11

Confirmatory clinical trials: Analysis of continuous efficacy data

11.1 Introduction

As we have seen, several summary measures of central tendency can be used for continuous outcomes. The most common of these measures is the mean. In clinical trials we calculate sample statistics, and these serve to estimate the unknown population means. When developing a new drug, the estimated treatment effect is measured by the difference in sample means for the test treatment and the placebo. If we can infer (conclude) that the corresponding population means differ by an amount that is considered clinically important (that is, in the positive direction and of a certain magnitude) the test treatment will be considered efficacious.

In Chapter 10 we saw that there are various methods for the analysis of categorical (and mostly binary) efficacy data. The same is true here. There are different methods that are appropriate for continuous data in certain circumstances, and not every method that we discuss is appropriate for every situation. A careful assessment of the data type, the shape of the distribution (which can be examined through a relative frequency histogram or a stem-and-leaf plot), and the sample size can help justify the most appropriate analysis approach. For example, if the shape of the distribution of the random variable is symmetric or the sample size is large (> 30) the sample mean would be considered a "reasonable" estimate of the population mean. Parametric analysis approaches such as the two-sample *t* test or an analysis of variance (ANOVA) would then be appropriate. However, when the distribution is severely asymmetric, or skewed, the sample mean is a poor estimate of the population mean. In such cases a nonparametric approach would be more appropriate.

It should be emphasized at this point that the term "nonparametric" is not a quality judgment compared with the term "parametric." The nomenclature simply serves to delineate two types of analyses. Nonparametric tests are not "less good" than parametric tests. Indeed, if it were appropriate to use a nonparametric approach in a certain circumstance, that test would have higher statistical power than a parametric approach. We respectfully feel that the differentiation between parametric and nonparametric approaches in many introductory Statistics textbooks is misleading, and does tend to imply that nonparametric tests are naturally inferior to the other: Nonparametric tests are commonly discussed separately, often toward the end of the book, leaving the reader feeling that the books' authors regarded these discussions as unwanted but obligatory. We encourage you as your first step to consider what valid and appropriate analyses there are for a given situation, and then to select the most efficient analysis method from among them. We have reinforced this notion by including nonparametric analysis approaches side by side with parametric approaches.

11.2 Hypothesis test of two means: Two-sample *t* test or independent groups *t* test

A common measure of central tendency of continuous outcomes is the mean. In clinical studies employing measurement of continuous variables such as blood pressure, the typical response among participants in a treatment group is represented by this summary descriptive statistic. As we have seen, sample statistics, by definition, vary from sample to sample. When developing new drugs we would like to make an inference about the magnitude of the difference between two population means, typically represented by the symbol μ, one for a test treatment and the other for a control. If the difference in means exceeds the typical variability that would be expected from sample to sample, we can conclude that the difference is unlikely to be due to chance. More specifically, when comparing two population means, we are interested in testing the null hypothesis,

$$H_0: \mu_1 - \mu_2 = 0.$$

If the null hypothesis is rejected the following alternate hypothesis is better supported by the data:

$$H_A: \mu_1 - \mu_2 \neq 0.$$

Treatment group 1 is represented by n_1 observations, $x_{11}, x_{12}, x_{13}, \ldots, x_{1n_1}$. Similarly, treatment group 2 has n_2 observations, $x_{21}, x_{22}, x_{23}, \ldots, x_{2n_2}$. For this statistical test these two groups must be independent. The population means, μ_1 and μ_2, are estimated by the sample means from each group, $\bar{x}_1$ and $\bar{x}_2$:

$$\bar{x}_1 = \frac{\sum_{i=1}^{n_1} x_{1i}}{n_1},$$

$$\bar{x}_2 = \frac{\sum_{i=1}^{n_2} x_{2i}}{n_2}.$$

The population variances, σ_1^2 and σ_2^2, are estimated by the sample variances:

$$s_1^2 = \frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2}{n_1 - 1},$$

$$s_2^2 = \frac{\sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2}{n_2 - 1}.$$

Assuming that the two populations have the same, albeit unknown, population variance, an average or pooled estimate of the sample variances is an estimator of the unknown population variance. The pooled variance, s_p^2, is obtained as:

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}.$$

Finally, the pooled standard deviation, s_p, is the square root of the variance:

$$s_p = \sqrt{s_p^2}.$$

The estimator for the difference in population means is the difference in sample means, that is, $\bar{x}_1 - \bar{x}_2$. The standard error of the estimator, $\mathrm{SE}(\bar{x}_1 - \bar{x}_2)$, is calculated as:

$$\mathrm{SE}(\bar{x}_1 - \bar{x}_2) = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

The test statistic for the two-sample t test is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\mathrm{SE}(\bar{x}_1 - \bar{x}_2)}.$$

Under the null hypothesis of equal population means, the test statistic follows a t distribution with $n_1 + n_2 - 2$ degrees of freedom (df), assuming that the sample size in each group is large (that is, > 30) or the underlying distribution is at least mound shaped and somewhat symmetric. As the sample size in each group approaches 200, the shape of the t distribution becomes more like a standard normal distribution. Values of the test statistic that are far away from zero would contradict the null hypothesis and lead to its rejection. In particular, for a two-sided test of size α, the critical region (that is, those values of the test statistic that would lead to rejection of the null hypothesis) is defined by $t < t_{\alpha/2,n1+n2-2}$ or $t > t_{1-\alpha/2,n1+n2-2}$. Note that as t distributions are symmetric, $|t_{\alpha/2}| = t_{1-\alpha/2}$. If the calculated value of the test statistic is in the critical region, the null hypothesis is rejected in favor of the alternate hypothesis. If the calculated value of the test statistic is outside the critical region, the null hypothesis is not rejected.

As there are an infinite number of t distributions there is no concise way to display all possible values that may be encountered. However, as can be seen in Table 11.1, the value of the t distribution that cuts off the upper 2.5%

area of the distribution becomes smaller with increasing sample sizes (and therefore increasing df). Tabled values in Appendix 2 are provided for other values of α.

Table 11.1 Sample values from t distributions for a two-sided test of $\alpha = 0.05$

Degrees of freedom ($n_1 + n_2 - 2$)	$t_{0.975}$
10	2.2281
30	2.0423
50	2.0086
100	1.9840
200	1.9719

The use of the two-sample t test is illustrated here with sample data from a clinical trial of an investigational antihypertensive drug.

The research question

Does the test treatment lower SBP more than placebo?

Study design

In a randomized, double-blind, 12-week study, the test treatment, one tablet taken once a day, was compared with placebo (taken in the same manner). The primary endpoint of the study was the mean change from baseline SBP. The primary analysis will be based on a two-sample t test with $\alpha = 0.05$ (two-sided).

Data

In the placebo group (146 individuals) the mean change from baseline was −3.4 mmHg with a standard deviation of 17.4 mmHg. In the test treatment group (154 individuals) the mean change from baseline was −19.2 mmHg with a standard deviation of 16.9 mmHg.

Statistical analysis

The null and alternate statistical hypotheses can be stated as:

$$H_0: \mu_{\text{TEST}} - \mu_{\text{PLACEBO}} = 0.$$

$$H_A: \mu_{\text{TEST}} - \mu_{\text{PLACEBO}} \neq 0.$$

The pooled sample variance is calculated as:

$$s_p^2 = \frac{17.4^2(145) + 16.9^2(153)}{146 + 154 - 2} = 293.95.$$

It follows from this that the pooled standard deviation is:

$$s_p = \sqrt{293.95} = 17.1.$$

The estimate of the difference in mean change from baseline is:

$$\bar{x}_{\text{TEST}} - \bar{x}_{\text{PLACEBO}} = -19.2 - (-3.4) = -15.8.$$

The standard error of the difference is calculated as:

$$\text{SE}(\bar{x}_{\text{TEST}} - \bar{x}_{\text{PLACEBO}}) = 17.1\sqrt{\frac{1}{146} + \frac{1}{154}} = 1.98.$$

The test statistic is then calculated using these values:

$$t = \frac{-15.8}{1.98} = -7.98$$

Under the null hypothesis of no difference in population means, and assuming somewhat symmetric distributions, the test statistic follows a t distribution with 298 (that is, 146 + 154 − 2) df. Therefore the critical region (values of the test statistic that lead to rejection) is defined as $t < -1.968$ and $t > 1.968$. Note that this particular entry is not in Appendix 2, but the closest is for 300 df.

Interpretation and decision-making

As $-7.98 < -1.968$, the null hypothesis is rejected in favor of the alternate one. The mean change from baseline for the test treatment group is significantly different from the placebo group's at the $\alpha = 0.05$ level. To determine the p value associated with this test, we need statistical software or an extensive look-up table. Given the large sample size in this example, we can use the percentiles of the standard normal distribution to approximate the p value. These study results allow us to conclude that the test treatment is efficacious. The difference between treatments in the magnitude of the change in SBP was unlikely to be the result of chance. Therefore the sponsor can submit these data as substantial statistical evidence of the test treatment's efficacy.

11.3 Hypothesis test of the location of two distributions: Wilcoxon rank sum test

The two-sample *t* test is useful on many occasions, but there are occasions when its use is not justified. One reason is that the sample size is small (< 30 per group). Although small studies are certainly encountered frequently in clinical research, most confirmatory efficacy studies are sizable, and so this reason is not applicable here. A second reason is, however, applicable. The most common reason why a two-sample *t* test would not be appropriate is a heavily skewed distribution, whether or not the sample size is large.

The sample mean is a poor measure of central tendency when the distribution is heavily skewed. Despite our best efforts at designing well-controlled clinical trials, the data that are generated do not always compare with the (deliberately chosen) tidy examples featured in this book. When we wish to make an inference about the difference in typical values among two or more independent populations, but the distributions of the random variables or outcomes are not reasonably symmetric, nonparametric methods are more appropriate. Unlike parametric methods such as the two-sample *t* test, nonparametric methods do not require any assumption about the shape of a distribution for them to be used in a valid manner. As the next analysis method illustrates, nonparametric methods do not rely directly on the value of the random variable. Rather, they make use of the rank order of the value of the random variable.

It is appropriate to note here that performing an analysis on an assigned rank instead of on the raw data results in a loss of information. Think of the related example of receiving a grade A on an assignment. If a grade A is given for any mark between 90% and 100%, the grade alone does not tell you how well you have actually done on the assignment: A score of 91% is assigned the same grade as a score of 100%. If the mark for this assignment is the first one of several in a course that will ultimately be combined to yield your final grade in some manner, you may very legitimately be interested in your actual (raw) score. Nevertheless, in clinical trials there can be a sound rationale for not using raw data in certain circumstances.

When rather extreme departures from required assumptions are noted, our choice of an appropriate statistical method should be one of first **validity** and second **efficiency**. The difference between an extreme departure from required assumptions and any departure from required assumptions is again a matter of judgment. It should be noted that many of the parametric methods in this book are robust to departures from distributional assumptions, meaning that the results are valid under a number of conditions. This is especially true with the larger sample sizes encountered in therapeutic exploratory and confirmatory trials. We should also note that all the methods described in this book require that observations in the analysis are independent. There are statistical methods to be used for dependent data, but they are not described in this book.

In our opinion, therefore, nonparametric methods should be chosen when assumptions (such as normality for the *t* test) are clearly not met and the sample sizes are so small that there is very little confidence about the properties of the underlying distribution. The nonparametric method discussed in this section is a test of a shift in the distribution between two populations with a common variance represented by two samples, and it will always be valid when comparing two independent groups.

The two-sample *t* test was based on the assumption that the two samples were drawn from an underlying normal population with the same (assumed) population variance. A rejection of the null hypothesis in the setting of the two-sample *t* test would imply that the two populations from which the samples were drawn were represented by two normal distributions with the same variance (shape), but with different means. The Wilcoxon rank sum test does not require the assumption of the normal distribution, but does require that the samples be drawn from the same population. The Wilcoxon rank sum test tests a similar hypothesis such that, if it is rejected, the two populations from which the samples were drawn had the same shape (not necessarily normal or otherwise symmetric), but differed by some distance. That is, a rejection of the null hypothesis in the setting of Wilcoxon's rank sum test would imply that the two population distributions were shifted, that is, not overlapping.

Although this approach has its advantages, one disadvantage is that no single numerical estimate, either a point estimate or an interval estimate, can convey the extent to which the populations differ because the test of the location shift is based on relative rank and not the original scale.

Using the Wilcoxon rank sum test, interest is in a location shift between two population distributions so the following null hypothesis is tested:

> H_0: The location of the distribution of the random variable in population 1 does not differ from the location of the random variable for population 2.

If the null hypothesis is rejected the following alternate hypothesis is better supported by the data:

> H_A: The location of the distribution of the random variable in population 1 is different from the location for population 2.

Treatment group 1 (representing population 1) is represented by n_1 observations measured on a continuous scale, $x_{11}, x_{12}, x_{13}, \ldots, x_{1n_1}$. Similarly, treatment group 2 (representing population 2) has n_2 observations measured in a continuous scale, $x_{21}, x_{22}, x_{23}, \ldots, x_{2n_2}$. The total sample size of the two groups is $n_1 + n_2$. The first step in calculating the test statistic is to order the values of all observations from smallest to largest, without regard to the treatment group. Then, a rank is assigned to each observation, starting with 1 for the smallest value after sorting, then 2, and so on for all $n_1 + n_2$ observations. If two or more observations are tied, the assigned rank will be the average of the ranks that would have been assigned if there were no ties. For example, if the third, fourth, and fifth sorted observations were all tied, the assigned rank for each of the three observations would be $[3 + 4 + 5]/3 = 4$. The next largest value would then be assigned a rank of 6.

At this stage, we now have n_1 ranks for treatment group 1, $r_{11}, r_{12}, r_{13}, \ldots, r_{1n_1}$. Similarly, treatment group 2 has n_2 ranks, $r_{21}, r_{22}, r_{23}, \ldots, r_{2n_2}$. The test statistic for the Wilcoxon rank sum test is the sum of the ranks in group 1:

$$S_1 = \sum_{i=1}^{n_1} r_{1i}\,.$$

Only the ranks from group 1 are required because, if the values from group 1 tend to be smaller than those from group 2, the sum of ranks will be small, leading to rejection of the null hypothesis. Similarly, if the values from group 1 tend to be larger than those from group 2 the sum or ranks will be a large number and will also lead to rejection.

The null hypothesis will be rejected if the test statistic is less than or equal to or greater than or equal to cut points obtained from a table (which need not be provided here) – that is, the null hypothesis will be rejected if $S_1 \leq W_L$ or $S_1 \geq W_U$. Other authors (Schork and Remington, 2000) have suggested a large sample approximation, which is possible because the test statistic, S_1, is approximately normally distributed with mean $[n_1(n_1 + n_2 + 1)]/2$ and variance $[n_1n_2(n_1 + n_2 + 1)]/12$. The derivation of these two parameters is beyond the scope of this text. Applying a familiar mathematical operation (standardization of a normally distributed random variable), we obtain an alternate test statistic, which has an approximate standard normal distribution:

$$Z = \frac{S_1 - \dfrac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\dfrac{n_1 n_2(n_1 + n_2 + 1)}{12}}}.$$

Values of this test statistic can then be compared with the more familiar critical values of the standard normal distribution.

To illustrate this method, consider the following example that (deliberately) has a small dataset.

The research question

Does the test treatment lower SBP more than placebo?

Study design

In a randomized, double-blind, 6-week study, the test treatment (one tablet taken once a day) was compared with placebo. The primary endpoint of the study was the mean change from baseline SBP. Given the small sample size of the study, the primary analysis is based on the Wilcoxon rank sum test with $\alpha = 0.05$ (two-sided).

Data

Each value listed below represents change from baseline SBP for a participant in a clinical trial comparing a new antihypertensive treatment with placebo. Lower values indicate a greater reduction in blood pressure from baseline, the favored outcome.

Test treatment ($n = 10$):
−8, −1, 0, 2, −20, −18, −12, −17, −14, −11.
Placebo ($n = 10$):
−9, 0, −4, −4, −3, 1, −7, 1, 2, −3.

Statistical analysis

After ordering all observations from highest to lowest within the two groups, we have the following:

Test	−20	−18	−17	−14	−12	−11	−8	−1	0	2
Placebo	−9	−7	−4	−4	−3	−3	0	1	1	2

Then ranking each observation across groups, accounting for ties as described above, we obtain the following ranks:

Test	1	2	3	4	5	6	8	14	15.5	19.5
Placebo	7	9	10.5	10.5	12.5	12.5	15.5	17.5	17.5	19.5

The test statistic is computed as the sum of the ranks for the test treatment group:

$$S_1 = 1+2+3+4+5+6+8+14+15.5+19.5 = 78.$$

When testing the null hypothesis at the two-sided $\alpha = 0.05$ level and a sample size of 10 in each group, the critical region is any value of $S_1 \leq 78$ or ≥ 132.

