CHAPTER 2

STATISTICAL ASPECTS OF CONTROLLED DIET STUDIES

JANICE A. DERR, PHD

Planning the Study
- Experimental Designs
  - Crossover and Parallel-arm Designs
  - Designs to Avoid
  - Control Groups
- Sample Size and Power
  - Selection of Design
  - Size of Experimental Effect
  - Statistical Power and Significance
  - Variation
  - Resources
  - Dropout Rate and Compliance

Implementing the Study: Statistical Issues
- Randomization
- Data Management

Analyzing the Data
- Exclusion of Data
- The Analytic Approach
- Multiple Tests of Significance

Conclusion

The proper planning of a research study and its data management and analysis involve many decisions. What type of experimental design should be used? How many participants should be studied and for how long? How should test diets be assigned to participants? How should the data be analyzed when a participant drops out or fails to comply with a protocol? A statistician can help investigators in the planning of the study by addressing these questions. This chapter provides information about the advantages and disadvantages of different experimental designs and discusses issues related to sample size and power, study implementation, and data analysis.

PLANNING THE STUDY

A good experimental study is organized around a set of research questions originating from the scientific objectives of the project. The purpose of the initial project-planning sessions should be to identify and prioritize the research questions that the study will be designed to answer. A statistician should be involved even at this early stage. By participating in the project-planning sessions, the statistician gains a better understanding of the scientific issues motivating the study.

Several issues that have an impact on study design and data analysis can be clarified at the project-planning stage. These include the major objectives of the study, the variables and comparisons of primary interest, secondary or exploratory variables and comparisons, policy on the exclusion of data, and the population to which generalizations about the study are to be extended. Clarifying these issues during the planning phase will provide a good basis for making statistical decisions throughout the remainder of the study.

Experimental Designs

Once the research questions are clarified, it is time to consider possible experimental designs that could be used to achieve the study objectives. The statistician should be able to present the relative advantages and disadvantages of several options and discuss them with the investigators. To do so requires a good understanding of the research setting and any practical limitations it may have. A statistician gains this understanding through regular contact with the scientist and exposure to the study setting—the laboratories, the kitchens, and the areas where participants and dietitians will be interacting.

Crossover and Parallel-arm Designs

An example of the way in which a statistician might discuss different options for the experimental design of a study is given by the comparison between a crossover design and a parallel-arm design. In a *crossover design,* each participant receives all test diets in a randomized order. In a *parallel-arm design,* each participant is assigned at random to only one test diet; different groups of participants receive different test diets. Table 2-1 illustrates these two designs. The advantages and disadvantages associated with each design are summarized in Table 2-2 and discussed here.

The main advantage of a crossover design is that the sample size required to detect a given experimental effect is smaller than with a parallel-arm design. The reduced sample size is feasible because each participant receives each test diet; the statistical comparison among test diets is made using the within-participant error. In a parallel-arm design, the comparison between test diets is made using the between-

eat right. American Dietetic Association

TABLE 2-1

Crossover and Parallel-arm Designs

Crossover design: Each subject receives all test diets in randomized order. This example shows three diets and three periods.

Subjects	Period 1	Period 2	Period 3
Subject 1	Diet A	Diet B	Diet C
Subject 2	Diet B	Diet C	Diet A
Subject 3	Diet C	Diet A	Diet B
Subject 4	Diet A	Diet C	Diet B
Subject 5	Diet C	Diet B	Diet A
Subject 6	Diet B	Diet A	Diet C
⋮	⋮	⋮	⋮
Subject n	Diet C	Diet A	Diet B
Subject Totals	**Period 1**	**Period 2**	**Period 3**
Diet A	n/3	n/3	n/3
Diet B	n/3	n/3	n/3
Diet C	n/3	n/3	n/3

Parallel-arm design: Each subject receives only one test diet. Only one time period is required.

