5

Data, central tendency, and variation

5.1 Introduction

Selecting an appropriate study design to best address the study objectives is just the first step towards answering the questions of interest. When most people think about Statistics they are probably thinking about data. Unfortunately, statisticians are not infrequently assigned the nickname "number crunchers," a name that accentuates the numerical aspects of the use of statistics but completely ignores the design, methodology, and interpretation aspects of the discipline of Statistics. Number crunching (and computational accuracy) is certainly a necessary component of Statistics, but it is important to bear in mind that it is far from sufficient.

Having reviewed the concepts of study design we now turn our attention to data. We are interested in various questions relating to data, such as: What are data? How might we classify different types of data? How are data used to answer questions arising during clinical trials? This last question is, perhaps, the most important one for this book. We start to answer it first in conceptual terms before turning our attention to more specific points.

5.2 Populations and samples

It is of considerable interest in new drug development to assess the effects of a drug in a particular population, the population containing individuals who may be prescribed the drug if and when it is approved. This population is known as the target population. Not all the adults in the USA and the UK would be ideal candidates for a therapeutic confirmatory trial because of the presence of other conditions or the use of other drugs, or for logistical reasons because they do not live close enough to a center that participates in clinical studies. Therefore, another population of interest is all adults in the USA and the UK who meet the specific eligibility criteria (including a precise definition of hypertensive) of a study. This group of individuals is considered the study population.

As study populations are often very large, however, it is not possible to administer the drug to every member of the population, so a sample from the study population is chosen and the effects of the drug in that sample are determined in a clinical study. In clinical trials, samples are typically considered or assumed to be simple random samples from the study population. A simple random sample is a sample in which each observational unit (for example, study participant) has the same probability of selection from the population. In other fields in which Statistics are used (most notably population surveys) samples need not be selected in this manner.

A clinical trial provides numerical statements of the drug's effects in the specific sample employed, but the investigator and the regulatory agency are really interested in the drug's (likely) effect in the whole population. Therefore, statistical procedures have been developed to allow numerical assessments of the likely effects in the study population based on the evidence collected from the sample that participated in the trial.

There are important limitations to the usefulness of generalizing the effects from a series of clinical trials to the patient population as a whole. The population from which clinical trial participants are sampled, the study population, may not truly be representative of the population (the target population) about which we would like to make conclusions. The target population may be sicker, have greater needs for concomitant medications, and have more chronic illnesses than the relatively homogeneous population from which the study sample arose. This point is well expressed by Senn (1997, p 28):

> In a clinical trial the primary formal objective is to assess what effects the treatments *did* have on the patients studied in order to say what effects they *may* have. To say what effects the treatments *will* have or even *will probably* have, requires arguments which go well beyond any formal examination of the data.

The following discussions address statistical methods that are applied to data from a sample of study participants with the objective of making an inference about the study population. By including relevant populations in studies and carefully documenting the methodology that gave rise to the study sample in regulatory documents and clinical communications, reviewers and physicians can judge for themselves the extent to which the results from the study can be inferred to the clinical situation.

5.3 Measurement scales

Data are anything that is measured. Examples of data encountered in clinical studies include height, weight, plasma concentration of a drug in a sample, days from the start of a study to a particular adverse event, the presence or absence of a characteristic of interest, and the gender of a study participant. Some of these examples may be surprising because we often think of data as numbers, but data may also be non-numeric.

Data can generally be classified into one of the following scales of measurement: nominal, ordinal, interval, or ratio.

5.3.1 Nominal scale

Nominal measurement scales involve names of characteristics. Characteristics frequently encountered in clinical studies that are measured on the nominal scale include gender (female or male), occurrence or not of an adverse event, a coded adverse event (for example, headache, asthenia, nausea), and race or ethnicity. Data measured on a nominal scale cannot be operated on arithmetically. We could not, for example, compare the values of females and males and come up with a meaningful result. An important caution is worth noting at this point. It is not uncommon to encounter data measured on the nominal scale to be represented as numbers or codes in electronic databases. An example would be when, in a database of a clinical study, the presence or absence of an adverse event (for example, headache) is represented as 0 (absent) or 1 (present). Before we undertake a statistical analysis of any sort it is necessary to understand fully the nature of the data.

5.3.2 Ordinal scale

This scale is best defined as one in which an ordering of values can be assigned. Examples of data from clinical studies measured on an ordinal scale include: severity of an adverse event classified as mild, moderate, or severe; age categorized as < 65, 65–70, 71–75, and > 75 years. The ordinal nature of the measurement scales means that we can say that a mild headache is less severe than a moderate headache, which is less severe than a severe headache. However, we cannot say that the difference between mild and moderate is the same as the difference between moderate and severe.

5.3.3 Interval scale

In contrast, differences between any two values measured on the interval scale do have meaning. Temperature measured on the Celsius or Fahrenheit scale is an example of an interval scale. For example, the difference between 32°F and 64°F is the same as the difference between 64°F and 96°F. On the interval scale, a value of zero is not a true zero (meaning absence of heat) because a value of -1°F is colder still. We can perform addition and subtraction on interval scaled data but, because the value of zero is meaningless, we cannot perform multiplication or division and obtain a meaningful result.

5.3.4 Ratio scale

Data measured on the ratio scale have all of the characteristics of interval scaled data with the exception that, in this case, a value of zero does represent a true zero. Height, which is measured on the ratio scale, has a true zero. A height of zero centimeters or inches means that there is no height. Likewise, a weight of zero kilograms or pounds means that there is no weight. An important characteristic of ratio scaled data is that the ratio of two values can be computed. For example, a study participant who weighs 220 pounds weighs twice as much as one who weighs 110 pounds.

The importance of identifying these scales of measurement is that not all statistical analysis approaches are appropriate for each of them. It is important to note that, although a particular characteristic may be measured on one scale, it may be reported using another. For example, age at the time of study entry may be measured on a ratio scale, but reported using an ordinal scale (for example, < 25, 25–64, > 64 years).

5.4 Random variables

Many individuals are involved in clinical trials and a number of characteristics of these participants are recorded. As characteristics such as age, systolic blood pressure, and gender can vary from individual to individual, they are generally classified as random variables (or, simply, variables). A common convention in statistics is to represent a particular random variable as a letter, such as x. A particular realization, or value, of a random variable for a particular individual (participant i in this case) is often denoted using a subscript such as x_i. We use these conventions in this chapter and throughout the text.

5.5 Displaying the frequency of values of a random variable

Since a random variable such as age can take on a number of values for a group of study participants it is of interest to know something about the relative frequency of each value. The relative frequency is the count of the number of observations with a specific value (for example, the number of 30-year-old participants) divided by the total number in the sample. An informative first step in a statistical analysis is to examine characteristics of the relative frequency of values of the random variable of interest, which can also be called the empirical distribution of the random variable. This knowledge is an essential part of selecting the most appropriate statistical analysis. Statistical software packages offer a number of methods to describe the relative frequency of values including tabular frequency displays, dot plots, relative frequency histograms, and stem-and-leaf plots.