Interpretation and decision-making

As the value of the test statistic is in the rejection region (only just, but still in it), the null hypothesis is rejected. The conclusion is that the distributions of the two populations from which the samples were selected differ in their location. The test treatment is associated with a greater reduction in SBP than placebo.

Alternately, if we were to use the test statistic based on a normal approximation, it would be:

$$Z = \frac{78 - \dfrac{10(10+10+1)}{2}}{\sqrt{\dfrac{10 * 10(10+10+1)}{12}}} = -2.007.$$

Under the null hypothesis, this test statistic follows a standard normal distribution. The null hypothesis is rejected because the test statistic falls in the rejection region for a two-sided test of $\alpha = 0.05$ based on the standard normal distribution ($Z < -1.96$ or $Z > 1.96$).

11.4 Hypothesis tests of more than two means: Analysis of variance

The t tests are extremely helpful, commonly used tests, but they do have one noteworthy limitation: They can address only the equality of two means. In the present context, they can compare only the results from two treatment groups. Situations that require us to test the equality of more than two means occur quite frequently, and so a test that can be used with two or more groups is needed.

In many instances in drug development, two or more doses may seem to be promising based on results from earlier phases of clinical development. The question of interest therefore becomes: Of all the doses studied, which has the greatest beneficial effect? Confirmatory efficacy studies aim to answer this question. As in other study designs that we have discussed, the sponsor would like to minimize the chance of committing a type I or II error. We therefore need an appropriate statistical method that can identify the best dose (among a number of them), while accounting for the inherent variability in the data and limiting the chances of committing an error in the final decision-making process. Analysis of variance (ANOVA) is well suited to this task.

Assume that there are k independent groups ($k > 2$), each of which represents populations of interest, for example, individuals given a particular treatment. An important objective of many clinical trials is to determine if there is any difference among the treatments administered with regard to the underlying population means. The null hypothesis for such an objective is:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k.$$

If there is sufficient evidence to conclude that the null hypothesis should be rejected, the alternate hypothesis that would be favored is that there was at least one difference among all $[k(k - 1)]/2$ pairs of population means:

$$H_A: \mu_1 \neq \mu_2 \text{ or } \ldots \mu_1 \neq \mu_k \text{ or } \ldots \mu_{k-1} \neq \mu_k.$$

Each treatment group j ($j = 1, 2, \ldots, k$) is represented by n_j observations, $x_{1j}, x_{2j}, x_{3j}, \ldots, x_{n_j j}$. The sample sizes for each of the groups need not be equal. For each group the population mean, μ_j, is estimated by the sample mean, $\bar{x}_j$:

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}.$$

We can calculate the mean of all values across the k groups, the grand mean, as:

$$\bar{x}_{.} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{n},$$

where

$$n = \sum_{j=1}^{k} n_j,$$

the overall sample size. The total variability across all $n = n_1 + n_2 + \ldots n_k$ observations is the sum of the squared difference between each observation and the grand mean divided by the number of df:

$$V_T = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{.})^2}{n - 1}.$$

The sum of the squared deviations of each observation from the overall mean (the numerator) is also called the "total sums of squares."

The population variance for each group, σ_j^2, is estimated by the sample variance:

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n_j - 1}.$$

While the notation here is a little more complicated than we have seen before (because of the addition of the subscript j) the basic principle is exactly the same. All we have done to this point in this example is to calculate the sample means and variances for each group in the study.

An estimate of the average variance over all k groups represents the "typical" spread of data over the entire study or experiment. This variability is often referred to as random variation or noise. In the ANOVA strategy this number is called the within-group variance (or mean square error), and is calculated as a weighted average of the sample variances:

$$\text{Within-group variance } (V_w) = \frac{\sum_{j=1}^{k} (n_j - 1)s_j^2}{n - k}.$$

The denominator – that is, the df – in this calculation may be puzzling at first, but, again, the principle is the same as we have seen before. Recall that, when estimating the sample variance, the df value is $n - 1$. This is because the sum of deviations has to equal 0. Given knowledge of $n - 1$ observations in the sample, we can determine the last observation: It is the value that will ensure that the sum of all deviations adds to 0. In this case, the "minus 1" is applied for all k groups. This leads to:

$$(n_1 - 1) + (n_2 - 1) + \ldots (n_k - 1) =$$
$$(n_1 + n_2 + \ldots n_k) - k = n - k.$$

As there are also k sample means, each representing an estimate of the typical value of the population (that is, the population mean), those estimates may also vary from sample to sample. The variance of the means across all groups is called the among-group variance (or the mean square among groups), and is calculated as a weighted average (weighted by the sample size) of the squared differences of each sample mean from the grand mean:

$$\text{Among-group variance } (V_{\mathrm{A}}) = \frac{\sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x}.)^2}{k-1}.$$

where

$$\bar{x}. = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{n},$$

the grand mean, as before. The total variability in the data can be split or partitioned as the within-group variability (the background variability) and the among-group variability of means (how much the sample means vary from the overall mean):

$$V_{\mathrm{T}} = V_{\mathrm{A}} + V_{\mathrm{W}}.$$

As we have seen with a number of methods so far (most notably, the two-sample t test) the extent to which point estimates differ is measured against the typical variability of means from sample to sample. In the case of an ANOVA, we have an analogous method by which we can evaluate the extent to which the means differ. If the variance among the samples greatly exceeds the typical variance of the data in general there is an indication that the typical difference in means is not the result of random variation, but of systematic variation. If the variance among the samples is similar to the variance of the data in general such a result suggests that, whatever the difference in means, it is just like what happens by chance alone.

The test statistic (F) for this comparison in the ANOVA takes the form of a ratio of the among-sample variance to the within-sample variance:

$$F = \frac{V_{\mathrm{A}}}{V_{\mathrm{W}}}.$$

This test statistic is not well defined in all cases, which means that a rejection region is not automatically defined from a known distribution. However, if some assumptions are made about the distribution of the random variable X, the distribution of the test statistic can be defined. The following assumptions are required for an appropriate use of ANOVA:

- Each group represents a simple random sample from each of k populations and the observations are statistically independent.
- The random variable, X, is normally distributed within each population.
- The variance of the random variable, X, is equal among all k populations.

Given these assumptions the test statistic, F, follows an F distribution with $(k - 1)$ numerator df and $(n - k)$ denominator df. This is written in shorthand as $F_{k-1,n-k}$. Although we do not describe this distribution in detail, its essential characteristics are that it is a two-parameter distribution (that is, the numerator and denominator df) and it is asymmetric. As you might imagine, this distribution is not nearly as convenient to work with as the standard normal distribution. Defining the critical region for a given situation is best accomplished using statistical software because there are countless F distributions, each requiring a table. Similarly, calculating the sums of squares is best left to software (it can certainly be done by hand, but the required calculations are tedious).

ANOVA can be extended to situations where the experimental units (in our context, study participants) are classified on a number of factors. When they are classified on the basis of one factor, it is referred to as a one-way ANOVA. The result of partitioning the total variance into its components, in this case among and within samples defined by one factor, is displayed in Table 11.2.

The F distribution with $(k - 1)$ numerator df and $(n - k)$ denominator df is used to define the rejection region for a test of size α. The critical region may be obtained from a table of values or provided by statistical software. Tabled F values

Table 11.2 General one-way ANOVA table

Source	Sum of squares	Degrees of freedom	Mean square	*F*
Among samples	$\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x}.)^2$ $= \text{SSA}$	$k - 1$	$\dfrac{\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x}.)^2}{k - 1}$ $= V_\text{A}$	$\dfrac{V_\text{A}}{V_\text{W}}$
Within samples	$\sum_{j=1}^{k} (n_j - 1)s_j^2$ $= \text{SSW}$	$n - k$	$\dfrac{\sum_{j=1}^{k} (n_j - 1)s_j^2}{n - k}$ $= V_\text{W}$	
Total	$\sum_{j=1}^{k}\sum_{i=1}^{n_j} (x_{ij} - \bar{x}.)^2$ $= \text{SST}$	$n - 1$	$\dfrac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} (x_{ij} - \bar{x}.)^2}{n - 1}$ $= V_\text{T}$	

for a number of combinations of α, numerator and denominator df are provided in Appendix 4. The null hypothesis of no difference among means will be rejected only if the value of the test statistic, *F*, is larger than the cut point specified from the parameters of the distribution. Therefore, the test is inherently one sided – that is, the rejection region is any value $F \geq F_{(k-1),(n-k)}$.

Rejection of the null hypothesis means only that there is at least one difference among all pairwise comparisons of means. This conclusion is hardly satisfactory in the world of drug development because the decisions to be made typically require the selection of a dose or treatment regimen for purposes of designing another study or proposing a dose for marketing approval.

11.5 A worked example with a small dataset

Since, as noted, the calculations involved in ANOVA are fairly tedious, we illustrate the method using an overly simplistic example with a small dataset. This example is for illustrative purposes: In reality, datasets for which ANOVA is most appropriate have large sample sizes and are analyzed using statistical software. However, once you have a conceptual understanding of ANOVA you can interpret ANOVA tables for a wide variety of study designs.

The research question

Does the reduction in SBP differ among three doses of an investigational antihypertensive drug?

Study design

A clinical study was conducted to investigate three doses of an investigational antihypertensive drug. Fifteen participants were recruited (five per group), and randomized to three treatment groups: 10 mg, 20 mg, and 30 mg. Each treatment was taken once a day. SBP was measured 5 min before the administration of the drug (baseline) and again 30 min after. A "change from baseline score" was calculated for each participant by subtracting the baseline value from the post-treatment value.

Data

The change from baseline scores for the 15 participants are displayed below:

- 10 mg treatment group: −6, −5, −6, −7, −6
- 20 mg treatment group: −8, −9, −8, −9, −6
- 30 mg treatment group: −10, −8, −10, −8, −9.

Statistical analysis

A one-factor ANOVA is the appropriate analysis here assuming that the data are normal: The only factor of interest is the dose of drug given. There are three levels of this factor: 10, 20, and 30 mg. Following convention, the results of an ANOVA are displayed in an ANOVA summary table such as the model in Table 11.3. In the following calculations the values are presented without their units of measurement (mmHg) simply for convenience. At the end of the calculations, however, it is very important to remember that the numerical terms represent values measured in mmHg. The calculations needed are as follows.

1. Calculate the group means and the grand mean:
   - 10 mg group mean $= \bar{x}_{10} = \dfrac{-30}{5} = -6$
   - 20 mg group mean $= \bar{x}_{20} = \dfrac{-40}{5} = -8$
   - 30 mg group mean $= \bar{x}_{30} = \dfrac{-45}{5} = -9$
   - grand mean $= \bar{x} = \dfrac{(-6) + (-8) + (-9)}{3} = -7.67.$
2. Calculate the group sample variances:
   - 10 mg group sample variance =

$$s^2_{10} = \frac{((-6)-(-6))^2 + ((-5)-(-6))^2 + ((-6)-(-6))^2 + ((-7)-(-6))^2 + ((-6)-(-6))^2}{4} = 0.50$$

   - 20 mg group sample variance =

$$s^2_{20} = \frac{((-8)-(-8))^2 + ((-9)-(-8))^2 + ((-8)-(-8))^2 + ((-9)-(-8))^2 + ((-6)-(-8))^2}{4} = 1.50$$

   - 30 mg group sample variance =

$$s^2_{30} = \frac{((-10)-(-9))^2 + ((-8)-(-9))^2 + ((-10)-(-9))^2 + ((-8)-(-9))^2 + ((-9)-(-9))^2}{4} = 1.00.$$

3. Calculate the total sums of squares (SST): The total sums of squares is the variability of observations across all three groups. It is calculated by summing the squared difference of each observation (in this case 15 of them) from the grand mean, − 7.67. For brevity, the calculation is not written out here. We suggest that you verify the calculations with software:

- $SST = 35.33$.

4. Calculate the among-sample sums of squares (SSA):

   $SSA = 5((-6) - (-7.67))^2 + 5((-8) - (-7.67))^2 + 5((-9) - (-7.67))^2 = 23.33.$

5. Calculate the within-sample sums of squares (SSW):

- $SSW = (4)(0.50) + (4)(1.50) + (4)(1.00) = 12.$

As expected, the total sums of squares is the sum of the among-sample sums of squares and the within-sample sums of squares.

6. Calculate the df:
   - Total: We started with 15 scores. To get the same grand mean, 14 of these can vary, but number 15 cannot. Therefore, there are $(n - 1)$ df:

     df (total) = 15 − 1 = 14.
   - Among samples: There are three groups, and thus three sample means. These must also average to the grand mean. Once two have been determined, the third can be only one value (that is, it cannot vary). Again, therefore, there are $(k - 1)$ df:

df (among) = 3 − 1 = 2.

- Within samples: By exactly the same logic that we saw for the within-groups sums of squares, we can calculate these df as:

  df (within) = df (total) − df (among) = 14 − 2 = 12.

  (Note: There is also another way to think of this. Within each sample there are five values. Therefore, there are four df per sample. There are three samples. The total within-samples df is the total of the df within each sample, or 4 + 4 + 4 = 12.)

7. Construct the ANOVA table: Having calculated the total sums of squares from all sources of variation, along with their degrees of freedom, we can now start to construct the ANOVA table. The only other calculations required are the mean squares for among-samples and within-samples (divide each sums of squares by its associated df) and the test statistic, *F* (divide among-samples mean square by within-samples mean square). All of this information is shown in the partial ANOVA table presented as Table 11.3.
8. Determine if the test statistic is in the rejection region: As always, we need to determine if the test statistic *F* falls in the rejection region. So far, we have not determined the rejection region for this test. As noted earlier, the *F* distribution has two parameters that determine its shape and, therefore, the *F* values that cut off tail areas of the distribution. The two parameters are the numerator df (associated with the numerator of the *F* ratio or the among-sample source of variation) and the denominator df (associated with the denominator of the *F* ratio or the within-samples source of variation). In this case, the numerator df is 2 and the denominator df is 12. This is written as:

   $F(2,12) = 11.67$.

Tables with values of *F* for several distributions are used to determine the significance of this result, or the critical values can be obtained from statistical software. We have provided a table in Appendix 4. For a test of size $\alpha = 0.05$, the critical value associated with 2 numerator df and 12 denominator df that cuts off the upper 5% of the distribution is 3.89. Although tabled values are helpful at identifying nominal *p* values (for example, ≤ 0.01) statistical software is required to report the specific *p* value. Using statistical software, you will find that the actual *p* value is 0.002. Table 11.4 shows the completed ANOVA table for this example. You will see that the *p* value is commonly included in a complete ANOVA table.

Table 11.3 One-way ANOVA table for the SBP study (partially complete)

Source	Sum of squares	Degrees of freedom	Mean square	*F*
Among samples	23.33	2	11.67	11.67
Within samples	12.00	12	1.00	
Total	35.33	14		

Table 11.4 Completed one-way ANOVA table for the SBP study

Source	Sum of squares	Degrees of freedom	Mean square	*F*	*p* value
Among samples	23.33	2	11.67	11.67	0.002
Within samples	12.00	12	1.00		
Total	35.33	14			

It is important to recognize that the actual p value, not simply $p < 0.05$, is stated in the table. Regulatory reviewers and journal editors prefer this practice, because the actual value provides more information than simply a statement that the value is less than 0.05.

Interpretation and decision-making

As the value of the test statistic, 11.67, is in the rejection region for this test of size $\alpha = 0.05$ (that is, $11.67 > 3.89$), the null hypothesis is rejected in favor of the alternate, which means that at least one pair of the population means is not equal.

Recall the original research question: Does the reduction in SBP differ among three doses of a new antihypertensive? The results of the one-way ANOVA that we have conducted so far are interpreted in the following manner:

- There is evidence at the $\alpha = 0.05$ level that the levels of the factor "dose of drug" differ. Therefore, there is a statistically significant difference in SBP change scores between the groups. (The p value of 0.002 indicates that the null hypothesis would also have been rejected at smaller α levels, for example, at the $\alpha = 0.01$ level.)

The above statement by itself does not, however, tell us anything about which group showed the greatest change score, or indeed how any specific group compared with any of the other groups. Consideration of the group means is necessary to do this. These means, with the associated units of measurement reinserted, are:

- 10 mg group = −6 mmHg
- 20 mg group = −8 mmHg
- 30 mg group = −9 mmHg.