Diet A	Diet B	Diet C
Subject 1	Subject 2	Subject 3
Subject 4	Subject 5	Subject 6
Subject m-2	Subject m-1	Subject m

TABLE 2-2

Crossover and Parallel-arm Designs: Advantages (A) and Disadvantages (D)

Study Feature	Crossover Design	Parallel-arm Design
Sample size	A Smaller	D Larger
Duration of study	D Longer	A Shorter
Use of facilities and resources	A More evenly distributed across time	D Effort concentrated in a shorter time period
Expectation of subjects	D Requires greater commitment	A Requires less commitment
Design considerations	A Balanced randomization	A Balanced randomization
	D Not suitable when carryover effects are expected	A No adverse consequences of carryover
	D Low % dropouts required	A Moderate % dropouts acceptable
	D Susceptible to confounding from Period X Test Diet interactions	A Free of confounding from Period X Test Diet interactions
Data analysis	D More complex	A Less complex

participant error, which is generally larger than the within-participant error. Table 2-3 shows an example of sample size calculations for a diet study having either a parallel-arm or a crossover design. This will be discussed in more detail later in Selection of Design.

Counterbalancing the crossover design's advantage of reduced sample size are several considerations that add to its complexity. Because each participant must be given all of the test diets, the participants in a crossover design must be enrolled for a much longer period of time than in a parallel-arm design. In addition, the crossover design relies on *balance* to partition the effects of time period (see Table 2-1) from the effects of test diet. When a participant drops out before the study is finished, that balance is threatened. Therefore, well-worked-out strategies for participant retention should be incorporated in the protocol of a crossover design.

If the crossover design is to retain its increased efficiency relative to the parallel-arm design, the response to a diet given in one test period should not affect the response to diets given in subsequent test periods (referred to as a *carryover effect*). Data from prior or pilot studies can identify the length of time required for the measures of interest to stabilize under test diet conditions. One strategy often

TABLE 2-3

Example of Sample Size Requirement for a Parallel-arm Design Compared to a Crossover Design

Minimum Detectable Difference Total Cholesterol (mg/dL)[1,2]	Two-group parallel-arm design Total number of subjects	Two-period crossover design Total number of subjects
6.5[3]	156[3]	51[4]
8.0	102	34
10.0	71	22
16.0	27	9

[1]For 80% power with one-tailed $\alpha = 0.05$:
the variance calculation for the parallel-arm designs uses: $\sigma^2_{among} + \sigma^2_{within} = 89.3 + 173.7 = 263.0$; the variance calculations for the crossover designs uses: $\sigma^2_{within} = 173.7$.
[2]These estimates were derived from data reported in Kris-Etherton et al (12).
[3]For number of subject per group, divide total in column by 2. Each group is assigned one test diet.
[4]Each subject experiences both test diets in random order.

employed to minimize the carryover effect is to include interim periods, known as *washout periods* (Figure 2-1), between test periods. The purpose of the washout period is to allow each participant's measurements to return to a baseline level before the participant begins the next test diet. Measurements taken at baseline can be used to assess carryover effects. Jones and Kenward (1) offer a more technical treatment of crossover designs and carryover effects.

Washout periods can be designed in one of two ways. This period can serve as a ''break'' from the experimental regimen during which time the participant's diet is not under experimental control. Alternatively, the participants may all be fed a standard diet such as one following the Dietary Guidelines for Americans (2), or the National Cholesterol Education Program (NCEP) guidelines (3). Another strategy for minimizing carryover effects does not make use of a washout period. Rather, the test diets follow each other in sequence without a break, and the test diet periods are long enough so that the endpoint measurements are not influenced by the previous diet. This strategy would be employed when there is no interest in the values of variables at the beginning of each test diet period.

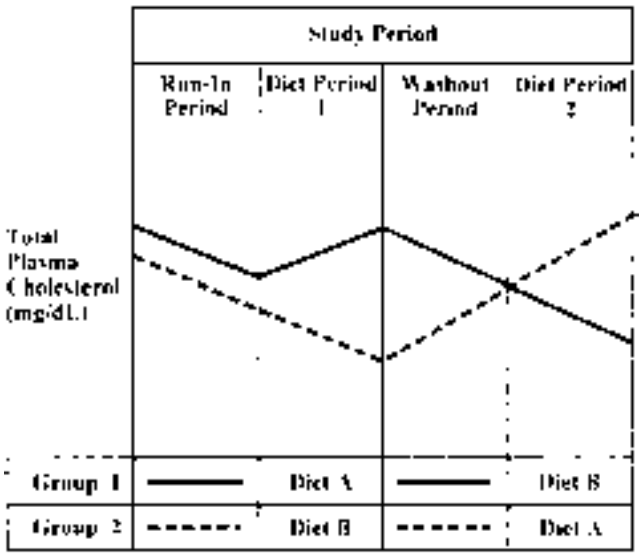

FIGURE 2-1. Schematic of a two-period crossover design. Two groups of subjects are treated with hypothetical diets A and B. Total plasma cholesterol is measured at the beginning and/or end of each study period.

