9

Confirmatory clinical trials: Safety data II

9.1 Introduction

Chapter 8 focused on adverse event (AE) data, a large component of the overall safety data collected in clinical trials. Although AE data are often presented descriptively, we demonstrated that it is indeed possible to conduct inferential statistical analyses using AE data. This chapter discusses other safety data, including laboratory data, vital signs, and an assessment of cardiac safety that involves investigation of the cardiac QT interval (the QT interval can be identified on the ECG, as seen in Figure 9.2). In each of these cases, descriptive statistics, including measures of central tendency and dispersion, and categorical data are common forms of assessment.

9.2 Analyses of clinical laboratory data

Safety monitoring in clinical studies can be both data and labor intensive. In the context of later-stage therapeutic exploratory and therapeutic confirmatory trials, the collection of laboratory data is no exception. Typically, participants in clinical trials provide blood or urine samples at every clinic visit. There is an expansive range of clinical chemistry tests that can be conducted using these samples.

Samples may be analyzed by laboratories associated with each site (sometimes called local labs), each with its own handling procedures, assays, and reporting conventions, but this is not an optimal strategy. The use of a site's own laboratory poses no difficulties when the emphasis is on medical care, that is, the values obtained for a single individual. However, when conducting clinical research the emphasis is on using data from a group of individuals to make optimally informed conclusions and decisions.

Differences from local lab to local lab may preclude a sponsor from meaningfully combining data from all participants across a number of investigative sites. A statistical approach to standardizing laboratory values from a number of different labs (each potentially with their own reference ranges) has been described by Chuang-Stein (1992). However, standardization is time-consuming and the use of a number of local labs can introduce unwanted sources of variability that are neither easily quantified nor accounted for.

To overcome the difficulties with using local labs the use of central laboratories (central labs) is desirable. The advantages of using a central lab are that the samples are handled in a similar fashion, the assays used are consistent over time and across individuals, and the reporting conventions (for example, units of measurement) are uniform. Techniques for proper sample collection, storage, and handling, including shipment to a central lab, should be included in study protocols. Once the samples have been obtained by the central lab they are analyzed and the data recorded in a database that includes participant identifiers, study visit, date and time of sample collection, test name, result, reporting units, and the value of the reference ("normal") range.

The determination of values for reference ranges is based on the distribution of test values in large samples. Reference ranges are determined using large databases from a general population and typically represent "2σ" limits, assuming that the values are normally distributed in the general population. The lower limit

of the reference range is the value that cuts off the lowest 2.5% of values from individuals in the general population ($\mu - 2\sigma$). Likewise, the upper limit of the reference range is the value that cuts off the highest 2.5% of values from individuals in a general population ($\mu + 2\sigma$). Reference ranges for certain parameters (for example, hematocrit) may be defined specific to age and gender. Whichever approach is employed, local or central labs, the reference ranges are provided with lab values themselves to gauge the extent to which an individual's value is considered within an expected range or extreme.

In ICH Guidance E3 (1995), several analyses of clinical laboratory data are recommended. The approaches to describing clinical laboratory data include:

- measures of central tendency (for example, means or medians) for all groups at all time points examined
- shift analyses that classify laboratory values at baseline and later time points as normal, low or high relative to a reference range
- description of the number and proportion of participants for whom a change of a specified magnitude or more was reported at a particular time point. This is typically called a responders' analysis
- graphical displays of each subject's baseline value plotted against an on-treatment and/or end-of-study value
- identification of individual values that are so extreme that they would be considered clinically significant.

9.2.1 Measures of central tendency at each time point

Laboratory values are summarized descriptively for continuous measures by displaying the sample size, measures of central tendency (including the mean and median), the standard deviation, and the minimum and maximum values. A sample of such a descriptive display is provided in Table 9.1.

As the primary comparison is among or between treatment groups, the groups are displayed in the columns. Values of each test over time are of secondary interest and, therefore, are placed on the rows of the table. Reading between columns, we can see if the typical value (for example, the mean) for a parameter differs between groups. It is also possible to read down the column (that is, across time within a group) to see how the typical values vary over time. Provision of the minimum and maximum values allows the reviewer to identify any extreme values that might be considered out of the normal range. On occasion, similar analyses may also be presented for change from baseline values (typically calculated as endpoint value minus baseline value). If there are consistent and systematic changes from the start of the study, they may be apparent by examining the mean values and looking for values that deviate considerably from zero.

It may be of interest to provide a confidence interval for the change from baseline value within a group where an interval estimate that

Table 9.1 Summary of hemoglobin values (g/dL)

		Treatment group	
Visit	Statistic	Placebo	Active
Baseline	*n*	20	20
	Mean (SD)	13.78 (1.97)	14.61 (2.05)
	Median	13.5	14.6
	Min., Max.	11.0, 17.3	11.2, 17.7
Endpoint (last visit)	*n*	20	20
	Mean (SD)	13.41 (2.07)	13.75 (2.00)
	Median	13.3	13.5
	Min., Max.	10.6, 16.9	10.4, 17.2

excludes zero represents evidence of a change in mean value that exceeds what might be observed by chance alone. Similarly, confidence intervals may be calculated to provide an estimate of the between-group difference in a laboratory parameter. Comparison with a control group can be especially important when there is a laboratory test that changes as a result of study procedures (for example, decreases in hematocrit or hemoglobin as a consequence of frequent blood sampling). A summary of the change from baseline at the last visit is provided in Table 9.2.

9.2.2 A confidence interval for a mean with unknown variance

For a sample size of n observations of a random variable, the sample mean, an estimator of the population mean, is calculated as:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

and the sample standard deviation, s, an estimator of the population standard deviation is calculated as:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}.$$

The standard error of the sample mean is then calculated as:

$$\mathrm{SE}(\bar{x}) = \frac{s}{\sqrt{n}}.$$

Finally, assuming that the random variable is normally distributed (or at least symmetrically distributed with a sample size ≥ 30), a $(1-\alpha/2)\%$ confidence interval is:

$$\bar{x} \pm t_{1-\alpha/2,n-1}\mathrm{SE}(\bar{x}),$$

where $t_{1-\alpha/2,n-1}$ represents the reliability factor and is the value of the t distribution with $n-1$ degrees of freedom (df) to the left of which is $(1-\alpha/2)\%$ of the area under the curve. These values are provided in Appendix 2.

As an example, let us calculate a 95% two-sided confidence interval for the mean hemoglobin value at the end of the study for the active group using data in Table 9.1.

Data

The data are 20 hemoglobin values from individuals treated with the active drug, obtained from blood samples collected at the last visit of the study. The mean and standard deviation were calculated as 13.75 and 2.00, respectively, and these values serve as the basis of the confidence interval.

Statistical analysis

As the population variance is unknown and is therefore being estimated by the sample variance, we use the t distribution for a reliability factor. The use of the t distribution requires us to assume that the underlying distribution of hemoglobin values is approximately normally distributed, or at least symmetrically distributed. The standard error of the sample mean is calculated as:

Table 9.2 Summary of change from baseline hemoglobin values (g/dL)

		Treatment group	
Visit	Statistic	Placebo	Active
Endpoint (last visit)	*n*	20	20
	Mean (SD)	−0.37 (1.47)	−0.86 (1.67)
	Median	−0.4	−0.8
	Min., Max.	−3.8, 2.1	−4.5, 1.4

$$\mathrm{SE}(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{2.00}{\sqrt{20}} = 0.44.$$

As we are interested in a 95% two-sided confidence interval, the value of the variate that cuts off the upper 2.5% of area from the t distribution with 19 df is 2.093. Therefore, the 95% confidence interval for the mean hemoglobin is calculated as follows:

$$13.75 \pm 2.093\ (0.44) = (12.83,\ 14.67).$$

Interpretation and decision-making

From the confidence interval we can conclude with 95% confidence that the true population mean hemoglobin is in the interval (12.83, 14.67). Assuming that the reference range is 12–15 g/dL for females and 14–17 g/dL for males, we can proceed with development of the new drug with some degree of assurance although gender-specific intervals would be more informative.

9.2.3 A confidence interval for the difference in two means with equal unknown variance

Within-group confidence intervals can be informative, but usually the primary interest in a clinical trial is to compare the effect of one treatment with that of another. Therefore, a confidence interval for the difference in two means can better address the goals of the research.

For two independent groups 1 and 2, a sample size of n_1 observations of a random variable from group 1 and n_2 observations of a random variable from group 2, the sample means from each group are:

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

$$\bar{x}_2 = \frac{\sum_{j=1}^{n_2} x_{2j}}{n_2}, \text{ respectively.}$$

The within-group sample variances are estimated as:

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

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

As before, these sample statistics are estimates of the unknown population parameters, the population means, and the population variances. If the population variances are assumed to be equal, each sample statistic is a different estimate of the same population variance. It is then reasonable to average or "pool" these estimates to obtain the following:

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

The standard error of the difference in sample means is:

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

Calculation of a confidence interval for the difference in means requires an assumption of normal data (or, alternately, symmetrical distributions with sample sizes of 30 or more). If the population variances are assumed to be equal, a two-sided $(1 - \alpha)\%$ confidence interval is:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{1-\alpha/2, n_1+n_2-2} \mathrm{SE}(\bar{x}_1 - \bar{x}_2),$$

where $t_{1-\alpha/2, n_1+n_2-2}$ represents the reliability factor and is the value of the t distribution with n_1+n_2-2 df to the left of which is $(1 - \alpha/2)\%$ of the area under the curve.

To illustrate this methodology we use the data from Table 9.2 to calculate a between-group difference in the mean change from baseline hemoglobin at the end of the study.

Data

The description of the data for this analysis is provided in Section 9.2.1.

Statistical analysis

For this analysis we are required to use the t distribution, and therefore to make the assumption that the distribution of change from baseline values is normally or approximately normally distributed. When calculating a between-group confidence interval, it is very important to understand how the difference is being calculated and what the interpretation of the interval is, given the direction of the difference.

In this case, each change from baseline value is calculated as "endpoint minus baseline." Therefore, a mean value of change from baseline that was > 0 would imply an increase from baseline, whereas a mean change from baseline value that was < 0 would imply a decrease from baseline. In this instance we are interested in the between-group difference in mean change from baseline. We interpret the calculated confidence interval accordingly.

To start, the point estimate for the between-group (active minus placebo) difference in mean change from baseline is $(-0.86) - (-0.37) = -0.49$. To calculate the standard error we first need to obtain an estimate of the pooled variance, which is calculated as follows:

$$s_p^2 = \frac{(20-1)1.67^2 + (20-1)1.47^2}{(20+20-2)} = \frac{(19)2.79 + (19)2.16}{38} = 2.47.$$

The pooled standard deviation is calculated as:

$$\sqrt{2.47} = 1.57.$$

The standard error of the difference in means is calculated as:

$$\text{SE}(\bar{x}_1 - \bar{x}_2) = 1.57\sqrt{\frac{1}{20} + \frac{1}{20}} = 0.50.$$

The final component is to obtain the reliability factor from the t distribution with 38 df, which is 2.02. The 95% confidence interval for the difference in means is therefore calculated as:

$$-0.49 \pm 2.02(0.50) = (-1.5, 0.52).$$

Interpretation and decision-making

On the basis of this confidence interval, there does not appear to be much of a difference between the groups with respect to a change in hemoglobin from baseline to the end of the study, particularly because the confidence interval includes the value 0.