Therefore, we can now state that the 30 mg group showed the greatest mean decrease in SBP, the 20 mg group the second greatest mean decrease, and the 10 mg group the least mean decrease. However, a full answer to the research question has still not been supplied, at least not in terms of determining possible statistical differences between specific pairs of dose levels.

The ANOVA test statistic revealed that, overall, the groups differed statistically significantly, but, as there are more than two groups, it cannot reveal the precise pattern of statistical significance. For any three groups (call them D, E, and F) there are $C_2^3 = 3$ possible comparisons between pairs of groups: D can be compared with E; D can be compared with F; and E can be compared with F. These three comparisons can lead to the following patterns of outcomes:

- All groups differ statistically significantly from each other.
- None of the groups differs statistically significantly from any other group.
- D and E both differ statistically significantly from F, but do not differ statistically significantly from each other.
- D and F both differ statistically significantly from E, but do not differ statistically significantly from each other.
- E and F both differ statistically significantly from D, but do not differ statistically significantly from each other.

To determine which pattern of outcomes occurred in any given situation, an additional statistical test is needed. In situations such as this, where we have a partial answer to our original research question, multiple comparisons are performed. These are tests that allow us to compare the means of each pair of groups to see which pairs (if any) differ statistically significantly from each other. Multiple comparisons therefore provide a more detailed understanding of our data than the overall test (referred to as the omnibus test) provided by the ANOVA. If the omnibus test yields a nonsignificant result, multiple comparisons are not necessary, because, in fact, none of them would be significant. In the case of a significant omnibus test, the second option above is not actually a possible outcome, whereas all of the others are. This means that we need a method of determining which of the other possibilities is the case – that is, we need a statistical methodology that will allow us to conduct multiple comparisons, and to use this methodology before we can provide the full answer to our original research question. The full answer is provided in Section 11.10, but first we need to look at another issue.

11.6 A statistical methodology for conducting multiple comparisons

In clinical studies, the probability of declaring a treatment efficacious when in reality it is not efficacious is termed α. This is the probability of detecting a false positive, or committing a type I error. As we have seen, the probability of committing a type I error should be limited to a specific value so that erroneous conclusions are not made very often. For a sponsor, committing a type I error could result in investing significant amounts of money on a drug that really does not work. For a regulatory agency, committing a type I error (approving a drug that is not efficacious) could result in many people taking a drug that does not offer a meaningful treatment benefit and may carry some risk (every drug has a side-effect profile). It is therefore important to constrain the probability of committing a type I error to an acceptable level. Traditionally, this acceptable level has been and is still regarded as the $\alpha = 0.05$ level, but, as noted before, we can choose other values when we consider them appropriate.

The important point to note here is that the $\alpha = 0.05$ level is deemed appropriate when a single test is being conducted. Multiple comparisons, by definition, mean that more that one test is being conducted. When testing a number of pairwise comparisons – for example, after an ANOVA where the null hypothesis has been rejected – it is not acceptable to test each pairwise comparison at the $\alpha = 0.05$ level because of the potential inflation of the overall type I error rate.

When three treatment groups are evaluated in a clinical study, there are three possible pairwise comparisons of means (*D* vs *E*, *D* vs *F*, and *E* vs *F*). If each mean is tested at the $\alpha = 0.05$ level, and assuming that they are mutually exclusive, the probability of declaring at least one of the pairs significantly different is equal to 1 minus the probability of accepting all three (by the complement rule). Assuming that the comparisons are independent, the probability of accepting all three null hypotheses is the probability of accepting the first null hypothesis multiplied by the probability of accepting the second multiplied by the probability of accepting the third. When testing each at the $\alpha = 0.05$ level, this probability becomes:

$$P\,(\text{incorrectly rejecting at least one hypothesis}) = 1 - (0.95)(0.95)(0.95) = 1 - 0.95^3 = 0.14.$$

That is, instead of a type I error rate of $\alpha = 0.05$, this analysis has resulted in a higher probability of committing a type I error, just by chance alone.

In fact, the comparisons made here cannot be thought of as independent because each group is compared with two others in this case. It is more correct to use an inequality sign to say that the probability is no more than 0.14, that is, ≤ 0.14. However, this technicality is of little comfort because, to make sound decisions, we would really like to limit that probability to a reasonable level. In general, if *C* comparisons are each made at the α level, the probability of rejecting at least one by chance alone is:

$$P\,(\text{rejecting at least one of } c \text{ hypotheses}) \le (1 - (1 - \alpha)^c)$$

Table 11.5 lists the probability of rejecting at least one hypothesis for a number of values of *C*, the number of hypothesis tests performed at the conventional $\alpha = 0.05$ level.

Table 11.5 Maximum probability of committing a type I error when each hypothesis is tested at $\alpha = 0.05$

C: No. of hypotheses tested at $\alpha = 0.05$	Maximum probability of type I error
1	0.050
2	0.098
3	0.143
4	0.185
5	0.226
6	0.265
7	0.302
8	0.337
9	0.370
10	0.401
15	0.537
20	0.642

Suppose that a clinical trial has to evaluate four doses of a test treatment and a placebo (a total of five groups) on relieving headache pain. The study was carefully designed and conducted, and the data are now ready for the statistical analysis. A one-way ANOVA is conducted, and the conclusion from the omnibus F test (comparison of the among-sample variance with the within-sample variance) is that the population means are not all equal. Five treatment groups give rise to $C_2^5 = 10$ pairwise group comparisons. Suppose that one of the researchers failed to get input from the trial statistician, and hurriedly (and mathematically correctly) analyzed all 10 pairwise comparisons of means performed using 10 two-sample t tests. The researcher takes his or her results to the study director and the rest of the study team and points out with tremendous excitement that the pairwise comparison of the lowest dose with the placebo yielded a p value of 0.023, a statistically significant result at the $\alpha = 0.05$ level. A surge of positive energy fills the room as everyone but the statistician declares, "We have found our lowest effective dose! On to the confirmatory trial!"

As you have probably realized by now, there would actually be little reason for enthusiasm, as the study statistician would very soon point out. The problem is this: While each of the 10 two-sample t tests had been conducted mathematically correctly, it is not appropriate statistical methodology to use 10 two-sample t tests in this setting. The analytic strategy employed did not limit the type I error rate to 0.05. Rather, as seen in Table 11.5, when 10 such pairwise comparisons are made – that is, 10 hypotheses are tested – the probability of rejecting at least one of the hypotheses is limited to 0.401, a value considerably greater in magnitude than 0.05. In other words, use of this naïve analytic strategy has resulted in an inflated type I error. There is up to a 40% chance of being misled by one test with a nominal p value ≤ 0.05.

The issue of type I error inflation caused by multiple testing appears in many guises in the realm of new drug development. This issue is of great importance to decision-makers, and we discuss this topic again later in the chapter. For now, we have not yet provided a full answer to our research question; our description of analysis of variance is incomplete without a discussion of at least one analysis method that controls the overall type I error rate when evaluating pairwise comparisons from an ANOVA.

11.7 Bonferroni's test

Bonferroni's test is the most straightforward of several statistical methodologies that can appropriately be used in the context of multiple comparisons. That is, Bonferroni's test can appropriately be used to compare pairs of means after rejection of the null hypothesis following a significant omnibus F test. Imagine that we have c groups in total. Bonferroni's method makes use of the following inequality:

$$P(R_1 \text{ or } R_2 \text{ or } R_3 \text{ or } \ldots \text{ or } R_c)$$
$$\leq P(R_1) + P(R_2) + P(R_3) + \ldots + P(R_c).$$

This means that the probability of rejecting at least one of c hypotheses is less than or equal to (thus the term "inequality") the sum of the probabilities of rejecting each hypothesis. This inequality is true even if the events, in this case rejecting one of c null hypotheses, are not independent. Recall from Section 6.2 that, when events are not independent, the probability of intersecting events should be subtracted. Using Bonferroni's method, testing each pair of means with an α level of $\alpha_B = \frac{\alpha}{c}$ will ensure that the overall type I error rate does not exceed the desired value of α. It follows that the probability of rejecting at least one of c null hypotheses can be expressed as follows:

$$p(\text{rejecting at least one of } c \text{ hypotheses at } \alpha_B \text{ level}) \leq c\left(\frac{\alpha}{c}\right) = \alpha.$$

It is important to note that the researcher in our scenario in Section 11.6 who hurriedly conducted 10 pairwise comparisons using 10 two-sample t tests and rejoiced in one particular finding was not completely out of line in the analytic strategy chosen. It is indeed possible to approach this situation (the need for 10 pairwise comparisons) with the intent to conduct 10 two-sample t tests. However, a correction must be made to the α level used to determine statistical significance. In the scenario as told in Section 11.6 the researcher did not perform this critical step.

In practice, then, we can carry out each of a series of pairwise comparisons of means using a two-sample t test for each comparison, but the α level must be modified accordingly. When deciding whether or not to reject the null hypothesis associated with each comparison, we need to use an α level of $\alpha_B = \frac{\alpha}{c}$ instead of the naïve choice of α. Note that this is equivalent to defining a rejection region for each test as:

$$t < t_{\alpha/2c,n-k} \text{ or } t > t_{1-(\alpha/2c),n-k}$$

which makes sense as the tail areas in the left and right of the t distribution are smaller than those obtained using the two-sided test of size α.

Consider the two-sample t-test statistic again:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

In an ANOVA involving more than two groups, we estimate the underlying variability from more than two samples, and yet we are interested in the extent to which (only) two of the means differ from each other. Therefore, when comparing the means of two samples, the pooled standard deviation from the two-sample case, s_p, is replaced by an estimate that captures the variability across all groups in the analysis – the mean square error or the within-samples mean square. Recall from Section 11.4 that this quantity has the same interpretation as the pooled standard deviation, the typical spread of data across all observations.

When using Bonferroni's method, the null hypothesis associated with a pairwise comparison is rejected if the calculated test statistic, that is,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{V_w \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

is in the rejection region defined as $t < t_{(\alpha/2c),n-k}$ or $t > t_{1-(\alpha/2c),n-k}$.

Remember that V_w comes from the ANOVA table and it is the mean square error, which has also been referred to as the within-samples variability or, more informally, the background noise. This is analogous to s_p^2 in the two-sample case. As we assume equal variances, we use the estimator that uses the most data and therefore gives the most precise estimate.

The critical value can be determined from a table or software (using a two-sided test of size α/c). The estimate of the underlying variability, V_w, comes from the ANOVA table, and the sample sizes for each group are known *and equal*. Then we can define a quantity, the minimally significant difference (MSD), which is the smallest difference (in absolute value) between any two sample means that could be considered statistically significant at the α level. (Note that when sample sizes are not equal the MSD is not defined, but there are other methods available.)

$$\text{MSD} = t_{1-(\alpha/2c),n-k} \sqrt{V_w \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Once the value of MSD has been determined, the absolute value of the difference in means will be compared with the MSD. If the absolute value of the difference in means, $|(\bar{x}_1 - \bar{x}_2)|$, is greater than or equal to the MSD the null hypothesis will be rejected.

11.8 Employing Bonferroni's test in our example

Having introduced Bonferroni's test, we can now return to our earlier example to see how to apply Bonferroni's method to our pairwise comparisons of treatment group means.

Statistical analysis

The significant result of the omnibus F test led to the rejection of the null hypothesis of no significant differences, thereby revealing the presence of a significant difference between at least one pair of means. It is now of interest to determine precisely which pair or pairs of means are significantly different.

Given that the decisions made from this trial could result in sizeable further investment in the development of the investigational antihypertensive drug, the company would like to minimize its chances of committing a type I error. That is, it would like to maintain an overall type I error of 0.05. As we have just seen in Section 11.7, one analysis that will maintain this desired type I error of 0.05 is Bonferroni's method.

In our example of three treatment groups there are three pairwise comparisons of interest. Therefore, each pairwise comparison will be tested at an α level of 0.05/3 = 0.01667. This α level will require defining a critical value from the *t* distribution with 12 (that is, 15 − 3) df that cuts off an area of 0.00833 (half of 0.01667) in the right-hand tail. Use of statistical software reveals that the critical value is 2.77947. From inspection of the ANOVA table presented as Table 11.4 the within-samples mean square (mean square error) can be seen to be 1. The final component needed for the MSD is:

$$\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \sqrt{0.4} = 0.632.$$

Then the MSD is equal to:

$$\text{MSD} = (2.77947)(1)(0.632) = 1.757.$$

The mean values for each group are −6 mmHg (10 mg), −8 mmHg (20 mg), and −9 mmHg (30 mg). The absolute values of the three differences in means are displayed in Table 11.6.

Table 11.6 Absolute values of differences in means

	20 mg treatment group	30 mg treatment group
10 mg treatment group	2	3
20 mg treatment group		1

Each cell represents the differences in means for the groups represented by each row and column. The differences between the 10 mg and 20 mg groups and the 10 mg and 30 mg groups were both greater than the MSD (1.757). Therefore, these differences are considered statistically significant at the α = 0.05 level. The difference between the 20 mg and 30 mg groups was not significant, however, because it was less than the MSD.

Interpretation and decision-making

We are now in a position to provide a full answer to our research question of interest as expressed at the start of Section 11.5: Will the reduction in SBP differ among three doses of an investigational antihypertensive drug?

The first step in our analytical strategy was to conduct an ANOVA. This ANOVA tested the null hypothesis that there were no differences among the three means. The null hypothesis was tested at an α level of 0.05, and was rejected on the basis of the significant omnibus *F* test.

The second step in our analytical strategy was to determine which of the pairs of means were significantly different from each other. Testing each of the three hypotheses at an α level of 0.05 would have resulted in a probability of committing a type I error possibly > 0.05 (the desired level). Bonferroni's inequality was therefore used to test each of the three hypotheses at an α level of 0.05/3 = 0.01667. Using the critical value for this α level resulted in two pairs of means being declared significantly different at the 0.05 level.

The full interpretation of the study, therefore, is that the magnitude of the reduction in SBP does indeed differ according to different dose levels. The 20 mg and 30 mg doses both resulted in a statistically significantly greater SBP reduction than the 10 mg dose. There was insufficient evidence to claim that there is a statistically significant difference between the 20 mg and 30 mg doses.

What are the implications of this interpretation? First, if we decided that it would be useful to continue the clinical development program with another trial, it would be salient to note that, in terms of efficacy, the 10 mg dose was inferior to the other two. Therefore, if continuing, it is likely that we would not include the 10 mg dose in further trials. What else would help us to decide to continue with the clinical development program? The safety and tolerability of the 20 and 30 mg doses would need to be examined and deemed acceptable. Examining the safety and tolerability data from the participants in these two treatment groups would provide the evidence on which to base this decision (the safety and tolerability data from participants in the 10 mg treatment group would not be informative at this point). If there were no safety or tolerability concerns with the 20 or 30 mg doses, the next stage in development could be to continue to investigate both of these doses. Another possible interpretation is discussed in Section 11.10.

11.9 Tukey's honestly significant difference test

Bonferroni's method that we have just discussed is perhaps one of the most easily understood methods to maintain an overall type I error, which is one of its advantages. In addition, Bonferroni's method does indeed control the overall type I error rate well, such that it is guaranteed to be ≤ α. However, like many items that we discuss in this book, it has its disadvantages as well as its advantages.

Bonferroni's test is overly conservative, in that the critical values required for rejection need not be as large as they are. In other words, using a less conservative method may result in more null hypotheses being rejected. The reason that Bonferroni's method is so conservative is that it does not in any way account for the extent of correlation among the various hypotheses being tested. If a method could take into account the overlap, or lack thereof, of the various hypotheses, the critical values would not need to be defined as narrowly as with Bonferroni's. In this section, we therefore discuss another analytical strategy for multiple comparisons, Tukey's honestly significant difference (HSD) test.

Bonferroni's method for testing pairs of means (maintaining an overall type I error rate of α) involved comparing the absolute differences in means to the MSD, which was defined as a function of:

- the critical value from a *t* distribution with a combined area of α/*c* in the tails of the distribution
- the within-samples variability
- the sample sizes in each group.

Once a value of the MSD was determined each difference in means was calculated and compared with the MSD. Any difference that was equal to or greater than the MSD was considered statistically significant. Tukey's HSD test is carried out in a similar manner. A value called the honestly significant difference is determined as a function of three things:

1. The critical value from the studentized range statistic
2. The within-samples variability
3. The sample sizes in each group.