Prior to the first test period, feeding studies can also incorporate a *run-in period* (Figure 2-1) during which the participants experience the protocol of the study. A run-in period helps investigators achieve several objectives important to the successful conduct of the study: (1) familiarizing participants with the feeding protocol used during the study; (2) allowing participants who discover they cannot tolerate the protocol to drop out prior to randomization; and (3) allowing participants to achieve a baseline value of the measurements of interest while on a common diet. The specific diets fed to the participants during the run-in period and the washout period, and the possibility that participants' diets are not under experimental control during these periods, are important issues that should be discussed among scientists and the project statistician.

In comparison with a crossover design, the parallel-arm design offers a relatively straightforward means of comparing the response to a set of test diets. However, there are two important issues that must be addressed with the parallel-arm design. First, the random assignment of participants to test diets must be done with care in order to ensure that the groups have a similar profile with respect to key variables prior to the test period. This is discussed in more detail later in Randomization. Second, because of the larger sample size requirement of the parallel-arm design relative to the crossover design, some parallel-arm designs will require more participants than can be processed at once at any one research facility. It thus may be necessary to conduct the parallel-arm design in *blocks,* or replicates of the design. The composition of the blocks must be carefully considered. Each block should contain a complete replicate of the design (that is, all of the test diets under consideration in all of the

different orders) and the randomization of participants to test diets should be balanced within each block. Blocks of the design can be conducted by different facilities, as is done in a multicenter study, or by the same facility over a period of time.

Designs to Avoid

Every statistician's nightmare is the *confounded* design. This is a design in which it is not possible to distinguish between a response to treatments, such as a set of test diets, and some other factor in the design. An example of a confounded design is a crossover design in which all participants receive Diet A during test period one, Diet B during test period two, and Diet C during test period three. In this design, the effect of test period is indistinguishable from the effect of diet. A second example of a confounded design is a parallel-arm design in which females receive Diet A and males receive Diet B. In this second example, it is not possible to distinguish between a diet effect and a gender effect. Confounded designs produce uninterpretable results, and there is no miracle of data analysis that can remedy the problem. It therefore is in everyone's best interest to discuss potential confounding factors with the statistician on the research team and to make sure that the critical factors are identified and accounted for in the design.

Control Groups

In feeding studies, the definition of a control is problematic. What is a *control diet?* Is it the participant's free-living diet? Is it the NCEP Step One diet? (3) The only way a statistician can help to answer this question is to enter the discussion about the objectives of a proposed feeding study. Do the scientists want to examine the response of a test diet as a *difference* from a baseline value? If so, then perhaps a standard reference diet can be used to establish a baseline. Are the scientists interested only in comparing a set of test diets, using as measurements only the values at the end of each test diet period? If so, then perhaps there is no need for a reference diet.

However, the strategy of not using a reference diet should be adopted with care. Seasonal effects, if there are any, become critically important for crossover designs as well as for parallel-arm designs conducted in blocks over periods of time. For example, effects of increasing sunlight exposure from winter to summer must be addressed in studies that examine calcium and vitamin D metabolism. Designs can become unbalanced because of dropouts or slightly uneven numbers in the original demographic categories. Including a reference diet in these designs can permit a seasonal adjustment in the comparison among test diets.

Including a control group in the design is important when participants are recruited from population extremes. For example, consider a study in which participants in the 95th to 99th percentile for total cholesterol are recruited. In this hypothetical study, there is a test diet that is thought to lower total cholesterol. Plasma total cholesterol will be measured at baseline and after participants consume the test diet for a specific time period. To provide a valid estimate of the cholesterol-lowering effect of the test diet, a proportion (usually half) of the eligible participants should be assigned at random to a control diet. This is because participants whose first blood sample puts them in an extreme percentile group will tend to have a second measurement closer to the population mean even with *no* intervention at all. This effect is known as the *regression to the mean.* Without a control group, the regression to the mean effect can be misinterpreted to be a treatment effect.