An example of a frequency table is provided in Table 5.1, in which the frequency of age values in a sample of 100 study participants is displayed. The left-hand column is the value of age for which frequency information is provided. The column labeled “Frequency” is the count of the number of participants with the particular value of age. The column “Percentage” is the count of the number of participants with the particular value of age divided by the total number of observations in the sample and multiplied by 100 to express this figure as a percentage of the total. The next column “Cumulative frequency” represents the total count of age values less than or equal to the age value on a certain row. Similarly, “Cumulative percentage” is the cumulative frequency count of age values as a percentage of the total. As seen in Table 5.1 there is one 40-year-old individual (1% of the total) and there are five who are 40 and younger (5% of the total). A frequency table allows us to see how common all values are, but it can be difficult to see whether or not certain values tend to cluster together.

Another helpful way of displaying the relative frequency of observed values is to group values into equally spaced intervals and display the

Table 5.1 Frequency table of age values

Age (years)	Frequency	Percentage	Cumulative frequency	Cumulative percentage
31	1	1.00	1	1.00
35	1	1.00	2	2.00
39	2	2.00	4	4.00
40	1	1.00	5	5.00
43	3	3.00	8	8.00
45	1	1.00	9	9.00
48	2	2.00	11	11.00
49	4	4.00	15	15.00
50	4	4.00	19	19.00
51	2	2.00	21	21.00
52	2	2.00	23	23.00
53	2	2.00	25	25.00
55	3	3.00	28	28.00
57	3	3.00	31	31.00
58	3	3.00	34	34.00
59	5	5.00	39	39.00
60	2	2.00	41	41.00
61	6	6.00	47	47.00
62	4	4.00	51	51.00
63	2	2.00	53	53.00
64	1	1.00	54	54.00
65	1	1.00	55	55.00
66	3	3.00	58	58.00
67	4	4.00	62	62.00
68	3	3.00	65	65.00
69	2	2.00	67	67.00
70	3	3.00	70	70.00
71	4	4.00	74	74.00
72	3	3.00	77	77.00
73	4	4.00	81	81.00
74	1	1.00	82	82.00
75	3	3.00	85	85.00
76	1	1.00	86	86.00
77	3	3.00	89	89.00
78	3	3.00	92	92.00
79	1	1.00	93	93.00
80	2	2.00	95	95.00
81	2	2.00	97	97.00
82	1	1.00	98	98.00
83	1	1.00	99	99.00
88	1	1.00	100	100.00

resulting frequency in a histogram. There is no single width of each interval, or bin, that can be recommended. However, one might consider the quantity W as a starting point for the width:

$$W = \frac{\text{Maximum value} - \text{Minimum value}}{n}$$

It is typically desirable to have at least 5 bins and no more than 10, although less or more may be informative. Once the number and width of the bins have been determined the next step is to count the number of observations that fall into each interval and display the frequency of each grouping with contiguous bars. It is important that the intervals or bins are defined such that each observation can be assigned to only one interval. Using the 100 age values in the previous example, a histogram, displayed in Figure 5.1, has been constructed from the following categories: 30–39, 40–49, 50–59, 60–69, 70–79, 80–89. Note that each bar is centered over the interval midpoint. For example, the bar centered at 54.5 represents the relative frequency of age values in the interval 50–59.

By grouping the 100 age values (that is, the 100 participants in the indicated age groups) into categories, much of the detail evident in Table 5.1 has been lost. A display that retains the graphical nature of the histogram and the detail of the tabular frequency is a stem-and-leaf plot. A stem-and-leaf plot displays the first significant digit of the value of random variable as a "stem" and the subsequent significant digit as a "leaf." The stems are ordered from lowest to highest so that the relative frequency of each value can be surmised in one concise display. A stem-and-leaf plot of 100 individuals' age values is provided in Figure 5.2. To assist in your interpretation of this display, the youngest participant in this study was 31, the oldest was 88, and there were four 50 year olds.

The shape of the overall distribution in this case could be called somewhat bell shaped, as characterized by relatively fewer observations at either extreme than in the middle. Some distributions are symmetric, whereas others are asymmetric. Those that are asymmetric are said to be skewed. If fewer observations are at the upper end of the distribution (that is, the long tail is toward the right or higher values) the distribution's shape is called positively skewed. If the long tail is pointing toward the left, or lower values, the distribution is called negatively skewed. In the case of this particular example, turning Figure 5.2 on its side so it has lower values on the left reveals that the distribution of age values is somewhat negatively skewed. Although the stem-and-leaf display in Figure 5.2 has more detail (that is, more bins) than the histogram in Figure 5.1, the histogram retains the basic shape elements of the stem-and-leaf display.

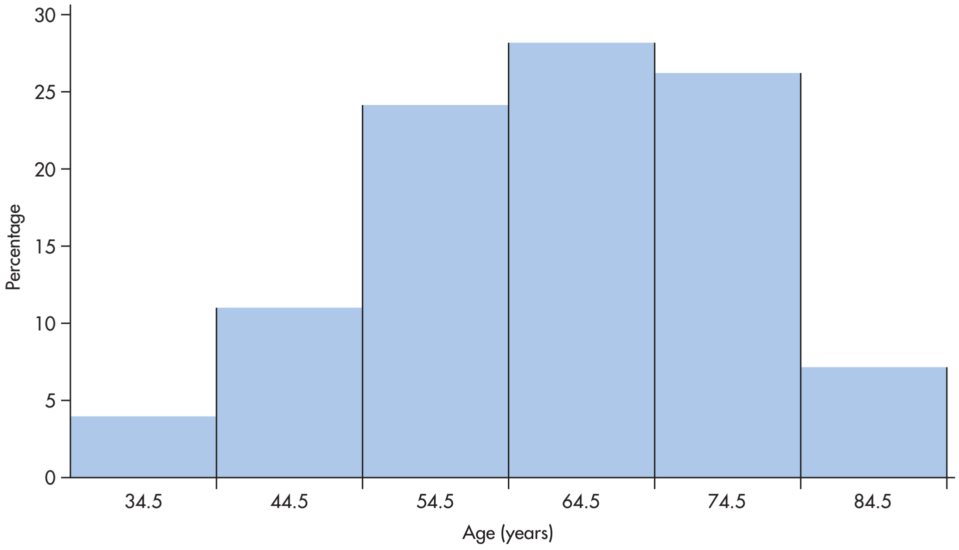

Figure 5.1 Histogram of 100 age values

3	1
3	599
4	0333
4	5889999
5	0000112233
5	55577788899999
6	001111112222334
6	5666777788899
7	000111122233334
7	55567778889
8	001123
8	8

Figure 5.2 Stem-and-leaf display of 100 age values

5.6 Central tendency

One fundamental idea in the development of new pharmaceutical products is that pharmaceutical companies (sponsors) would like to demonstrate that participants who receive a test treatment tend to fare better than those who receive some alternate therapy. This alternate therapy could be an inactive control (a placebo) or some other approved therapy (an active control). We said "tend to fare better" because participants will not all respond in the same way to the same test treatment. It is also true that, if and when the drug is approved for marketing and prescribed for patients, some patients will do better on the drug than others, but it is still very useful to clinicians to know how patients will tend to respond.