The studentized range statistic, called *q* in the following description of the test, has a limited use for us now and we shall not spend any additional time characterizing it, except to say that the value of *q* does account for the relative size of differences among the normalized means, resulting in a test with an overall type I error of exactly 0.05. The value, *q*, is often provided in tables and to look it up we need to know the number of groups (*k* from the ANOVA description), and the number of df associated with the within-samples mean square ($n - k$). Statistical software packages also supply this number. The HSD (or, equivalently, the MSD_T for minimum significant difference – Tukey) is defined as:

$$\text{HSD} \equiv \text{MSD}_T = q\sqrt{\frac{V_\text{W}}{n}}.$$

In this expression *n* represents the per-group sample size which, for the moment, we require to be equal.

Once the value of HSD has been determined, the absolute value of the difference in means is compared with it. If the absolute value of the difference in means, $|(\bar{x}_1 - \bar{x}_2)|$, is greater than or equal to the HSD the null hypothesis is rejected.

The quantity represented by the letter "*q*" is determined from a table of values used just for this test. Two characteristics are needed to determine the appropriate value of *q* each time that it is used. These characteristics are represented by the letters "*a*" and "*v*." The letter *a* represents the number of groups, which in this example is 3. The letter *v* represents the df, which in this test is the df associated with the within-samples mean square. In this case, the value of *v* is 12, as calculated for and shown in the ANOVA summary table in Table 11.4. From the table of *q* values for Tukey's test (provided in Appendix 5) the value of *q* associated with an

(a, v) value of (3, 12) is 3.77. HSD is then calculated as follows:

$$\text{HSD} = 3.77\sqrt{\frac{1}{5}} = 1.686.$$

The absolute values of the three differences in means were displayed in Table 11.6. The differences between the 10 mg and 20 mg groups and the 10 mg and 30 mg groups were both greater than the HSD (1.686). Therefore, these differences are considered statistically significant at the 0.05 level. The difference between the 20 mg and 30 mg groups was not significant, however, because it was less than the HSD.

Although Tukey's method does not require equal sizes among the groups, imbalanced group sizes do require a different calculation of HSD. When the sample sizes are unequal among all groups being compared, there is not one common value of HSD because this value relies on the sample size per group. For the comparison of any two means with group sample sizes of n_1 and n_2, the value of HSD corresponding to that particular comparison is:

$$\text{HSD} = \frac{q}{\sqrt{2}}\sqrt{V_w\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

In the case that n_1 and n_2 are equal this expression simplifies to the one we originally presented.

Interpretation and decision-making

Having gone through the calculations necessary for Tukey's test, we can look at how these results would lead to decision-making, and also compare the interpretation and decision-making with those that followed from using Bonferroni's methodology on the same dataset.

The statistical interpretations of these results are the same as with Bonferroni's method. The 20 and 30 mg doses both resulted in a statistically significantly greater SBP reduction than the 10 mg dose. There was insufficient evidence to claim that there is a statistically significant difference between the 20 mg and the 30 mg doses.

11.10 Implications of the methodology chosen for multiple comparisons

The most important lesson to be learned from our discussions of various analytic methodologies for multiple comparisons is that the method chosen can have a major impact on the risk of making incorrect decisions.

Consider the absolute difference in any two means that was required to reject a null hypothesis of H_0: $\mu_1 - \mu_2 = 0$ after rejection of the omnibus F test. In the case of the naïve approach, which was to test each pair of means separately and use an α level of 0.05 in each case, the minimum significant difference would be 1.126, but the overall type I error could be guaranteed to be bounded only by 0.143 (see Table 10.5). The use of Bonferroni's method resulted in a minimum significant difference of 1.757, but it is overly conservative and the overall type I error rate would be guaranteed to be < 0.050. Tukey's method, which accounts for the actual distribution of differences through q, resulted in a minimum significant difference of 1.686 and guaranteed that the overall type I error rate $= 0.050$, resulting in a more powerful test than Bonferroni's method. Given their importance, these characteristics are summarized in Table 11.7.

Lastly, it is important to note that differences such as these underscore the importance of declaring the primary analysis approach in a study protocol or statistical analysis plan. Committing to the most appropriate analysis from first principles is not only good scientific discipline, it is also necessary to withstand regulatory scrutiny.

It should be noted that these are not the only acceptable methods applicable to multiple comparisons from an ANOVA. In each individual case, the choice among possible approaches is largely dependent on the study design. For example, Dunnett's test can be used when the only comparisons of interest are each test treatment versus a control (for example, in a placebo-controlled, dose-ranging study). Like Tukey's test, Dunnett's method is more powerful than Bonferroni's. In general, other methods gain power compared with Bonferroni's method by

Table 11.7 Characteristics of the methods to test the three pairwise comparisons of means in the ANOVA example

Method	Minimally significant difference	Overall type I error rate: *P*(rejecting at least one null hypothesis)
Naïve approach (incorrect)	1.126	≤ 0.143
Bonferroni's test (correct but conservative)	1.757	< 0.050
Tukey's HSD test (correct, and more powerful than Bonferroni's)	1.686	$= 0.050$

using methods that account for the correlation of tests (for example, Tukey's HSD test) or by reducing the number of tests about which we would like to make an inference (for example, Dunnett's test). When conducting these types of analyses, it is theoretically possible (although not common) to report a significant overall *F* test, but not declare any pairwise comparison as statistically significant as a result of the multiple comparison procedure.

Consideration of the possible clinical interpretation of these results is also worthwhile. The interpretations given in the above sections are the full statistical interpretations from the statistical analyses that were performed on the data collected in this study. In real clinical trials, these results are also interpreted clinically, that is, their clinical significance is discussed. Making these clinical efficacy interpretations is the province of the clinicians on the study team. As we emphasized earlier in this book, we are not clinicians, and these "hypothetical comments" concerning the potential clinical significance of hypothetical data must be regarded in this light.

First, the clinical significance of a decrease in SBP of 6 mmHg versus a decrease of 8 or 9 mmHg would need to be considered. As these numerical values are all relatively close, let us create some hypothetical values that conform to the same overall pattern of significance but are more different from each other. Suppose that these mean decreases in SBP were observed using the same doses of a different antihypertensive drug:

- 10 mg group mean = -6 mmHg
- 20 mg group mean = -18 mmHg
- 30 mg group mean = -19 mmHg.

Suppose also that Tukey's test provided evidence of the same pattern of statistical significance:

- 10 versus 20 mg = $-6-(-18) = 12$; $p < 0.05$
- 10 versus 30 mg = $-6-(-19) = 13$; $p < 0.05$
- 20 versus 30 mg = $-18-(-19) = 1$; not significant (ns).

In this scenario, the clinical significance of a decrease in SBP of 6 mmHg versus a decrease of 18 or 19 mmHg would need to be considered. Suppose that decreases of 18 and 19 mmHg are both considered to be much more clinically significant than a decrease of 6 mmHg. Suppose also that the 20 mg dose had a good (and therefore acceptable) safety profile, whereas the safety profile of the 30 mg dose was not so good. Of relevance in this scenario is that there was not a statistically significant difference in efficacy between these two dose groups. It is true that the mean decrease in the 20 mg group was numerically less than the mean decrease in the 30 mg group, but it was not statistically significantly less. Therefore, it might be the case that, when input had been received from all members of the study team, including statisticians and clinicians, a decision would be made to progress only the 20 mg dose to further trials: The 10 mg dose is statistically significantly less effective, and the 30 mg dose has a less desirable safety profile while also not being statistically significantly more effective (see Turner, 2007).

This scenario illustrates several key points:

- Decision-making is not necessarily straightforward.
- The empirical evidence from our clinical trials provides the basis for rational decision-making.

- In most cases many members of the study team, including statisticians and clinicians, are needed to make the optimum decision.

In real life, clinical interpretations are vital to balance the relative weight of safety and efficacy considerations. If a higher dose of a given drug is considerably more efficacious than a lower dose and leads to only a minimal increase in very mild side-effects, a clinician may decide that, on balance, it is worth recommending the higher dose. Conversely, if a higher dose of a given drug is only minimally more efficacious than a lower dose and leads to a considerable increase in moderate or severe side-effects, a clinician may recommend the lower dose.

11.11 Additional considerations about ANOVA

Before completing our discussions of ANOVA, there are several additional points that we would like to address, because these questions may have occurred to you as you have read the preceding descriptions of the use of ANOVA and multiple comparisons in this chapter.

11.11.1 ANOVAs with only two groups

A one-way ANOVA containing three levels was used as the worked example in this section because a *t* test cannot address a design with more than two levels. However, the one-way ANOVA can certainly be used in situations involving only two levels. A reasonable question, therefore, is: In situations involving only two levels, where the only possible comparison is between one level and the other, is there any advantage in using the one-way ANOVA instead of the *t* test?

The answer is no. In cases where there are only two levels, either test is applicable. The values obtained in the calculations of the respective tests will be different, but the tests will give precisely the same answer in terms of the degree of statistical significance obtained by the respective test statistic. That is, the *t* value and *F* value will not be the same (the *F*-test statistic will be square of the *t*-test statistic), but the associated *p* values will be identical. The advantage of the ANOVA lies with its ability to address situations involving more than two levels, which are very common in clinical research.

11.11.2 Only collect data that you intend to analyze

Consider a scenario where a series of possible comparisons exists, but the investigator is genuinely interested only in one of these comparisons. Such a hypothetical scenario might involve a study employing four groups, with participants in each group receiving one of four dose levels (1, 2, 3, and 4) of a particular drug, and primary interest lay with comparing dose levels 1 and 4 – that is, out of the possible six comparisons, interest lay only with the comparison of dose levels 1 and 4. A question that arises here is: Is it possible to argue that this one comparison could be made without having to adopt a more conservative approach? The correct answer from a purely statistical computational viewpoint is yes, this argument can successfully be made. The individual test may be undertaken using a *t* test at the $\alpha = 0.05$ level, that is, without adopting a more conservative approach, because this one particular comparison of interest was specified from first principles. However, this is not the final answer here.

Although this argument is perfectly satisfactory from a purely computational view, another question begs to be asked: If the investigator was interested only in comparing dose levels 1 and 4, why were dose levels 2 and 3 included in the study? This question is pertinent in several ways. It costs a lot of time and money to collect such clinical data, and the costs associated with participants in two of the four experimental groups would be wasted. Much more important than the unnecessary costs, however, would be that the participants in the dose level 2 and 3 treatment groups would have taken part in the study for no useful reason, a gross violation of experimental ethics.

A much more realistic scenario is one in which four doses are included in such a study because the investigator does not have clear logical ideas (hypotheses) about the relative merits (perhaps relative efficacy) of the doses. In this case, an original omnibus analysis such as the one-factor ANOVA provides a very efficient initial test for differences among the groups. If a statistically significant result is given by the ANOVA, the investigator can then proceed to comparing pairs of groups in formal (and appropriate) multiple comparison testing.

11.12 Nonparametric analyses of continuous data

There are times when the required assumptions for ANOVA, a parametric test, are not met. One example would be if the underlying distributions are non-normal. In these cases, nonparametric tests are very useful and informative. For example, we saw in Section 11.3 that a nonparametric analog to the two-sample *t* test, Wilcoxon's rank sum test, makes use of the ranks of observations rather than the scores themselves. When a one-factor ANOVA is not appropriate in a particular case a corresponding nonparametric approach called the Kruskal–Wallis test can be used. This test is a hypothesis test of the location of (more than) two distributions.

11.13 The Kruskal–Wallis test

All that is required for this test to be employed is that the observations classified into *k* groups are independently sampled from populations and the random variable is continuous with the same variability across the populations represented by the samples. Importantly, no assumption about the shape of the underlying distribution is required, making this test suitable for non-normal underlying distributions.

In the Kruskal–Wallis test the original scores are first ranked and an ANOVA analysis is then carried out on the ranks. As with Wilcoxon's rank sum test, ranking of the observations must deal with ties. The sums of squares are based on these ranks, and the test statistic is based on a ratio of the among-samples variability in ranks and the within-samples variability in ranks.

All observations, x_{ij}, are assigned ranks, r_{ij}, and therefore the usual sums of squares can be calculated for the rank scores, r_{ij}. For brevity, the expressions for each are provided in Table 11.8, a general one-way ANOVA table, on the basis of ranks.

The quantities in the ANOVA table based on ranks represent similar quantities as the ANOVA table based on the original scores:

r_{ij} is the rank for individual *i* in group *j*
n_j is the sample size for group *j*

$$n = \sum_{j=1}^{k} n_j \text{ is the total sample size}$$

$\bar{r}_j$ is the average rank for group *j*

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (r_{ij} - \bar{r}_j)^2}{n_j - 1} \text{ is the variance of ranks in group } j.$$

$\bar{r}$ is the average rank over all groups (the grand mean rank), which can be simplified as:

$$\bar{r}_{.} = \frac{n+1}{2}.$$

The omnibus test statistic, X^2, follows a χ^2 distribution with $k - 1$ df. If the omnibus test is rejected the pairs of groups can be evaluated using a Bonferroni-type approach. This requires the assumption that the ranks are normally distributed. As with the parametric one-way ANOVA, a minimally significant difference in ranks can be calculated for this purpose as:

$$\text{MSD} = z_{1-(\alpha/2c)} \sqrt{V_{\text{W,ranks}} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

For the sake of this example, we use the data from the parametric ANOVA example to illustrate the Kruskal–Wallis test. If it seems at all strange to use the same data for both examples, a parametric analysis and a nonparametric analysis, it is worth noting that a nonparametric analysis is always appropriate for a given dataset meeting the requirements at the start of the chapter. Parametric analyses are not always appropriate for all datasets.

Table 11.8 General one-way ANOVA table for ranks (Kruskal–Wallis test)

Source	Sum of squares	Degrees of freedom	Mean square	X^2
Among samples	$\sum_{j=1}^{k} n_j(\bar{r}_j - \bar{r}.)^2$ $= \mathrm{SSA}_{\mathrm{ranks}}$	$k - 1$	$\dfrac{\sum_{j=1}^{k} n_j(\bar{r}_j - \bar{r}.)^2}{k - 1}$ $= V_{\mathrm{A,\ ranks}}$	$V_{A,ranks}/V_{W,ranks}$
Within samples	$\sum_{j=1}^{k}(n_j - 1)s_j^2$ $= \mathrm{SSW}_{\mathrm{ranks}}$	$n - k$	$\dfrac{\sum_{j=1}^{k}(n_j - 1)s_j^2}{n - k}$ $= V_{\mathrm{W,\ ranks}}$	
Total	$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(r_{ij} - \bar{r})^2$ $= \mathrm{SST}_{\mathrm{ranks}}$	$n - 1$	$\dfrac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(r_{ij} - \bar{r})^2}{n - 1}$ $= V_{\mathrm{T,\ ranks}}$	

Statistical analysis

The analysis begins with ordering all 15 observations. Note that statistical software packages order and rank the observations and do the ANOVA for you. The ordered observations from lowest to highest across the three groups are as follows:

10 mg										−7	−6	−6	−6		−5
20 mg				−9	−9			−8	−8					−6	
30 mg	−10	−10	−9			−8	−8								

Then ranking each observation, accounting for ties as described for the one-way ANOVA, the following ranks are obtained:

10 mg										10	12.5	12.5	12.5		15
20 mg				4	4			7.5	7.5					12.5	
30 mg	1.5	1.5	4			7.5	7.5								

The within-samples average ranks are:

$$\bar{r}_{10} = 12.5$$
$$\bar{r}_{20} = 7.1$$
$$\bar{r}_{30} = 4.4.$$

And the grand mean rank:

$$\bar{r}_{.} = 8.$$

The within-samples variances (of ranks) are:

$$s_{10}^2 = 4.63$$
$$s_{20}^2 = 12.18$$
$$s_{30}^2 = 9.05.$$

The among-samples mean square is calculated as:

$$V_{\mathrm{A,ranks}} = \frac{5(12.5 - 8)^2 + 5(7.1 - 8)^2 + 5(4.4 - 8)^2}{2} = 85.05.$$

The within-samples mean square (mean square error) is calculated as:

$$V_{\mathrm{W,ranks}} = \frac{4(3.13) + 4(12.18) + 4(9.05)}{12} = 8.12.$$

Finally, the test statistic is the ratio of these two:

$$85.05/8.12 = 10.48.$$

We note that the test statistic is greater than the critical value of 5.991 (2 df with an α level of 0.05), so the null hypothesis is rejected.

The next step is to decide which groups (three comparisons) are different with respect to their location. For this purpose the MSD is calculated as:

$$\mathrm{MSD} = z_{0.992}\sqrt{8.12\left(\frac{1}{5} + \frac{1}{5}\right)} = 2.41\sqrt{8.12\left(\frac{1}{5} + \frac{1}{5}\right)} = 4.34.$$

The differences in mean ranks are displayed in Table 11.9. Differences in mean ranks that are greater than the MSD are considered significantly different.