The true effect of the test diet is the *net* difference between change in total cholesterol for the group given the test diet and the change in total cholesterol for the control group. (See Figure 2-2 for an illustration of this net difference.) Davis (4) discusses the regression to the mean effect in lipid studies and gives suggestions for ways in which this effect can be reduced, such as using the mean of several baseline blood samples to classify participants prior to a dietary intervention.

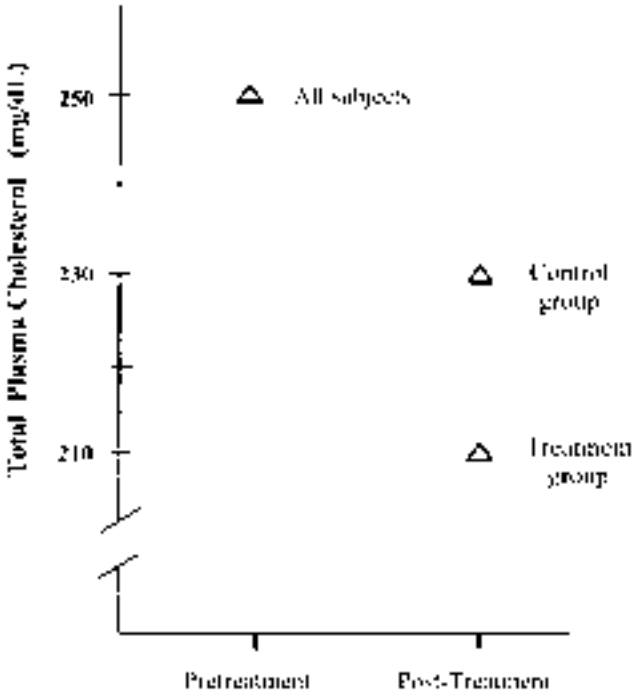

FIGURE 2-2. The regression-to-the-mean effect. In this hypothetical example, total plasma cholesterol of a treatment group and a control group is measured at baseline and at the end of a diet period. Participants are considered eligible for the study if their baseline total cholesterol is in an extreme percentile range relative to the population. The regression-to-the-mean effect causes the post-treatment value of the control group to come closer to the population mean than to the baseline value with no intervention at all. The effect of the treatment is the net difference between the treatment group response and the control group response. The final estimated effect is: (250–210) − (250–230) = 40 − 20 = 20 mg/dL.

Sample Size and Power

How many participants should be used in the research study? This is a crucial issue that will be closely examined by the institutional review board and the agency providing the funds for the research study. The institutional review board must consider two mandates that influence the choice of sample size: (1) the benefits of the research study must justify the potential risk to the participants, and (2) human participants should not be exposed to an excessive amount of risk. On one hand, a study with an inadequate number of participants exposes these participants to an unjustifiable amount of risk, because the study will fail to meet its objectives. On the other hand, a study with more participants than are needed to address the research questions exposes some participants to an unnecessary amount of risk. Therefore, considerations of human safety require the investigators to justify that they are including an adequate number of participants, but no more, to meet their research objectives.

From the perspective of the funding agency, the cost of processing each participant through a feeding trial also suggests that the number of participants should be the minimum that is adequate to address the research objectives. With these pressures in mind, the estimation of sample size should be carried out with great care and with the highest quality of information available.

The best number of participants for a feeding study is influenced by several factors: (1) the study design; (2) the size of the experimental effect that the investigators wish to detect; (3) the desired level of statistical power and significance; (4) the amount of variation within and among participants; and (5) the number of dropouts and the level of compliance for participants in the study. The statistician on the research team should compute and illustrate the statistical properties of the range of sample sizes and circumstances under consideration.

The statistician can also use previous data from similar studies in order to determine sample size requirements for different candidate designs. If no such data are available from either the research team or the published literature, it may be necessary to conduct a pilot study.