When we flip a fair coin ten times, we do not always expect to observe five heads and five tails. If we do several series of ten flips, we know that, by chance, we will observe six heads and four tails sometimes, and even more lopsided results would not be all that surprising. The same phenomenon happens with the response to test treatments in clinical studies. When doctors prescribe a new medicine to a patient it would be helpful to know what kind of response could be expected. Although we might expect that a fair coin flipped ten times will result in five heads, we also would expect that four or three heads could be observed. The determination of values that might be expected is the next topic in this chapter, that is, measures of central tendency.

Once we have assembled individual observations in a sample from a clinical study, our ability to understand the nature of those observations as a whole is limited by our ability to synthesize several disparate pieces of observation into an overall impression. Imagine that you have observed the following 10 observations of age of study participants in an early exploratory therapeutic clinical trial: 45, 62, 32, 38, 77, 28, 25, 62, 41, and 50.

Regulatory authorities are concerned about how well study participants match those in the general population of patients with the condition. How might such a question be answered? There are several strategies here.

5.6.1 The mode

One possible way to answer this question is to report that the most common value of age is 62. There are two such observations with this value of age. This measure of central tendency is known as the mode. The mode is most commonly used with non-numeric data (for example, most of the study participants were female), but it may also be useful for numeric data if there are only a few unique values. Unfortunately, the choice of the mode as the typical value in this case is a little misleading. Although there are two 62-year-olds in the study, most study participants (seven of them) are younger than that.

A question that comes to mind here is: What would the mode have been if all values of age were unique? The answer is that there would have been no mode – all values occurred equally as frequently. Likewise, suppose that there had also been two observations with the value of age of 32: In this case there would not be one value of the mode, but two. These two properties of the mode – that is, it is undefined in some instances and it may have multiple values in others – are considerable drawbacks to its use.

5.6.2 The median

Another reasonable choice for the typical value of age would be the value of age that is right in the middle of all values. This middle value is called the median. Citing the median would ensure that there were as many participants younger than the typical value as there were participants older. In this case, as there are 10 values there is no single middle value because there are an even number of values. The fifth and sixth greatest values of age are 41 and 45. To obtain the median value with an even number of observations, we simply split the difference between the two middle values. In this case the median is calculated as:

$$\frac{(41 + 45)}{2} = 43.$$

A quick check to know that we got it right is to see that exactly 5 observations are less than 43 and exactly 5 observations are greater than 43. When there is an odd number of observations, the median is the value of the middle observation after ordering them from the smallest to the largest. Unlike the mode, the median for a set of observations is unique: There is only one value and it is always defined.

5.6.3 The arithmetic mean

The last measure of central tendency that we consider is the most commonly encountered. The arithmetic mean is the sum of the individual observations divided by the total number of observations. Using mathematical notation the mean is calculated as:

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

where Σ stands for the addition of the values of each observation in the sample (*n* of them), that is:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n.$$

For our sample of 10 values of age, the mean is 46 (verification of this calculation is left to you). The arithmetic mean, commonly called the average, is the value that balances the weight of the distribution.

Some noteworthy characteristics of the mean are that, like the median, it is unique and always defined for a set of observations. However, the mean is sensitive to extreme observations – that is, if there is a single observation that is much higher or lower than the rest, the mean will be heavily influenced by that single observation.

One of the primary goals of Statistics is to use data from a sample to estimate an unknown quantity from an underlying population, called a population parameter. In general, we typically use the arithmetic mean as the measure of central tendency of choice because the sample mean is an unbiased estimator of the population mean, typically represented by the symbol μ. The main conceptual point about unbiased estimators is that they come closer to estimating the true population parameter, in this case the population mean, than biased estimators. When extreme observations influence the value of the mean such that it really is not representative of a typical value, use of the median is recommended as a measure of central tendency.

Returning to the query posed by the regulatory authorities, we came up with the following responses. The typical value of age in our sample using the mode is 62, using the median is 43, and using the mean is 46. Suppose that the authorities are satisfied with that response initially and then pose the following question. "So was your study among middle-aged adults with the condition?" You refer once again to the list of 10 observations and realize that it is not that simple. There actually were some younger adults in your study and it would be ideal to quantify the extent to which the mean does not tell the whole story. It is no surprise that not all values of age in the study are the same. Fortunately there are ways to quantify the extent to which they vary from participant to participant.

5.7 Dispersion

Dispersion refers to the variety or "spread" of individual observations in a sample. As for central tendency there are various measures of dispersion.

5.7.1 The range

A quick way to reflect the variety of values in a sample is to cite the lowest and highest values, the minimum and maximum. Calculating the difference between these two (calculated as maximum minus minimum) yields a value called the range:

$$Range = x_{\max} - x_{\min}.$$

Although the range is informative in that it conveys the difference between the two most extreme values, it does have a deficiency: It really does not adequately reflect the extent to which observations are similar or dissimilar. Imagine a study in which 99 of the 100 participants are aged between 20 years and 29 years, and one is 60 years old. The range is quite large (40 years), but the value of this range does not give any indication about how close together most age values in the sample are to each other.

5.7.2 The variance

In contrast, the variance of a sample does indicate how close together most values in a sample are. The sample variance is calculated as the sum of squared deviations of each observation from the sample mean divided by the sample size minus 1:

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

A calculation of this sort ensures that the measure of dispersion is positive (squaring the deviations ensures that) and dividing by $(n - 1)$ results in a quantity that represents an average of sorts. The sample variance is the "typical" or "average" squared deviation of observations from the sample mean. The use of the $(n - 1)$ in the denominator may seem confusing, but the reason why this is done is that calculating the sample variance in this manner yields an unbiased estimator of the population variance, which is represented by the symbol σ^2. (The exact mathematical demonstration that s^2 is an unbiased estimator of σ^2 is beyond the scope of this text.)

5.7.3 The standard deviation

Although very useful in some ways, the sample variance has the unfortunate characteristic that it is expressed in terms of squared units that are typically nonsensical. From our earlier example of 10 ages, we would calculate the sample variance as 282 "squared years." To overcome the significant drawback of squared units we can take the square root of the sample variance to obtain the standard deviation (s):

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

The standard deviation represents an average (of sorts) deviation of each observation from the sample mean. Again, the only reason why we do not call this quantity the average deviation without qualification is that there really are n deviations from the sample mean, but the standard deviation is calculated using the denominator of $(n - 1)$ instead of n. The sample standard deviation is an unbiased estimator of the population standard deviation. For our previous example of 10 age values, the value of the sample standard deviation is 16.8 years (we leave confirmation of this to you).

