

Introduction



1.1 The Multiplicity Problem.....	2
1.1.1 Basic Statistical Concepts.....	3
1.2 Examples of Multiplicity in Practice.....	5
1.2.1 Multiple Comparisons in a Marketing Experiment.....	5
1.2.2 Multiple Endpoints in a Clinical Trial	6
1.2.3 Subgroup Analysis in a Clinical Trial	6
1.2.4 Analysis of a Sociological Survey	7
1.2.5 An Epidemiology Example: Data Snooping Versus Data Mining	7
1.2.6 Industrial Experimentation and Engineering.....	8
1.2.7 Identifying Clinical Practice Improvement Opportunities for Hospital Surgeries.....	9
1.2.8 Genomics Data and Large-Multiple Testing.....	9
1.3 When Are Multiple Comparisons/Multiple Testing Methods Needed?	10
1.4 Selecting an MCP Using This Book	10
1.4.1 Statistical Modeling Assumptions.....	10
1.4.2 Multiple Comparisons/Multiple Testing Objectives	13
1.4.3 The Set (Family) of Elements to Be Tested	14
1.5 Controversial Aspects of MCPs.....	15
1.6 Chapter 1 Problems.....	18
1.7 Solutions to Chapter 1 Problems	19

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 an 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. Controversial examples of associations that have been published in the popular press include cellular phones with brain tumors, power lines with leukemia, vitamins with IQ, season of the year with low mental performance, genetics with homosexuality, abortions with breast cancer, remarriage with cancer, electric razors with cancer, and on and on. Many such claims have failed to replicate in further studies: a study by Ioannidis (2005) finds a surprisingly large number of replication failures among “influential” medical studies, Bofetta et al. (2008) discuss false positives in cancer epidemiology, and Bertram et al. (2007) discuss false positives in genetic association 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 $p \leq 0.05$, as a “real” effect. (Section 1.1.1 below reviews the definition of the “ p -value” and related statistical concepts.) On the surface, this practice seems innocuous. After all, isn't that the rule you learned in statistics classes—to report results where “ $p \leq 0.05$ ” as “real”?

The problem, briefly stated, is that when multiple tests are performed, “ $p \leq 0.05$ ” 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 $p \leq 0.05$ outcome was observed, then you have two options:

- Option 1: Because data this extreme are unlikely when there is no true effect, you may choose to believe the observed effect is real. (You commit a *Type I error* if indeed there is no true effect.)
- Option 2: You can say that such data aren't unlikely *enough* and decline to decide that the effect is real. (You commit a *Type II error* if there really is an effect).

Because the 1 in 20 chance of a Type I error 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 $p \leq 0.05$ outcome, even when none of the effects is real. Further, the probability of *at least one* incorrect $p \leq 0.05$ outcome is, with k independent tests, $1 - 0.95^k$, which equals 64% when $k=20$, 92% when $k=50$, and 99.4% when $k=100$! Thus, with multiple tests, 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 mistaken 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 examples in Section 1.2 indicate.

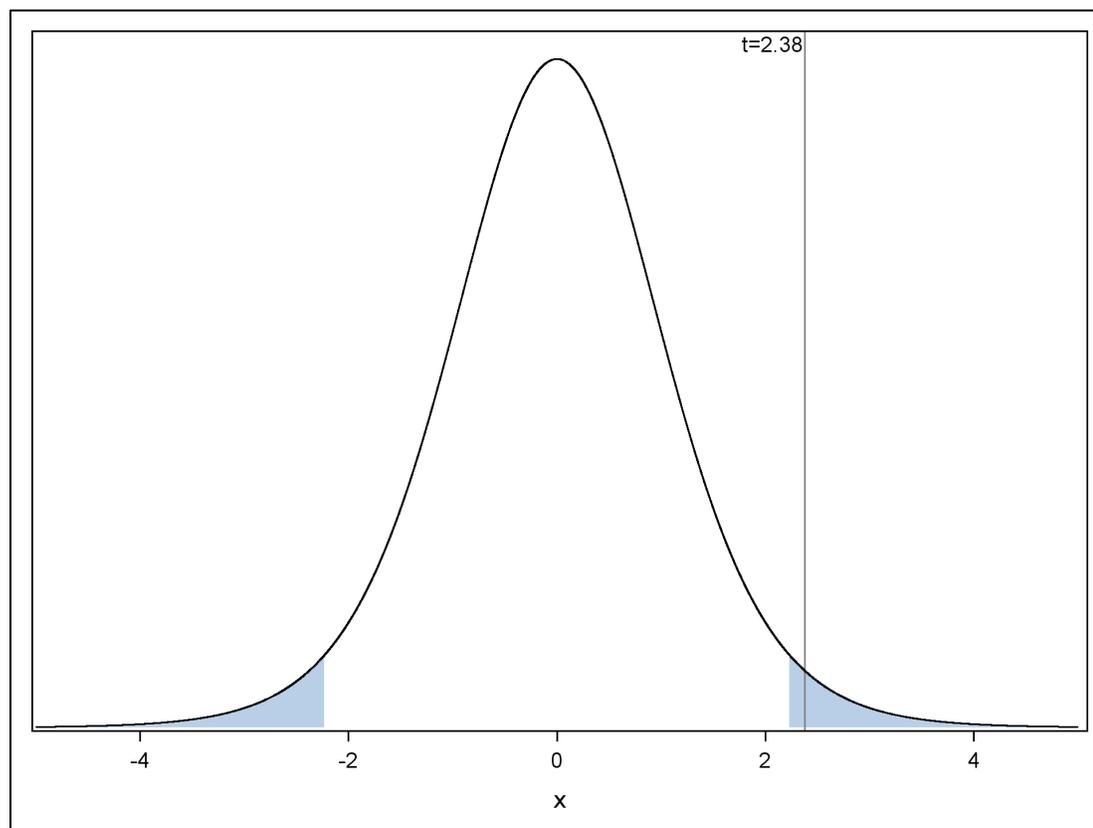
Before proceeding, here is a review of some basic statistical concepts.