Selection of Design

The decision about sample size is inseparable from the decision about design. For example, a crossover design will generally require fewer participants but greater time commitment per participant than will a parallel-arm design. (See the earlier discussion in Experimental Designs.) Table 2-3 shows sample sizes that were computed during the planning stages of a feeding study. These estimates were calculated so that the investigators could make an informed decision between using a two-period crossover design and a parallel-arm design. Using estimates of within- and among-participant error from a previous study of a similar participant population, the calculations showed that a total of 102 participants would be required for a parallel-arm design (51 participants in each of two groups) in order to detect a minimum difference of 8.0 mg/dL total cholesterol between the two test diets at 80% power and 5% significance. The same statistical characteristics could be achieved in a crossover design with 34 participants, with each participant receiving both test diets in randomized order.

Size of Experimental Effect

Sample size requirements also depend on the size of the effect that the investigators want to detect with high probability. This effect size, often called the *minimum detectable difference,* is the smallest difference between means (eg, the mean response on each test diet) that the investigators would consider important. The criteria of ''importance'' must be determined from the clinical or research perspective. For example, consider the study discussed in Selection of Design, in which 8.0 mg/dL total cholesterol was determined to be the smallest difference between test diets that was important from a clinical perspective. A parallel-arm design with 156 participants would be able to detect a difference of 6.5 mg/dL total cholesterol with high probability. This difference, although statistically significant with 156 participants, might not be considered clinically important in the context of this hypothetical study.

The objective in computing sample size is to provide a match between an effect size that is meaningful to the investigators and the effect size that can be detected in the data analysis as statistically significant. It is not possible to estimate sample size without a criterion for effect size; the p-values generated from a study without any criteria for sample size are meaningless.

Statistical Power and Significance

Power refers to the probability of detecting a significant effect if one exists. Having high power, such as 80% or 90%, to detect a significant effect means that a correct conclusion is likely to be made about whether a dietary variable causes a change in an outcome variable such as blood pressure or blood lipids. The statistical properties of power and significance are components in the calculation of sample size. These properties determine the probability that the study results are truly representative of the entire reference population. Hypothesis testing is the basis of power and significance, but a detailed description is beyond the scope of this chapter. A very clear explication of these concepts can be found in Zar (5), Meinert (6) and Friedman, Furberg, and DeMets (7).

Variation

A reliable estimate of sample size depends on good estimates of variation for the response variables of interest. Data from previous studies or from a pilot study can be used to estimate variation. The participant population from previous studies should be as similar as possible to the participant population in the proposed study.

It also is critically important to choose which estimate of variance is the correct one for each sample size estimation.

For example, the variation *among* participants is used in a parallel-arm design, and the variation *within* participants is used for a crossover design. If the value of an outcome variable will be estimated from an average of several samples (for example, the mean total cholesterol from blood samples taken on three consecutive days), then the sample size formula should incorporate the variance of this mean.

Resources

Computational formulas for sample sizes vary according to the design and the type of variables to be measured. Reliable formulas and tables for sample size calculations can be found in Kraemer and Thiemann (8) and Cohen (9). Computer software can be purchased to automate the computations. One product is PASS® (10), which computes sample size and power for a broad range of experimental designs and types of response variables. Users of computer software are strongly cautioned to compare the computer output with hand-calculated and tabled values in order to make sure that the program is being used and interpreted correctly.

Dropout Rate and Compliance

When participants drop out of a study or do not completely follow the protocol for a test diet, the statistical power of the design is reduced. Investigators should estimate dropout rate and compliance from previous similar studies and inflate the sample size estimates accordingly. The question of whether and how to use data from dropouts or noncompliant participants will be discussed in Analyzing the Data.

Implementing the Study: Statistical Issues

Randomization

One of the statistician's tasks is ensuring that a valid randomization procedure is used to assign participants to test diets in a way that protects against selection bias. A randomization can also provide balance in the design so that the main effects of interest will not be confounded with other factors. For example, in a parallel-arm design, the randomization scheme should provide balance across gender, race, and age group for each test diet. In a crossover design, the test diet sequences should be balanced across time period and carryover effects from one diet to the next.