The sample standard deviation captures a great deal of information about the spread of the data. The value of the standard deviation is helpful across a number of datasets because of the results of what is called Tchebysheff's theorem. A simple way of thinking of Tchebysheff's theorem is that most values lie close to the sample mean. According to this theorem, no matter what the shape of the distribution is:

- 25% ($\frac{1}{4} = \frac{1}{2^2}$) or less of observations lie outside of 2 standard deviations away from the mean
- 11% ($\frac{1}{9} = \frac{1}{3^2}$) or less of observations lie outside of 3 standard deviations away from the mean
- 6% ($\frac{1}{16} = \frac{1}{4^2}$) or less of observations lie outside of 4 standard deviations away from the mean.

Applying Tchebysheff's theorem to our sample of 10 ages, we can say to the regulatory agency that the study really is not just among middle-aged adults.

5.7.4 Variability and the coefficient of variation

A commonly asked question among investigators is: How do I know if I have a lot of or a little variability in my study results? There is no straightforward answer to this question: The magnitude of the variance (or synonymously the standard deviation) can be called a lot or a little only when it is compared with some other quantity – that is, it is relative.

In Chapter 11 we discuss an analytical strategy called analysis of variance (ANOVA) in which one variance is compared with another. For now, another useful measure of relative dispersion is the coefficient of variation (CV), calculated as the ratio of the sample standard deviation to the sample mean:

$$\mathrm{CV} = \frac{s}{\bar{x}}.$$

The coefficient of variation is useful when comparing the magnitude of variability between two or more different random variables.

To illustrate the coefficient of variation, consider the following (extremely simple and artificial) example. Imagine that there are two random variables in an early therapeutic exploratory clinical trial. One random variable is pulse (ranging from 50 to 80) and the other is age, which in this case is pulse minus 20. We can see that, from this example, values of pulse and age are just as disperse, but what differs between them is the mean. Hence, when we calculate the standard deviation, one random variable will appear to have more or less dispersion, but, after re-scaling the standard deviation with the sample mean, the measure of dispersion is the same.

5.7.5 Percentiles

Another descriptive measure of variability or dispersion is the percentile. The Pth percentile is the value of the random variable, $X = X_{\frac{P}{100}}$, such that:

- $P\%$ of values of X are $\leq X_{\frac{P}{100}}$
- $100 - P\%$ of values of X are $> X_{\frac{P}{100}}$.

For example, the 75th percentile is the value of X below which 75% of the values lie and above which 25% lie. The 50th percentile is synonymous with the median. Likewise the 25th percentile is the value of X below which 25% of the values lie and above which 75% lie. The difference between the 75th and 25th percentiles is called the interquartile range, which can be a useful measure of dispersion when the distribution of the random variable is heavily skewed or asymmetric.

5.8 Tabular displays of summary statistics of central tendency and dispersion

As we discuss in more detail in Chapter 6, one of the primary goals of studies in a clinical development program is to describe the effect that the test treatment had on study participants so that some inference can be made about the drug's effects on patients who may receive the drug in the future. Summary descriptive statistics of central tendency and dispersion give us better understanding of the typical effect of the test treatment and how varied participants' responses were.

In our experience the mean and the standard deviation are the most commonly used summary statistics for these purposes. However, other measures can be useful to reviewers when interpreting data from clinical studies. We encourage researchers to present the following statistics: The sample size, the mean, the median, the standard deviation, and the minimum and the maximum. Presenting all these values for a given random variable provides a reviewer with two measures of central tendency and two measures of dispersion. For clinical studies that are comparative in nature, such as therapeutic confirmatory trials, it is our recommendation that summary statistics – for example, the mean and standard deviation – be formatted in a report so that the primary comparison of interest is read across columns (left to right). In clinical studies this is typically treatment groups or dose groups. Secondary comparisons of interest – for example, time points of observation – should be arranged as separate rows.

5.9 Review

1. What scale are each of the following participant characteristics measured on:

   (a) eye color
   (b) body mass index (kg/m^2)
   (c) number of cerebrovascular events diagnosed in the past 5 years
   (d) days from study entry to last follow-up visit
   (e) concentration of test drug in plasma (ng/mL)
   (f) blood pressure classification: Normal; prehypertension; stage 1 hypertension; stage 2 hypertension.

2. Using the histogram in Figure 5.1 and the stem-and-leaf plot in Figure 5.2, comment on the appropriateness of each of the following measures of central tendency:

   (a) mean
   (b) median
   (c) mode.

3. From the frequency table of age values in Table 5.1, calculate:

   (a) the median or 50th percentile
   (b) the 25th percentile
   (c) the 75th percentile.

5.10 Reference

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

6

Probability, hypothesis testing, and estimation

6.1 Introduction

A common goal of pharmaceutical clinical trials is to establish with some high degree of confidence that the test treatment is superior to a control with respect to some measurable effect. If we are able to say that the expected effect of the test treatment tends to be superior (by some amount) to the expected effect of the control, we could conclude that the test treatment was superior to the control.

To accomplish this objective, sponsors design studies that allow them to attribute any difference in the response of interest to the test treatment itself. This is accomplished through the use of randomization, a carefully selected study population, treatment blinding, careful data collection, and other measures that minimize the possibility that other factors may have influenced the outcome of the study. However, if too few study participants are studied any difference observed might have been caused by chance. A chance, or spurious, result is one that may not be repeatable or, to put it another way, a chance result is not reliable. Provision of a high degree of confidence that a new drug is beneficial requires sponsors to demonstrate that effects observed from a new treatment are reliable. This chapter discusses the statistical concepts that allow researchers to make the conclusion that the effect seen in a study was unlikely to be the result of chance.

6.2 Probability

The statements at the end of the previous section can be expressed differently, and more quantitatively, in the language of Statistics. We noted that provision of substantial evidence that a new drug is beneficial requires sponsors to demonstrate that effects observed from a new treatment are reliable. This chapter discusses the statistical concepts that allow researchers to make the conclusion that the effect seen in a study is reliable, that is, it is unlikely to be the result of chance.

The statistical techniques that can be used to rule out chance events require us first to consider some concepts of probability. Many outcomes in life are inherently uncertain, and others can be considered certain. If you play the lottery, it is uncertain whether you will win on any given occasion (it is also incredibly unlikely). If you drop an apple, it is certain that it will fall to the ground. Other outcomes fall in the middle of the range. It is useful to be able to quantify the degree of certainty, and conversely the degree of uncertainty, associated with a particular occurrence. This is the realm of probability.

Like the word significance, the concept of probability is used in everyday language as well as in the discipline of Statistics. As Turner (2007) noted, the statement “I’ll probably be there on Saturday” involves a probabilistic statement, but there is no degree of quantification (if you know the individual making this statement, past experience may lead you to have an informed opinion concerning the relative meaning of “probably,” but this is a subjective judgment).

As for many aspects of statistical analysis, there are axioms in probability that make it a very useful tool. In the context of Statistics, probability can be defined in quantifiable terms. A probability is a numerical quantity between zero and one that expresses the likely occurrence of a future event. A probability of 0 denotes that the event will not occur. A probability of 1 denotes that the event will undoubtedly occur. Any numerical value between 0 and 1 expresses a relative likelihood of an event occurring.