1.1.1 Basic Statistical Concepts

One of the most widely used tests is the two-sample comparison. For example, say there are two groups, drug and placebo, and the goal is to see whether the drug is any different from the placebo. Participants are randomly divided into two groups, with n_1 participants in one group and n_2 in the other; the total number of participants is $n = n_1 + n_2$.

The hypotheses described here are for the two-sample t -test, a common test for comparing two groups. The basic elements of hypothesis testing and error rates are quite similar for other testing applications. The assumptions of the two-sample t -test are important: random, independent samples from the two groups, common variances, and normally distributed data. These assumptions can be relaxed in some cases that are discussed later in this book.

- The *null hypothesis* is $H_0 : \mu_1 = \mu_2$; that is, the hypotheses that the population means are equal.
- The *alternative hypothesis* is $H_A : \mu_1 \neq \mu_2$; that is, the hypotheses that the population means are not equal.
- The *test statistic* is $T = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, where $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$.
- The *decision rule* is to reject H_0 if $|T| \geq t_{1-\alpha/2, n-2}$, where $t_{1-\alpha/2, n-2}$ is the *critical value*.
- The p -value is the probability of observing a test statistic as large as or larger than the $|T|$ that was observed in the study, assuming the null hypothesis is true. See Figure 1.1.

Figure 1.1 Hypothesis Testing Graph

The p -value for the graph of Figure 1.1 is the area under the t distribution beyond the observed test statistic value 2.38, plus the area under the t distribution beyond the negative of the observed test statistic value, -2.38 . The shaded region is the rejection region, with critical value approximately 2.2 from the graph, and its area is 0.05. So the p -value is less than 0.05 in Figure 1.1.

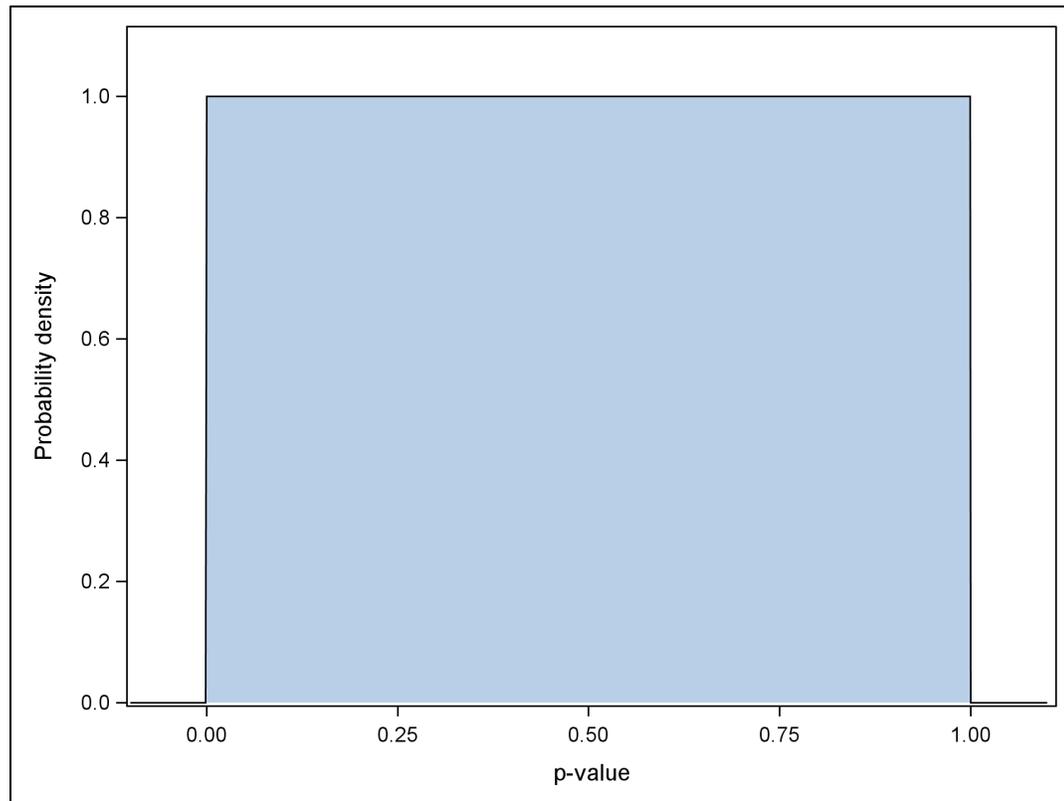
By construction, the p -value is found $\leq \alpha$ wherever $|T| \geq t_{1-\alpha/2, n-2}$. Thus, when all of the assumptions are satisfied,