9.2.4 Shift analysis

Another method used to analyze clinical laboratory data is called a shift analysis. For this analysis the data themselves are not the actual numeric values of the laboratory test, but a categorical ordinal variable that indicates whether the value was within the reference range (normal), low relative to the reference range (low), or high relative to the reference range (high). With these classifications on observations from baseline and some other post-randomization time point (for example, end of study), the primary interest is in the proportion of individuals who shifted from normal to high or normal to low. Depending on the parameter being investigated, a shift from high to low or low to high may also be of interest.

A typical summary table representing this kind of analysis is provided in Table 9.3. As seen there 25% of participants in the placebo group who had normal values at baseline had low values at the last visit. In the active group 20% of participants experienced this shift from baseline to last visit.

9.2.5 Responders' analysis

We noted earlier in this book that there is no such thing as an effective drug without some associated risks. Some drugs may be known to be

Table 9.3 Shift analysis of hemoglobin values

	Baseline value					
	Placebo (n = 20)			Active (n = 20)		
Last visit	Low	Normal	High	Low	Normal	High
Low	3 (15%)	5 (25%)	0	1 (5%)	4 (20%)	1 (5%)
Normal	2 (10%)	7 (35%)	2 (10%)	1 (5%)	11 (55%)	0
High	0	1 (5%)	0	0	0	2 (10%)

associated with a small but consistent change in a clinical laboratory parameter (for example, treatment with hydrochlorothiazide is often associated with increases in blood glucose). Imagine a scenario in which a small change is not troubling in itself. A concern may then be: What is the chance that an individual who receives the test treatment will have a change in the lab test above a certain threshold, one that would no longer be trivial?

An analysis approach that may be informatively used here is to calculate a change from baseline for each observation and then categorize the change from baseline value as either a responder (that is, someone whose change from baseline was less or greater than a specified value) or a non-responder (that is, someone whose change from baseline was within the tolerable values of change). Whether or not a decrease or increase in the lab value is indicative of harm depends on the laboratory test itself. The descriptive analysis for this type of data includes the presentation of counts and percentages (recall that these can be represented as proportions) of responders in each group. As there usually are a number of visits at which the lab test is performed, the analysis may be presented for all post-baseline visits, the last visit, or both.

An extension of the responder analysis described above would be to categorize the change from baseline values into several (> 2) categories (for example, no change, increase $\leq X$, increase $> X$).

9.2.6 Graphical displays of end-of-study values plotted against baseline

One common element shared by a number of the analyses of laboratory data that we have described is that the magnitude of change from the start to the end of the study is important, but so is the final value itself. In addition, the relative frequency of such outcomes is of vital interest when gauging the overall risk of treatment with a new drug. One descriptive approach to address several of these issues is a graphical one.

A scatter plot of each individual's baseline value plotted against his or her end-of-study value enables us to see how many individuals (in the absolute or relative sense) had end-of-study values beyond a normal level or changes from baseline to end-of-study that represent a significant health risk. As an example, hemoglobin values at the end of the study have been plotted against the baseline value for two treatment groups (placebo and active) in Figure 9.1.

Note the diagonal line in each plot that connects all points for which the baseline value is equal to the end-of-study value. With the end-of-study value on the y axis, points above the diagonal line represent an increase from baseline and points below represent a decrease from baseline. Larger vertical deviations from the diagonal line represent larger changes from baseline values. Thus, the need to interpret a number of quantities at once is satisfied by one graphical display.

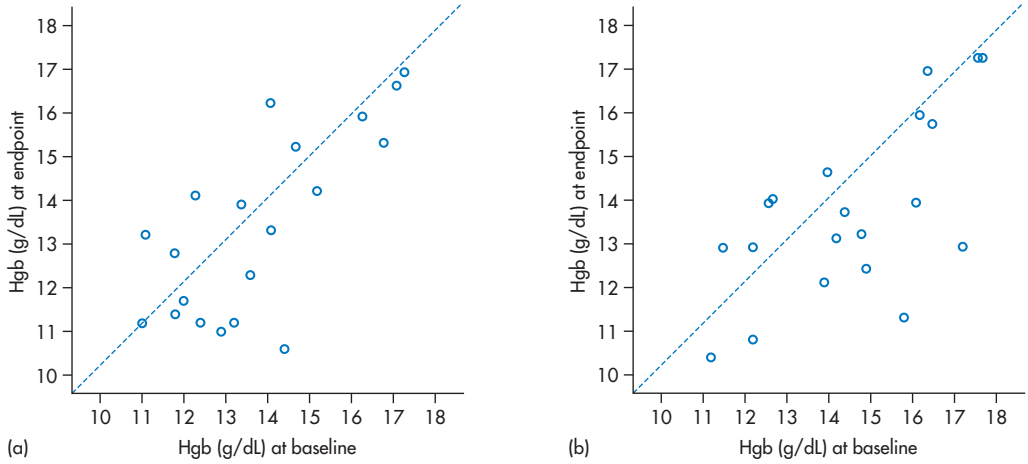

Figure 9.1 Scatterplot of hemoglobin values at baseline and end of study: (a) Placebo; (b) active

9.2.7 Clinically significant laboratory values

A graphical display such as the one in Figure 9.1, or a table of summary descriptive statistics including the minimum and maximum, may reveal values that are so extreme that they merit additional scrutiny. This is typically accomplished with the use of a listing that provides all values of the laboratory test, the dates and times of the sample collection, and the characteristics of the participant. Such an analysis is not based on aggregate information but rather on an individual observation. If a clinically significant observation were noted a medical reviewer would look to see if the participant's values returned to normal levels or remained abnormal, and if there were any accompanying AEs.

9.3 Vital signs

Vital signs typically measured in clinical trials are blood pressure (both systolic or SBP and diastolic or DBP) and heart rate, often measured as pulse rate, in beats per minute. Weight might also be of interest. In our ongoing scenario of the development of a new antihypertensive drug, blood pressure measurements are efficacy measurements, not safety measures as such. However, we discuss the use of blood pressures in safety assessment here because this is so common in the development of non-antihypertensive drugs.

As for laboratory data, both continuous and categorical data analytical methods can be employed here. Measures of central tendency and dispersion are appropriate for continuous data. Categories of interest, and the associated categorical data, can take various forms. Imagine a trial in which the treatment phase is 12 weeks and participants visit their investigational site every 2 weeks – that is, a baseline value taken before treatment commences is followed by six values measured during the treatment phase. It may be of interest to know how many individuals show clinically significant vital sign changes during the treatment period. In this case a precise definition of clinically significant must be provided in the study protocol. The following hypothetical changes in vital signs might be considered of clinical significance by clinicians on the study team if they occurred at any of the six measurement points in the treatment phase:

- an increase from baseline in SBP ≥ 20 mmHg
- an increase from baseline in DBP ≥ 12 mmHg
- an increase from baseline in SBP ≥ 15 mmHg and an increase in DBP ≥ 10 mmHg
- a pulse rate ≥ 120 beats/min and an associated increase from baseline of at least 15 beats/min.

The clinicians on the study team might also be interested in sustained changes in vital signs. Hypothetical examples of definitions of sustained changes might be:

- an increase from baseline in SBP ≥ 15 mmHg at each of three consecutive visits
- an increase from baseline in DBP ≥ 10 mmHg at each of three consecutive visits
- an increase from baseline in pulse rate ≥ 10 beats/min at each of three consecutive visits.

Appropriate categorical analyses could then be used with these data.

9.4 QT interval prolongation and torsades de pointes liability

The ECG is a very recognizable pattern of biological activity. The ECG consists of the P wave, the QRS complex, and the T wave. These components, represented in Figure 9.2, are associated with different aspects of the cardiac cycle: Atrial activity, excitation of the ventricles, and repolarization of the ventricles, respectively. Modern computerized systems not only display these electrophysiological signals but also concurrently digitize them and store them for later examination.

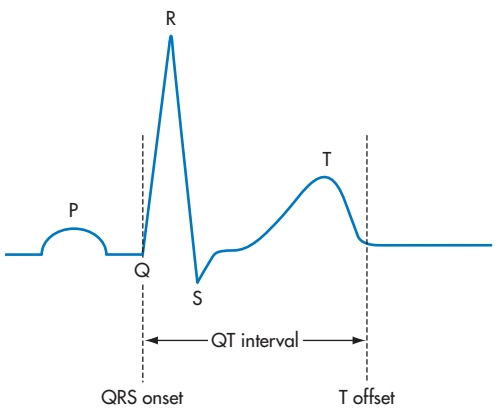

Figure 9.2 Stylized representation of the ECG, showing the QT interval

The QT interval is highlighted in Figure 9.2. This interval is of particular interest in assessing cardiac safety in drug development, because QT interval prolongation is one potentially informative surrogate biomarker available for very serious cardiac events including sudden cardiac death. (This section focuses on cardiac safety assessment in all systemically available drugs being developed for uses other than the control of cardiac arrhythmias: QT/QT_c interval prolongation – an occurrence deemed highly undesirable in all other drugs – can occur with antiarrhythmic drugs as a consequence of their mechanism of clinical efficacy [ICH Guidance E14, 2005].) The ICH Guidance E14 addresses the evaluation of QT intervals in clinical development programs.

The time interval between the onset of the QRS complex and the offset of the T wave is defined as the QT interval. Consider an individual with a steady heart rate of 60 beats/min, a number chosen to make the math easy in this example. This represents one heart beat/second, and so the total length (in the time domain) of all ECG segments during one beat would add up to 1 second, represented in this research field as 1000 milliseconds (ms). Each component of the ECG can therefore be assigned a length, or duration, in milliseconds. The length of the QT interval can be obtained by measurement from inspecting the ECG and identifying the QRS onset and the T-wave offset.

As the heart beats faster (heart rate increases), the duration of an individual cardiac cycle decreases, because more cardiac cycles now occur in the same time. Therefore, as the cardiac cycle shortens, so do each of the components of the cardiac cycle. This means that the QT interval will tend to be shorter at a higher heart rate. As it is of interest to examine the QT interval at various heart rates, the interval can be "corrected" for heart rate. This leads to the term QT_c, which is calculated (by one of several methods including two corrections attributed to Bazett and Fridericia), taking into account the actual QT and the heart rate (the duration of

the entire cardiac cycle, sometimes referred to as the RR interval) at that point. The title of ICH Guidance E14 uses the term "QT/QT_c interval" to indicate that both QT and QT_c are of interest: In this book, the term "QT interval" represents both QT and QT_c.