For example, Table 2-4 shows a set of test diet sequences that were used in a feeding study having a four-period crossover design. In this example, the statistician discussed the study with the investigator and learned that dropouts were most likely to occur within the first week of the first test period (these were participants who discovered they could not tolerate the protocol). Therefore, the statistician devised the randomization in two steps: First, participants were assigned to test diets for Period 1. The scientist then reported which participants had dropped out by the end of Period 1. Once the dropouts were eliminated from Period 1, the sequence of test diets for Periods 2, 3, and 4 were then computed for the remaining participants. This strategy achieved the best balance of diets in each period and of pairs of diets across the whole design.

A great variety of randomization schemes can be devised to meet the requirements of a feeding study. Meinert (6) and Friedman, Furberg, and DeMets (7) provide detailed descriptions of methods for assigning participants at random to groups.

Data Management

The development of a data management system should begin during the early project-planning stages. Data management encompasses the entire process from information gathering to data analysis: (1) the design and testing of all forms used to gather data; (2) the development of systems for labeling samples, identifying participants, and masking (ie, ''blinding'') certain processes; (3) data entry and error checking; (4) online storage plus offline archiving and retrieval of files; (5) the production of summary statistics at interim stages of the project; and (6) the development of data files to be used for analysis. A good data management system is essential to producing high-quality information that can be readily analyzed, and to ensure the security, safety, and confidentiality of the data.

The first responsibility of the data management team is to develop and test the forms that will be used in data collection. In feeding studies this generally involves forms that track variables stemming from many activities, such as participant characteristics and participant responses during recruitment and throughout the study, laboratory assays of biologic samples, and nutrient analysis of the diets. The data management team needs to work with the investigators and the project staff to develop a system that accommodates not just data management but also clinic, laboratory, and kitchen procedures. The timing of each measurement and the nature of each variable should be planned in advance. Adequate identification and labeling systems need to be developed to facilitate the collection of data. The data management team should develop a system for tracking laboratory samples and data as well as a coding system for information that must be kept masked. All forms must be tested (and revised) before the study begins. Meinert (6) provides a comprehensive reference to good practices in data management.

Another task of the data management team is to create a central database with a high level of security, protection, and quality control. For the data to be useful not just during the life of the project but also for a wide range of future research investigations, the issues of quality control, storage, security, access, and reporting are of paramount importance. Appropriate system security features need to be implemented to ensure that access to data, forms, and reports is restricted to authorized personnel and investigators. Frequent backups of data protect the database from possible data loss or corruption caused by electronic or power irregularities.

TABLE 2-4

Randomization Scheme for a Four-Period Crossover Design[1]

Subject ID	Period 1	Period 2	Period 3	Period 4
1	D	B	C	A
2	C	A	D	B
3	C	A	B	D
4	D	C	B	A
5	A	B	C	D
6	A	C	D	B
7	D	A	B	C
8	C	B	D	A
9	A	D	B	C
10	B	A	C	D
11	D	C	A	B
12	C	D	A	B
13	C	B	A	D
14	B	—[1]	—	—
15	A	—[1]	—	—
16	B	D	C	A
17	B	D	A	C
18	A	—[1]	—	—

Totals	Period 1	Period 2	Period 3	Period 4
A	3[2]	4	4	4
B	3[2]	4	4	4
C	5	3	4	3
D	4	4	3	4

Individual ordered diet pairs

— A	AB	AC	AD	— B	BA	BC	BD	— C	CA	CB	CD	— D	DA	DB	DC
4	5	3	3	4	3	4	4	5	5	3	4	4	4	4	3

[1]Diet treatment groups are indicated by letters A–D.
[2]For Period 1 there were originally 5 subjects in Diet Group A and 4 subjects in Diet Group B. Several subjects dropped out before Period 2.

Data quality control should include review and error checking at a number of stages of data entry and management. All forms should be reviewed for unusual events or missing information before data are entered. Programs for identifying out-of-range values can be executed once the data are entered. It is necessary to develop a system for querying missing information and out-of-range data that keeps the project staff in communication with the data management team. At regular intervals, the data management team should produce summary reports with descriptive information from the database. The data management team is also responsible for producing files in the appropriate format for data analysis.