$$P(p \leq \alpha | H_0 \text{ is true}) = \alpha.$$

This leads to an important point that you will see repeatedly throughout this book:

- ✓ When the null hypothesis is true and when all assumptions are satisfied, the p -value has a *uniform distribution*.

The uniform distribution is a continuous distribution between 0 and 1 with all values equally likely, as shown in Figure 1.2.

Figure 1.2 Uniform Distribution of p -Value When H_0 Is True

1.2 Examples of Multiplicity in Practice

As indicated above, statistical methods that correctly take multiplicity effects into account will generally make you more cautious and conservative about declaring observed effects as real. The need for this conservatism in multiple comparisons and multiple tests arises 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 the following:

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

To answer these questions, researchers must perform many comparisons of advertisements, both within and between sexes. Determining that advertisement C is better than E will launch a multimillion-dollar nationwide campaign. So in this case, lack of repeatability in the population of any putative effect seen in the small-scale study means that all this money is wasted.

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, real studies like this usually involve multiple questions about preferences, such as “Did this ad make you want to purchase the product?” When all such questions are analyzed, the data analysis contains many more comparisons than the comparisons between ad types.

Thus, even in this simple example, there might be dozens of multiple tests or comparisons. 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 Multiple Endpoints in a Clinical Trial

In pharmaceutical development, there are multiple phases of clinical trials in which the effects of therapies (often pharmaceutical compounds) on subjects are evaluated for safety and efficacy. Typically, there are many ways to measure both safety and efficacy. In Chapter 2 and later in the book, a hypothetical clinical trial to evaluate a remedy for the common cold is discussed. There are many symptoms of the common cold, including coughing, sneezing, runny nose, and itchy eyes. The medicine might relieve all, some, or none of these symptoms. Each of these symptoms, when measured, is called an “endpoint.” Because there are four of these endpoints, the study is said to have “multiple endpoints.”

Regulatory agencies such as the United States Food and Drug Administration are charged with ensuring the safety and efficacy of pharmaceutical products, such as our hypothetical cold remedy. When there are multiple endpoints, there are more ways to “win” in the clinical trial; that is, there are more ways to obtain a statistically significant result by chance alone. Multiple comparisons procedures are used to control the probability of such chance occurrences to an acceptably low level.

For safety, it is common practice to perform hundreds and sometimes even thousands of tests for disproportionality of adverse events, concomitant medications, laboratory measurements, and vital signs across treatment and control groups. Accurate understanding of the tradeoff of false positives and true positives in these tests and the determination of appropriate critical values are critical to the success of the treatment and its risk assessment for the public at large.

There are many sources of multiplicity in clinical trials other than multiple endpoints, including multiple dose comparisons, subgroup analysis, and interim analysis; these are discussed throughout the book.

1.2.3 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, often using a single predefined outcome of interest (e.g., whether the patient was cured). Assuming there are not multiple primary endpoints as discussed in Section 1.2.2 above, there is no multiplicity problem, because 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 with the multiplicity problem in mind. If the data are 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, it actually has happened! A first example 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 a statistically significant ($p \leq 0.05$) improvement in a particular subgroup. The trial's conclusions were then revised to recommend the therapy not only for the subgroup, but for the entire patient population! 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 Blue Goo, 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.

1.2.4 Analysis of a Sociological Survey

Blazer et al. (1985) reports 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.

1.2.5 An Epidemiology Example: Data Snooping Versus Data Mining

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.

“Data mining”—the exploitation of large databases to uncover new opportunities—has proven to be very profitable for business enterprises. However, the term “data mining” has historically had a very negative connotation for academic researchers, who consider “data mining” to be synonymous with “data snooping”—turning up nuggets of fool’s 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”? Protecting oneself against the problems of multiple inferences is a first step. 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 mitigate the problem. Procedures for fitting tree-based classification models often use multiplicity-adjusted rules to choose the splitting points. Finally, “hold-out samples” are commonly used to ensure that items mined from training samples are indeed replicable in external data.