A probability value can be represented as a fraction or as a decimal value. In addition, it is common in some aspects of Statistics to multiply the decimal expression of a particular probability by 100 to create a percentage statement of likelihood. A probability of 0.5 would thus be expressed as a 50% chance that an event would occur. Percentage statements of likelihood are a central component of hypothesis testing, which is introduced later in the chapter.

The probability of an event (E) can be represented as $P(E)$ and we use this notational convention. In general, the probability of either of two events (A or B) occurring is calculated as:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$

In other words, the probability of either event occurring is the sum of the probabilities of each event minus the probability of both occurring together (or jointly).

Consider the cross-tabulation of the gender and age of participants in a clinical trial as presented in Table 6.1. As seen there were 200 participants, 100 of whom were male and 100 female. There were 65 participants aged 45 years or younger, 90 between 46 and 64 years, and 45 who were aged 65 years or older. We illustrate several of the axioms of probability using Table 6.1. For example, the probability of selecting at random a participant from this group who was male or aged 65 years or older:

$$P(\text{male or} \geq 65) = P(\text{male}) + P(\geq 65) - P(\text{male and} \geq 65) =$$

$$P(\text{male or } \geq 65) = \frac{100}{200} + \frac{45}{200} - \frac{15}{200} = \frac{130}{200} = 0.65.$$

In the special case that the events A and B cannot occur at the same time, they are said to be mutually exclusive, meaning that $P(A \text{ and } B) = 0$. Hence, for mutually exclusive events A and B:

$$P(A \text{ or } B) = P(A) + P(B).$$

A randomly selected participant cannot be both "≤ 45" and "≥ 65." The events of selecting a participant aged 45 years or younger and one 65 years or older are mutually exclusive. This result is generalizable to more than two events of interest.

Table 6.1 Cross-tabulation of age and gender

Age (years)	Male	Female	
≤ 45	35	30	65
46–64	50	40	90
≥ 65	15	30	45
	100	100	200

If one or more events, $E_1, E_2, \ldots E_n$, represent all unique and mutually exclusive outcomes in a particular circumstance, the probability of observing at least one of the events sums to one:

$$P(E_1 \text{ or } E_2 \text{ or} \ldots \text{or } E_n) = P(E_1) + P(E_2) + \ldots + P(E_n) = 1.$$

This result can be used to calculate the probability of one or more events of interest. For any event E_1 among n mutually exclusive and exhaustive events:

$$P(E_1) = 1 - \{P(E_2) + \ldots + P(E_n)\}.$$

This expression is called the complement rule and will be referenced throughout this book.

The probability of selecting a male at random can be calculated by adding the probabilities for the events "male ≤ 45 years," "male 46–64 years," and "male ≥ 65 years," because these are all mutually exclusive events. The probability can be calculated as follows:

$$\begin{aligned} P(\text{male}) = {} & P(\text{male} \leq 45 \text{ years}) + \\ & P(\text{male 46–64 years}) + P(\text{male} \geq 65 \text{ years}) \\ = {} & \frac{35}{200} + \frac{50}{200} + \frac{15}{200} \\ = {} & \frac{100}{200} = \frac{1}{2}. \end{aligned}$$

The probability of an event B given that A has been observed is called a conditional probability and is defined as:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)},$$

where the vertical bar signifies "given."

The probability of selecting a participant ≤ 45 years of age, given that a male has been selected, is:

$$P(\leq 45 \text{ years} \mid \text{male}) = \frac{P(\leq 45 \text{ years and male})}{P(\text{male})}$$

$$\frac{P(\leq 45 \text{ years and male})}{P(\text{male})} = \frac{\frac{35}{200}}{\frac{100}{200}} = \frac{35}{100}.$$

It follows that the probability of two events occurring jointly is calculated as:

$$P(A \text{ and } B) = P(A)P(B|A).$$

One important use of conditional probabilities occurs in Bayes' theorem. The conditional probability of an event A given an event B is:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

Note that throughout this book we have adopted a standard mathematical notation for the product of two or more terms. In the expression above the numerator is the product of the two terms, $P(B|A)$ and $P(A)$, that is, these two quantities are multiplied. Please keep this standard in mind when you encounter other mathematical expressions.

It is also possible to state the probability of an event, A, as a function of two or more conditional events. If the events B and C are mutually exclusive and exhaustive – for example, they represent male and female – the probability of event A can be expressed as:

$$P(A) = P(A \mid B)P(B) + P(A \mid C)P(C).$$

This expression can be extended to more than two conditional events.

A common application of Bayes' theorem is in estimating the probability of a participant having a disease, given a positive test for that disease. These concepts are important in their own right with regard to the development of diagnostic tests. As the clinical trials discussed in this book are for the purposes of developing new pharmaceutical interventions rather than testing for the existence of a disease or condition, this issue may not seem directly relevant. However, these concepts are discussed in Chapter 12 in a different light, and we would therefore like to establish these concepts at this earlier stage.

For simplicity, in the notation used in this example we define the following events using the symbol "≡" which means "is equivalent to":

- $D+$ ≡ participant has the disease of interest
- $D-$ ≡ participant does not have the disease of interest
- $T+$ ≡ participant tests positive for the disease
- $T-$ ≡ participant tests negative for the disease.

When developing a diagnostic test, investigators identify two groups: One is known (by some gold standard testing procedure) to have the disease; the other is known (also by a gold standard testing procedure) not to have the disease. Then all participants in both these groups are given the new diagnostic test. The accuracy of a new diagnostic test is measured by two criteria:

1. Sensitivity is the probability that a new test will have a positive result among those who are known to have the disease. This is denoted by: $P(T+|D+)$.
2. Specificity is the probability that a new test will have a negative result among those who are known not to have the disease. This is denoted by: $P(T-|D-)$.

Once a new diagnostic test has been developed it may be considered for a public health screening program. Evaluating the utility of a proposed new diagnostic test in a population involves the following two criteria:

1. The true positive rate is the probability that a participant has the disease given that she or he has tested positive. This is denoted by $P(D+|T+)$ and is also referred to as predictive value positive. The complement, $1 - P(D+|T+) = P(D-|T+)$, is the false-positive rate.
2. The true-negative rate is the probability that a participant does not have the disease given that she or he has tested negative. This is denoted by $P(D-|T-)$ and is also referred to as predictive value negative. The complement, $1 - P(D-|T-) = P(D+|T-)$, is the false-negative rate.

If the true-positive rate is low (or the false-positive rate high) a number of participants will needlessly incur the expense and anxiety of further medical investigations. If the true-negative rate is low (or the false-negative rate high) a number of them will carry on undiagnosed. The goal would be to adopt a screening tool that had high rates of true positives and true negatives. Bayes' theorem can be used to show that the rates of true positives and true negatives are a function of the sensitivity and specificity of the diagnostic test itself and the prevalence of the disease in the population of interest.