It is of considerable interest in drug development to determine whether the investigational drug under development leads to prolongation of the QT interval: Although QT interval prolongation can be congenital, it can also be acquired, for example, induced by drug therapy. QT prolongation, which represents delayed cardiac repolarization of the myocardial cells, is regarded as a potentially very informative surrogate marker for certain dangerous cardiac arrhythmias, namely polymorphic ventricular tachycardia and torsades de pointes, and sudden cardiac death. Extensive ECG monitoring during preapproval clinical trials is therefore a critical part of clinical development programs, and results from this testing must be presented to a regulatory agency to obtain marketing approval. One of the biggest causes of delay in getting a new drug approved by a regulatory agency, or failure to be given marketing approval, is cardiac safety issues, and therefore the choice of the correct study design, appropriate methodology for collecting optimum quality data, and appropriate statistical analyses are of tremendous importance.

Although ICH Guidance E14 (2005) provides guidance on each of these considerations, we focus here on the statistical approaches that should be taken in the investigation of QT prolongation. As this guidance noted (p 9), "The QT/QT_c interval data should be presented both as analyses of central tendency (for example, means, medians) and categorical analyses. Both can provide relevant information on clinical risk assessment." The effect of the investigational drug on the QT intervals is most commonly analyzed using the largest time-matched mean difference between the drug and placebo (adjusted for baseline) over the data collection period.

Categorical analyses are based on the number and the percentage of individuals who meet or exceed a predefined upper limit. Such limits can be stated in the study protocol in terms of either absolute QT interval prolongation values or changes from baseline. At this time, there is no consensus concerning what is the "best" choice of these upper limit values. ICH Guidance 14 therefore suggested that multiple analyses using several predefined limits is a reasonable approach in light of this lack of consensus. For absolute QT interval data, the guidance suggests providing absolute numbers and percentages of individuals whose QT intervals exceed 450, 480, and 500 ms. For change-from-baseline QT interval data, the same information might be provided for increases exceeding 30 ms and those exceeding 60 ms. The design and analysis of studies intended to evaluate changes in QT can be rather difficult to implement. Some of the difficulties and areas for further research brought to light by ICH Guidance E14 are discussed in a recent paper (Pharmaceutical Research and Manufacturers of America QT Statistics Expert Working Team, 2005).

For further discussion of QT/QT_c interval prolongation and other cardiac safety assessments for noncardiac drugs, see Morganroth and Gussak (2005) and Turner and Durham (2008).

9.5 Concluding comments on safety assessments in clinical trials

In this chapter we have seen that the goal of safety analyses is to cast a wide net in the hopes of identifying any events that may be attributable to treatment with the new drug. Such a broad search, however, also has a significant disadvantage: If we look at so many outcomes, we might find one that looks problematic just by chance alone. Rather than rely solely on statistical approaches to limit the chance of this occurring, a sensible approach is to substantiate such a finding with additional data, either a similar result in a different study or some data on the medical explanation for the event (the mechanism of action).

The analysis tools that we have described in this chapter provide ways to evaluate the risk of the new drug, given the constraints of sample sizes obtainable in clinical development. The limitations of relatively little human experience

before marketing approval have to be considered, especially when reviewing clinical safety data. Thus far, regulatory agencies have not required pharmaceutical companies to increase the sizes of their studies to find the best way to uncover safety risks that would otherwise be hard to find. Rather, the emphasis has been to use more modern tools (for example, genetics and candidate screening) to identify potentially dangerous drugs before there are a large number of participant exposures (US Department of Health and Human Services, FDA, 2004). The role of postmarketing surveillance will continue to be important (see also ICH Guidance E2E, 2004; Strom, 2005; Mann and Andrews, 2007). This is especially true when we think of the relative homogeneity of participants in clinical trials compared with patients in the real world and the implications of the law of large numbers (recall discussions in Chapter 6).

As a final note to this chapter, any potential risks to individuals treated with a new drug have to be considered and cannot automatically be considered trivial. The acceptability of the magnitude of the risk depends largely on a statistical demonstration of the expected benefit of the new treatment, which is the topic of Chapters 10 and 11.

9.6 Review

1. What are some advantages and disadvantages of the various analytical approaches cited from ICH Guidance E3 listed in Section 9.2?
2. Refer to the data in Table 9.1:
   (a) Calculate a two-sided 90% confidence interval for the difference in mean hemoglobin value at endpoint (last visit).
   (b) Calculate a two-sided 95% confidence interval for the difference in mean hemoglobin value at endpoint (last visit).
   (c) Calculate a two-sided 99% confidence interval for the difference in mean hemoglobin value at endpoint (last visit).
   (d) What is the statistical interpretation of these results?
3. What is the statistical and clinical interpretation (or relevance) of the following 95% confidence intervals for the between-group difference (for example, test group minus placebo) in mean change from baseline hemoglobin (g/dL) at endpoint (last visit)?

   (a) (−1.2, 2.6)
   (b) (1.7, 3.4)
   (c) (−6.2, −2.3).

9.7 References

Chuang-Stein C (1992). Summarizing laboratory data with different reference ranges in multi-center trials. *Drug Information J* **26**:74–84.

ICH Guidance E2E (2004). *Pharmacovigilance Planning*. Available at: www.ich.org (accessed July 1 2007).

ICH Guidance E3 (1995). *Structure and Content of Clinical Study Reports*. Available at: www.ich.org (accessed July 1 2007).

ICH Guidance E14 (2005). *The Clinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential for Non-Antiarrhythmic Drugs*. Available at: www.ich.org (accessed July 1 2007).

Mann R, Andrews E, eds (2007). *Pharmacovigilance*, 2nd edn. Chichester: John Wiley & Sons.

Morganroth J, Gussak I, eds (2005). *Cardiac Safety of Noncardiac Drugs: Practical guidelines for clinical research and drug development*. Totowa, NJ: Humana Press.

Pharmaceutical Research and Manufacturers of America QT Statistics Expert Working Team (2005). Investigating drug-induced QT and QTc prolongation in the clinic: a review of statistical design and analysis considerations: report from the Pharmaceutical Research and Manufacturers of America QT Statistics Expert Team. *Drug Information J* 39:243–266.

Strom BL, ed. (2005). *Pharmacoepidemiology*, 4th edn. Chichester: John Wiley & Sons.

Turner JR, Durham TA (2008). *Integrated Cardiac Safety: Assessment methodologies for noncardiac drugs in discovery, development, and postmarketing surveillance*. Hoboken, NJ: John Wiley & Sons, in press.

US Department of Health and Human Services, Food and Drug Administration (2004). *Challenge and Opportunity on the Critical Path to New Medical Products*. Available from www.fda.gov (accessed July 1 2007).

10

Confirmatory clinical trials: Analysis of categorical efficacy data

10.1 Introduction: Regulatory views of substantial evidence

When thinking about the use of statistics in clinical trials, the first thing that comes to mind for many people is the process of hypothesis testing and the associated use of p values. This is very reasonable, because the role of a chance outcome is of utmost importance in study design and the interpretation of results from a study. A sponsor's objective is to develop an effective therapy that can be marketed to patients with a certain disease or condition. From a public health perspective, the benefits of a new treatment cannot be separated from the risks that are tied to it. Regulatory agencies must protect public health by ensuring that a new treatment has "definitively" been demonstrated to have a beneficial effect. The meaning of the word "definitively" as used here is rather broad, but we discuss what it means in this context – that is, we operationally define the term "definitively" as it applies to study design, data analysis, and interpretation in new drug development.

Most of this chapter is devoted to describing various types of data and the corresponding analytical strategies that can be used to demonstrate that an investigational drug, or test treatment, is efficacious. First, however, it is informative to discuss the international standards for demonstrating efficacy of a new product, and examine how regulatory agencies have interpreted these guidelines. ICH Guidance E9 (1998, p 4) addresses therapeutic confirmatory studies and provides the following definition:

> A confirmatory trial is an adequately controlled trial in which the hypotheses are stated in advance and evaluated. As a rule, confirmatory trials are necessary to provide firm evidence of efficacy or safety. In such trials the key hypothesis of interest follows directly from the trial's primary objective, is always pre-defined, and is the hypothesis that is subsequently tested when the trial is complete. In a confirmatory trial it is equally important to estimate with due precision the size of the effects attributable to the treatment of interest and to relate these effects to their clinical significance.

It is common practice to use earlier phase studies such as therapeutic exploratory studies to characterize the size of the treatment effect, while acknowledging that the effect size found in these studies is associated with a certain amount of error. As noted earlier, confidence intervals can be helpful for planning confirmatory studies. The knowledge and experience gained in these earlier studies can lead to hypotheses that we wish to test (and hopefully confirm) in a therapeutic confirmatory trial, for example, the mean reduction in systolic blood pressure (SBP) for the test treatment is 20 mmHg greater than the mean reduction in SBP for placebo. As we have seen, a positive result from a single earlier trial could be a type I error, so a second study is useful in substantiating that result.

The description of a confirmatory study in ICH Guidance E9 (1998) also illustrates the importance of the study design employed. The study should be designed with several important characteristics:

- It should test a specific hypothesis.
- It should be appropriately sized.
- It should be able to differentiate treatment effects from other sources of variation (for

example, time trends, regression to the mean, bias).
- The size of the treatment effect that is being confirmed should be clinically relevant.

The clinical relevance, or clinical significance, of a treatment effect is an extremely important consideration. The size of a treatment effect that is deemed clinically relevant is best defined by medical, clinical, and regulatory specialists.

Precise description of the study design and adherence to the study procedures detailed in the study protocol are particularly important for confirmatory studies. Quoting again from ICH Guidance E9 (1998, p 4):

> Confirmatory trials are intended to provide firm evidence in support of claims and hence adherence to protocols and standard operating procedures is particularly important; unavoidable changes should be explained and documented, and their effect examined. A justification of the design of each such trial, and of other important statistical aspects such as the principal features of the planned analysis, should be set out in the protocol. Each trial should address only a limited number of questions.

Confirmatory studies should also provide quantitative evidence that substantiates claims in the product label (for example, the package insert) as they relate to an appropriate population of patients. In the following quote from ICH Guidance E9 (1998, p 4), the elements of statistical and clinical inference can be seen:

> Firm evidence in support of claims requires that the results of the confirmatory trials demonstrate that the investigational product under test has clinical benefits. The confirmatory trials should therefore be sufficient to answer each key clinical question relevant to the efficacy or safety claim clearly and definitively. In addition, it is important that the basis for generalisation . . . to the intended patient population is understood and explained; this may also influence the number and type (e.g. specialist or general practitioner) of centres and/or trials needed. The results of the confirmatory trial(s) should be robust. In some circumstances the weight of evidence from a single confirmatory trial may be sufficient.

The terms "firm evidence" and "robust" do not have explicit definitions. However, as clinical trials have been conducted and reported in recent years, some practical (operational) definitions have emerged, and these are discussed shortly.