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 levels 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.”

Unlike randomized clinical trials, with epidemiological studies, there is no control group; hence, conclusions from the data analyses can be misleading for a variety of reasons. Ioannidis (2005) finds a surprisingly high number of replication failures in epidemiological studies. While many of these might be related to the problems of not having appropriate control groups or appropriate control variables, multiplicity is also a likely culprit.

1.2.6 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 easily be statistically 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 because 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 Type II errors also are important (see Chapter 18–20 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 that is subject to global (familywise) Type I error control.

1.2.7 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 suggest controlling 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 among surgeons.

1.2.8 Genomics Data and Large-Multiple Testing

With gene expression studies, it is common for thousands of genes to be evaluated simultaneously for possible association with a disease or special condition. Typically, the goal is to identify which genes are of interest, then identify the action of those genes, and finally to develop a therapy that inhibits (or, in some cases, promotes) the actions of those genes in order to benefit the individual. The problem is that, with thousands of genes studied, the usual “1 in twenty” type I error rate is completely unacceptable, because thousands of genes might be incorrectly flagged. In genetics studies, millions of single nucleotide polymorphisms (SNPs) can now be screened for association with disease or any phenotype of interest. This adds another order of magnitude to the multiple testing problem, and makes determination of suitable biomarkers even more difficult. The cost in wasted follow-up testing can be significant. As a result, it has become standard to apply some sort of multiple comparisons correction for these applications.

The information revolution has led to similar applications in many fields. Nowadays, it is commonplace to sift through large databases in search for actionable anomalies, from unusual patterns in astronomical databases to unusual patterns suggesting fraud or perhaps even

terrorism. When thousands, or perhaps even millions, of potential items are screened for anomalous behavior, the likelihood of false signals is very high.

1.3 When Are Multiple Comparisons/Multiple Testing Methods 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. This book describes methods for overcoming the problem and calls 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 acronym MCP will refer generically to any simultaneous inference procedure, although sometimes the acronym MTP, or “multiple testing procedure,” is used.

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 situations 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% level of confidence.
- You are prepared to perform much data manipulation to find a statistically 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

After you have identified 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.

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, regardless of whether you wish to perform multiple comparisons. 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 typically conservative when there are correlations. Generalized Bonferroni methods and standard false discovery rate controlling methods are in this group. See Chapters 2, 13, and 19.

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. 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 Chapters 4 and 14.

Unbalanced One-Way ANOVA and 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 5, 6, 9, and 14.

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 Chapters 8, 14, and 15.

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 10.

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.

Normality (or at least approximate normality) remains an important assumption for these models. See Chapters 11 and 12.

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 Chapters 11 and 23.

Independent Observations from Parametric Nonnormal Distributions

As an example, suppose you know that the observations are counts of defects on a manufactured item, and you wish to compare shifts A, B, and C. The model used may be Poisson, and you still wish to perform multiple comparisons. In this case, you can use any of several SAS procedures to fit the Poisson model, and can perform adjustments for multiple comparisons easily using the fitted results from such models. See Chapters 12 and 15.

Dependent Observations from Parametric Nonnormal Distributions

Following the previous example, suppose you know that the counts of defects on manufactured items are associated with different machines. You still wish to compare shifts A, B, and C, but you want to account for the machine effect. In this case, you may model the observations on a common machine as dependent, using a random effects model, where the machine effect is considered random. Again the model may be Poisson, but with a repeated measures component. In this case, you can use PROC GLIMMIX both to perform the repeated measures modeling and to perform the multiple comparisons. See Chapters 12 and 15.

Nonnormally Distributed (Continuous) Data from General (Unspecified) Distributions

If the distributions are nonnormal and unspecified, you can still make inferences with multiplicity adjustment, using bootstrap and permutation methods. The assumed structure of the data is that the observation vectors are independent, and the covariance matrices are constant. See Chapter 16.

Binary Data

If your observations are binary (or more generally, if your distributions used for testing are discrete distributions), then there are dramatic gains in power that may be achieved using resampling-based multiple testing methods that are not achievable using parametric modeling. An example of such binary data was given previously in Section 1.2.3, where the observation vectors indicate presence or absence of a number of psychiatric conditions. See Chapter 17.

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 17), or which uses large-sample approximations from a proportional-hazards model or a parametric survival analysis model. See Chapters 15 and 17.

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. One 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. Another decision is which error rate you want to control, a decision that must be considered carefully. 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.

Confidence Interval-Based Methods

These methods are useful for providing an explicit range of values for each parameter of interest. Such intervals are also useful for determining directional relationships and statistical significance. Confidence intervals are discussed in chapters 3 through 12. Further interval-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 13; directional error rates are considered in Chapter 18.

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 are discussed in detail in Chapter 13. Further applications are given in Chapters 14 through 17. False discovery rate controlling methods discussed in Chapter 19 fall in this category as well.