Table 11.9 Absolute differences in mean ranks

	20 mg	30 mg
10 mg	5.4	8.1
20 mg		2.7

Interpretation and decision-making

As the difference in mean ranks exceeds the MSD for the comparison of 10 vs 20 mg and 10 vs 30 mg, we can conclude that these distributions differ in location. This testing procedure ensured that the overall type I error did not exceed 0.05. To interpret the clinical relevance of the differences detected by the test requires some additional point estimates. As the initial procedure was a nonparametric one, the differences in sample means are not appropriate. A more reasonable choice would be to compare the medians as an estimate of the treatment effect.

The nonparametric one-way ANOVA can be quite useful in a number of settings. The most obvious is when reasonable judgment does not allow you to conclude that the distributional assumptions for the one-way parametric ANOVA will hold. Another instance is when the data available for analysis are only ordinal (for example, like a rank) such that the difference between two values does not hold the same meaning as an interval scaled random variable.

There are a number of nonparametric analysis methods dealing with continuous data. The last statistical method included in this chapter is to be used when the continuous outcome is time to an event.

11.14 Hypothesis test of the equality of survival distributions: Logrank test

In Chapter 8 we described analyses to estimate the survival distribution of time to an adverse event. The survival function is the probability that an individual survives (that is, does not experience the event) longer than time t:

$$S(t) = P(\text{individual survives longer than } t).$$

In Chapter 10 the use of this method was discussed in terms of estimating the median survival time for participants in a clinical trial. The median survival time can be helpful as a single summary statistic that defines a typical survival time. However, survival distributions may deviate at various points in time. In this section we present the logrank test, which can be used to test the equality of two or more survival distributions. This is not the only test that can be used for this purpose, but it is a natural extension of a method that we have already described and so we have chosen to discuss it.

A test of the equality of two survival distributions would be expressed in terms of the null hypothesis:

$$H_0: S_1(t) = S_2(t).$$

If there is sufficient evidence to reject the null hypothesis the alternate hypothesis would be favored:

$$H_A: S_1(t) \neq S_2(t).$$

If, in the context of the survival distribution we consider all of the times at which an event occurred and index them as $t(1) < t(2) < t(3) \ldots < t(H)$, it is possible to create a 2 × 2 classification table for event times t(h), where $h = 1, 2, 3, \ldots, H$ in which the numbers of individuals with and without the event of interest are displayed for each group. Table 11.10 is a sample cross-classification table for time h.

Table 11.10 Cross-classification table of treatment and event at time h

Event?	Group 1	Group 2	Total
Yes	m_{1h}	m_{2h}	m_h
No	$n_{1h} - m_{1h}$	$n_{2h} - m_{2h}$	$n_h - m_h$
	n_{1h}	n_{2h}	n_h

Given the familiar set-up of this contingency table it may not surprise you that we can use the methods of the stratified (Mantel–Haenszel) χ^2 test to define a test statistic. Each of the distinct event times is treated as a stratum, just as we treated investigative centers as strata earlier. The test statistic for the logrank test is:

$$X^2_{LR} = \frac{\left(\sum_{h=1}^{H} \frac{n_{h1}\, n_{h2}}{n_h}(\hat{p}_{h1} - \hat{p}_{h2})\right)^2}{\sum_{h=1}^{H} \frac{n_{h1}\, n_{h2}}{n_h - 1}\bar{p}_h \bar{q}_h}.$$

As before, the proportion of observations with the characteristic of interest at time h for the two independent groups is denoted by $\hat{p}_{h1}$ and $\hat{p}_{h2}$, respectively. The overall proportion of individuals with the characteristic of interest within each time h is denoted by $\bar{p}_h$. The overall proportion of individuals without the characteristic of interest within each time h is denoted by $\bar{q}_h = 1 - \bar{p}_h$.

When the sample size is reasonably large ($n > 30$), the test statistic X^2_{LR} follows a χ^2 distribution with 1 df. Values of the test statistic that lie in the critical region are those with $X^2_{LR} > \chi^2_{1,1-\alpha}$, that is, values of χ^2 with 1 df that cut off the upper tail area of α.

To illustrate an example, we use the data from Chapter 9 with some modifications. Although the event of interest in that case was an adverse event, a safety parameter, we can treat it this time as an efficacy parameter.

Event of interest

The event of interest is return to a state of normal blood pressure (by some measure). The treatment administered to the group demonstrating earlier event times would be considered the better treatment.

Design

In this 10-day study of a novel antihypertensive, hypertensive study participants were randomly assigned to test treatment or placebo (10 in each group). They were monitored once a day (in the evening) to measure their resting SBP. The primary endpoint of the study was the time (days) to return to a normal blood pressure.

Data

The event times for the placebo and active groups are provided below ("C" indicates a censored observation):

Placebo:	3(C)	4	5	8	8	8	10(C)	10(C)	10(C)	10(C)
Active:	2	3	3	4	4	4	10(C)	10(C)	10(C)	10(C).

The unique times at which events occurred (not censored observations) are on days 2, 3, 4, 5, and 8. Table 11.11 represents the required contingency tables for the logrank test.

Table 11.11 Contingency table of treatment by event at each event time

Day 2: Normal SBP?	Active	Placebo	Total
Yes	1	0	1
No	9	10	19
	10	10	20
Day 3: Normal SBP?	**Active**	**Placebo**	**Total**
Yes	2	0	2
No	7	10	17
	9	10	19
Day 4: Normal SBP?	**Active**	**Placebo**	**Total**
Yes	3	1	4
No	4	8	12
	7	9	16
Day 5: Normal SBP?	**Active**	**Placebo**	**Total**
Yes	0	1	1
No	4	7	11
	4	8	12
Day 8: Normal SBP?	**Active**	**Placebo**	**Total**
Yes	0	3	3
No	4	4	8
	4	7	11

Note that on day 2 there were 10 participants at risk for the event in the active group. On day 2 one participant in the active group had the event of interest and is therefore removed from the number at risk at later time points. At day 3 there were nine remaining in the active group, two of whom experienced the event, leaving seven in the "risk set" for later times. On day 3 one placebo participant was censored, meaning that day 3 was the last known time at which the participant had not experienced the event. This person is removed from the risk set for later times. The tables are filled out in a similar manner for all times at which the events occurred. The important thing to remember with these contingency tables is that the number in each group decreases for later time points when the individual either had the event or was censored.

The test statistic can be computed by hand, but software is the ideal method, especially for more than a handful of event times. The numerator part of the test statistic would be calculated as:

$$\left[\frac{(10)(10)}{20}(0.10 - 0) + \frac{(9)(10)}{19}(0.22 - 0) + \ldots + \frac{(4)(7)}{11}(0 - 0.43)\right]^2$$
$$= 1.90.$$

The denominator would be calculated as:

$$\frac{(10)(10)}{19}(0.05)(0.95) + \frac{(9)(10)}{18}(0.11)(0.89) + \ldots + \frac{(4)(7)}{10}(0.27)(0.73)$$
$$= 2.286.$$

The test statistic is calculated as the ratio of the two:

$$\chi^2_{LR} = \frac{1.90}{2.286} = 0.831.$$

Interpretation and decision-making

As we saw in Table 10.5, the critical value for the test at an α level of 0.05 is 3.841. As the value of the test statistic $0.831 < 3.841$ there is not enough evidence to reject the null hypothesis.

Small studies such as this can be difficult to interpret. There is a suggestion that the times to response may be shorter with the active treatment, but the hypothesis test did not suggest that the variation seen was attributable to anything but chance given the sample size.

11.15 Review

1. In a therapeutic exploratory trial comparing a single dose of a new analgesic to placebo, 17 individuals were treated with the new analgesic (test treatment) and 15 were treated with the placebo (control). The participants reported the severity of their pain 6 hours after dental surgery using a visual analog scale (VAS). Pain scores on this scale range from 0 to 100, where 0 = "no pain" and 100 = "very severe pain." The mean (SD) pain score in the test treatment group ($n = 17$) was 18 (7). The mean (SD) score in the control group ($n = 15$) was 24 (8). Investigators would like to know if the mean VAS pain score is different between the two populations assumed to be represented by the two samples of study participants.
   - (a) What are the null and alternate hypotheses?
   - (b) Assume α = 0.05. What are the values of the rejection region?
   - (c) What assumptions are necessary for the use of the *t* test?
   - (d) What is the value of the test statistic?
   - (e) What is your interpretation of the hypothesis test?

2. This ANOVA table represents data from a study of an analgesic. The variable of interest is a pain score (higher values mean greater pain).

Source	SS	df	MS	*F*
drug	99.89459	2	*	*
Error	*	30	*	
Total	338.57355	32		

   - (a) Write in the missing values of the ANOVA table (denoted with *).
   - (b) In this study, how many treatments were tested?
   - (c) What are the null and alternate hypotheses?
   - (d) How many individuals were studied?
   - (e) What is the critical region for a test with α = 0.05?
   - (f) What is the statistical conclusion and interpretation of the hypothesis test?

3. Consider an ANOVA with four treatment groups (30 participants in each), placebo, and three doses of an investigational drug: Low, medium, and high.

(a) What are the null and alternate hypotheses?
(b) What assumptions must be made for the ANOVA?
(c) If the omnibus *F* test is significant, what are the pairwise comparisons that would be of interest?
(d) Why would Tukey's test be useful to evaluate the pairwise comparisons in (b)?
(e) Assume the mean square within-samples is 20. What is the value of the minimum significant difference – Tukey (MSD_T) that would determine whether pairs of treatments were significantly different?

4. In what situations would the Kruskal–Wallis test be appropriate?

11.16 References

Schork MA, Remington RD (2000). *Statistics with Applications to the Biological and Health Sciences*, 3rd edn. Upper Saddle River, NJ: Prentice-Hall.

Turner JR (2007). *New Drug Development: Design, methodology, and analysis*. Hoboken, NJ: John Wiley & Sons.

12

Additional statistical considerations in clinical trials

12.1 Introduction

The previous chapters have provided you with an introduction to statistical methods and analyses that are commonly used in pharmaceutical clinical trials, with an emphasis on therapeutic confirmatory trials. Although we certainly have not covered all of the analyses that can be conducted in these trials, those that we have discussed have given you a solid foundation that will also enable you to understand the basics of other analyses.

Throughout our discussions we have illustrated the importance of selecting the appropriate analytical strategy that best serves the objective of a given trial. There is hardly a single statistical method that always applies to a given study design or type of data: Rather, the choice of the analytical strategy for a given trial is the result of statistical considerations, clinical judgments, and regulatory standards.

In this chapter we highlight additional statistical considerations relevant to therapeutic confirmatory trials, and other study designs that also provide important information upon which to base decision-making. These additional insights and information build upon the material presented so far. As this chapter is largely conceptual rather than computational, we have included a number of references to guide your further reading.

12.2 Sample size estimation

An important part of study design is the "determination" of the required sample size. Before starting, we should note that we prefer the term "estimation" to the terms "determination" and "calculation" of a sample size. Although a mathematical calculation is certainly performed here, the values that are put into the appropriate formula are chosen by the researcher.

It is also appropriate to note that not all clinical trials utilize formal sample size estimation methods. In many instances (for example, FTIH studies) the sample size is determined on the basis of logistical constraints and the size of the study thought to be necessary to gather sufficient evidence (for example, pharmacokinetic profiles) to rule out unwanted effects. However, when the objective of the clinical trial (for example, a superiority trial) is to claim that a true treatment effect exists while at the same time limiting the probability of committing type I or II errors (α and β), there are computational methods used to estimate the required sample size. The use of formal sample size estimation is required in therapeutic confirmatory trials, this book's major focus, and strongly suggested in therapeutic exploratory trials.

12.2.1 Sample size for continuous outcomes in superiority trials

Consider the simple case of a superiority trial of an investigational drug (the test treatment) being compared with placebo with respect to a continuous outcome (for example, change from baseline SBP). The null hypothesis typically tested in such a trial and its complementary alternate hypothesis are:

$$H_0\text{: } \mu_{\text{TEST}} - \mu_{\text{PLACEBO}} = \Delta$$

$$H_A\text{: } \mu_{\text{TEST}} - \mu_{\text{PLACEBO}} \neq \Delta.$$

There are a number of values of the treatment effect (delta or Δ) that could lead to rejection of the null hypothesis of no difference between the two means. For purposes of estimating a sample size the power of the study (that is, the probability that the null hypothesis of no difference is rejected given that the alternate hypothesis is true) is calculated for a *specific* value of Δ. In the case of a superiority trial, this specific value represents the minimally clinically relevant difference between groups that, if found to be plausible on the basis of the sample data through construction of a confidence interval, would be viewed as evidence of a definitive and clinically important treatment effect.

Another way of stating this is that, if the true difference in population means is as large as a specific value of Δ proposed as clinically important, we would like to find the sample size such that the null hypothesis would be rejected $(1 - \beta)\%$ of the time. The sample size must also be chosen so that α is maintained at an acceptably low value.

The sample size formula required to test (two-sided) the equality of two means from random variables with normal distributions is:

$$n \text{ per group} = \frac{2\sigma^2(Z_{1-\alpha/2} + Z_\beta)^2}{\Delta^2}.$$

In this equation:

- n is the sample size per group
- σ^2 is the assumed variance
- $Z_{1-\alpha/2}$ is the value of the Z distribution that defines an area of size $\alpha/2$ in the upper tail of the Z distribution
- Z_β is the value of the Z distribution that defines an area of size β in the lower tail of the Z distribution
- Δ is the difference in means that we would like to detect, if it exists, by virtue of rejecting the null hypothesis.

Both α and β are design parameters, and are chosen at the discretion of those designing the trials. In confirmatory trials, α is 0.05 and β is typically 0.10 or 0.20 (meaning that the study has 90% or 80% power, respectively). The choices of σ and Δ are not quite as straightforward, because the range of possible values is outside the direct control of the study planner. The standard deviation σ must be estimated using (any) available data, and the value of the treatment effect Δ is determined using clinical judgment.

It is important to note that, all other things being equal, the following statements are true:

- The required sample size increases as the variance increases.
- The required sample size increases as the size of the treatment effect decreases.
- The required sample size increases as α decreases.
- The required sample size increases as the power $(1 - \beta)$ increases.

This sample size formula can be illustrated with the following example. Suppose that, in exploratory therapeutic trials of a new antihypertensive, the standard deviation for the between-treatment difference in mean change from baseline SBP was estimated to be 50 mmHg. After reviewing the literature and consulting with regulatory authorities, it is agreed that a between-treatment group difference in mean change from baseline (that is, the treatment effect) of at least 20 mmHg would be considered a clinically important benefit of a new drug to treat hypertension. The study sponsor is planning a confirmatory trial comparing the test drug with a placebo and would like to have an excellent chance (90%) of claiming that the treatment effect is not zero if the drug is as efficacious as they believe. From the expression above, the sample size required per group is:

$$n = \frac{2(50)^2(1.96 + 1.645)^2}{20^2}$$

= 133 per group for a total of 266 individuals.

This sample size estimate would be described in the study protocol in this manner:

> A total of 266 participants (133 per group) will be randomized in this study in a 1:1 ratio to test and placebo. Assuming a common standard deviation of 50 mmHg, this sample size will provide 90% power to detect a between-group difference in mean change from baseline of at least 20 mmHg using a two-sided test of size α = 0.05.

The power of the study is the probability of rejecting the null hypothesis of no difference in means, assuming that the true difference is at least 20 mmHg and the estimated variance is correct. As we have seen in this book, all estimates have sampling variation associated with them. Therefore, it can be helpful to see how the power to detect a difference of 20 mmHg varies as a function of sample size using three different values of the standard deviation. The impact of these two factors on the power can be seen in Figure 12.1, a graphical display called a power curve. Figure 12.1 is a compelling illustration of the importance of the assumed value of the standard deviation. Consider that, in the design of the study in our worked example, the assumed standard deviation of 50 mmHg led to a sample size of 133 per group for a power of 90% to detect the important difference of 20 mmHg. If the standard deviation was underestimated such that it was really 70 mmHg, the study would really only have 64% power to detect the difference that was considered important. Of course, this cannot be known in advance of a trial, but a post hoc examination of the study data, and a possible re-estimation of the standard deviation, can better inform future trials and increase the probability that they will be successful.