In its guidance document *Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products*, the US Food and Drug Administration (US Department of Health and Human Services, FDA, 1998) describes the introduction of an effectiveness requirement according to a standard of "substantial evidence" in the Federal Food, Drug, and Cosmetic Act (the FDC Act) of 1962:

> *Substantial evidence* was defined in section 505(d) of the Act as "evidence consisting of adequate and well-controlled investigations, including clinical investigations, by experts qualified by scientific training and experience to evaluate the effectiveness of the drug involved, on the basis of which it could fairly and responsibly be concluded by such experts that the drug will have the effect it purports or is represented to have under the conditions of use prescribed, recommended, or suggested in the labeling or proposed labeling thereof."
>
> US Department of Health and Human Services, FDA (1998, p 3)

The phrase "adequate and well-controlled investigations" has typically been interpreted as at least two studies that clearly demonstrated that the drug has the effect claimed by the sponsor submitting a marketing approval. Furthermore, a type I error of 0.05 has typically been adopted as a reasonable standard upon which data from clinical studies are judged. That is, it was widely believed that the intent of the FDC Act of 1962 was to state that a drug could be concluded to be effective if the treatment effect was clinically relevant and statistically significant at the $\alpha = 0.05$ level in two independent studies.

The ICH Guidance E8 (1998, p 4) clarified this issue:

> The usual requirement for more than one adequate and well-controlled investigation reflects the need for *independent substantiation*

of experimental results. A single clinical experimental finding of efficacy, unsupported by other independent evidence, has not usually been considered adequate scientific support for a conclusion of effectiveness. The reasons for this include the following:

- Any clinical trial may be subject to unanticipated, undetected, systematic biases. These biases may operate despite the best intentions of sponsors and investigators, and may lead to flawed conclusions. In addition, some investigators may bring conscious biases to evaluations.
- The inherent variability in biological systems may produce a positive trial result by chance alone. This possibility is acknowledged, and quantified to some extent, in the statistical evaluation of the result of a single efficacy trial. It should be noted, however, that hundreds of randomized clinical efficacy trials are conducted each year with the intent of submitting favorable results to the FDA. Even if all drugs tested in such trials were ineffective, one would expect one in forty of those trials to "demonstrate" efficacy by chance alone at conventional levels of statistical significance. It is probable, therefore, that false positive findings (that is, the chance appearance of efficacy with an ineffective drug) will occur and be submitted to FDA as evidence of effectiveness. Independent substantiation of a favorable result protects against the possibility that a chance occurrence in a single study will lead to an erroneous conclusion that a treatment is effective.
- Results obtained in a single center may be dependent on site or investigator-specific factors (for example, disease definition, concomitant treatment, diet). In such cases, the results, although correct, may not be generalizable to the intended population. This possibility is the primary basis for emphasizing the need for independence in substantiating studies.
- Rarely, favorable efficacy results are the product of scientific fraud.

Although there are statistical, methodologic, and other safeguards to address the identified problems, they are often inadequate to address these problems in a single trial. Independent substantiation of experimental results addresses such problems by providing consistency across more than one study, thus greatly reducing the possibility that a biased, chance, site-specific, or fraudulent result will lead to an erroneous conclusion that a drug is effective.

This guidance further clarified that the need for substantiation does not necessarily require two or more identically designed trials:

> Precise replication of a trial is only one of a number of possible means of obtaining independent substantiation of a clinical finding and, at times, can be less than optimal as it could leave the conclusions vulnerable to any systematic biases inherent to the particular study design. Results that are obtained from studies that are of different design and independent in execution, perhaps evaluating different populations, endpoints, or dosage forms, may provide support for a conclusion of effectiveness that is as convincing as, or more convincing than, a repetition of the same study.
>
> ICH Guidance E8 (1998, p 5)

Regulatory agencies have traditionally accepted only two-sided hypotheses because, theoretically, one could not rule out harm (as opposed to simply no effect) associated with the test treatment. If the value of a test statistic (for example, the Z-test statistic) is in the critical region at the extreme left or extreme right of the distribution (that is, < -1.96 or > 1.96), the probability of such an outcome by chance alone under the null hypothesis of no difference is 0.05. However, the probability of such an outcome in the direction indicative of a treatment benefit is half of 0.05, that is, 0.025. This led to a common statistical definition of "firm" or "substantial" evidence as the effect was unlikely to have occurred by chance alone, and it could therefore be attributed to the test treatment. Assuming that two studies of the test treatment had two-sided p values < 0.05 with the direction of the treatment effect in favor of a benefit, the probability of the two results occurring by chance alone would be 0.025×0.025, that is, 0.000625 (which can also be expressed as 1/1600).

It is important to note here that this standard is not written into any regulation. Therefore,

there may be occasions where this statistical standard is not met. In fact, it is possible to redefine the statistical standard using one large well-designed trial, an approach that has been described by Fisher (1999).

Whether the substantial evidence comes from one or more than one trial, the basis for concluding that the evidence is indeed substantial is statistical in nature. That is, the regulatory agency must agree with the sponsor on several key points in order to approve a drug for marketing:

- The effect claimed cannot be explained by other phenomena such as regression to the mean, time trends, or bias. This highlights the need for appropriate study design and data acquisition.
- The effect claimed is not likely a chance outcome. That is, the results associated with a primary objective have a small p value, indicating a low probability of a type I error.
- The effect claimed is large enough to be important to patients, that is, clinically relevant. The magnitude of the effect must account for sampling during the trial(s).

A clinical development program contains various studies that are designed to provide the quantity and quality of evidence required to satisfy regulatory agencies, which have the considerable responsibility of protecting public health. The requirements for the demonstration of substantial evidence highlight the importance of study design and analytic strategies. Appropriate study design features such as concurrent controls, randomization, standardization of data collection, and treatment blinding help to provide compelling evidence that an observed treatment effect cannot be explained by other phenomena. Selection of the appropriate analytical strategy maximizes the precision and efficiency of the statistical test employed. The employment of appropriate study design and analytical strategies provides the opportunity for an investigational drug to be deemed effective if a certain treatment effect is observed in clinical trials.

10.2 Objectives of therapeutic confirmatory trials

Table 10.1 provides a general taxonomy of the objectives of confirmatory trials and specific research questions corresponding to each. Confirmatory trials typically have one primary objective that varies by the type of trial. In the case of a new antihypertensive it may be sufficient to demonstrate simply that the reduction in blood pressure is greater for the test treatment than for the placebo. A superiority trial is appropriate in this instance. However, in other therapeutic areas – for example, oncology – other designs are appropriate. In these therapeutic areas it is not ethical to withhold life-extending therapies to certain individuals by randomizing them to a placebo treatment if there is already an existing treatment for the disease or condition.

In such cases, it is appropriate to employ trials with the objective of demonstrating that the clinical response to the test treatment is equivalent (that is, no better or worse) to that of an existing effective therapy. These trials are called

Table 10.1 Taxonomy of therapeutic confirmatory trial objectives

Objective of trial	Example indication	Example research question
Demonstrate superiority	Hypercholesterolemia	Is the magnitude of LDL reduction for the test treatment greater than for placebo?
Demonstrate equivalence	Oncology	Is the test treatment at worst trivially inferior to and at best slightly better than the active control with respect to the rate of partial tumor response?
Demonstrate noninferiority	Anti-infective	Is the microbial eradication rate for the test treatment at least not unacceptably worse than for the active control?

equivalence trials. A question that arises here is: Why would we want to develop another drug if there is already an existing effective treatment? The answer is that we believe the test treatment offers other advantages (for example, convenience, tolerability, or cost) to justify its development. Another type of trial is the noninferiority trial. These trials are intended only to demonstrate that a test treatment is not unacceptably worse (noninferior) than an active control. Again, the test treatment may provide advantages other than greater therapeutic response such as fewer adverse effects or greater convenience.

Equivalence and noninferiority trials are quite different from superiority trials in their design, analysis, and interpretation (although exactly the same methodological considerations apply to collect optimum quality data in these trials). Superiority trials continue to be our focus in this book, but it is important that you are aware of other designs too. Therefore, in Chapter 12 we discuss some of the unique features of these other design types.

10.3 Moving from research questions to research objectives: Identification of endpoints

There is an important relationship between research questions and study objectives, and it is relatively straightforward to restate research questions such as those in Table 10.1 in terms of study objectives. As stated in ICH Guidance E9, a confirmatory study should be designed to address at most a few objectives. If a treatment effect can be quantified by an appropriate statistical measure, study objectives can be translated into statistical hypotheses. For example, the extent of low-density lipoprotein (LDL)-cholesterol reduction can be measured by the mean change from baseline to end-of-treatment, or by the proportion of study participants who attain a goal level of LDL according to a treatment guideline. The efficacy of a cardiovascular intervention may be measured according to the median survival time after treatment. For many drugs, identification of an appropriate measure of the participant-level response (for example, reported pain severity using a visual analog scale) is not difficult. However, there may be instances when the use of a surrogate endpoint can be justified on the basis of statistical, biological and practical considerations. Measuring HIV viral load as a surrogate endpoint for occurrence of AIDS is an example.

Identification of the endpoint of interest is one of the many cases in clinical research that initially seem obvious and simple. We know exactly what disease or condition we are interested in treating, and it should be easy to identify an endpoint that will tell us if we have been successful. In reality, the establishment of an appropriate endpoint, whether it is the most clinically relevant endpoint or a surrogate endpoint, can be difficult. Some of the statistical criteria used to judge the acceptability of surrogate endpoints are described by Fleming and DeMets (1996), who caution against their use in confirmatory trials. One might argue that the most clinically relevant endpoint for a antihypertensive is the survival time from myocardial infarction, stroke, or death. Fortunately, the incidence of these events is relatively low during the typical observation period of clinical trials. The use of SBP as a surrogate endpoint enables the use of shorter and smaller studies than would be required if the true clinical endpoint had to be evaluated. For present purposes, we assume the simplest scenario: The characteristic that we are going to measure (blood pressure) is uncontroversial and universally accepted, and a clinically relevant benefit is acknowledged to be associated with a relative change in blood pressure for the test treatment compared with the control.

Common measures of the efficacy of a test treatment compared with a placebo include the differences in means, in proportions, and in survival distributions. How the treatment effect is measured and analyzed in a clinical trial should be a prominent feature of the study protocol and should be agreed upon with regulatory authorities before the trial begins. In this chapter we describe between-group differences in general terms. It is acceptable to calculate the difference in two quantities, A and B as "A minus B" or "B minus A" as long as the procedure chosen is identified unambiguously.

10.4 A brief review of hypothesis testing

We discussed hypothesis testing in some detail in Chapter 6. For present purposes, the role of hypothesis testing in confirmatory clinical trials can be restated simply as follows:

> Hypothesis testing provides an objective way to make a decision to proceed as if the drug is either effective or not effective based on the sample data, while also limiting the probability of making either decision in error.

For a superiority trial the null hypothesis is that the treatment effect is zero. Sponsors of drug trials would like to generate sufficient evidence, in the form of the test statistic, to reject the null hypothesis in favor of the alternate hypothesis, thereby providing compelling evidence that the treatment effect is not zero. The null hypothesis may be rejected if the treatment effect favors the test drug, and also if it favors the placebo (as discussed, we have to acknowledge this possibility).