Analyzing the Data

Exclusion of Data

In any clinical study, the investigators must decide which data from which participants should be included in the analysis. For example, there may be reason to believe that not everybody complied fully with the protocol. Some responses may appear atypical. Some participants may have dropped out before the study was finished. A well-planned feeding study will include a discussion of these issues and a statement of policy in advance of the trial.

It is likely that the statistician on the research team will bring to this discussion the *intent-to-treat paradigm.* This paradigm, well established in the clinical trials literature, directs the investigators to analyze data from all participants that were randomized into the study. Excluding participants according to compliance, errors in delivering the test diet, or other criteria can lead to an unknown amount of bias in the results. However, the intent-to-treat paradigm relates most directly to clinical trials in which noncompliance and inaccurate delivery of the treatment are considered valid aspects of the treatment regimen as it may be applied to the population at large. Friedman, Furberg, and DeMets (7) provide a discussion of the intent-to-treat paradigm from the clinical perspective.

There is room for discussion about the way in which unusual observations or departures from treatment will be handled in feeding studies. This discussion should be held in advance of any data analysis and should result in a carefully documented policy on criteria for excluding participants or data from analysis. This policy should be fully described in the publications resulting from the study. Regardless of the approach taken, the policy should be sufficiently well-considered and valid to withstand the scrutiny of peer review.

The Analytical Approach

Before any analysis, the statistician inspects the data through various graphical, descriptive, and diagnostic routines to ensure that the assumptions of the proposed analytical approach are satisfied. For linear models such as analysis of variance and linear regression, these assumptions include normality of errors and constant error variance. For crossover designs the structure of the covariance matrix should also be evaluated. Data *outliers,* or extreme values that do not appear to belong to the distribution of the majority of the data, will be evaluated. Some response measures may require a transformation, such as a logarithmic transformation for skewed data, before they are analyzed.

It is the responsibility of the statistician to select the most powerful statistical analysis compatible with the nature of the data. This is because the use of human participants requires that exposure to risk be balanced with maximal benefit from the acquired data. Several aspects of feeding studies suggest that the best statistical analysis is likely to be complex:

1. Many experimental designs have more than one test diet. A simple two-sample procedure such as a t-test comparison of one test diet to a control diet will generally be less powerful than an analysis of variance, which includes data from all of the test diets with follow-up comparisons of pairs of test diets.
2. A study design may involve a number of factors such as population subgroups, replicates, or centers that should be represented in the analysis.
3. Covariates such as baseline values may need to be incorporated in the analysis.
4. Crossover studies require the investigation of carryover effects and ''test diet by time period'' interactions that are not of direct interest but affect the analytical approach and results.
5. It may be necessary to provide a seasonal adjustment for designs that include blocks of time.
6. The occurrence of dropouts in the study generates incomplete data, which adds to the complexity of the analysis.

There are a wide variety of statistical approaches that can incorporate complex information, and more refinements in methodology are always appearing in the literature. It is a good idea to identify the probable analytical model at the project planning stages and to nominate alternatives to use if key assumptions are not met.

Multiple Tests of Significance

Most feeding studies involve a multiplicity of response measurements and comparisons of interest. For example, several plasma lipid measurements such as total cholesterol, LDL cholesterol, and HDL cholesterol may be included in the study. There may be several test diets in the study and an interest in comparing all possible pairs of test diets. It is important to determine in advance of the data analysis how this multiplicity will be managed because each statistical test carries with it an error rate given by the α level of the test. The error rates of each test within the same study are additive, which means that if the error rates for statistical tests are uncontrolled, there is a high chance that one or more false conclusions will result from the data analysis of the entire study.

A well-planned study provides not only an investigation of the major research questions of interest but also an exploratory analysis of other factors that could lead to the next research study. A good way to manage multiplicity in tests of significance is to make a well-defined distinction between these two phases. At the project planning stage, the study team should define the measurements and comparisons of interest comprising the primary aims of the study. Other measurements and comparisons should then be designated as ''secondary'' or exploratory. One approach to multiplicity is for each primary measurement (such as total cholesterol, LDL cholesterol, and HDL cholesterol) to be analyzed without mutual adjustment for multiplicity. Within each primary measurement, a multiple comparison procedure with good statistical properties should be used to adjust the significance level of comparisons of primary importance, such as comparisons in the level of total cholesterol between pairs of test diets. The best multiple comparison procedure will depend on the design and the structure of the comparisons. Neter, Wasserman, and Kutner (11) describe several frequently used procedures for making multiple comparisons.

