte's stock fell 68% . . . " As in the case of the preoperative radiation treatment, it seems likely that the statistically significant result was an incorrect conclusion caused by the multiplicity effect.

1.2.3 Analysis of a Sociological Survey

Blazer et al. (1985), report results of a survey of residents of North Carolina who were distributed nearly equally between urban and rural counties. Psychiatric interviews and questionnaires were given to a randomly selected set of about 3,900 people, one per household. Each person was classified dichotomously (yes/no) as agoraphobic, alcohol-dependent, antisocial, cognitive deficient, dysthymic, major depressive, obsessive-compulsive, and schizophrenic. These classifications result in eight-dimensional binary vectors, one for each subject. For example, the vector $(0, 0, 1, 0, 0, 0, 1, 0)$ denotes a person who was diagnosed as antisocial and obsessive-compulsive.

One goal of the study was to relate the diagnoses to the demographic variables age, sex, race (white and non-white), marital status (married with spouse, separated/divorced, widowed, nonmarried), education (non-high school, high school), mobility (moved in last year, did not move), and location (rural, urban). With eight diagnoses and seven demographic classifications, there are a total of $7 \times 8 = 56$ tests, all of which are interesting comparisons. Without considering the effect of multiplicity, it is clear that erroneously significant results might be claimed. Our point here is not to quibble with the claims of Blazer et al., but merely to point out (1) how easy it is for multiple tests to arise with survey data, and (2) that the multiplicity effect should be carefully considered in any such analysis.

1.2.4 An Epidemiology Example: Data Snooping

With the advent of the Information Revolution, researchers have access to ever larger databases. Methods have been developed to "mine" such databases for otherwise hidden information. However, it's all too easy for such "data mining" to become "data snooping"—turning up nuggets of fools' gold (to continue with the metaphor) which are artifacts of excessive data manipulation rather than indicators of real lodes of useful information.

How do researchers keep “data mining” from becoming “data snooping”? Recognition of the problems of multiple inference can be the key. Many data mining procedures have built-in safeguards against such problems. For example, in fitting complex statistical models, data mining procedures often use a “penalty function” to avoid sample-specific overfitting problems. Similarly, procedures for fitting tree-based classification models often use multiplicity-adjusted rules to choose the splitting points.

The following example illustrates the potential dangers of data snooping. Needleman et al. (1979), claimed that lead in drinking water adversely affected IQs of school children. While high levels of lead are indisputably toxic, the study aimed to prove that variations in levels of lead below the accepted “safe” level were in fact associated with mental performance. Ernhart et al. (1981), in a critical review of their finding, claimed that the statistically significant conclusions were “probably unwarranted in view of the number of nonsignificant tests.” Ernhart, et al. essentially repeated the study and found no evidence for a decrease in IQ.

The analyses of Needleman et al. can be considered a classic case of “data snooping.” In their analysis, various covariates and subgroup analyses were performed in an effort to find statistical significance. It was only after such analyses that significant lead and IQ associations were found. As reported in Palca (1991), “the printouts show[ed] that Needleman’s first set of analyses failed to show a relationship between lead level and subsequent intelligence tests.”

1.2.5 Industrial Experimentation and Engineering

In industry the first phase of experimentation often begins with a screening experiment, where many factors are studied using only a few experimental runs. Since many factors are tested, there is a multiplicity problem: factors that are truly inert can be easily called “significant.”

As with any decision problem, errors of various types must be balanced against costs. In screening designs, there are costs of declaring an inactive factor to be active (Type I error), and costs of declaring an active effect to be inactive (Type II error). Type II errors are troublesome as addressed in Lin (1995). However, when there are enough runs in the experiment, linear regression and the usual *t* tests on the parameters provide sufficient protection against Type II errors; for saturated or nearly saturated designs, various other procedures have been devised (Box and Meyer, 1986; Lenth, 1989).