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.

Each item in the list so far provides weaker inference than the ones above it, according to a classification first made by Hsu (1996). 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.

Graphical Methods

Perhaps you don't care about formal inference and just want a quick graph to suggest which leads to pursue next. There are several such methods, including p -value plots, diffograms, side-by-side box plots, histograms, and volcano plots. These are discussed throughout the book; just look for the pictures!

False Discovery Rate Controlling Methods

When you have massive multiplicity in your data, as in the case of genomics, proteomics, and astrophysics, where the number of hypotheses can easily be in the thousands or millions, you usually do not expect that every significant result is real and replicable. Rather, you just want to ensure that a controlled high proportion (e.g., 0.95 or more) of the significant results is real and replicable. In these cases, you may wish to control the false discovery rate (which is essentially the expected fraction of erroneous rejections among all rejected hypotheses), rather than the familywise error rate. Chapter 19 is devoted to this topic, with further applications in Chapter 22.

Bayesian Methods

Historically, there has been a large gulf between Bayesians and frequentists as regards multiple comparisons procedures. Nowadays, the gulf is not so wide, as there are Bayesian ways of viewing the problem that provide results that are not too different from frequentist methods. These Bayesian methods include simultaneous Bayesian credible intervals, posterior probabilities of meaningful differences, and posterior probabilities of null hypotheses. If you wish to take advantage of the benefits that Bayesian methods can offer for multiple comparisons procedures, see Chapter 20.

Decision-Theoretic Methods

Related to Bayesian methods are decision-theoretic methods. These methods aim for a simple and natural objective: choose the method that provides the best results. Here, “best results” means the results that minimize loss, or, equivalently, maximize benefit. You can supply a cost/benefit function to your decisions, and choose the decision that optimizes this function. See Chapter 20.

- ✓ Note: There is rarely, if ever, one and only one correct method. You should select a method only after careful consideration of the relative consequences of choosing that method versus the alternative methods.

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 that 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 initially in Chapter 4, with additional discussion in most 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 initially in Chapter 4, with additional discussion throughout the book.

Multiple Comparisons with the Best

If your interest only concerns comparing treatment means with the (unknown) “best” or “worst” (the highest mean or the lowest mean, depending on the application), see Chapter 23.

Comparisons with the Average Mean (“Analysis of Means,” or ANOM)

You may wish to identify “outlier” groups—that is, those that differ significantly from the overall average. The ANOM method is ideally suited for this analysis and is introduced in Chapter 5.

General Contrasts or Linear Functions

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 Chapters 3 and 7, and 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 and are introduced in Chapter 7.

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 11, 12, 16, 17, and 23.

Infinitely Many Comparisons

Although this category sounds like “data snooping,” it is actually permissible when done properly. See Chapters 9 and 23.

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, 13, and 19.

Confidence Bounds for Regression Functions

These applications are discussed in Chapter 9.

1.5 Controversial Aspects of MCPs

It would be wrong to suggest that all multiple testing inference issues are resolved by using an appropriate MCP, selected as suggested in the outline above. With MCPs, as with any statistical inference method, there is never a single technique 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 briefly discusses some of the controversies.

Size of the Family

For classical multiple comparisons methods, 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 historical line of research that suggests not to perform multiplicity-adjustments on 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 as 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 can cause scientists and experimenters to completely ignore the multiplicity problem. Appropriate use of multiple testing can be 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, you may note that multiplicity effects are real, and that Type I errors can and do occur, as has been demonstrated in the literature. 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.

Composite Inferences versus 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 make up the difference, then the composite inference is usually better (more powerful) than the individual, multiplicity-adjusted inferences. An example illustrates the difference.

EXAMPLE: Extrasensory Perception

While controversial, testing for extrasensory perception (ESP) has attracted interest in the scientific and government communities, particularly because 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.)

False Discovery Rate (FDR) versus Familywise Error Rate (FWE)

With the Information Revolution and the attendant tests of thousands, perhaps even millions of hypotheses, FDR-controlling procedures have gained stature as they are often more appropriate for such applications. However, the question naturally arises, if FDR controlling methods are

good for testing myriad hypotheses, why not use them for testing just a few? This brings up the question of which error rate to control. Like many controversies in statistics, there is no simple answer to the question. However, the first step is to understand clearly what these error rates mean, and to understand the consequence of each error rate for your practical application. Then you can make an informed choice.

Bayesian and Decision-Theoretic Methods

(This section is written for Bayesians; if you are not 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, the aim is to explain the commonly used tools for the analysis of multiple inferences, and since these methods are mostly frequentist, the 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 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).

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 (FWE-controlling) and Bayesian methods “need not be grossly disparate.”

If you are in the Bayesian camp, please follow the frequentist developments, keeping in mind that frequentist and Bayesian conclusions need not be grossly disparate when there is prior doubt about many of the hypotheses tested. Methods that have Bayesian rationale are presented in Chapters 19 and 20 of this book.