The decision to reject the null hypothesis depends on the value of the test statistic relative to the distribution of its values under the null hypothesis. Rejection of the null hypothesis means one of two things:

1. There really is a difference between the two treatments, that is, the alternate hypothesis is true.
2. An unusually rare event has occurred, that is, a type I error has been committed, meaning that we reject the null hypothesis given that it is true.

Regulatory authorities have many reasons to be concerned about type I errors. As a review at the end of this chapter, the reader is encouraged to think about the implications for a pharmaceutical company of committing a type I or II error at the conclusion of a confirmatory efficacy study.

The test statistic is dependent on the analysis method, which is dependent on the study design; this, in turn, is dependent on a precisely stated research question. By now, you have seen us state this fundamental point several times, but it really cannot be emphasized enough. In our experience, especially with unplanned data analyses, researchers can be so anxious to know "What's the p value?" that they forget to consider the possibility that **the study that generated the data was not adequately designed to answer the specific question of interest**. The steps that lead toward optimally informed decision-making in confirmatory trials on the basis of hypothesis testing are as follows:

1. State the research question.
2. Formulate the research question in the form of null and alternate statistical hypotheses.
3. Design the study to minimize bias, maximize precision, and limit the chance of committing a type I or II error. As part of the study design, prespecify the primary analysis method that will be used to test the hypothesis. Depending on the nature of the data and the size of the study, consider whether a parametric or nonparametric approach is appropriate.
4. Collect optimum-quality data using optimum-quality experimental methodology.
5. Carry out the primary statistical analysis using the prespecified method.
6. Report the results of the primary statistical analysis.
7. Make a decision to proceed as if the drug is either effective or ineffective:
   (a) If you decide that it is effective based on the results of this study, you may choose to move on to conduct the next study in your clinical development plan, or, if this is the final study in your development plan, to submit a dossier (for example, NDA [new drug application], MAA [marketing authorisation application]) to a regulatory agency.
   (b) If you decide that it is ineffective based on the results of this study, you may choose to refine the original research question and conduct a new study, or to abandon the development of this investigational new drug.

10.5 Hypothesis tests for two or more proportions

The research question of interest in some studies can be phrased: Does the test treatment result in a higher probability of attaining a desired state than the control? Examples of such applications include:

- survival after 1 year following a cardiovascular intervention
- avoiding hospitalization associated with asthmatic exacerbations over the course of 6 months
- attaining a specific targeted level of LDL according to one's background risk.

In a confirmatory trial of an antihypertensive, for example, a sponsor might like to know if the test treatment results in a higher proportion of hypertensive individuals (which can be interpreted as a probability) reaching an SBP < 140 mmHg.

10.5.1 Hypothesis test for two proportions: The *Z* approximation

In the case of a hypothesis test for two proportions the null and alternate statistical hypotheses can be stated as follows:

$$H_0: p_1 - p_2 = 0$$
$$H_A: p_1 - p_2 \neq 0$$

where the population proportions for each of two independent groups are represented by p_1 and p_2.

The sample proportions will be used to estimate the population proportions and, as in Chapter 8, are defined as:

$$\hat{p}_1 = \frac{\text{number of observations in group 1 with the event of interest}}{\text{total number of observations in group 1 at risk of the event}}$$

and

$$\hat{p}_2 = \frac{\text{number of observations in group 2 with the event of interest}}{\text{total number of observations in group 2 at risk of the event}}.$$

The estimator for the difference in the two sample proportions is $\hat{p}_1 - \hat{p}_2$ and the standard error of the difference $\hat{p}_1 - \hat{p}_2$ is:

$$\text{SE}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}},$$

where $\hat{q}_1 = 1 - \hat{p}_1$ and $\hat{q}_2 = 1 - \hat{p}_2$. The test statistic for the test of two proportions is equal to:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\text{SE}(\hat{p}_1 - \hat{p}_2)}.$$

Use of a correction factor may be useful as well, especially with smaller sample sizes. A test statistic that makes use of the correction factor is:

$$Z = \frac{|\hat{p}_1 - \hat{p}_2| - \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\text{SE}(\hat{p}_1 - \hat{p}_2)}.$$

For large samples (that is, when $\hat{p}_1 n_1 \geq 5$ and $\hat{p}_2 n_2 \geq 5$), these test statistics follow a standard normal distribution under the null hypothesis. Values of the test statistic that are far away from zero would contradict the null hypothesis and lead to 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 $Z < Z_{\alpha/2}$ or $Z > Z_{1-\alpha/2}$. 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 an illustration of this hypothesis test, consider the following hypothetical data from a confirmatory study of a new antihypertensive. In a randomized, double-blind, 12-week study, the test treatment was compared with placebo. The primary endpoint of the study was the proportion of participants who attained an SBP goal < 140 mmHg. Of 146 participants assigned to placebo, 34 attained an SBP < 140 mmHg at week 12. Of 154 assigned to test treatment, 82 attained the goal. Let us look at how these results can help us to make a decision based on the

information provided. We go through the steps needed to do this.

The research question

Is the test treatment associated with a higher rate of achieving target SBP?

Study design

As noted, the study is a randomized, double-blind, placebo-controlled, 12-week study of an investigational antihypertensive drug.

Data

The data from this study are in the form of counts. We have a count of the number of participants in each treatment group, and, for both of these groups, we have a count of the number of participants who experienced the event of interest. As the research question pertains to a probability, or risk, we use the count data to estimate the probability of a proportion of participants attaining the goal SBP.

Hypotheses and statistical analysis

The null and alternate statistical hypotheses in this case can be stated as:

$$H_0: p_{TEST} - p_{PLACEBO} = 0$$
$$H_A: p_{TEST} - p_{PLACEBO} \neq 0$$

where the population proportions for each group are represented by p_{TEST} and $p_{PLACEBO}$. As the response is attaining a lower SBP, the group with the greater proportion of responses will be regarded as the treatment with a more favorable response. The difference in proportions is calculated as "test minus placebo." Positive values of the test statistic will favor the test treatment.

As the samples are large according to the definition given earlier, the test of the two proportions using the Z approximation is appropriate. For a two-sided test of size 0.05 the critical region is defined by $Z < -1.96$ or $Z > 1.96$. The value of the test statistic is calculated as:

$$Z = \frac{\hat{p}_{TEST} - \hat{p}_{PLACEBO}}{SE(\hat{p}_{TEST} - \hat{p}_{PLACEBO})}.$$

The difference in sample proportions is calculated as:

$$\hat{p}_{TEST} - \hat{p}_{PLACEBO} = \frac{82}{154} - \frac{34}{146} = 0.5325 - 0.2329 = 0.2996.$$

The standard error of the difference in sample proportions is calculated as:

$$SE(\hat{p}_{TEST} - \hat{p}_{PLACEBO}) = \sqrt{\frac{(0.5325)(0.4675)}{154} + \frac{(0.2329)(0.7671)}{146}} = 0.0533.$$

Using these calculated values, the value of the test statistic is:

$$Z = \frac{0.2996}{0.0533} = 5.62.$$

The test statistic using a correction factor is obtained as:

$$Z = \frac{0.2996 - \frac{1}{2}\left(\frac{1}{154} + \frac{1}{146}\right)}{0.0533} = 5.50.$$

Interpretation and decision-making

As the value of test statistic – that is, 5.62 – is in the critical region ($5.62 > 1.96$), the null hypothesis is rejected in favor of the alternate hypothesis. Note that the value of the test statistic using the correction factor was also in the critical region. The probability of attaining the SBP goal is greater for those receiving the test treatment than for those receiving placebo.

It is fairly common to report a p value from such an analysis. As we have seen, the p value is the probability (under the null hypothesis) of observing the result obtained or one that is more extreme. In this analytical strategy we refer to a table of Z scores and the tail areas associated with each to find the sum of the two areas (that is, probabilities) to the left of -5.62 (a result as extreme as the observed or more so) and to the right of 5.62 (the result observed and those more extreme). A Z score of this magnitude is way out in the right-hand tail of the distribution, leading to a p value < 0.0001.

The results of this study may lead the sponsor to decide to conduct a second confirmatory trial, being confident that the drug is efficacious. Alternately, if the entire set of clinical data are

satisfactory, the sponsor may decide to apply for marketing approval.

10.5.2 Hypothesis test for two (or more) proportions: χ^2 test of homogeneity

An alternative method to the Z approximation for the comparison of two proportions from independent groups is called the χ^2 test, which is considered a goodness-of-fit test; this quantifies the extent to which count data (for example, the number of individuals with and without the response of interest) deviate from counts that would be expected under a particular mathematical model. The mathematical model used in clinical studies for goodness-of-fit tests is that of homogeneity. That is, if a particular response is homogeneous with respect to treatment, we would expect all the responses of interest to be proportionally distributed among all treatment groups. The assumption of homogeneity will allow us to calculate the cell counts that would be expected. These will then be compared with what was actually observed. The more the expected counts under the particular model of interest (for example, homogeneity) deviate from what is observed, the greater the value of the test statistic, and therefore the more the data do not represent goodness of fit. The χ^2 test is useful because it can be used to test homogeneity across two or more treatment groups. We first describe the case of two groups and the more general case is described in Section 10.5.3.

If there are two independent groups of interest (for example, treatment groups in a clinical trial) each representing an appropriate population, the proportions of participants with the characteristic or event of interest are represented by $\hat{p}_1 = m_1/n_1$ and $\hat{p}_2 = m_2/n_2$. The counts of participants with events and nonevents can be displayed in a contingency table with two columns and two rows, representing the numbers of observations with (m_1 and m_2) and without ($n_1 - m_1$ and $n_2 - m_2$) the characteristic of interest. The marginal total of individuals with events (the sum across the two groups) is denoted by $R = m_1 + m_2$. The marginal total of individuals without the events (sum across the two groups) is denoted by $S = (n_1 + n_2) - (m_1 + m_2)$. Finally, the total sample size (sum across the two groups) is denoted by $N = n_1 + n_2$. The overall proportion of responses of interest across both groups is $\hat{p} = R/N$. The complementary proportion of responses is $\hat{q} = S/N$. A sample contingency table displaying the observed counts is represented in Table 10.2.

Table 10.2 Sample contingency table for two groups and two responses (2 × 2)

	Group		
Event or characteristic?	1	2	Total
Yes	m_1	m_2	R
No	$n_1 - m_1$	$n_2 - m_2$	S
	n_1	n_2	N

The null hypothesis for the χ^2 test of homogeneity for two groups is stated as:

> H_0: The distribution of the response of interest is homogeneous with respect to the two treatment groups. Equivalently, the proportion of "yes" responses is equal across the two groups.

The alternate hypothesis is:

> H_A: The distribution of the response of interest is not homogeneous with respect to the two treatment groups.