Type I errors also are troublesome, as they cause unnecessary experimental cost in the follow-up experiments, but are typically seen as having less importance than Type II errors in screening designs. Nevertheless, Type I errors are not necessarily free of cost. In particular, they can increase the cost of follow-up experimentation by including more factors than are really needed. Controlling Type I errors is a problem in multiple inference of the type considered in this book. While we consider Type II errors also to be important (see Chapter 7 in particular), the primary emphasis of most multiple comparisons and multiple testing procedures (including those in this book) is to find the most powerful method possible subject to global (familywise) Type I error control.

1.2.6 Identifying Clinical Practice Improvement Opportunities for Hospital Surgeries

As discussed by Pearce and Westfall (1997), health care has entered into the evidence-based decision making era. In no field is that more evident than cardiac surgery as evidenced by the publication of surgeon “report cards” of raw mortality data in New York and Pennsylvania newspapers (Green and Wintfeld, 1995). A principal reason for using such data is to identify continuous quality improvement (CQI) opportunities in clinical practice.

Hospital death, perioperative myocardial infarction, reoperation for bleeding, surgical wound infection, cerebrovascular accident, pulmonary complications, and renal failure are examined on a quarterly basis in these reports. Each of these adverse events is measured as a percentage of the total surgical procedures performed (individually and in total), and quarterly evaluations are made at the individual surgeon level. These examinations consist of testing the multiple hypotheses that each individual surgeon's outcomes for each adverse event do not differ significantly from the remainder of the group.

In order to drive out fear in the CQI process, the probability of declaring a false significance must be controlled. Without adjustment, the probability of declaring one surgeon worse than the others for at least one adverse outcome can approach 88 percent, even when the surgeons are identical in all respects except for patient assignment (assumed random). Such a high probability can cause fear and mistrust of the statistical methods. Pearce and Westfall (1997) used PROC MULTTEST to control this false significance probability at levels no higher than 5 percent, so that positive determinations could be viewed safely as a need for the improvement of a particular surgeon, and not as a spurious determination of differences between surgeons.

1.3 When Are Multiple Comparisons/Multiple Testing Methods (MCPs) Needed?

The previous examples show that multiple tests and multiple comparisons arise often in practice, and that improper conclusions can arise easily from such studies. In this book, we describe methods for overcoming the problem, and call such methods "MCPs," short for *Multiple Comparisons Procedures*, even though at times "MCP" will refer to a multiple testing method, or perhaps a simultaneous confidence interval method. Throughout this book, the term "MCP" will refer generically to *any* simultaneous inference procedure.

In general, then, when should you use an MCP? If *any* of the following apply to your multiple inferences, then you should be concerned about the multiple inference problem, and you should consider using an MCP. (Several of these are adapted from Westfall and Young 1993, p. 21.)

- It is plausible that many of the effects studied might truly be null.
- You want to ensure that any effects you claim are real, or reproducible, with the standard 95 percent level of confidence.
- You are prepared to perform much data manipulation to find a "significant" result. (For example, you perform many tests and play "pick the winner.")
- Your analysis is planned to be exploratory in nature, yet you still want to claim that any significant result is in fact real.
- Your experiment or survey is expensive and is unlikely to be repeated before serious actions are taken.
- There is a cost, real or implicit, that is associated with incorrectly declaring effects or differences to be "real."

1.4 Selecting an MCP Using This Book

Before deciding which test or procedure to use, you need to identify the three main components of your problem:

1. the assumptions of the statistical model that you are using
2. the comparison or testing objectives of your study
3. the collection of items that you want to test.

Once you have determined these three elements, you can identify an appropriate method of inference. What follows is a brief overview of the elements of each, with sections in the book where each item is discussed. Note that the chapters of this book are primarily arranged around 1, the assumptions of the model, with elements of 2 and 3 filling the subsections.