Decision-theoretic methods are related to Bayesian methods, and offer, in some ways, the best hope for resolution to the question “which method should I use?” From the decision-theoretic perspective, the answer is very simple: “choose the method that is best!” In Chapter 19, you will find practical ways to do this; however, as always, there is a catch: you must supply very subjective and specific loss functions that reflect the relative severity of Type I to Type II errors.

1.6 Chapter 1 Problems

Reviewing Hypothesis Testing and Confidence Interval Concepts

1. A researcher wants to know if taking zinc reduces the length of time that cold symptoms are present. She randomly assigns 50 people who recently contracted a cold to two groups, one of which will receive zinc in tablet form daily. The other will receive an identical tablet but without zinc (the placebo).
 - a) What is the null hypothesis?
 - b) What is the alternative hypothesis?
 - c) What is a Type I error in this study?
 - d) What is a Type II error in this study?
 - e) What is X , the measured data?
 - f) If the p -value satisfies $p \leq 0.05$, what do you conclude?
 - g) If the p -value satisfies $p \leq 0.05$, did a Type I error occur?
 - h) If the p -value satisfies $p > 0.05$, did a Type II error occur?
 - i) If the p -value satisfies $p > 0.05$, what can you say about the 95% confidence interval for $\mu_1 - \mu_2$?
2. Suppose that the null hypothesis is true, that the data are continuously distributed, and that all model assumptions are satisfied.
 - a) What is the probability that the p -value will be less than 0.025?
 - b) What is the probability that the p -value will be greater than 0.025?
 - c) What is the probability that the p -value will be between 0.025 and 0.975?
 - d) What happens in cases a)-c) if the model assumptions are *not* satisfied?

A Case Study in Multiple Comparisons

3. You wish to study how vitamins affect people's strength. You randomly divide 100 people into five groups of 20, asking each person to take a daily vitamin pill. One group (the control) takes a dummy pill with no vitamins (a placebo). The remaining four groups take, respectively, a low dose of vitamin brand A, a high dose of vitamin brand A, a low dose of vitamin brand B, and a high dose of vitamin brand B.
 - a) List the comparisons of interest. There should be several. State why you are interested in each of the stated comparisons.
 - b) Consider the bullet points at the beginning of Section 1.3 concerning when multiple comparisons procedures are needed. State how each bulleted point applies to your collection of comparisons noted in problem a.

- c) Look at Section 1.4. State your
 - i) statistical modeling assumptions that apply to this situation (examine the list in Section 1.4.1; which model(s) apply in this example?),
 - ii) testing objectives (consider the list in Section 1.4.2; which objectives apply in this case?), and
 - iii) the family of comparisons of interest (consider the list in Section 1.4.3; which objectives apply in this case).
- d) See Section 1.5. What are the controversial aspects of MCPs as they apply to this particular case setting?
- e) Attempt to identify costs (perhaps in a \$ sense, perhaps in a pain sense, or perhaps in some other sense) for the situation where there is one or more Type I errors in your family of tests in this case study.
- f) Attempt to identify costs (perhaps in a \$ sense, perhaps in a pain sense, or perhaps in some other sense) for the situation where there is one or more Type II errors in your family of tests in this case study.

Another Case Study in Multiple Comparisons

4. You are a geneticist, screening thousands of particular genotypes (that is, specific genetic sequences) for association with a particular disease. Each genotype gives rise to a test for genetic association, which is simply a comparison of percentages of genotypes with and without the disease. For example, if 90% of the diseased people have that particular genotype while only 15% of the non-diseased people have it, this is potentially strong evidence that the particular genotype is associated with the given disease.

You do not intend to make a firm determination of genotype/disease association from this initial screening study. Rather, you will only identify a collection of genotypes to study further using a new sample of diseased and non-diseased individuals.

Repeat problems 3a-3f above, but with reference to this case study. Make sure to highlight differences between these two cases, particularly with regard to the multiple comparisons issue.

1.7 Solutions to Chapter 1 Problems

1.
 - a) $H_0 : \mu_1 = \mu_2$
 - b) $H_A : \mu_1 \neq \mu_2$ (Comment: The alternative might be stated as one-sided because the product cannot be approved if it is significant but in the wrong direction, making people sicker! However, it is common practice for the evaluation of pharmaceutical interventions to perform two-sided tests at $\alpha=0.05$, with approval only if the effect is in the right direction. This practice allows for equivalence with the usual 95% two-sided interval; it also allows the data to provide the scientifically interesting conclusion that the intervention has an effect in the “wrong direction.”)
 - c) A rejection of the null hypothesis when the effect of zinc is no different from the effect of placebo.
 - d) A failure to reject the null hypothesis when zinc has a different effect than placebo.