12.2.2 Sample size for binary outcomes in superiority trials

We have encountered a number of statistical methods used to test the difference between two population proportions. Suppose that we are interested in estimating the sample size for a superiority trial of an investigational drug (the test treatment), which will be compared with placebo with respect to a binary outcome, for example, proportion of individuals attaining a goal SBP. The null hypothesis and its complementary alternate hypothesis typically tested in such a trial are:

$$H_0\colon p_{\mathrm{TEST}} - p_{\mathrm{PLACEBO}} = 0.$$

$$H_A\colon p_{\mathrm{TEST}} - p_{\mathrm{PLACEBO}} \neq 0.$$

As in Chapter 10, the population proportions for each of two independent groups are represented by p_{TEST} and p_{PLACEBO}. Just as for the case of continuous outcomes, the power of the study is calculated for a *specific* value of $\Delta = p_{\mathrm{TEST}} - p_{\mathrm{PLACEBO}}$, a value that is considered the minimally clinically relevant difference (CRD).

The sample size formula required to test (two-sided) the equality of two population proportions used here is cited from Fleiss et al. (2003).

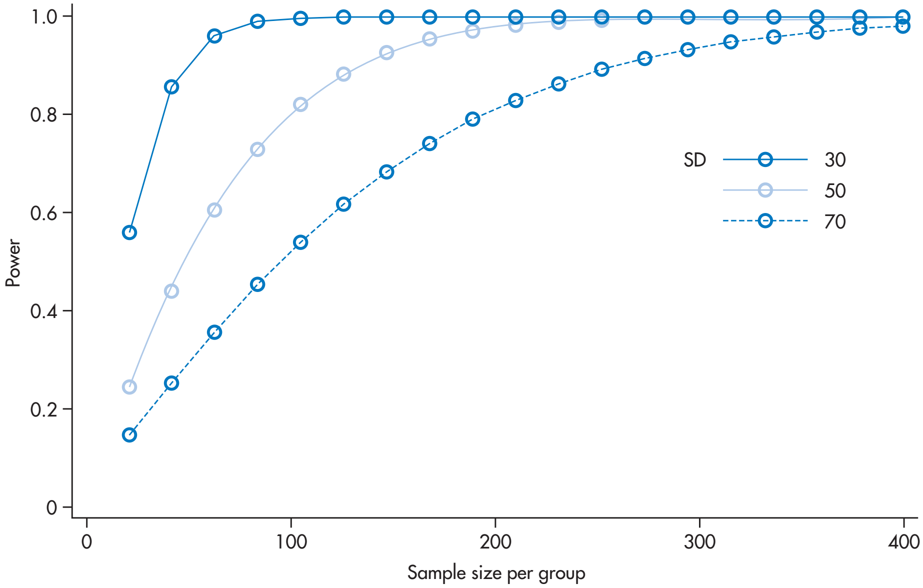

Figure 12.1 Power curve (Δ = 20 mmHg) as a function of sample size (n) for σ = 30 mmHg, 50 mmHg, and 70 mmHg

The calculation involves two parts. The first makes use of a normal approximation:

$$n' = \frac{(Z_{1-\alpha/2}\sqrt{2\bar{p}\bar{q}} + Z_{\beta}\sqrt{p_{\mathrm{TEST}}q_{\mathrm{TEST}} + p_{\mathrm{PLACEBO}}q_{\mathrm{PLACEBO}}})^2}{(\Delta)^2},$$

where:

$$\Delta = p_{\mathrm{TEST}} - p_{\mathrm{PLACEBO}},$$

$$\bar{p} = \frac{p_{\mathrm{TEST}} + p_{\mathrm{PLACEBO}},}{2},$$

$$\bar{q} = 1 - \bar{p},$$

$$q_{\mathrm{TEST}} = 1 - p_{\mathrm{TEST}},$$

and

$$q_{\mathrm{PLACEBO}} = 1 - p_{\mathrm{PLACEBO}}.$$

Note that the sample size depends not only on the value of Δ, but also on the individual proportions themselves. The implication of this is that the sponsor must make a reasonable estimate of the response in the placebo group (that is, p_{PLACEBO}) and then postulate a value of Δ that is clinically relevant. The corresponding value of p_{TEST} can be obtained by subtraction. This first sample size estimate (n') can be improved through the use of a continuity correction, which gives more accurate results when a discrete distribution (in this case the binomial distribution) is used to approximate a continuous distribution (in this case the normal). The sample size formula with continuity correction is:

$$n \text{ per group} = \frac{n'}{4}\left(1 + \sqrt{1 + \frac{4}{n'|\Delta|}}\right)^2.$$

In a confirmatory efficacy trial the study sponsor would like to evaluate a test treatment (an antihypertensive) versus placebo with respect to a binary outcome of attaining a goal SBP ≤ 140 mmHg. After reviewing several sources of data the sponsor estimates that the placebo response will be around 0.20 (that is, 20% of individuals will attain the goal without medical therapy). The sponsor would like to estimate the sample size required to detect a difference in response rates of 0.20 – that is, the postulated value of the response for test treatment is 0.40. As the study is a confirmatory trial, 90% power is recommended and the test will be a two-sided test with α = 0.05.

Substituting these values into the formula for the per-group sample size, we obtain:

$$n' = \frac{(1.96\sqrt{2(0.30)(0.70)} + 1.645\sqrt{(0.40)(0.60) + (0.20)(0.80)})^2}{(0.20)^2} = 306.$$

With a continuity correction the result is:

$$n = \frac{306}{4}\left(1 + \sqrt{1 + \frac{4}{306(0.20)}}\right)^2 = 316 \text{ individuals per group.}$$

This sample size estimate would be described in the study protocol in this manner:

> A total of 632 individuals (316 per group) will be randomized in this study in a 1:1 ratio to the test treatment and the placebo treatment. Assuming a placebo response rate of 20%, this sample size will provide 90% power to detect a between-group difference in response rates of 20% using a two-sided test with α = 0.05.

As for continuous data, a power curve can be generated for a number of scenarios for binary outcomes. As seen in Figure 12.2, the power of a test of proportions (for a fixed value of Δ) is quite sensitive to the particular assumed value of the response rate in the control (for example, placebo) group.

12.2.3 α and β reconsidered using Bayes' theorem

After a study has been completed, a statistical analysis provides a means either to reject or to fail to reject the null hypothesis. The statistical conclusion will, in part, be used to justify whether or not further investment is made in the development of a test product. A sound business strategy would dictate that further investment be made only if objective information from the study suggests it. Inferential statistics

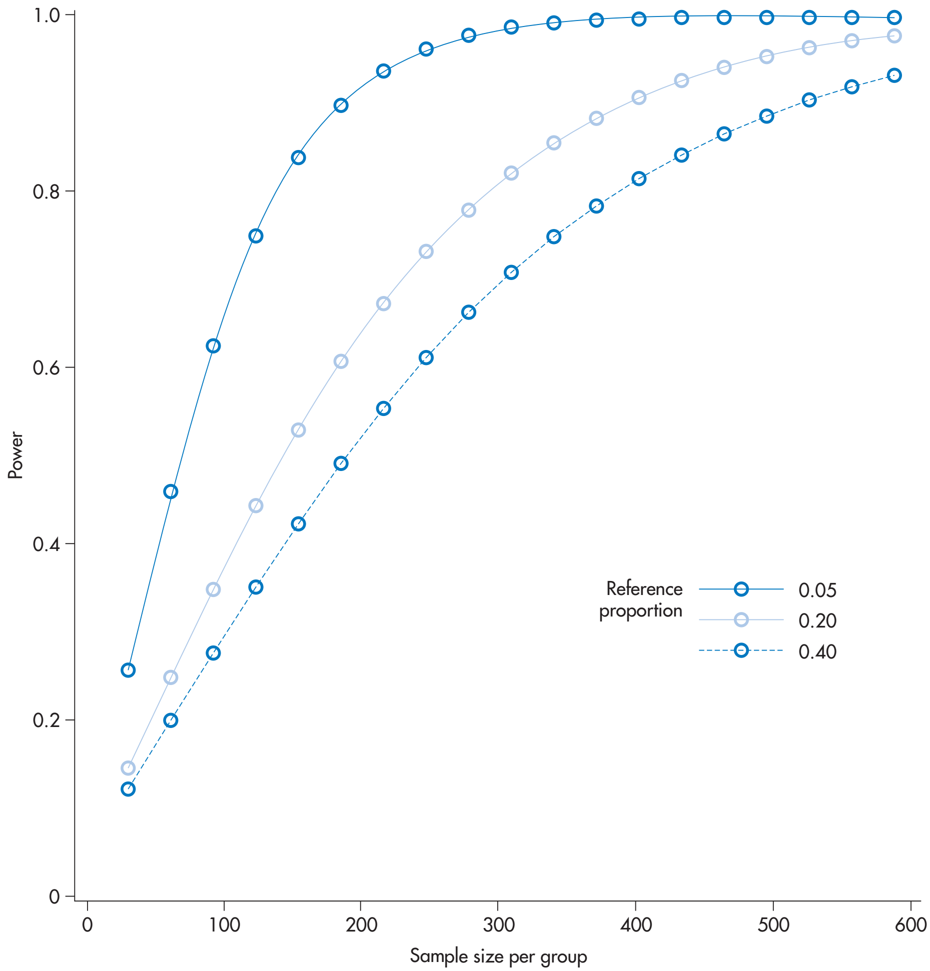

Figure 12.2 Power curve ($\Delta = 0.10$) as a function of sample size (n) for $p_{\text{TEST}} = 0.05$, $p_{\text{TEST}} = 0.20$, and $p_{\text{TEST}} = 0.40$

(for example, a hypothesis test) is the appropriate means to differentiate a real effect from a chance effect. For the remainder of this section we investigate the wisdom of adopting such a policy, using an approach similar to that described by Lee and Zelen (2000). In particular, the remainder of this section will address the following two questions:

1. How likely is the sponsor to be misled by the result of the statistical conclusion from a hypothesis test with design parameters α and β?
2. What is the role of accumulating evidence about the true treatment effect on the credibility of results from a hypothesis test?

Throughout this book we have emphasized the role of Statistics in designing and analyzing studies that enable sponsors to make decisions about the future development of new drugs. When developing a new drug, information is accumulated over time, with each step informing the next. As studies are completed through various stages of clinical development (FTIH studies, therapeutic exploratory studies, and one or more therapeutic confirmatory studies) evidence is gathered that supports the efficacy of the new drug. This is true only if new studies are planned because such a promise exists. Hence we assume over time, with the accumulation of new information, that various scientists involved in the development program *could* make an informed guess about the probability that the drug works (that is, that the alternate hypotheses considered in Sections 12.2.1 and 12.2.2 really represent the truth). Let us call this probability, τ (tau).

$$\tau = \text{probability that } \Delta \neq 0.$$

However, as additional studies would be conducted only if the accumulating evidence suggested that there was a treatment benefit, τ represents the probability that the treatment is truly effective. We can think of $\tau = 0$ as representing a molecule that has just been discovered, for which no evidence has been generated about its ultimate effect on a clinical outcome of interest. At the other extreme a value of $\tau = 0.8$ represents a drug for which a great deal of information has been collected and most of the data support a beneficial effect of the treatment. Values of τ around 0.5 may represent a drug for which some (or limited) data support a treatment benefit.

Consider the following probabilities, which express the likelihood of the true state of affairs given the statistical conclusion at the end of the study:

$$\alpha^* = P(\text{Null is true}|\text{Reject null})$$

$$\beta^* = P(\text{Alternate is true}|\text{Fail to reject the null}).$$

The value α* is the probability that a rejected null hypothesis (for example, p value $\leq \alpha$) is misleading. That is, it represents the chance that, having rejected the null hypothesis of no effect, the treatment is not efficacious. Its complement, $(1 - \alpha^*)$, is the probability that the rejected null hypothesis is consistent with the truth (that is, the treatment is efficacious).

Similarly, the value β* is the probability that failure to reject the null hypothesis (for example, p value $> \alpha$) is misleading. It represents the chance that, having failed to reject the null hypothesis of no effect (and acting as if the null is true), the treatment really is efficacious. The complement, $(1 - \beta^*)$, is the probability that our inability to reject the null hypothesis is consistent with the truth (that is, the treatment is not efficacious).

If we are to adopt a policy of using inferential statistics to make decisions in the light of uncertainty, we would like to minimize these probabilities, α* and β*, as they directly lead to wasted investment in the former case or a lost commercial opportunity in the latter.

Using Bayes' theorem (recall our discussions in Chapter 6), the probability,

$$\alpha^* = P(\text{Null is true}|\text{Reject null}),$$

can be written as:

$$\alpha^* = P(\text{Reject null} \mid \text{Null is true})\,\frac{P(\text{Null is true})}{P(\text{Reject null})}.$$

Recall also from Chapter 6 that the marginal probability of an event can be expressed as a series of conditional probabilities as long as the conditional events are mutually exclusive and exhaustive. This allows us to express the probability,

$$P(\text{Reject null}) = P(\text{Reject null}) \mid \text{Null is true})\,P(\text{Null is true}) + P(\text{Reject null}) \mid \text{Alternative is true})P(\text{Alternative is true}).$$

Finally, putting this entire expression together we have:

$$\alpha^* = \frac{P(\text{Reject null}) \mid \text{Null is true})P(\text{Null is true})}{P(\text{Reject null}) \mid \text{Null is true})P(\text{Null is true}) + P(\text{Reject null}) \mid \text{Alternative is true})P(\text{Alternative is true})}.$$

This probability can then be expressed as a function of the design parameters, α and β, and the estimated probability that the alternative is true, τ:

$$\alpha^* = \frac{\alpha(1-\tau)}{\alpha(1-\tau)+(1-\beta)\tau}.$$

Bayes' theorem and algebra can be used in a similar fashion to obtain the following expression for β^*:

$$\beta^* = \frac{\beta\tau}{\beta\tau+(1-\alpha)(1-\tau)}.$$

We can use these two expressions to answer the questions posed at the beginning of this section. The first of these is:

> How likely is the sponsor to be misled by the result of the statistical conclusion from a hypothesis test?

The short answer is that it depends on the power (and therefore the sample size) of the study. To illustrate this, assume that the value of the design parameter α is dictated by regulatory concerns, which is reasonable especially in confirmatory trials. Further, before a new study is completed there is still some doubt as to whether the new treatment is efficacious, such that the value of τ is conjectured to be 0.5. Resulting values of the error rates, α^* and β^*, are presented in Table 12.1 as a function of the power (or, equivalently, β) of the study.

The key message from Table 12.1 is that the probability of both errors decreases with increases in statistical power. A study planned with power of 0.5 and a statistical decision to reject the null (for example, because p value $\leq$ 0.05) yields a probability of 0.09 that the two treatments are not significantly different. In contrast, a study with power 0.9 and the same outcome (to reject the null) yields a probability of 0.05 that the two treatments are really significantly different. Even though the statistical test has indicated that further investment should be considered because the test treatment appears to be efficacious, the underpowered study leads to an unwise decision 1.8 times (0.09/0.05) more often than the conventionally powered study. Similar statements can be made about unwisely abandoning an efficacious product by examining the values of β^*. Another way of stating this is that the greater the statistical power for a study, the more reliable the decisions made as a result.

Now consider the second question:

> What is the role of accumulating evidence about the true treatment effect on the credibility of results from a hypothesis test?

To address this question, the error rates, α^* and β^*, are presented in Table 12.2 as a function of τ (a measure of the likelihood the treatment is efficacious) with power 0.9 and $\alpha = 0.05$, typical values for highly powered studies. An examination of a couple of cases will help to answer this question.

When $\tau = 0$ there is a great deal of uncertainty about the probability that the treatment is efficacious. This situation may apply when there is no experience with the test treatment or some experience with mixed or poor results. When a statistically significant result leading to rejection of the null hypothesis has been observed in this situation, the sponsor will be misled into thinking that the drug is effective when it really is not with probability 0.33. On the other hand,

Table 12.1 Error rates α^* and β^* as a function of β ($\alpha = 0.05$ and $\tau = 0.5$)

β	$1-\beta$ (power)	α^*	$1-\alpha^*$	β^*	$1-\beta^*$
0.5	0.5	0.09	0.91	0.34	0.66
0.4	0.6	0.08	0.92	0.30	0.70
0.3	0.7	0.07	0.93	0.24	0.76
0.2	0.8	0.06	0.94	0.17	0.83
0.1	0.9	0.05	0.95	0.10	0.90

Table 12.2 Error rates α^* and β^* as a function of τ ($\alpha = 0.05$ and $1 - \beta = 0.9$)

τ	α^*	$1 - \alpha^*$	β^*	$1 - \beta^*$
0.1	0.33	0.67	0.01	0.99
0.3	0.11	0.89	0.04	0.96
0.5	0.05	0.95	0.10	0.90
0.7	0.02	0.98	0.20	0.80
0.9	0.01	0.99	0.49	0.51

failure to reject the null hypothesis in this situation will mislead the sponsor who abandons development with probability 0.01.