1.4.1 Statistical Modeling Assumptions

The choice of a statistical model is a completely separate issue from multiple tests and multiple comparisons, and is a choice that you must make before using any statistical procedure. Failure to identify an appropriate model invalidates MCPs, just as it invalidates any statistical procedure. Also, failure to use the structure of the data completely can result in inefficient methods. For example, methods that assume independence of comparisons or tests usually are valid, in the sense of controlling error probabilities, but are inefficient when compared to methods that fully utilize correlation information.

The following list contains major statistical model classes covered in this book:

Unstructured Models (or Models with Little Structure)

These are models where little is assumed about distributions, correlations, etc. Nonparametric procedures fall in this class. The models for the actual data in this case may be quite complicated, but the assumption is that the analysis has been distilled down to a collection of p-values. Multiple inference methods in this class consist essentially of adjusting these p-values for the purposes of making tests. Such methods work reasonably well for a variety of models, and if you have a model that is not contained in one of the major classes given below, then you can choose an MCP that assumes little structure. In particular, these methods are valid, though somewhat conservative, for all correlation structures, and can be termed “Generalized Bonferroni Methods.” See Chapter 2.

Balanced One-way Analysis-of-Variance (ANOVA)

These are models for data from experiments where several groups are compared, and where the sample sizes are equal for all groups. Independence of data values is a crucial assumption for these models; and if they are not independent, then you might be able to use one of the alternatives listed below. Other assumptions strictly needed for these models are homogeneity of error variance and normality of the observations within each group, but these are not as important as the independence assumption (unless severely violated). See Chapter 3.

Unbalanced One-way ANOVA, or Analysis-of-Covariance (ANCOVA)

These data are similar to the balanced ANOVA except that sample sizes may be unbalanced, or the comparisons between means might be done while controlling one or more covariates (e.g., confounding variables, pre-experimental measurements). The distributional assumptions are identical to those of the ANOVA, with the exception that for ANCOVA, the normality assumption must be evaluated by using residuals and not actual data values. See Chapters 4 through 6.

Two-way and Higher-Way ANOVA

In these cases, you consider the effects of two or more factors, with possibly unbalanced sample sizes and/or covariates. The distributional assumptions are the same as for the unbalanced one-way ANOVA or ANCOVA (if there are covariates). See Chapter 9.

Repeated Measures ANOVA Data

When there are repeated measures on the same experimental unit, the crucial independence assumption that is used for the previous models no longer applies. For example, the data may contain repeated measures on blood pressure for an individual. In such cases, you can model the dependence of blood pressure measurements by using a variety of possible dependence structure models, and perform multiplicity-adjusted analyses within the context of such models. See Chapter 10.

Multivariate Responses with Normally Distributed Data

In these models, there are multiple measurements on the same individual. While repeated measures models usually assume that the measurements are taken on the same characteristic (like blood pressure), the multivariate response models allow completely different scales of measurement. For example, blood pressure and self-rated anxiety level form a multivariate response vector. Multiple inferences from such data are improved by incorporating the correlations among such measurements. In addition to the normality assumption, the multivariate observation vectors also are assumed independent, with constant covariance matrices. Our suggested method of analysis will allow covariates as well, so you can perform multiple comparisons with multivariate analysis of covariance (MANCOVA) data. See Chapter 10.

Nonnormally Distributed (but Continuous) Data

If the distributions are nonnormal, you still can make approximate inferences with multiplicity adjustment, using bootstrap and permutation methods. The general structure of the data is that the observation (vectors) are assumed independent, and the covariance matrices are assumed constant. However, the distributional form is not specified. The methods described herein also are valid if the distributions are normal. See Chapter 11.

Binary and Discrete Data

If your observations are binary (or more generally, if your distributions used for testing are discrete distributions), then there are fantastic gains in power that may be achieved for the multiple testing methods. An example was given previously in Section 1.2.3, where the observation vectors indicate presence or absence of a number of psychiatric conditions. In Chapter 14 we also give an application of large-sample multiple inferences from a logistic regression model for a binary outcome. See Chapters 12 and 14.

Heteroscedastic Responses

If the error variances are not constant, then the ordinary methods might be biased (in the sense of providing higher error rates than advertised) or inefficient (in the sense that the method lacks power to detect real differences). See Chapter 11.