We illustrate this concept for the true-positive rate, which is:

$$= P(D+ \mid T+)$$

$$= \frac{P(T+ \mid D+)P(D+)}{P(T+)} \text{ by Bayes' theorem.}$$

Bayes's theorem is applied again to obtain:

$$= \frac{P(T+ \mid D+)P(D+)}{P(T+ \mid D+)P(D+) + P(T+ \mid D-)P(D-)}.$$

Noting that $P(T+ \mid D-) = 1 - P(T- \mid D-)$ we have the desired result.

$$= \frac{P(T+ \mid D+)P(D+)}{P(T+ \mid D+)P(D+) + [1 - P(T- \mid D-)]P(D-)}.$$

Note that $P(D+)$ is often called the prevalence of the disease in a population. Its complement is $P(D-) = 1 - P(D+)$. Prevalence of a disease is estimated through the use of epidemiologic studies and not clinical trials. Thus, to fully evaluate the utility of the new diagnostic test, we must have an estimate of the prevalence of the disease, the sensitivity of the test, and the specificity of the test.

Two events are said to be statistically independent if the probability of one occurring does not depend on the other. If A and B are independent events the joint probability is given by:

$$P(A \text{ and } B) = P(A)P(B).$$

If we sample from our 200 study participants "with replacement" – that is, after each selection the participant is available for selection again – the probability of selecting a male does not depend on previous selections. The probability of selecting two males in a row is given by:

$$P(\text{two males in a row}) = P(\text{male})P(\text{male})$$

$$= \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$$

$$= \frac{1}{4} = 0.25.$$

The probability of selecting four males in a row is:

$$P(\text{four males in a row}) = P(\text{male})P(\text{male})P(\text{male})P(\text{male})$$

$$= \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$$

$$= \frac{1}{16} = 0.0625.$$

These basic principles and characteristics of probability are referred to throughout subsequent chapters.

6.3 Probability distributions

In Chapter 5 we described a number of ways to examine the relative frequency distribution of a random variable (for example, age). An important step in preparation for subsequent discussions is to extend the idea of relative frequency to probability distributions. A probability distribution is a mathematical expression or graphical representation that defines the probability with which all possible values of a random variable will occur. There are many probability distribution functions for both discrete random variables and continuous random variables. Discrete random variables are random variables for which the possible values have "gaps." A random variable that represents a count (for example, number of participants with a particular eye color) is considered discrete because the possible values are 0, 1, 2, 3, etc. A continuous random variable does not have gaps in the possible values. Whether the random variable is discrete or continuous, all probability distribution functions have these characteristics:

- All possible values of the random variable must be represented by the distribution function.
- The probability of each value of the random variable occurring is bounded by 0 and 1, inclusive.
- The probabilities of values of the random variable occurring must sum to 1 (in the case of a discrete random variable) or integrate to 1 (in the case of a continuous random variable).

A simple example of a discrete probability distribution is the process by which a single participant is assigned the active treatment when the event "active treatment" is equally likely as the event "placebo treatment." This random process is like a coin toss with a perfectly fair coin. If the random variable, X, takes the value of 1 if active treatment is randomly assigned and 0 if the placebo treatment is randomly assigned, the probability distribution function can be described as follows:

$$P(X = x) = \frac{1}{2}, \text{ where } x = 0 \text{ or } 1.$$

This probability distribution function has the characteristics defined previously:

- The random variable can take on only values of 0 or 1.
- The probability distribution function is defined for both values.
- The probability of each value is between 0 and 1, inclusive. It is, in fact, half for both values.
- Finally, the sum of the probability of all mutually exclusive outcomes is equal to one, that is, $P(X = 0)+P(X = 1) = 1$.

6.4 Binomial distribution

The first probability distribution function that we discuss in detail is the binomial distribution, which is used to calculate the probability of observing x number of successes out of n observations. As the random variable of interest, the number of successes, is discrete (as are all counts), the binomial distribution is called a discrete random variable distribution. The binomial distribution is applicable when the following conditions apply:

- Each of n observations results in only one of two outcomes (one is typically called a success and the other failure).
- The probability of a success, p, is the same from observation to observation.
- Each observation is independent of the others.

The probability of observing x successes out of n observations under these conditions (called a Bernoulli process) can be expressed as:

$$P(X = x; p, n) = C^n_x p^x (1 - p)^{n-x}.$$

The left part of this expression can be read as "the probability of the random variable, X, taking on a particular value of x, given parameters p and n." The quantity $(1 - p)$ is the probability of failure for any trial. The notation C^n_x is shorthand to represent the number of combinations of taking x successes out of n observations when ordering is not important. This quantity can be calculated as:

$$C^n_x = \frac{n!}{x!(n - x)!}.$$

The expression $n!$ is read as "n factorial" and is calculated as $n(n - 1)(n - 2) \ldots (1)$.

The mean of the binomial distribution function is:

$$\text{Mean} = np.$$

The variance of the binomial distribution is:

$$\text{Variance} = np(1 - p).$$

A simple example of the use of the binomial distribution is the result of four random assignments to either the active or the placebo treatment group when each outcome is equally likely. What is the probability of observing 0, 1, 2, 3, or 4 assignments to the active treatment group out of 4 random treatment assignments when the probability of assigning to active or placebo is equally likely? We must assume that the outcome of one assignment does not impact the outcome on subsequent assignments, that is, they are independent. There are only two possible outcomes on any given trial: Assignment to active or placebo. The probability of each outcome, the number of "successes" or assignments to active, can then be calculated using the binomial probability distribution function.

The probability of each outcome of four random treatment assignments is displayed in Table 6.2. In some instances, we may be interested in knowing what the probability of observing x or fewer successes would be, that is, $P(X \leq x)$. This cumulative probability is also displayed for each outcome in Table 6.2. For a discrete random variable distribution, the sum of probabilities of each outcome must sum to 1, or unity.

As you might expect, the most probable outcome is 2 actives (probability 0.375) and the least probable outcomes are 0 and 4 actives (each with a probability of 0.0625). We can use the cumulative probability distribution to answer other probability questions of interest. For example, what is the probability of observing 3 or fewer actives? This probability is denoted as $\text{P}(X \leq 3) = 0.9375$. We can use the complement rule from Section 6.2 to calculate the probability of observing 2 or more actives, $\text{P}(X \geq 2)$, as:

$$1 - P(X \leq 1) = 1 - 0.3125 = 0.6875.$$

The binomial distribution is discussed later in the chapter to illustrate concepts of hypothesis testing.

6.5 Normal distribution

Similar probability models can be used for continuous random variables. The most common, and arguably the most important of these in Statistics, is the normal distribution. As it is encountered so frequently in this book, we spend some time describing its characteristics and uses.

The normal distribution is a particular form of a continuous random variable distribution. The relative frequency of values of the normal distribution is represented by a normal density curve. This curve is typically described as a bell-shaped curve, as displayed in Figure 6.1.