If the null hypothesis is true – that is, the proportion of participants with the event of interest is similar across the two groups – the expected count of responses in groups 1 and 2 would be in the same proportion as observed across all groups. That is, the expected cell count in row 1 (participants with events of interest) for group 1 is:

$$E_{1,1} = \hat{p}n_1.$$

Likewise, the expected cell count in row 1 (participants with events of interest) for group 2 is:

$$E_{1,2} = \hat{p}n_2.$$

Similarly, the expected cell count in row 2 (participants without the event of interest) for group 1 is:

$$E_{2,1} = \hat{q}n_1.$$

Lastly, the expected cell count in row 2 (participants without the event of interest) for group 2 is:

$$E_{2,2} = \hat{q} n_2.$$

The corresponding observed counts in Table 10.2 are:

$$O_{1,1} = m_1,$$
$$O_{1,2} = m_2,$$
$$O_{2,1} = n_1 - m_1,$$

and

$$O_{2,2} = n_2 - m_2.$$

The test statistic χ^2 is calculated as the sum of squared differences between the observed and expected counts divided by the expected count for all four cells (two groups and two responses) of the contingency table:

$$X^2 = \sum_{i=1}^{2} \sum_{r=1}^{2} \frac{(O_{r,i} - E_{r,i})^2}{E_{r,i}}.$$

Under the null hypothesis of homogeneity, the test statistic, X^2, for two groups and two responses (for example, interest is in the proportion) is approximately distributed as a χ^2 with 1 degree of freedom (df). Only large values of the test statistic are indicative of a departure from the null hypothesis. Therefore, the χ^2 test is implicitly a one-sided test. Values of the test statistic that lie in the critical region are those with $X^2 > \chi_1^2$.

The notation in this section tends to be more complex than we have encountered in previous chapters. A worked example using the data from Section 10.5.1 may clarify the description. In a randomized, double-blind, 12-week study, the test treatment was compared with placebo. The primary endpoint of the study was the proportion of participants who attained an SBP goal $<$ 140 mmHg. Of 146 participants assigned to placebo, 34 attained an SBP $<$ 140 mmHg at week 12. Of 154 assigned to test treatment, 82 attained the goal.

The research question

Are participants who take the test treatment more likely than placebo participants to attain their SBP goal?

Study design

The study is a randomized, double-blind, placebo-controlled, 12-week study of an investigational antihypertensive drug.

Data

The data from the study are represented as the contingency table displayed in Table 10.3.

Table 10.3 Contingency table for individuals attaining goal SBP

Attained SBP $<$ 140?	Placebo	Test	Total
Yes	34	82	116
No	112	72	184
	146	154	300

Statistical analysis

The null and alternate statistical hypotheses can be stated as follows:

H_0: The proportion of individuals who attained SBP $<$ 140 mmHg is homogeneous (equal) across the two treatment groups.

H_A: The proportion of individuals who attained SBP $<$ 140 mmHg is not homogeneous across the two treatment groups.

In cases where there are only two categories, such as in this one, we need to know only how many individuals are in the "yes" row, because the number in the "no" row can be obtained by subtraction from the sample size within each group.

To calculate the test statistic, we first need to know the expected cell counts. These can be calculated as the product of the marginal row total and the marginal column total divided by the total sample size. The expected cell counts under the null hypothesis of homogeneity are provided in Table 10.4. The expected cell count for the placebo group in the first row ("Yes") was calculated as: (146)(116)/300 = 56.453. The expected cell count for the test treatment group in the second row ("No") was calculated as: (154)(184)/300 = 94.453. You are encouraged to

verify the remaining two cell counts using the same methodology.

Table 10.4 Expected cell counts for χ^2 test of homogeneity

Attained SBP < 140?	Placebo	Test	Total
Yes	56.453	59.547	116
No	89.547	94.453	184
	146	154	300

Now that we have calculated the expected cell counts, we can calculate the test statistic using these expected cell counts in conjunction with the observed cell counts:

$$X^2 = \frac{(34 - 56.453)^2}{56.453} + \frac{(82 - 59.547)^2}{59.547} + \frac{(112 - 89.547)^2}{89.547} + \frac{(72 - 94.453)^2}{94.453}$$
$$= 28.3646$$

Tabled values to determine critical regions are not as concise as those for the standard normal distribution, because there is not just one χ^2 distribution but many of them. However, the χ^2 distribution with 1 df is quite frequently encountered as 2×2 contingency tables. Hence, for reference, values of the χ^2 distribution for 1 df that cut off various areas in the right-hand tail are provided in Table 10.5. Additional values of $\chi^2_{1-\alpha}$ are provided in Appendix 3.

Table 10.5 Critical values for the χ^2 distribution with 1 degree of freedom

α (one sided)	$\chi^2_{(1-\alpha),1}$
0.10	2.706
0.05	3.841
0.01	6.635
0.001	10.38

For a test of size 0.05 the value of the test statistic, 28.3646, is much greater than the critical value of 3.841.

Interpretation and decision-making

Just as the hypothesis test using the Z approximation resulted in a rejection of the null hypothesis, so does this χ^2 test. We can also tell from the critical values in Table 10.5 that the p value must be < 0.001 because less than 0.001 of the area under the 1 df χ^2 distribution lies to the right of the value 10.38 and the calculated test statistic, 28.3646 lies to the right of that value.

10.5.2.1 Odds ratio as a measure of association from 2 × 2 contingency tables

Many articles published in medical journals cite a measure of association called an odds ratio, which is an estimate of the relative risk of the event or outcome of interest, a concept that was introduced in Chapter 8. If the probability of an outcome of interest for group 1 is estimated as $\hat{p}_1$ the odds of the event are:

$$\text{Odds of the event for group 1} = \frac{\hat{p}_1}{1 - \hat{p}_1}.$$

Similarly:

$$\text{Odds of the event for group 2} = \frac{\hat{p}_2}{1 - \hat{p}_2}.$$

Then the estimated odds ratio is calculated as:

$$\text{Odds ratio} = \frac{\hat{p}_1(1 - \hat{p}_2)}{\hat{p}_2(1 - \hat{p}_1)}.$$

Note that an equivalent definition of the odds ratio using the observed counts from the 2×2 contingency table in Section 10.5.2 is:

$$\text{Odds ratio} = \frac{O_{1,1}O_{2,2}}{O_{1,2}O_{2,1}}.$$

A standard error may be calculated for purposes of constructing a confidence interval for the odds ratio, but it requires an iterative solution. Statistical software is useful for this purpose. Interested readers will find a wealth of information on the odds ratio in Fleiss et al. (2003).

If the estimated probabilities of the event are the same (or similar) between the two groups, the odds ratio will have a value around 1 (unity). Thus an assumption of no association in a 2×2 table implies that the odds ratio is equal to 1.

This also means that the χ^2 test for binary outcomes from Section 10.5.2 can be considered a test of the null hypothesis that the population odds ratio = 1. Values of the odds ratio appropriately < 1 or appropriately > 1 are suggestive of an association between the group and the outcome.

Using the data from Table 10.3 as presented and using the formula for observed cell counts, the estimated odds ratio is calculated as:

$$\text{Odds ratio} = \frac{(34)(72)}{(112)(82)} = 0.27.$$

Interpreting this value as an estimate of the relative risk of attaining the target SBP level, we would say that patients treated with placebo are 0.27 times as likely as patients treated with the active drug to attain the SBP goal. This statement may seem awkward (we would not disagree), which points out a potentially difficult aspect of the odds ratio. As the name implies it is a ratio scaled quantity so the odds ratio can be expressed as *a/b* or *b/a*. Keeping in mind that the odds ratio is an estimate of the relative risk, selecting the more appropriate method will aid the clinical interpretation of the result. In this case the response of interest is a favorable outcome, so a relative risk > 1 would imply that a favorable outcome was more likely after treatment with the active drug than the placebo. Similarly, if the response of interest is a bad outcome (for example, serious adverse event) a relative risk < 1 would suggest that the probability of a bad outcome was less for the active drug than the placebo.

Hence we can make more sense of this calculated value by taking its inverse as 1/0.27 = 3.75. This expression is more appealing and an accurate interpretation in that patients treated with the test drug are 3.75 times more likely to attain the SBP goal than those treated with placebo. One can also obtain this result by switching the order of the columns in Table 10.3 and performing the calculation as:

$$\text{Odds ratio} = \frac{(82)(112)}{(72)(34)} = 3.75.$$

Odds ratios are one of the most common statistics cited from logistic regression analyses. Logistic regression is an advanced topic and therefore not included in this book. An overly simple description is that it is an analysis method by which binary outcomes are modeled (or explained) using various predictor variables. The proper interpretation of odds ratios from logistic regression models will depend on the way in which the predictors were used in the statistical model. However, the general concept is the same as in this example. The odds ratio represents the relative increase in risk of a particular event for one group versus another. We recommend two excellent texts on logistic regression by Hosmer and Lemeshow (2000) and Kleinbaum and Klein (2002).

10.5.2.2 Use of the χ^2 test for two proportions

The χ^2 test of homogeneity is useful for comparing two proportions under the following circumstances:

- The groups need to be independent.
- The responses need to be mutually exclusive.
- The expected cell counts are reasonably sized.

With regard to the last of these requirements, we need to operationally define "reasonably sized." A commonly accepted guideline is that the χ^2 test is appropriate when at least 80% of the cells have expected counts of at least five. In the case of the worked example, the use of the χ^2 test is appropriate on the basis of independence (no participant was treated with both placebo and test treatment) and sample size. If a participant can be counted in only one response category the responses are considered mutually exclusive or non-overlapping, as was the case here.

The χ^2 test of homogeneity is a special case because it can be used for any number of groups. The more general case is discussed in the following section.

10.5.3 Hypothesis test for *g* proportions: χ^2 test of homogeneity

If there are *g* independent groups of interest (for example, treatment groups in a clinical trial) each representing relevant populations, the

proportion of individuals with the characteristic or event of interest is represented by:

$$\hat{p}_i = \frac{m_i}{n_i}$$

for $i = 1, 2, \ldots g$, where g represents the number of groups. The counts of individuals with events and nonevents can be displayed in a contingency table with g columns and two rows representing the numbers of observations with (m_i) and without $(n_i - m_i)$ the characteristic of interest. The marginal total of individuals with events (the sum across the g groups) is denoted by:

$$R = \sum_{i=1}^{g} m_i .$$

The marginal total of individuals without the events (sum across the g groups) is denoted by:

$$S = \sum_{i=1}^{g} n_i - m_i .$$

Finally, the total sample size (sum across the g groups) is denoted by:

$$N = \sum_{i=1}^{g} n_i .$$

The overall proportion of responses across all groups is:

$$\hat{p} = \frac{R}{N}.$$

A sample contingency table displaying the observed counts in this more general case is represented in Table 10.6.

As before with the case of two groups, the null hypothesis is stated as:

> H_0: The distribution of the response of interest is homogeneous with respect to the g treatment groups. Equivalently, the proportion of "yes" responses is equal across all g groups.

The alternate hypothesis is:

> H_A: The distribution of the response of interest is not homogeneous with respect to the g treatment groups.