Time-to-Event or Survival Data

If your data consist of time until an event (like death), with many censored observations, you can perform the multiple comparisons in a way that accounts for finite-sample discreteness of the observations (Chapter 12), or which uses large-sample approximations from a proportional-hazards model or a parametric survival analysis model. See Chapters 12 and 14.

1.4.2 Multiple Comparisons/Multiple Testing Objectives

Different MCPs may address different inferential objectives, so which procedure you should choose depends on which kinds of inferences you want to make. Perhaps the major distinction is whether you want to simply assess mean equality or whether you want to go further and construct confidence intervals for mean differences. A related decision is the choice of which error rate you want to control, though this is a decision to be approached cautiously. Or you might want to use an informal, graphically based method, rather than any formal error-rate-controlling method at all.

The following list contains major types of multiple inference methods, along with sections in the book where they are described. The types of inference are ordered from strongest to weakest, in a sense to be defined below, according to a classification first made by Hsu (1996).

Confidence Interval-Based Methods

These methods are useful for providing an explicit range of values for each parameter of interest. Such intervals are useful also for determining directional relationships and statistical significance. Confidence intervals are discussed throughout the book. Sections 2.3.1 and 2.3.2 define the concept, and Chapters 3 through 7 are devoted primarily to confidence interval applications. Further intervals-based applications are found later in the book, side-by-side with testing applications.

Confident Directions Methods

These methods allow you to assert inequalities involving parameters of interest—for example, that the mean for one group is less than the mean for another—without being able to give a likely range of values. Confident directions methods are introduced in Chapter 8, primarily in the context of one-sided stepwise testing methods.

Testing-Based Methods

You would use these methods if you just want to make yes/no decisions concerning hypotheses of interest. Many such methods are conveniently discussed within the context of "closed testing procedures," which we discuss in detail. Chapter 2 and Chapter 8 contain the fundamental ideas and applications of multiple testing. Further applications are given in Chapters 9 through 14.

Tests of Homogeneity

With these methods, all you can say is whether or not the hypotheses of interest are all true, without identifying which ones might be false. Such methods only control Type I errors in the "weak" sense, not in the more appropriate "strong" sense. Frankly, methods in this class are usually applied erroneously, with the mistaken idea that they provide the same type of inference as the stronger methods. Therefore, we will discuss these methods mainly in order to discourage their use.

Each item in this list provides weaker inference than the ones above it. For example, simultaneous confidence intervals for differences between means can be used to infer equality or inequality, but multiple tests for inequality cannot always be converted into confidence intervals. Conversely, methods that provide stronger inferences are often less powerful than those tailored specifically for less ambitious results. For example, if the goal of your study is just to make yes/no decisions concerning mean equality, then you can use a testing-based method with much greater power than interval-based procedures, while maintaining error rate control.

As far as error rates are concerned, the standard methods are those that control the "Familywise Error Rate" (or FWE) in the "Strong" sense (defined in Section 2.3.3). However, you might choose an alternative error rate to control, such as the "False Discovery Rate", discussed in Section 2.3.5. Also, sometimes tests of homogeneity are viewed as providing another, "weak" alternative to strong control of the FWE. **Note:** you should select a nonstandard error rate *only after careful consideration of the consequences* of choosing an alternative to the strong control methods, which should be considered the "gold standard" of MCPs.

In some cases, the results of multiple inferences can be displayed nicely in graphs. For confidence interval applications of graphical display, see Section 3.3.2; for hypothesis testing applications see Section 2.6.

1.4.3 The Set (Family) of Elements to Be Tested

The type of MCP that is best for your data also depends on the set of elements which you want to compare. To control error rates, this set of items must be stated in advance, and strictly adhered to. Otherwise, the analysis is called “data snooping,” as discussed in Section 1.2.4.