- e) X = length of time cold symptoms are present (probably measured in days).
 - f) Reject H_0 .
 - g) Unknown. It is possible, but considered to be unlikely, because the probability of a Type I error is only 0.05.
 - h) Unknown. Again, it is possible. If the effect of zinc is truly different from the effect of placebo, but only slightly different, then Type II errors are very likely (approaching 95% as the difference between zinc and placebo approaches zero).
 - i) The interval includes 0, showing that $\mu_1 - \mu_2$ could plausibly be zero, or less than zero, or greater than zero. The effect of zinc relative to the effect of placebo cannot be confidently determined in this case.
2. All of a)-c) are answered using the fact that the p -value is uniformly distributed under the given conditions. The probabilities are easily computed without specialized functions, but the SAS cumulative distribution function (cdf) forms are given here for convenience. SAS cdf and quantile functions are used frequently in the book.
- a) $P(p < 0.025) = 0.025 = \text{cdf}('uniform', .025)$.
 - b) $P(p > 0.025) = 1 - 0.025 = 0.975 = 1 - \text{cdf}('uniform', .025)$.
 - c) $P(0.025 < p < 0.975) = 0.95 = 0.975 - 0.025 = \text{cdf}('uniform', .975) - \text{cdf}('uniform', .025)$.

When all the assumptions are satisfied, the p -value has the uniform distribution. If the assumptions are not satisfied, then the distribution of the p -value is not necessarily uniform, the probability calculations might be wrong, and the true probabilities can differ by unknown amounts from the values calculated in a)-c). Simulation studies show that the difference between true and calculated probabilities is small when the assumptions are only mildly violated.

3. a) A possible listing (not exhaustive—there are other comparisons you might be interested in as well).

Comparison	Why Interesting
Placebo vs. High dose of A	To determine if A has an effect (if there is an effect, it should show up at least in the high dose)
Placebo vs. Low Dose of A	To determine if Low Dose of A has an effect
Placebo vs. High dose of B	To determine if B has an effect
Placebo vs. Low Dose of B	To determine if Low Dose of B has an effect
High dose of A vs. High Dose of B	To determine if the A effect differs from the B effect

- b) The need for multiple comparisons methods:
 - i) First bullet: It is indeed plausible that all of the effects are null in this example: vitamins might have absolutely no effect on strength.
 - ii) Second bullet: You would like any claims to stand up to the usual scientific standards—e.g., no more than a 0.05 chance of a Type I error (claiming a vitamin/strength association when none exists in reality)—or equivalently, 95% confidence in the conclusions.

- iii) Third bullet: If you were to perform the five tests and “pick the winner” (that is, claim a significant result as long as the most extreme p -value is less than 0.05), then multiple comparisons methods are needed.
 - iv) Fourth bullet: This bullet seems not to apply here because the design and analysis plans are very explicit; this is not an exploratory analysis. On the other hand, if the comparisons were to be decided after looking at the data, then the analysis must be called exploratory.
 - v) Fifth bullet: This particular study will take time, energy, and cost. It might well be replicated, however, should a statistically significant result be found. If this is the case, then this particular bullet might not be such a great concern.
 - vi) Sixth bullet: There are indeed costs; these are discussed further in the answers to e) and f).
- c) Section 1.4 considerations:
- i) Statistical modeling assumptions: This example seems to fit directly into the classic, balanced, one-way ANOVA. However, depending on the subject selection and the strength measurement, one might choose to analyze the data as nonnormally distributed.
 - ii) Testing Objectives: Confidence intervals are always desirable, so you'd be wise to state that method as a default. However, in later chapters you will find that you can get more power by using confident directions and testing-based methods. For example, the “confident directions” objective might be appropriate in this example, since it is easy to specify the directions of the alternatives of interest a priori.
 - iii) Family of comparisons: These comparisons fall within the umbrella family called “General Contrasts.”
- d) Controversial aspects:
- i) Size of a family: There are five comparisons above in a). However, you could have chosen more or fewer, and the controversial aspect is that the conclusions reached will depend heavily upon the number of elements in the family. If you include more contrasts in the family, then you include more scientific questions of interest, but sacrifice ability to claim significances. Conversely, if you include fewer elements in the family, then you have a better chance to claim significance, but might be excluding tests of interest.
 - ii) Composite versus individual: In this example, it might be of primary interest to know whether the vitamins have any effect at all, rather than which specific doses and brands do what. Thus, rather than formulate individual comparisons involving the subcomponents, it might be most interesting to formulate the study as a single test of vitamins (overall) versus no vitamins.
 - iii) False Discovery Rate versus Familywise Error Rate: As discussed in Section 1.5, false discovery rate controlling methods are usually considered more appropriate than familywise error rate controlling methods when there are many tests (say, in the thousands). In this example, there are only five comparisons, so you may choose a familywise error rate controlling method. However, the choice is not automatic, and you should carefully consider the consequences of each method. See Chapter 18.