If the null hypothesis is true, that is, the proportion of individuals with the event of interest is similar across the groups, the expected count of responses in group i will be in the same proportion as observed across all groups. That is, the expected cell count in row 1 (individuals with events of interest) for group i is:

$$E_{1,i} = \hat{p} n_i .$$

Similarly, the expected cell count in row 2 (individuals without the event of interest) for group i is:

$$E_{2,i} = \hat{q} n_i.$$

The expected cell counts are calculated in this manner for all $2g$ cells of the contingency table. The corresponding observed counts for groups $i = 1, 2, \ldots, g$, in Table 10.6 are:

$$O_{1,i} = m_i$$

and

$$O_{2,i} = n_i - m_i.$$

The test statistic X^2 is calculated as the sum of squared differences between the observed and expected counts divided by the expected count

Table 10.6 Sample contingency table for g groups and two responses ($g \times 2$)

	Group				
Event or characteristic?	1	2	. . .	g	Total
Yes	m_1	m_2	. . .	m_g	R
No	$n_1 - m_1$	$n_2 - m_2$	. . .	$n_g - m_g$	S
	n_1	n_2	. . .	n_g	N

for all $2g$ cells (g groups and 2 responses) of the contingency table:

$$X^2 = \sum_{i=1}^{g} \sum_{r=1}^{2} \frac{(O_{i,g} - E_{i,g})^2}{E_{i,g}}.$$

Under the null hypothesis of homogeneity, the test statistic, X^2, for g groups and two responses is approximately distributed as a χ^2 with $(g-1)$ df. Values of the test statistic that lie in the critical region are those with $X^2 > \chi^2_{g-1}$.

10.5.4 Hypothesis test for *r* responses from *g* groups

The χ^2 test can be applied to more general situations, including data with r response levels and g independent groups. When there are more than two response categories, however, the null and alternate hypotheses cannot be stated simply in terms of one proportion, but need to be stated in terms of the distribution of response categories.

One example containing more than two groups would be an evaluation of the following three categories of response: Worsening, no change, and improvement. It would not be sufficient to state the null hypothesis in terms of the proportion of individuals with a response of worsening because there are two other responses of interest. We highlight this point because the χ^2 test is used extensively in clinical research, and it can be correctly applied to multilevel responses and multiple groups. If we use the more general terminology, "distribution of responses is homogeneous with respect to treatment group," we are always correct no matter how many responses there were or how many groups.

The specific methodology associated with these more general cases is beyond the scope of our text. The most appropriate and efficient analyses of data of this type can depend on the hypothesis of interest and whether or not the response categories are ordered. Additional details can be found in two excellent texts by Stokes et al. (2001) and Agresti (2007).

10.5.5 Hypothesis test for two proportions: Fisher's exact test

The two methods described earlier, the Z approximation and the χ^2 test of homogeneity, are appropriate when the sample sizes are large enough. There are times, however, when the sample sizes in each group are not large enough or the proportion of events is low such that $n\hat{p} < 5$. In such cases another analysis method, one that does not require any approximation, is appropriate.

An alternate hypothesis test for two proportions is attributed to Fisher. Fisher's exact test is applicable to contingency tables with two or more responses in two or more independent groups. We consider one case, 2×2 tables, represented by counts of individuals with and without the characteristic of interest (two rows) in each of two treatment groups (two columns), for which the cell counts are small. For this test the row and column marginal totals are considered fixed. That is, one assumes that the total number of individuals with events is fixed as well as the number in each group. The extent to which the two groups are similar or dissimilar accounts for the distribution of events between the two groups. For any 2×2 table, the probability of the particular distribution of response counts, assuming the fixed marginal totals, can be calculated exactly via something called the hypergeometric distribution (we do not go into details here). Using slightly different notation from the examples above, the cell counts and marginal totals of a general 2×2 table are displayed in Table 10.7. The total number of

Table 10.7 Cell counts and marginal totals from a general 2×2 table

Event or characteristic of interest?	Group 1	Group 2	Total
Yes	Y_1	Y_2	$Y_\cdot$
No	N_1	N_2	$N_\cdot$
	n_1	n_2	n

"yes" responses is denoted by the symbol, $Y_.$, where the dot in the index means that the count is obtained by summing the responses over the two columns, that is, $Y_1 + Y_2$. Likewise, the total number of "no" responses is denoted by the symbol, $N_.$, the sum over groups 1 and 2.

Given the fixed margins as indicated in Table 10.7, the probability of the distribution of responses in the 2×2 table is calculated from the hypergeometric distribution as:

$$P(Y_1, Y_2, N_1, N_2 \mid Y_., N_., n_1, n_2, n) = \frac{Y_.!N_.!n_1!n_2!}{n!Y_1!Y_2!N_1!N_2!}.$$

The null and alternate hypotheses in this case are as follows:

H_0: The proportion of responses is independent of the group.
H_A: The proportion of responses is not independent of the group.

If the null hypothesis is rejected, the alternate hypothesis is better supported by the data.

For this test there is no test statistic as such, because this test is considered an exact test. Therefore, we need not compare the value of a test statistic to a distribution. Instead, the p value is calculated directly and compared with the predefined α level. Recall that a p value is the probability, under the null hypothesis, of observing the obtained results or those more extreme, that is, results contradicting the null hypothesis. The calculation of the p value for this exact test entails the following three steps:

1. Calculate the probability of the observed cell counts using the expression above.
2. For all other permutations of 2×2 tables with the same marginal totals, calculate the probability of observed cell counts in a similar manner.
3. Calculate the p value as the sum of the observed probability (from the first step) and all probabilities for other permutations that are less than the probability for the observed table.

As a consequence, the p value represents the likelihood of observing, by chance alone, the actual result or those more extreme. The calculated p value is compared with the value of α and we either reject or fail to reject the null hypothesis.

As an example of Fisher's exact test, we consider other data from the antihypertensive trial introduced in Section 10.5.1. These data are presented in Table 10.8.

Table 10.8 Contingency table for individuals attaining SBP < 120 mmHg

Attained SBP < 120?	Placebo	Test	Total
Yes	1	3	4
No	145	151	296
	146	154	300

The research question

Is there sufficient evidence at the $\alpha = 0.05$ level to conclude that the probability of attaining a SBP < 120 mmHg (a remarkable response for a hypertensive person!) is greater for people receiving the test treatment than for those receiving the placebo?

Study design

The study is a randomized, double-blind, placebo-controlled, 12-week study of an investigational antihypertensive drug.

Data

The data from the study are represented as a contingency table as displayed in Table 10.8. As seen in Table 10.8, only four individuals had the event of interest. Neither the Z approximation nor the χ^2 test would be appropriate given the small cell sizes of one and three.

Statistical analysis

The null and alternate statistical hypotheses can be stated as:

H_0: The proportion of individuals who attained SBP < 120 mmHg is independent of treatment group.
H_A: The proportion of participants who attained SBP < 120 mmHg is not independent of treatment group.

In this instance, independence means that the probability of the response is no more or less likely for one group versus the other. In his original paper, Fisher stated the null hypothesis slightly differently (although equivalent mathematically). The null hypothesis, after Fisher, can be stated in this form: The population odds ratio of response to nonresponse for one group versus the other is equal to one.

In Figure 10.1 all of the possible permutations of cell counts, given the marginal totals, are displayed. To be concise, the row and column labels are not included. The calculated probability from the hypergeometric distribution is provided to the right of each arrangement of cell counts. The probabilities in Figure 10.1 are included to illustrate the calculation. Note that by definition, 0! = 1. For this particular dataset it is manageable to calculate each probability with a calculator, but in many instances this particular test should be done using statistical software. When calculating these probabilities by hand it is helpful to re-write the factorial expressions in a way so that numerator and denominator terms "cancel out." For example, writing 154! as 154*153*152*151! allows us to cancel 151! from the numerator and denominator of the probability associated with the observed result.

The calculated p value is the probability from the observed result plus all probabilities less than the probability associated with the observed result. For this example the exact p value is:

$$p \text{ value} = 0.263453 + 0.236537 + 0.068119 + 0.054910 = 0.623019.$$

Rounding to three significant digits, this can be expressed as p value = 0.623.

Interpretation and decision-making

Comparing the p value of 0.623 to $\alpha = 0.05$, the statistical conclusion is not to reject the null hypothesis. There is insufficient evidence to conclude that the alternate hypothesis is true. If the goal of a new antihypertensive therapy were to

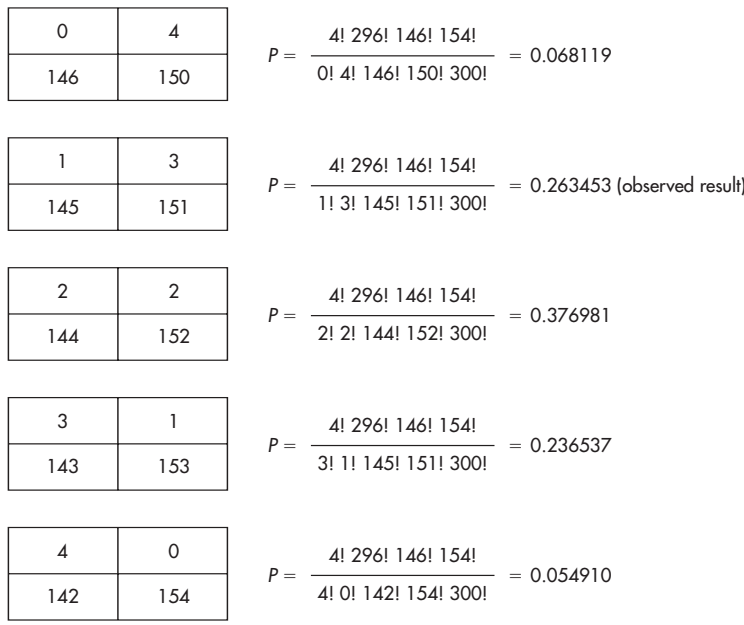

Figure 10.1 All permutations of response counts given fixed marginal totals and probabilities of each

reduce SBP to levels < 120 mmHg, such a result would be disappointing and may lead to a decision to halt the clinical development program. However, the study was not designed to answer such a question. In fact, the research question, having been formulated as an exploratory analysis, may not be well suited for the study that was actually conducted. Perhaps a greater dose or more frequent administration of the investigational antihypertensive drug would increase the rate of the desired response. In any case, as the analysis earlier in the chapter illustrated, the new drug does seem to lower SBP to levels that would be considered clinically important (< 140 mmHg).

10.5.6 Test of two proportions from stratified samples: The Mantel–Haenszel method

Confirmatory efficacy studies typically involve a number of investigative centers and, accordingly, are known as multicenter trials. Multicenter trials have a number of benefits, which are discussed later. A common analysis method used in multicenter trials is to account for differences from center to center by including them in the analysis. Stratifying the randomization to treatment assignment by investigative center ensures that there are approximately equal numbers of participants assigned to test or placebo within each center. Analyses from studies with this design typically account for center as it is conceivably another source of variation. This is accomplished by calculating a summary test statistic within each center and then pooling or calculating weighted averages of the within-center statistics across all centers, thereby removing the effect of the centers from the overall test statistic.