Here are some families of elements that you might want to test:

All Pairwise Comparisons in the ANOVA

Here, you decide to compare each mean value with every other mean value, which is useful to obtain a confident relative ranking of treatment means. This application is discussed primarily in Section 3.3, with additional applications in all remaining chapters.

All Pairwise Comparisons with the Control

If you decide, *a priori*, that your interest is in comparisons of individual groups against a standard (or control), and not against each other, then more power can be attained. This application is discussed primarily in Section 3.4, again with additional applications throughout the book.

Multiple Comparisons with the Best

If your interest only concerns comparing treatment means with the (unknown) “best” (highest or lowest, depending on the application) treatment mean, see Section 14.2.

General Contrasts

If your interest is in a general set of predefined contrasts, such as orthogonal contrasts, or cell means comparisons in a two-way ANOVA, see Section 3.5.2 and Chapter 6, with additional examples given throughout the book.

Dose-Response Contrasts

Sometimes the goal of multiple testing is to find the minimum effective dose. For this application, multiple dose-response comparisons are of interest; see Sections 6.2.2 and 8.5.

Comparisons of Multivariate Measures across Two or More Groups

The preceding applications generally presume multiple treatment groups and a univariate measure. If you have multivariate measures as well as multiple treatment groups, you might want to compare treatment groups for every one of the multivariate measures. This application is discussed in Chapters 10, 11, 12, and 13.

Infinitely many Comparisons

Although this category sounds like “data snooping,” it is actually permissible when done properly. See Sections 6.3 and 14.3.

General Comparisons or Tests, Unstructured

General methods can be recommended for cases where the family is specified, but does not fit precisely into any of the categories above. These are given in Chapter 2.

Confidence Bounds for Regression Functions

These applications are discussed in Sections 6.3.2 through 6.3.5.

1.5 Controversial Aspects of MCPs

We would be wrong to suggest that all multiple testing inference issues are resolved by selecting an appropriate MCP, as suggested in the outline above, and proceeding. With MCPs, as with any statistical inference method, there is never one and only one method that is “the one and only correct method” for the analysis of any data. However, with MCPs, this issue is greatly compounded in that there can be enormous differences between the results obtained either with or without multiplicity adjustment; and there can be dramatic differences also depending upon the approach that you take to analyzing the data. This section discusses briefly some of the controversies.

1.5.1 Size of the Family

The size of the discrepancy between multiplicity-adjusted and nonmultiplicity-adjusted analysis is largely determined by the size of the “family” of tests considered: if you allow more inferences into your family, then your inferences are dramatically altered. Specifically, the larger the family, the less significant the results become.

Therefore, critics of MCPs point out that it seems easy to “cheat”; if your goal is to prove significance, then you can pare the family down to a suitably small size until statistical significance is obtained. Conversely, if your goal is to prove insignificance, then you can increase the family size until no significances remain.

There is a line of research that suggests *not* to multiplicity-adjust statistical tests, see Saville (1990), Rothman (1990), Cook and Farewell (1996), and Bailer (1991), among others. There are several issues brought up by these authors. First, the choice of the “family” is somewhat arbitrary, and inferences are *extremely* sensitive to the choice. Therefore, these authors argue that the most objective choice of a family is the test itself. Second, all MCPs lose power relative to the unadjusted methods. Thus, when Type II errors are considered as important or more important than Type I errors, the authors argue that some Type I error control should be sacrificed for the sake of controlling Type II errors. Third, these authors argue for unadjusted methods, but with complete disclosure of data analysis procedures, so that users can decide for themselves whether some of the claimed results are false significances.

Taken to its extreme, this practice of not considering multiplicity may cause scientists and experimenters to ignore completely the multiplicity problem. Appropriate use of multiple testing is a difficult and controversial subject; however, ignoring the problem will make it much worse, as shown in the examples of Section 1.2. Also, ignoring the problem makes it difficult for reviewers of scientific manuscripts to separate facts from Type I errors.