More precisely, it is one specific kind of symmetrical curve. The precise nature of this curve can be described mathematically by a formula that contains both the mean, μ, and the standard deviation, σ, of the population that is being represented graphically by the normal curve:

$$f(x;\, \mu,\, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

The term "population" is defined in detail later in the chapter. Until then we can think of a population as the largest group of experimental units (for example, study participants) about which we would like to make a conclusion.

As we need to know two parameters – that is, the mean, μ, and the standard deviation, σ, to fully characterize this distribution – it is consid-

Table 6.2 Distribution of the number of assignments to active from four random assignments when the probability of assignment to active and placebo is equal (p=0.5)

Outcome (no. of actives, x)	Probability of the outcome $P(X = x)$	Cumulative probability $P(X \leq x)$
0	$C^4_0 0.5^0(0.5)^4 = 0.0625$	0.0625
1	$C^4_1 0.5^1(0.5)^3 = 0.2500$	0.3125
2	$C^4_2 0.5^2(0.5)^2 = 0.3750$	0.6875
3	$C^4_3 0.5^3(0.5)^1 = 0.2500$	0.9375
4	$C^4_4 0.5^4(0.5)^0 = 0.0625$	1.0000

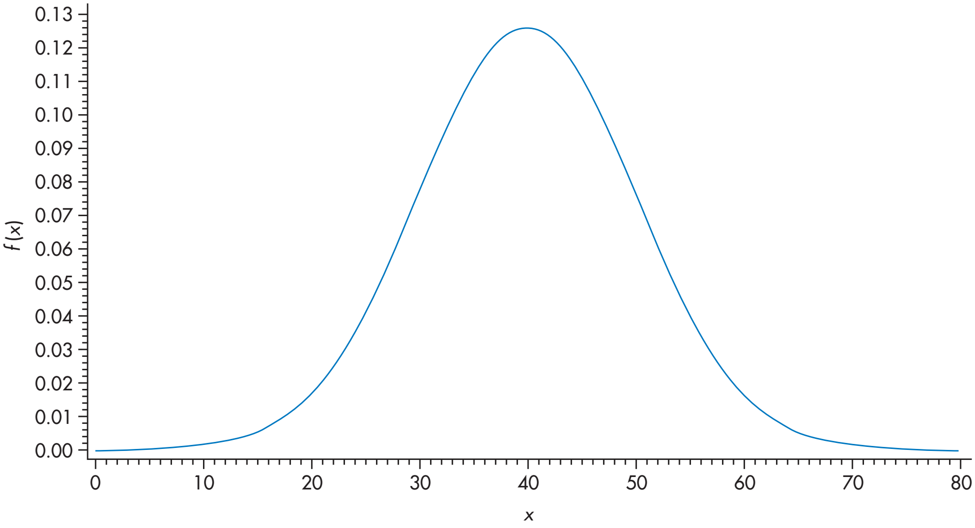

Figure 6.1 A normal density curve ($\mu = 40$, $\sigma = 10$)

ered to be a two-parameter distribution. This fact is also conveyed by the use of the symbols μ and σ on the left side of the expression. The mean specifies the distribution's location, whereas the standard deviation specifies the spread of the distribution. If a random variable X has a normal distribution with mean μ and variance σ^2, this is written as $X \sim N(\mu,\sigma^2)$. Note that most practical applications involve the use of σ rather than σ^2, but it is conventional to describe the normal distribution in terms of its mean and variance.

Figure 6.2 displays three normal density curves with the same mean (location) but different standard deviations (spread). Several characteristics of the normal distribution are very helpful in developing the statistical tests introduced in this book:

- The highest point of the normal curve occurs for the mean of the population, μ.
- The shape of the curve (relatively narrow or relatively broad) is influenced by the standard deviation, σ. The sides of the curve descend more gently as the standard deviation increases.
- At a distance of 2 standard deviations from the mean, the slope of the curve changes from a relatively smooth downward slope to a curve that technically extends out to infinity, that is, the curve technically never reaches (touches) the *x* axis of the graph. This concept is analogous to starting a certain distance away from a fence and taking steps that always cover half the distance between you and the fence. As your next step always covers only half the remaining distance, theoretically you never reach the fence. However, after a certain number of steps, you are, to all practical purposes, at the fence. In the same manner, the curve is regarded as intercepting the axis at a distance of 4 standard deviations from the mean.
- The area under the curve is 1.0. This can be demonstrated formally using integral calculus, which is beyond the scope of this book: A simpler demonstration is provided by Turner (2007, pp 94–5). That the area under the curve is equal to 1 is analogous to the statement that the probability of all mutually exclusive events must sum to 1.

The precise mathematics of the normal distribution allows quantitative statements of the area under the curve between any two points on the *x* axis. Of most interest here is the area under the curve between two points that are equidistant from the mean.

These points, equidistant from the mean on either side, can be represented by statements of the form:

$$\mu \pm \text{distance}.$$

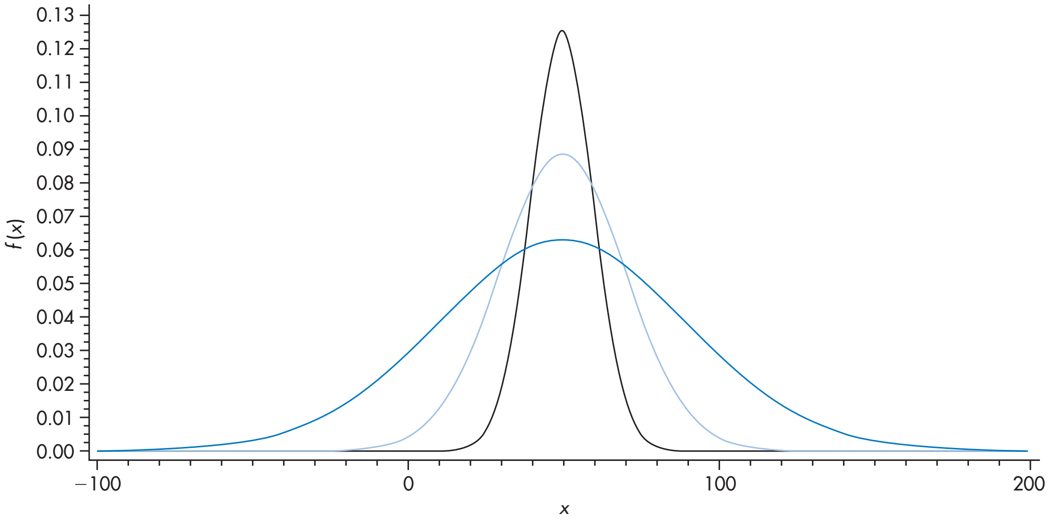

Figure 6.2 Three normal density curves with the same mean (location) but different standard deviations (spreads)

It can be shown for any normal distribution that:

- 68.3% of the area under the curve lies in the range μ ± σ
- 95.4% of the area under the curve lies in the range μ ± 2σ
- 99.73% of the area under the curve lies in the range μ ± 3σ.