The weights used in the analysis are chosen at the trial statistician's discretion, which provides a good example of the "art" of Statistics, because the statistician must make a well-informed judgment call. Some commonly employed choices of weights are as follows:

- equal weights for all centers
- weights proportional to the size of the center
- weights that are related to the standard error of the within-center statistic (for example, more precise estimates have more weight than less precise estimates).

One method applicable to the difference of two proportions, originally described by Mantel and Haenszel (1959) and well described by Fleiss et al. (2003), utilizes weights that are proportional to the size of each stratum (in this case, centers) to calculate a test statistic that follows approximately a χ^2 distribution.

Assume that there are h strata of interest, and within each of the strata ($h = 1, 2, \ldots, H$) there are n_{h1} observations for group 1 (for example, treatment group 1) and n_{h2} observations for group 2 (for example, treatment group 2). The proportion of observations with the characteristic of interest within each stratum for the two groups is denoted by $\hat{p}_{h1}$ and $\hat{p}_{h2}$, respectively. The overall proportion of participants with the characteristic of interest within each stratum is denoted by $\bar{p}_h$; the overall proportion without the characteristic of interest with each stratum is denoted by $\bar{q}_h = 1 - \bar{p}_h$.

The null hypothesis tested by the Mantel–Haenszel method is as follows:

> H_0: There is no overall association between response and group after accounting for the stratification factor.

If the null hypothesis is rejected, the data favor the following alternate hypothesis:

> H_A: There is an overall association between response and group after accounting for the stratification factor.

The test statistic for the Mantel–Haenszel method is:

$$X^2_{MH} = \frac{\left(\left| \sum_{h=1}^{H} \frac{n_{h1}\, n_{h2}}{n_h} (\hat{p}_{h1} - \hat{p}_{h2}) \right| - 0.5 \right)^2}{\sum_{h=1}^{H} \frac{n_{h1}\, n_{h2}}{n_h - 1} \bar{p}_h \bar{q}_h}.$$

Note that the differences in proportions, $\hat{p}_{h1} - \hat{p}_{h2}$, are weighted by the quantities $\frac{n_{h1} n_{h2}}{n_h}$.

This test statistic utilizes a continuity correction factor of 0.5 as well. As described by Fleiss et al. (2003), the test performs well when expected cell counts within each of H 2×2 tables differ by at

least 5 (maximum – minimum). The test statistic that is computed in this manner is approximately distributed as a χ^2 with 1 df.

A similar test statistic, Cochran's statistic, originally attributed to Cochran (1954), is described by Fleiss et al. (2003):

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

Note that Cochran's statistic does not use a correction factor and the denominator of the stratum weights is n_h instead of $(n_h - 1)$. We mention Cochran's statistic because it is used by some statistical software packages instead of the Mantel–Haenszel statistic. Fleiss points out that the difference between the Mantel–Haenszel statistic and Cochran's statistic is small when the sample sizes are large, but considerable when the sample sizes within each of the strata are small.

As an illustration of the Mantel–Haenszel method, we take the data from our example as detailed in Section 10.5.1 and separate them into data collected at each of three centers, which in this case represent the three strata.

The research question

Is there sufficient evidence at the $\alpha = 0.05$ level to conclude that the probability of attaining a goal SBP level is greater for individuals receiving test treatment than for those receiving the placebo after accounting for differences in response among centers?

Study design

The study is a randomized, double-blind, placebo-controlled, 12-week study of an investigational antihypertensive drug.

Data

The data from the study are represented as three contingency tables, one for each of the centers in Table 10.9.

Table 10.9 Contingency table for individuals attaining goal SBP by center

Center 1 Attained SBP < 140?	Placebo	Test	Total
Yes	12	24	36
No	34	21	55
	46	45	91
Center 2 **Attained SBP < 140?**	**Placebo**	**Test**	**Total**
Yes	15	31	46
No	29	19	48
	44	50	94
Center 3 **Attained SBP < 140?**	**Placebo**	**Test**	**Total**
Yes	7	27	34
No	49	32	81
	56	59	115
Overall **Attained SBP < 140?**	**Placebo**	**Test**	**Total**
Yes	34	82	116
No	112	72	184
	146	154	300

Statistical analysis

The null and alternate statistical hypotheses can be stated as:

> H_0: There is no overall association between the response (attaining SBP < 140 mmHg) and treatment group after accounting for center.
> H_A: There is an overall association between the response and treatment group after accounting for center.

For a test of size $\alpha = 0.05$, a χ^2 test with 1 df has a critical value of 3.841.

The differences in the proportions of interest (test minus placebo) are as follows:

- Center 1: (0.533 − 0.261) = 0.272
- Center 2: (0.620 − 0.341) = 0.279
- Center 3: (0.458 − 0.125) = 0.333.

The overall response rates for the event of interest and their complements are:

$$\text{Center 1: } \bar{p}_1 = \frac{36}{91} = 0.396 \text{ and } \bar{q}_1 = \frac{55}{91} = 0.604$$

$$\text{Center 2: } \bar{p}_2 = \frac{46}{94} = 0.489 \text{ and } \bar{q}_2 = \frac{48}{94} = 0.511$$

$$\text{Center 3: } \bar{p}_3 = \frac{34}{115} = 0.296 \text{ and } \bar{q}_3 = \frac{81}{115} = 0.704.$$

The test statistic is then computed as:

$$X^2_{MH} = \frac{\left|\left|\left(\frac{46 * 45}{91}\right)(0.272) + \left(\frac{44 * 50}{94}\right)(0.279) + \left(\frac{56 * 59}{115}\right)(0.333)\right| - 0.5\right|^2}{\left(\frac{46 * 45}{90}\right)(0.396)(0.604) + \left(\frac{44 * 50}{93}\right)(0.489)(0.511) + \left(\frac{56 * 59}{114}\right)(0.296)(0.704)}$$

$$= 27.21$$

Although the calculation details are not shown here, the value of Cochran's statistic for this example is 28.47, which is consistent with the result obtained for the Mantel–Haenszel statistic.

Interpretation and decision-making

The value of the test statistic is much greater than the critical value of 3.841. Hence the statistical decision is to reject the null hypothesis of no association after accounting for center differences. The proportion of responders is significantly higher among those receiving the test treatment. The p value associated with the test can be obtained from statistical software. However, we know from the sample of critical values in Table 10.5 that the p value must be < 0.001. As before, a pharmaceutical company would be encouraged by such results.

10.6 Concluding comments on hypothesis tests for categorical data

All of the methods described in this chapter are applicable to data that are in the form of "binary" events, that is, either the event or characteristic occurred for a given individual or it did not. For binary data, the summary statistic representing each treatment group is a sample proportion. To account for variation from sample to sample, hypothesis-testing methods allow a researcher to draw an inference about the underlying population difference in proportions. Although not covered in great detail, some of the methods can also be expanded to more than two categories.

In contrast, the methods described in Chapter 11 are applicable to data with outcomes that are continuous in nature. In those cases, other summary statistics are required to describe the typical effect in each group and the typical effect expected for the population under study.

10.7 Review

1. What constitutes "compelling evidence" of a beneficial treatment effect?
2. Consider a pharmaceutical company that has just completed a confirmatory efficacy study. What are the implications for the company of committing a type I error? What are the implications for the company of committing a type II error?
3. The equality of two proportions is being tested with the null hypothesis, H_0: $p_{TEST} - p_{PLACEBO} = 0$. Given that this is a two-sided test and using the following information, would the null hypothesis be rejected or not rejected?

   (a) $\alpha = 0.05$, Z approximation test statistic = 1.74
   (b) $\alpha = 0.10$, Z approximation test statistic = 1.74
   (c) $\alpha = 0.05$, Z approximation test statistic = 4.23
   (d) $\alpha = 0.01$, Z approximation test statistic = 4.23
   (e) $\alpha = 0.05$, χ^2 test statistic = 1.74
   (f) $\alpha = 0.10$, χ^2 test statistic = 1.74
   (g) $\alpha = 0.05$, χ^2 test statistic = 4.23
   (h) $\alpha = 0.01$, χ^2 test statistic = 4.23.

4. The equality of two proportions is being tested with the null hypothesis, H_0: $p_{TEST} - p_{PLACEBO} = 0$. Given that this is a two-sided test, what is the p value that corresponds to the following values of the Z approximation test statistic?

   (a) −1.56
   (b) −2.67

(c) 3.29
(d) 1.00.

5. The term "responders' analysis" was first introduced in Chapter 9 with regard to clinical laboratory data. A responders' analysis approach can be used in the context of efficacy data, as well. Consider a double-blind, placebo-controlled, therapeutic confirmatory trial of an investigational antihypertensive ("test drug"). Based on earlier experience, a period of 12 weeks is considered sufficient to observe a clinically meaningful treatment effect that can be sustained for many months. In this study, a participant whose SBP is reduced by at least 10 mmHg after 12 weeks of treatment is considered a responder. Similarly, a participant whose SBP is not reduced by at least 10 mmHg after 12 weeks is considered a non-responder. A total of 1000 participants were studied: 502 on placebo and 498 on test drug. Among the placebo participants, 117 were responders. Among those on the test drug, 152 were responders.

   (a) Summarize these results in a 2×2 contingency table.
   (b) The sponsor's research question of interest is: Are individuals treated with the test drug more likely to respond than those treated with placebo? What are the null and alternate statistical hypotheses corresponding to this research question?
   (c) What statistical tests may be used to test the null hypothesis? Are any more appropriate than others?
   (d) Is there sufficient evidence to reject the null hypothesis using a test of size $\alpha = 0.05$? Describe any assumptions necessary and show the calculation of the test statistic.
   (e) Calculate the odds ratio from the contingency table. What is the interpretation of the calculated odds ratio?

6. When would the Mantel–Haenszel χ^2 test be more useful than the χ^2 test?

10.8 References

Agresti A (2007). *An Introduction to Categorical Data Analysis*, 2nd edn. Chichester: John Wiley & Sons.

Cochran WG (1954). Some methods of strengthening the common χ^2 tests. *Biometrics* **10**:417–451.

Fisher LD (1999). One large, well-designed, multicenter study as an alternative to the usual FDA paradigm. *Drug Information J* **33**:265–271.

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

Fleming TR, DeMets DL (1996). Surrogate end points in clinical trials: are we being misled? *Ann Intern Med* **125**:605–613.

Hosmer DW, Lemeshow S (2000). *Applied Logistic Regression*, 2nd edn. Chichester: John Wiley & Sons.

ICH Guidance E8 (1997). *General Consideration of Clinical Trials*. Available at: www.ich.org (accessed July 1 2007).

ICH Guidance E9 (1998). *Statistical Principles for Clinical Trials*. Available at: www.ich.org (accessed July 1 2007).

Kleinbaum DG, Klein M (2002). *Logistic Regression: A self-learning text*, 2nd edn. New York: Springer.

Mantel N, Haenszel W (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Instit* **22**:719–748.

Stokes ME, Davis CS, Koch GG (2001). *Categorical Data Analysis using the SAS System*, 2nd edn. Chichester: Wiley & Sons.

US Department of Health and Human Services, Food and Drug Administration (1998). *Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products*. Available from www.fda.gov (accessed July 1 2007).