Once some studies have been completed and evidence has been gathered to support the efficacy of the new treatment, the value of τ may be around 0.5. This value represents at least moderate evidence that the treatment is truly efficacious. When a statistically significant result leading to rejection of the null hypothesis has been observed in this situation, the sponsor will be misled into thinking that the drug is effective when it really is not, with probability only 0.05. This reflects previous experience, which has shown that the treatment provides a benefit. Failure to reject the null hypothesis will mislead the sponsor who then abandons development as a result, with probability 0.10. Again, this probability reflects previous experience because the new evidence contradicts the prior belief that the treatment is efficacious so that acting on the new study result may be misleading.

The case where $\tau = 0.9$ represents nearly certain knowledge. It is hard to understand why any additional data would be required in this instance. However, an examination of the error rates α^* and β^* in this situation is illuminating. Rejection of the null hypothesis would come as no surprise so that such a result would rarely be misleading. Failure to reject the null hypothesis would come as a surprise because it is almost known with certainty that the null is false. Thus, this information is too bad to be true and acting on it is unwise.

The probability τ is analogous to the underlying prevalence of disease in a population. In the setting of diagnostic testing, α^* and β^* refer to the positive and negative predictive values of a test. As illustrated in Chapter 6, when evaluating a diagnostic test, even high values of sensitivity and specificity can lead to skepticism about a positive test when the prevalence of the underlying disease is low.

In a similar manner, τ should serve to temper the enthusiasm of study sponsors who have observed a new positive study result, especially early in development programs. It can be used to calibrate the credibility of statistical results. Without sufficient prior information about the treatment even a statistically significant result can lead to poor (and expensive) business decisions. When a sponsor desires either to continue or to discontinue development of a new drug as a result of a study, the results in this section point to the importance of power. Despite their other benefits, exploratory therapeutic trials, which tend to be small in size (and therefore have low power), are poor studies on which to make business decisions. Small, early clinical studies may provide some evidence on which to base future research, including τ. However, once that is done, there is no substitute for definitive, highly powered studies in appropriate populations, using acceptable clinically relevant endpoints. In short, power, a statistical design parameter, has a **direct** bearing on the quality of decision-making. We believe that recognition of this relationship is very much underappreciated, and that it has a profound bearing on the way sound business decisions should be made.

12.2.4 Importance of collaboration in sample size estimation

Sample size estimation requires the input of a number of specialists involved with the development of new drugs. The estimate of the standard deviation can be informed by exploratory therapeutic trials of the same drug or by literature reviews of similar drugs. Synthesis of these data from a number of sources requires statistical and clinical judgments. As was seen in Figure 12.1 the estimate of the standard deviation has an important effect on the sample size. Study teams should understand the sources of variability in the response variable and attempt to minimize unwanted variability.

The definition of the minimally clinically relevant difference of interest involves clinical, medical, and regulatory experience and judgments. The appropriate sample size formula depends on the test of interest and should take into account the need for multiple comparisons (either among treatments or with respect to multiple examinations of the data). The project statistician provides critical guidance in this area.

It is appropriate to note here that in some instances the sample size may not be completely dictated by the statistical requirements for a given power calculation. The ICH has published a guidance document (ICH Guidance E1, 1994) applicable to drugs given chronically. This guidance specifies the minimum number of individuals who should be exposed for certain periods of time so that potential adverse events (AEs) may come to light before the drug is marketed. The need for a larger safety database may supersede the sample size required to demonstrate a statistically significant and clinically relevant treatment effect.

In summary, sample size estimation requires the input of a number of disciplines involved in the design of clinical trials.

12.3 Multicenter studies

A certain number of participants need to be recruited for any given trial. In Section 12.2 we discussed sample size estimation, which takes into account a number of considerations that are important not only to the statistician but also to the clinical scientist and the regulator. Once determined, the value produced by this process of estimation is incorporated into the study protocol.

We have seen that relatively small numbers of participants are recruited for early phase trials (perhaps 20–80 in FTIH studies and 200–300 in early Phase II studies), and relatively larger numbers are recruited for therapeutic confirmatory trials (perhaps 3000–5000). It is relatively easy to recruit between 20 and 80 participants at a single investigational site. Indeed, as we noted in Section 7.3, conducting a FTIH study at a single center enhances consistency with respect to management of participants, study conduct, and assessment of AEs, and provides for frequent and careful monitoring of study participants. However, it is not feasible to recruit 3000–5000 participants at a single investigational site, so multicenter studies are typical at this stage of clinical development programs.

As for so many of the topics that we have discussed, multicenter studies have both advantages and disadvantages. Let us consider the disadvantages first and then focus on the advantages. The major disadvantage relates to the logistic demands of coordinating a multicenter trial. The rarer the medical condition of interest in the trial, the more sites that will probably be needed because fewer individuals will likely be available at each site. It is not unusual to have between 50 and 100 investigational sites participating in some trials. These sites may be scattered across a country and, increasingly, they may be scattered across several countries and continents. This occurrence has many consequences, including:

- All investigational sites must obtain approval from their investigational review board (IRB) to conduct their portion of the trial.
- The drug products used in the trial must be shipped to all sites, which may entail dealing with customs and import/export controls.
- If some sites speak different languages, all relevant issues must be addressed (for example, translating the informed consent form into each language).
- All principal investigators (one from each site) and certain members of their staff must receive training that will attempt to ensure consistency of all methodology used in the trial. Investigator meetings are held accordingly.
- Multicenter studies benefit from (rely on) the use of central labs to analyze certain samples taken during the trial (for example, blood samples). This is a complex shipping problem, especially when samples must be transported to the central laboratory under certain conditions and very quickly from distant locations.

Each of the above considerations adds considerably to the total cost of a multicenter trial.

From a statistical point of view, while every attempt is made to standardize the implementation of study methodology at all sites, perfect standardization is not a realistic expectation. Although simpler is usually better, study protocols often become complex during their development, and different investigational sites will likely differ in their implementation of some procedural aspects. This occurrence introduces extraneous variability into data collected and analyzed. Extensions of some of the analysis methods described in this book can be used to account for center-to-center variability, including multi-way ANOVA models for continuous variables, stratified χ^2 tests for categorical variables, and stratified log-rank tests for time-to-event analyses. Various other methodological controls can be introduced in an attempt to minimize such extraneous variability, but the success of any control strategy is unlikely to be perfect.

For these and other reasons, we believe that, if it were possible to conduct a trial requiring 3000–5000 participants at a single investigational site, sponsors would do so, even though this statement is at odds with the commonly cited major advantage of multicenter trials, which is that they enable greater generalization of results obtained from the trial. It is statistically possible to assess the treatment effect at each investigational site as well as assessing it using the data from all sites, although a given site needs to have reasonable enrollment for the treatment effect calculated from its participants to be meaningful. If similar treatment effects are observed at sites that tended to enroll relatively older individuals, relatively younger ones, ethnically homogeneous samples, ethnically heterogeneous samples, with less or more experience of treating the designated indication, and so forth, it is reasonable to have a certain degree of faith that the treatment effect is generalizable to the eventual patient population if and when the drug is approved for marketing.

12.4 Analysis populations

Various analysis populations for clinical trial data can be defined and are used in statistical analyses, including:

- The intent-to-treat (ITT) population: This comprises all participants in a clinical trial who were randomized to a treatment group, regardless of whether any data were actually collected from them.
- The safety population: This is a subset of the ITT population, defined as the population of participants who received at least one dose of a study treatment.
- The per-protocol population (also known as the efficacy or evaluable population): This is also a subset of the ITT population, and comprises individuals whose participation and involvement in the trial were considered to comply with significant requirements and activities detailed in the study protocol. Participants would typically be excluded from the per-protocol population if they exhibited poor dosing compliance, missed a number of clinic visits, or used prohibited medications that may interfere with the evaluation of the test treatment.

Both the ITT and the safety populations can be used in the analysis of safety data. The ITT and per-protocol population are typically used in the analysis of efficacy data.

12.4.1 Using both ITT and per-protocol populations in efficacy analyses

In therapeutic confirmatory trial efficacy, the same analyses are typically conducted twice, using data from the ITT population and data from the per-protocol population (see Turner, 2007). The analyses conducted using the ITT population are considered to be the primary analyses because ITT analysis provides a conservative strategy in the sense that it tends to bias against finding the results that the researcher "hopes" for, particularly in the case of superiority trials. The conservative nature of ITT analysis is deemed particularly appropriate when attempting to demonstrate the efficacy of an investigational drug because these data do not favor the desired outcome. Then, if there is compelling evidence of the drug's efficacy, this evidence will be particularly noteworthy. The ITT population is the most appropriate sample population from which to make inferences to the population of patients who may receive the drug if and when it receives marketing approval.

Having conducted primary analyses using the ITT population it is then appropriate to conduct secondary analyses using the per-protocol population, the subset of participants whose participation in the trial was compliant with the study protocol. This analysis is regarded as less conservative than ITT analysis because analysis of the per-protocol population may maximize the opportunity to demonstrate efficacy: The per-protocol population is the population in which the treatment is likely to perform best.

Regulatory authorities are encouraged if the results from the ITT efficacy analysis and the per-protocol efficacy analysis are similar, and their overall confidence in the trial results is increased. However, if they are not similar, questions may be raised as to why they are not. Some of these questions are (Turner, 2007):

- Is the per-protocol population a lot smaller than the ITT population (it will almost certainly be somewhat smaller)?
- If so, were there a lot of major protocol violations?
- Were a lot of participants removed for the same protocol violation?
- Were many of the participants with protocol violations enrolled at the same investigative site?
- Are there any systematic problems in the conduct of the trial?

All of the issues addressed by these questions can reduce the regulatory reviewers' overall confidence in the trial's findings.

12.4.2 Proper and improper subgroup analysis

Investigators may be interested to examine potential differences among groups of participants according to some characteristic. For example, there may be differences in the response to treatment according to age. An analysis to investigate such a phenomenon could involve separate analyses for participants aged 18–34, 35–54, 55–74, and 75 years and older. Similar analyses could be presented in which participants are grouped according to some measure of disease severity. Results such as these should be interpreted with caution because the more subgroups that are examined the greater the chance of discovering a false positive (recall our earlier discussions of multiple comparisons).

Although we have not discussed this topic, differences in treatment effects may be tested to see if they are homogeneous across the various subgroups. This test is called a test of the treatment-by-subgroup (for example, treatment by age) interaction. It is useful because it can rule out, using a hypothesis test, apparent differences among subgroups of subgroups that really represent random variation. Citing from the ICH Guidance E9 (1994, p 27):

> The treatment effect itself may also vary with subgroup or covariate – for example, the effect may decrease with age or may be larger in a particular diagnostic category of subjects. In some cases such interactions are anticipated or are of particular prior interest (e.g. geriatrics), and hence a subgroup analysis, or a statistical model including interactions, is part of the planned confirmatory analysis. In most cases, however, subgroup or interaction analyses are exploratory and should be clearly identified as such; they should explore the uniformity of any treatment effects found overall. In general, such analyses should proceed first through the addition of interaction terms to the statistical model in question, complemented by additional exploratory analysis within relevant subgroups of subjects, or within strata defined by the covariates. When exploratory, these analyses

should be interpreted cautiously; any conclusion of treatment efficacy (or lack thereof) or safety based solely on exploratory subgroup analyses are unlikely to be accepted.

These cautions having been noted, some unexpected subgroup findings may actually reveal important findings that should be further investigated in additional studies. This can be especially important when there is evidence of different safety profiles among subgroups. Matthews (2006) distinguished between two sorts of subgroup formation, and hence analysis:

- a limited number of subgroups identified *a priori* with an apparent biological/clinical reason for the difference of interest
- subgroups whose apparent importance is retrospective, and arises only as a result of doing analyses.

If the treatment effect appears to differ across subgroups identified in the first way, the phenomenon "should be taken much more seriously" than if the subgroups came to light via the second process (Matthews, 2006, p 171).

12.5 Dealing with missing data

For various reasons there are often participants in a trial for whom a complete set of data is not collected. This is the province of missing data. When conducting efficacy analyses we need to address this issue, and the way(s) in which it is addressed can influence the regulatory reviewers' interpretation of the analyses presented. The issue of missing data is problematic in clinical research because humans have complex lives. Human participants may choose to leave a study early or be unable to attend a specific visit, both situations leading to missing data. Nonclinical research involves tighter experimental control in which the subjects (animals) do not have the ability voluntarily to leave the study early.

Piantadosi (2005) observed that there are only three generic analytic approaches to addressing the issue of missing data:

1. Disregard the observations that contain a missing value.
2. Disregard a particular variable if it has a high frequency of missing values.
3. Replace the missing values by some appropriate value.

The last of these approaches is called imputation of missing values. As Piantadosi (2005, p 400) commented, although this approach sounds a lot like "making up data," when done properly it may be the most sensible strategy. While techniques for addressing missing data can be technically difficult, one commonly used, simple imputation method is called last observation carried forward (LOCF). In a study with repeated measurements over time, the most recent observation replaces any subsequent missing observations (Piantadosi, 2005).

An assumption of such an imputation strategy is that the future course of the individual's condition can reasonably be predicted by the last known state. If participants in the test group drop out of the study more often than those on placebo because the test treatment has failed, such an assumption may not be realistic. It is possible that participants who dropped out for treatment failure actually got worse than when they left the study. A commonly proposed strategy is to use a number of imputation methods and see how the analysis results change as a result. If the results of this sensitivity analysis suggest that the overall conclusion remains the same, it is less important how the missing data are managed.

Differential rates of loss to follow-up among groups or high rates in any single group complicate the management of missing data. Strategies that minimize the chance that participants will leave a study prematurely should be considered at the design and protocol writing stage, and incorporated in the protocol as appropriate.

A number of approaches to dealing with missing data are described by Molenberghs and Kenward (2007).

12.5.1 The importance of study conduct and study monitoring

While there are widely accepted methodologies for dealing with missing data, it is certainly preferable to have as many "actual" data as possible. This simple point underscores the critical nature of study conduct. All procedures detailed in the study protocol need to be followed, and all data required need to be collected to the greatest degree possible (there will always be occasional genuine reasons why this was not possible in a specific situation).

This need for as complete a dataset as possible underscores the importance of the clinical monitor. Two related and critical responsibilities of clinical monitors are to ensure that all sites in the trial follow the study protocol, and to check that the required data are recorded as and when they should be. A good monitor will spot the absence of recorded data sooner rather than later, which considerably increases the likelihood of locating and subsequently recording those data.

12.6 Primary and secondary objectives and endpoints

A given trial is conducted to collect optimum quality data with which to answer an identified and important research question. The data collected are intended to provide the most accurate answers to the research questions posed. A study protocol will often include both primary and secondary objectives, and also the associated primary and secondary endpoints.

12.6.1 The primary objective and endpoint

Turner (2007) noted that, in a very real sense, all the clinical studies that are conducted before a therapeutic confirmatory trial is undertaken have one purpose: To allow the primary objective in the therapeutic confirmatory trial to be stated as simply as possible. An objective that can be stated simply can be tested simply, that is, in a straightforward and unambiguous manner. This is a highly desirable attribute in a primary objective.

By the time a therapeutic confirmatory trial is appropriate it should be possible to state a single primary objective (or perhaps two if the sponsor really feels that this is appropriate) that is clinically relevant and biologically plausible. Having stated this primary objective, deciding upon the primary endpoint should be straightforward. Deciding on the appropriate study design and the associated statistical analyses should also be straightforward. Throughout this book we have focused on the development of a new antihypertensive drug. The primary objective of a therapeutic confirmatory trial in this therapeutic area will be to determine if the investigational drug does indeed lower blood pressure, and the associated endpoint(s) may be a certain magnitude reduction in systolic blood pressure (SBP), diastolic BP (DBP), or both.

At the analysis stage of the trial this endpoint provides the focus for rigorous statistical analysis and interpretation (Machin and Campbell, 2005). Formal hypothesis testing will be employed to determine the presence or absence of a statistically significant difference between the mean decrease seen in the drug treatment group and that seen in the control group. In addition, the clinical significance of the treatment effect will be addressed.

Having a single primary objective has an additional advantage in a study. It means that sample-size estimation can be based on that objective and the associated estimated treatment effect of interest (recall our discussion of sample-size estimation in Section 12.2). Having multiple primary endpoints requires adjustments for multiplicity and can be difficult to interpret if only one of multiple primary endpoints is found to have a statistically significant effect.

12.6.2 Secondary objectives and endpoints

In addition to the primary objective, a study may have a small number of secondary objectives. A secondary endpoint will be associated with each secondary objective. For example, assessments of quality of life may fall under the category of secondary objectives. (In some studies quality of life may be the primary objective: It is simply used here as a realistic example.) Quality of life (QoL) is an extremely important consideration, and particularly so in long-term pharmaceutical therapy. Even if a disease or condition cannot be cured, keeping the symptomatology at acceptable levels can be considered a tremendous success.