- iv) Bayesian methods: A Bayesian would say that the usual frequentist multiple confidence intervals and multiple testing methods are irrelevant. Instead, you should create a prior distribution for the vitamin effects, use the data to update the prior (making it a posterior distribution), and then draw all inferences from the posterior distribution. (Note that in this example, it would be prudent to place fairly large prior probabilities on or very near the null hypotheses of no effects, which can make Bayesian and frequentist FWE-controlling analyses roughly correspond.)
 - e) The costs of Type I errors include
 - i) consumers wasting their money to buy vitamins that do not improve strength,
 - ii) lost market share to the company (A or B) who came out “worse” than the opponent (by chance alone).
 - f) Costs of Type II errors include
 - i) not taking vitamins, when such would actually improve strength,
 - ii) taking too high a dose if the lower dose were effective (perhaps side effects from vitamin overdose),
 - iii) taking an inferior product (A or B) when one is actually superior.
- 4)
- a) In this case there are as many comparisons as genotypes: they are the comparisons of genotype frequency in the diseased group with the corresponding genotype frequency in the control group. Each comparison is interesting because it may help us predict which individuals are at risk for the given disease.
 - b) The need for multiple comparisons methods:
 - i) First bullet: It is indeed plausible that many or all of the effects are null in this example: there might be one genotype (out of the thousand or so) that is related to the disease, and all others unrelated.
 - ii) Second bullet: You would like to claim that a genotype/disease association exists only when one exists in reality. Perhaps you might require even more than the standard 95% confidence in this case; see also the answer to e regarding consequences of Type I errors.
 - iii) Third bullet: If you were to perform the thousand or so tests and “pick the winner”—that is, claim a significant result as long as the most extreme p -value is less than 0.05—then you are certain to “discover” at least one rather strong genotype/disease association, even when none exists in reality.
 - iv) Fourth bullet: This study is truly exploratory, as defined in the case description. However, you are not so concerned that all claimed effects are real, since there will be a follow-up study to screen false positives. Type I errors are not as great a concern in this case.
 - v) Fifth bullet: This particular study will take time, energy, and cost. However, it will be replicated, as stated in the case setting, and this issue is therefore not so much of a concern.
 - vi) Sixth bullet: There are indeed costs; these are discussed further in the answer to e) and f).
 - c) Section 1.4 considerations:
 - i) Statistical modeling assumptions: This case might fall in the category of “Binary and Discrete Data,” where genotypes are coded as 1/0.

- ii) Testing Objectives: In this example, the “Testing-based methods” might be preferred. A test of homogeneity is not appropriate, since that can tell you only whether there is some genotype/disease association, not which particular genotypes to evaluate further.
 - iii) Family of comparisons: These comparisons fall within the category of “Comparisons of Multivariate Measures across Two or more groups.” The multivariate measures are the binary genotype indicators, and the groups are people with and without the disease.
- d) Controversial aspects:
- i) Size of a family: This is a huge family. Since it is generally more difficult to detect significant results with large families, you might find no significant results if you use a typical multiple comparisons procedure. In this study, a replication is planned, and therefore you might avoid the usual multiple comparisons methods, hoping that any false positives are caught in the follow-up analysis. However, you also want to avoid wasting follow-up resources by following blind leads.
 - ii) Composite versus individual: In this example, it might be of primary interest to know whether there is any disease/genetic association at all. However, it is clear that if some overall significance were found, then one would want to follow it up with comparisons involving particular genotypes.
 - iii) False Discovery Rate versus Familywise Error Rate: As discussed in Section 1.5, false discovery rate controlling methods are usually considered more appropriate than familywise error rate controlling methods when there are many tests, say in the thousands. In this example, there are indeed thousands of comparisons, so you may choose a false discovery rate controlling method. However, the choice is not automatic, and you should carefully consider the consequences of each method. See Chapter 18.
 - iv) Bayesian methods: A Bayesian would say that the usual frequentist multiple confidence intervals and multiple testing methods are irrelevant. Instead, you should create a prior distribution for the genotype effects, use the data to update the prior (making it a posterior distribution), and then draw all inferences from the posterior distribution. (Note that in this example, it would be prudent to place fairly large prior probabilities on or very near the null hypotheses of no genotype effects, which can make Bayesian and frequentist FWE-controlling analyses roughly correspond.)
- e) The seriousness of Type I errors is a function of how much the follow-up testing for prospective associations costs, and how sensitive it is. If you determine that a particular disease is caused by a particular genotype, and then proceed to treat people who have that particular genotype, then at best you are wasting your (or their) money, and at worst you are causing them undue suffering from treatment side effects.
- f) Type II errors are more serious in this particular case study, since they won't be mitigated by follow-up testing. If you fail to identify an important genotype/disease association, then you may lose the ability to alleviate suffering for a portion of the population. If you are in the business of producing such cures, then you lose the potential revenue that such a cure might bring.