In response to these controversies, our view is that multiplicity effects are real, and that Type I errors can and do occur. You need to be aware of the various error rates to interpret your data properly. In answer to the issue concerning size of the family, our recommendation is to choose smaller, more focused families rather than broad ones, and that such a determination must be made *a priori* (preferably in writing!) to avoid the “cheating” aspect. Finally, assuming that you do decide to use a multiplicity adjustment method, you should use one that is as powerful as possible, subject to the appropriate error level constraint. In this book, you will find several examples of such methods.

1.5.2 Composite Inferences vs. Individual Inferences

Another controversial aspect of multiple testing is whether to analyze the data using a single composite inference (e.g., using meta-analytic procedures), or to require individual inferences. What is at issue is essentially the required strength of inference, as discussed in Section 1.4.2. You must make this choice on the basis of the subject matter under study, depending on what conclusions you want to be able to make. If your goal is to find whether there is a difference, overall, and you are not concerned with individual components that comprise the difference, then the composite inference is usually better (more powerful) than the individual, multiplicity-adjusted inferences. Here is an example to illustrate the difference:

EXAMPLE: Multiple Tests of ESP

While controversial, testing for extrasensory perception (ESP) has attracted interest in the scientific and government communities, particularly as it concerns possible application to international espionage (as discussed in Utts, 1995). While individual tests of significance of ESP might show marginal significance, such evidence usually disappears with appropriate definition of a family of tests and with analysis via an appropriate MCP. However, in this case it is perhaps more interesting to know whether ESP exists at all than whether ESP is found in a particular test, for a particular person. Utts (1991) discusses omnibus (meta-analytic) methods for such combined tests, finding convincingly significant evidence "for" the existence of ESP. (For discussions and rebuttals of the claims see the discussions following Utts' 1991 article.)

1.5.3 Bayesian Methods

(This section is written for Bayesians; if you are not a Bayesian, or if you don't know whether or not you are a Bayesian, then you may skip this section.)

We owe you (the Bayesian reader) an apology. Historically, the development of MCPs has been mostly along frequentist lines, and therefore, the methods that are commonly used are very non-Bayesian in flavor. In this book, our aim is to explain the commonly used tools for the analysis of multiple inferences, and since these methods are mostly frequentist, our discussions will largely follow the frequentist philosophy.

In simple inferences, there often are correspondences between frequentist and Bayesian methods that are comforting, and allow you to "compute as if a frequentist," but still to "act like a Bayesian." For example, the usual confidence intervals computed frequentist-style are Bayesian posterior intervals for suitable (usually flat) prior distributions. Similarly, p-values from one-sided tests of hypotheses that are calculated frequentist-style can be interpreted as Bayesian posterior probabilities, again with suitable priors (Casella and Berger, 1987). The correspondences break down somewhat in the case of two-sided tests as shown by Berger and Sellke (1987); nevertheless, there are broad correspondences that can be drawn even in that case.

Historically, there has been no such correspondence between frequentist and Bayesian methods in the case of multiple inferences that would allow you to take some comfort in the usual frequentist MCPs, should you be a Bayesian. It is, therefore, this issue of multiple comparisons that has, perhaps more than any other issue in statistics, polarized the Bayesian and frequentist communities, as recounted in Berry (1988) and Lindley (1990).

Recently, however, Westfall, Johnson and Utts (1997) demonstrated that some frequentist MCPs correspond roughly to Bayesian methods. The first list item in Section 1.3, which suggests that multiple inference methods are needed when it is suspected that many or all null hypotheses might be true, essentially refers to a Bayesian assessment of prior probabilities. If this condition holds, then, as noted by Westfall, Johnson, and Utts (1997), frequentist and Bayesian methods "need not be grossly disparate."

If you are in the Bayesian camp, we are sympathetic to your concerns. Please bear with us through the frequentist developments, keeping the idea in mind that frequentist and Bayesian conclusions need not be grossly disparate, when there is prior doubt about many of the hypotheses tested. We present methods that have Bayesian rationale in Chapter 13 of this book.