Formal hypothesis testing is less likely to occur for secondary endpoints. Descriptive statistics are more likely to be presented. It is also possible that findings of particular interest may lead to a primary objective in a subsequent trial. That is, these data are more suited to hypothesis formation than hypothesis testing. It is important to emphasize here that data leading to the formation of a hypothesis cannot be used to test that hypothesis: As just noted, a new dataset must be generated.

12.6.3 How many objectives should we list?

The number of objectives that should be incorporated in any clinical trial is often a topic of considerable debate among study teams (Turner, 2007). Some members will likely argue that, while taking all the trouble to conduct the trial, why not collect as much data as possible and ask as many questions as possible? This approach leads to a large number of study objectives, sometimes broken down into primary objectives, secondary objectives, and even tertiary objectives. It is certainly legitimate in some studies to be interested in more than one primary endpoint and possibly in several secondary endpoints. However, from a statistical point of view, increasing the number of objectives leads to serious problems, and it can compromise the weight of any particular piece of evidence that is eventually presented to regulatory agencies.

In Chapter 11 we discussed the issue of multiple comparisons and multiplicity in the context of pairwise treatment comparisons following a significant omnibus *F* test. When we adopt the 5% significance level ($\alpha = 0.05$), by definition it is likely that a type I error will occur when 20 separate comparisons are made. That is, a statistically significant result will be "found" by chance alone. The greater the number of objectives presented in a study protocol, the greater the number of comparisons that will be made at the analysis stage, and the greater the chance of a type I error. Machin and Campbell (2005) commented: "If there are too many endpoints defined, the multiplicity of comparisons then made at the analysis stage may result in spurious statistical significance."

The concern of multiplicity can also apply to studies in which data are examined during the study at interim time points. Interim analyses are discussed in Section 12.9.

12.7 Evaluating baseline characteristics

It is common practice in analyses of clinical data to inspect the distributions of baseline characteristics – for example, demographics and measures of disease severity – through the use of descriptive summary statistics. This is an important analysis because it helps to describe the sample representing the target population of interest. If the sample is representative of the target population the inferences drawn from the study will be considered relevant.

Sometimes the baseline homogeneity of these characteristics is assessed using a hypothesis test, for example, an omnibus *F*-test from a one-way ANOVA testing for differences in age. If a "significant" result is found, some researchers might offer this as evidence that something went awry with the randomization process. However, this view has two problems: One is that multiple hypothesis tests can lead to spurious findings or "false positives;" the second is that, on any given single instance, a proper randomization cannot ensure that this possibility does not occur. What randomization can ensure is that, on average, over all possible randomizations, distributions of baseline characteristics will be homogeneous across groups. This result is all that is required for proper statistical inferences. Senn (1997) emphasized that "inferential statistics calculated from a clinical trial make an allowance for differences between patients and that this allowance will be correct on average if randomization has been employed." It is worth noting that standard errors represent such allowances.

When there is evidence to suggest a baseline imbalance with respect to a characteristic that may influence an important outcome of the study, such as the primary efficacy endpoint, some investigators choose to examine the effect of this factor in additional statistical analyses. Possible approaches to this would include ANOVA or analysis of covariance (ANCOVA) in which a continuous variable (for example, age) is adjusted for in assessing the main effect of interest, that is, the treatment effect for the primary outcome variable.

Such a step is not required from a statistical point of view, as a result of the role of a properly executed randomization process, but it can be comforting if it supports the clinical relevance of the effect after adjustment for the baseline covariate. If there are specific explanatory factors that are suspected of having an effect on the outcome of interest at the start of a study, it is advisable to incorporate them into the overall study design (for example, through stratified randomization). A brief discussion of this topic has been published by Roberts and Torgerson (1999). The EMEA CPMP has also published a guidance document on baseline covariates (EMEA CPMP 2003).

12.8 Equivalence and noninferiority study designs

The goal of equivalence trials is to demonstrate that a new (test) drug (T) and an active comparator drug (C) are "equivalent" or have a similar effect. This means that, in the best-case scenario, the test treatment is trivially better than the reference treatment and, in the worst, it is tolerably worse.

Equivalence trials are important when it would be unethical to compare the test treatment with an inactive control, and when comparing the test with the control for equivalent efficacy with a superior safety profile for the test drug. The difference between groups that we believe to be "trivially better" or "tolerably worse" is called the equivalence margin. Defining the equivalence margin is not an easy task and requires input from regulatory authorities. The definition of the equivalence margin is required in estimating a sample size for such a study and it must be decided upon in advance of the study and detailed in the study protocol.

Noninferiority trials are very similar to equivalence trials in the manner of their statistical approach. In noninferiority trials the objective is to demonstrate that the test drug is no worse than – that is, not inferior to – the control. Assuming that the test drug had some other benefit, such as better tolerability or safety or cost, a claim of noninferiority could mean that the effect for the test drug is trivially worse than the control. The design, including the choice of the noninferiority margin, must be agreed to with regulatory authorities and provided in the study protocol. A guidance document published by the EMEA CPMP (2000) addresses issues related to interpreting data from superiority studies for noninferiority claims, although it is our opinion that such a practice is rarely justified.

12.8.1 Why the hypothesis-testing strategies are different in these designs

The research questions in equivalence and noninferiority trials are different from those used in superiority trials. Hypothesis testing strategies that are so frequently used in superiority trials do not serve the needs of these designs well. As Matthews (2006, p 199) commented: "Failing to establish that one treatment is superior to the other is not the same as establishing their equivalence." In other words, obtaining a nonsignificant p value in a superiority trial does not demonstrate that the two treatments are the same. As we shall see, conventional p values have no role in establishing equivalence or noninferiority.

12.8.2 Use of confidence intervals for inferences

Given that the research questions in these trials are different from those used in superiority trials, the formats of the null and alternate hypotheses are also different. The research question associated with an equivalence trial is: Does the test drug demonstrate equivalent efficacy compared with the comparator drug? The null hypothesis, stated in terms of differences in population means, is:

$$H_0: |\mu_{TEST} - \mu_{CONTROL}| \geq \delta_{equivalence}.$$

The alternate hypothesis is:

$$H_A: |\mu_{TEST} - \mu_{CONTROL}| < \delta_{equivalence}.$$

If the null hypothesis is rejected in this case, the conclusion would be that the two population means were within $\delta_{equivalence}$ units of each other. The equivalence margin would be selected such that the two treatments were considered equivalent. If two antihypertensive therapies were compared in this manner, an equivalence margin might be 5 mmHg (a trivial difference). The inferential statistical analysis for equivalence trials typically involves the calculation of a (1 −α)% confidence interval for the difference in population means. If the lower and upper bounds of the confidence interval are both within the equivalence margin, the conclusion is that we are (1 −α)% confident that the true difference in population means does not exceed $\delta_{equivalence}$. The conclusions that can be drawn from an equivalence trial are displayed in Figure 12.3.

The research question for a noninferiority trial is stated as: Is the test drug not inferior to the control? The null hypothesis to be tested in this study is:

$$H_0: \mu_{TEST} - \mu_{CONTROL} \geq \delta_{noninferiority}.$$

If the null hypothesis is rejected, the following alternate hypothesis will be favored:

$$H_A: \mu_{TEST} - \mu_{CONTROL} < \delta_{noninferiority}.$$

If the null hypothesis is rejected in this case, the conclusion would be that the population mean for the control treatment did not exceed that for the test group by more than $\delta_{noninferiority}$. The inferential statistical analysis for noninferiority trials typically involves the calculation of a one-sided (1 −α)% confidence interval for the difference in population means. If the upper bound of the confidence interval is within the noninferiority margin, the conclusion is that we are (1 −α)% confident that the true difference in population means is less than $\delta_{noninferiority}$. The conclusions that can be drawn from a noninferiority trial are displayed in Figure 12.4.

Equivalence and noninferiority trials may be the only viable means to test a new drug in

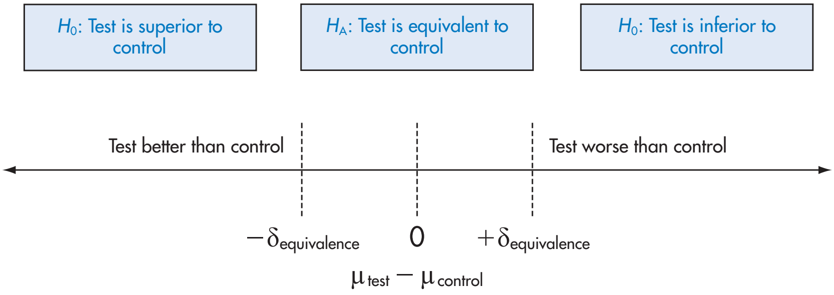
H_0: Test is superior to control
H_A: Test is equivalent to control
H_0: Test is inferior to control
Test better than control
Test worse than control
$-\delta_{equivalence}$
0
$+\delta_{equivalence}$
$\mu_{test} - \mu_{control}$

Figure 12.3 Conclusions to be drawn from the difference in population means from an equivalence trial

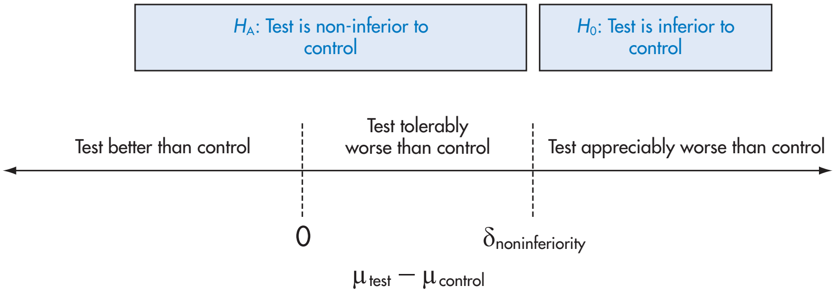
H_A: Test is non-inferior to control
H_0: Test is inferior to control
Test better than control
Test tolerably worse than control
Test appreciably worse than control
0
$\delta_{noninferiority}$
$\mu_{test} - \mu_{control}$

Figure 12.4 Conclusions to be drawn from the difference in population means from a noninferiority trial

certain circumstances. One important consideration in equivalence trials with a single active comparator is to consider what it means to conclude that the test drug is equivalent to the comparator. Not all marketed drugs are efficacious in every study. If the test drug were shown to be equivalent to the control, the test drug would be either efficacious or not efficacious. Which of these outcomes represents the truth depends on how the comparator would have performed had it been tested against a placebo. The ability to establish that a study can distinguish effective treatments from ineffective ones is called assay sensitivity. One way to establish this for equivalence trials is to select a comparator that had demonstrated consistent superiority to a placebo. Another option for equivalence trials in some instances is to include a third placebo arm. This is not possible when the ethics of the situation preclude this possibility. This is yet another illustration of the complexity of designing trials for which the outcomes have universally meaningful interpretations.

12.9 Additional study designs

Other appealing design features in new drug development include those that allow for monitoring of data while the trial is ongoing, and those that permit adaptations during a trial.

12.9.1 Interim analyses

Analyses conducted during a study are called interim analyses. Common uses of interim analyses are as follows:

- re-estimate the study sample size
- evaluate whether or not a study has accumulated sufficient data to stop early for definitive evidence of efficacy, for evidence of harm, or definitive evidence that the trial is unlikely to be successful in terms of its originally planned objectives.

A number of methodologies are available to assist in the quantification of evidence (accounting for type I and II errors) that enable early stopping of trials. Jennison and Turnbull (1999) provide a detailed description of sequential designs in which data are evaluated periodically for evidence of benefit, harm, or futility. Sequential designs typically involve the use of boundaries for the test statistic that define each of these outcomes.

One complicating factor of interim analyses is that they require the use of a data monitoring committee (DMC), which is independent of the study sponsor and others involved in the study. This is intended to protect the integrity of the clinical trial and to avoid any influence that knowledge of results may have on the future course of the trial. The work of the DMC is dictated by a specific protocol, or charter, written for the purpose of listing responsibilities of all parties and measures undertaken to protect the integrity of the trial. Ellenberg et al. (2003) have written a valuable reference outlining the complex issues associated with DMC involvement in trials.

12.9.2 Adaptive designs

Adaptive designs have become a topic of great interest, as evidenced by a recent Pharmaceutical Research and Manufacturers of America (PhRMA) working group convened to discuss adaptive designs methods. Dragalin (2006) provided an excellent overview of these studies. The ability to modify a study in midcourse may offer significant advantages to pharmaceutical companies, especially given the tremendous investment of time and money required for developing new drugs.

However, the logistical aspects of monitoring data at several points during a study are not trivial. An important concern with interim analyses is to ensure that knowledge of the results, however vague, does not unduly influence or bias the study. Hung et al. (2006, p 572) stated: "When the adaptation in confirmatory trials is extensive, the key hypothesis tested becomes unclear, protection of trial integrity is difficult, the infrastructure that is needed for logistics may be impossible to establish, and evaluation by regulatory agencies may be impossible." Summarizing the opportunities and the challenges of adaptive designs on behalf of the PhRMA working group, Gallo and Krams (2006, p 423) stated that: "We feel that the potential benefits for all involved parties suggested by adaptive designs are too enticing not to make every effort to find out if their promise can be realized." These designs represent an area for emerging research.

12.10 Review

1. Consider a design for an exploratory therapeutic trial of a new antihypertensive drug compared with placebo. It has been agreed that a between-treatment group difference in mean change from baseline (that is, the treatment effect) of at least 20 mmHg in SBP would be considered clinically meaningful. The primary hypothesis must be tested with $\alpha = 0.05$.

   (a) If the standard deviation for the between difference in mean change from baseline SBP is 40 mmHg, what is the required sample size for a test with power of 80%? What is the required sample size for a test with 90% power?

   (b) If the standard deviation for the between difference in mean change from baseline SBP is 60 mmHg, what is the required sample size for a test with power of 80%? What is the required sample size for a test with 90% power?

   (c) How is the estimate of the standard deviation obtained?

2. What are some advantages and disadvantages of using multiple investigational centers in clinical trials?

3. In what ways are noninferiority trials different from superiority trials?

12.11 References

Dragalin V (2006). Adaptive designs: terminology and classification. *Drug Information J* **40**:425–435.

Ellenberg SE, Fleming TR, DeMets DL (2003). *Data Monitoring Committees in Clinical Trials: A practical perspective*. Chichester: John Wiley & Sons.

EMEA Committee for Proprietary Medicinal Products (CPMP) (2000). *Points to Consider on Switching Between Superiority and Non-Inferiority*. London: EMEA.

EMEA CPMP (2003). *Points to Consider on Adjustment for Baseline Covariates*. London: EMEA.

Fleiss JL, Paik MC, Levin B (2003). *Statistical Methods for Rates and Proportions*, 3rd edn. Chichester: John Wiley & Sons.

Gallo P, Krams M (2006). PhRMA working group on adaptive designs: introduction to the full white paper. *Drug Information J* **40**:421–423.

Hung HMJ, O'Neill RT, Wang SJ, Lawrence J (2006). A regulatory view on adaptive/flexible clinical trial design. *Biometr J* **48**:565–573.

ICH Guidance E1 (1994). *The Extent of Population Exposure to Assess Clinical Safety for Drugs Intended for Long-Term Treatment of None-Life-Threatening Conditions*. Available at: www.ich.org (accessed July 1 2007).

Jennison C, Turnbull BW (2000). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, IL: Chapman & Hall/CRC.

Lee SJ, Zelen M (2000). Clinical trials and sample size considerations: another perspective. *Statist Sci* **15**: 95–110.

Machin D, Campbell MJ (2005). *Design of Studies for Medical Research*. Chichester: John Wiley & Sons.

Matthews JNS (2006). *Introduction to Randomized Controlled Clinical Trials*, 2nd edn. Boca Raton, FL: Chapman & Hall/CRC.

Molenberghs G, Kenward M (2007). *Missing Data in Clinical Studies*. Chichester: John Wiley & Sons.

Piantadosi S (2005). *Clinical Trials: A methodologic perspective*, 2nd edn. Chichester: John Wiley & Sons.

Roberts C, Torgerson DJ (1999). Understanding controlled trials: baseline imbalance in randomised controlled trials. *BMJ* **319**:185.

Senn S (1997). *Statistical Issues in Drug Development*. Chichester: John Wiley & Sons.

Turner JR (2007). *New Drug Development: Design, methodology, and analysis*. Hoboken, NJ: John Wiley & Sons.