The exploratory analysis of secondary measurements and comparisons can be treated in a number of ways. One option is to make a single Bonferroni adjustment to the p-values of all of the secondary tests. Another option is to report the unadjusted p-values of secondary tests in the literature without using the language of statistical inference. This means that secondary tests would be treated in a section on ''exploratory data analysis'' and discussed without using the terms *significantly different* or *not significantly different.*

There are many statistical approaches to multiplicity, and this topic is a matter of active debate in the statistical literature. As a guiding principle, the reader of the study's results should be able to determine how many statistical tests were conducted and what adjustments were made to account for multiplicity. Reporting only the unadjusted p-values of the few ''significant'' tests obtained from a vast search of

the database and hundreds or thousands of tests—known as *data-dredging*—not only misleads the reader but also is likely to lead to the embarrassment of researchers who produce unreplicable results.

When a careful approach to multiplicity is used, the manuscript can indicate future analysis plans suggested by exploratory approaches. The same cautious approach will allow the investigators to report their primary results with confidence.

Conclusion

The proper planning of a research study requires statistical insight into study design, implementation, and data analysis. Consultation with a statistician during the planning phase is essential for proper planning. The hypotheses must be specified a priori and are the keystone to the experiment and subsequent data analysis. The choice of experimental design is based on a number of factors, including the number of participants needed to show a prespecified effect size with a specified degree of confidence to detect that effect size and some assessment of likelihood of adherence and dropout. Inclusion of a control group is highly desirable and in many cases essential. Implementation of a well-designed study includes developing randomization procedures, study forms that have been pretested, methods for masking data collectors from knowing the treatment assignment of the participants they are measuring, plans for quality control of data collection and transmission, and methods for data management. The statistical analysis should be appropriate for the study design and should in advance address statistical issues such as adjusting significance levels for multiple comparisons, analytic approaches for dealing with dropouts, and conditions for excluding data.

Careful attention paid to the statistical aspects of design, implementation, and analysis in designing a human feeding study will result in a well-designed study with results that are readily interpretable.

References

1. Jones B, Kenward MG. *Design and Analysis of Cross-Over Trials.* London, UK: Chapman and Hall; 1989.
2. US Dept of Agriculture and US Dept of Health and Human Services. *Dietary Guidelines for Americans.* 4th ed. Washington, DC: US Government Printing Office; 1995. Home and Garden Bulletin No 232.
3. *The Second Report of the Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults.* Washington, DC: National Cholesterol Education Program; 1993. NIH publication 93–3095.
4. Davis CE. The effect of regression to the mean in epidemiologic and clinical studies. *Am J Epidemiol.* 1976; 104:493–498.
5. Zar JH. *Biostatistical Analysis.* 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, Inc; 1984.
6. Meinert CL. *Clinical Trials: Design, Conduct, and Analysis.* New York, NY: Oxford University Press; 1986.
7. Friedman LM, Furberg CD, DeMets DL. *Fundamentals of Clinical Trials.* 2nd ed. Littleton, Mass: PSG Publishing Co, Inc; 1985.
8. Kraemer HC, Thiemann S. *How Many Participants?* Newbury Park, Calif: SAGE Publications; 1987.
9. Cohen J. *Statistical Power Analysis for the Behavioral Sciences.* New York, NY: Academic Press; 1977.
10. *PASS: Power Analysis and Sample Size.* Kaysville, Utah: NCSS Statisticcal Software; 1991.
11. Neter J, Wasserman W, Kutner MH. *Applied Linear Statistical Models.* 2nd ed. Homewood, Ill: Richard D Irwin, Inc; 1985.
12. Kris-Etherton PM, Derr JA, Mitchell DC, et al. The role of fatty acid saturation on plasma lipids, lipoproteins, and apolipoproteins. I. Effects of whole food diets high in cocoa butter, olive oil, soybean oil, dairy butter, and milk chocolate on the plasma lipids of young men. *Metabolism.* 1993;42:121–129.