As the area under the entire density curve equals 1 the statements above also imply by the complementary rule that:

- 31.7% of the area under the curve lies outside μ ± σ
- 4.6% of the area under the curve lies outside μ ± 2σ
- 0.27% of the area under the curve lies outside μ ± 3σ.

Expressing a similar concept in terms of pertinent "round number" percentages:

- The central 90% of the area lies in the range μ ± 1.645σ
- The central 95% of the area lies in the range μ ± 1.960σ
- The central 99% of the area lies in the range μ ± 2.576σ.

You may recall from Chapter 5 that by using Tchebysheff's theorem we could estimate the probability with which observations fall within *k* standard deviations for *any* distribution. You are encouraged to compare the results from Tchebysheff's theorem and those cited above for the normal distribution. Although values of any percentage of interest can be determined from statistical tables of normal distributions, the 95% and 99% values are of particular importance in the context of this book.

It is important to note here that the areas under the curve of a continuous random variable distribution can be thought of as probabilities. Assume that we know that age in a population of study participants is normally distributed with a mean of 40 and variance of 100 (standard deviation of 10). This normal distribution is displayed in Figure 6.3 with vertical lines marking 1, 2, and 3 standard deviations from the mean.

It is then possible, using the results above and similar ones from statistical tables, to estimate the probability that a participant randomly selected from the population of study participants would be aged > 50 or < 30. The answer is 0.32 or 32% – that is, the proportion or

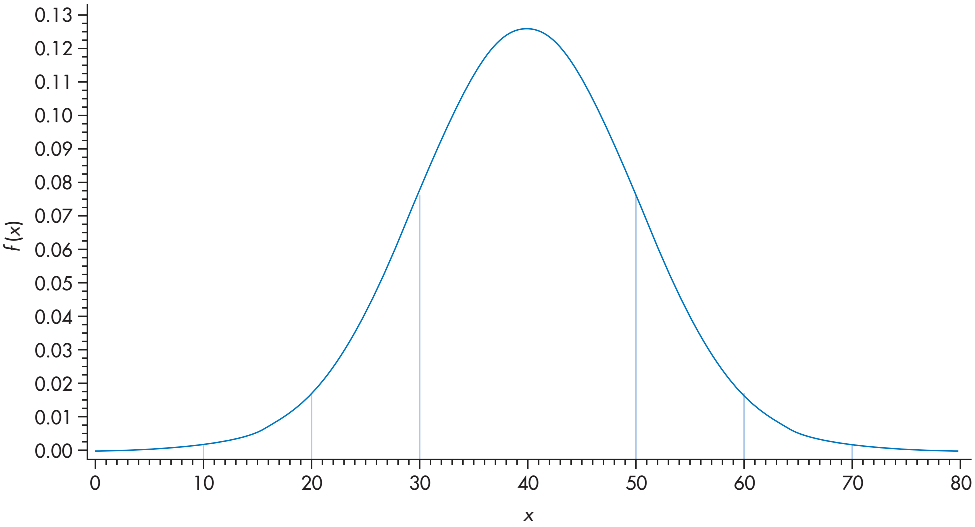

Figure 6.3 Normal distribution with mean of 40 and standard deviation of 10. Note that the vertical lines represent $\mu \pm \sigma$, $\mu \pm 2\sigma$, and $\mu \pm 3\sigma$

percentage of the area under the curve translates directly in to the percentage of participants (or other observational units) whose age values fall outside of the two identified points.

6.5.1 The standard normal (*Z*) distribution

One unique and important normal distribution is the standard normal distribution, or *Z* distribution, which has a mean of 0 and a variance of 1. If a random variable X is distributed as standard normal with mean 0 and variance 1, it is written as $X \sim N(0,1)$. To use some of the general results from normal distributions provided earlier, we can make the following statements for the standard normal distribution:

- The central 90% of the area lies between ± 1.645
- The central 95% of the area lies between ± 1.960
- The central 99% of the area lies between ± 2.576.

The standard normal or *Z* distribution is used extensively in Statistics and throughout this book. For later reference, the standard normal distribution is provided in Figure 6.4. Note that the area under the curve to the left of the value −1.96 is 0.025 (or 2.5%). As the distribution is symmetric, the area under the curve to the right of the value 1.96 is also 0.025. Another way of stating this is that, if we were to randomly select a value from the distribution, there is a 95% chance that the value would be between −1.96 and +1.96. One can also think of the values −1.96 and +1.96 as the 2.5th and 97.5th percentiles, respectively.

Values of the *Z* that define areas under the standard normal curve in the left tail, the right tail, and the symmetric central region are provided in Appendix 1.

6.5.2 Transforming a normal distribution to the standard normal distribution

One helpful method possible with a random variable that has a normal distribution with mean, μ, and variance, σ^2, is to transform values of the random variable so that they have the scale of the standard normal distribution. This

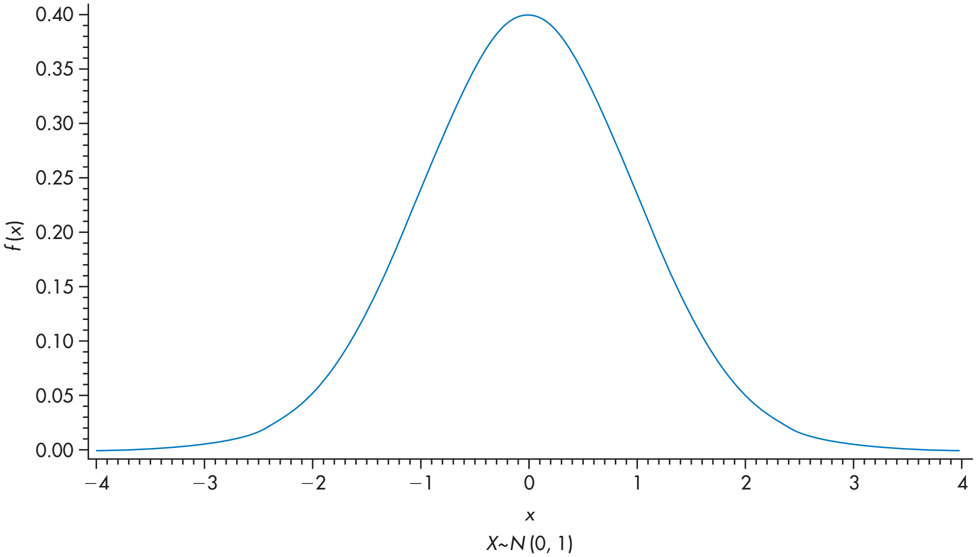

Figure 6.4 The standard normal (*Z*) distribution

makes it possible to answer a number of probability questions using statistical tables that provide the areas under the standard normal curve. In general, for a random variable $X \sim N(\mu,\sigma^2)$, the random variable:

$$Z = \frac{X - \mu}{\sigma}$$

is normally distributed with mean 0 and variance 1.

We can use the example from earlier in Section 6.5 to illustrate this method. If age in a population of study participants is normally distributed with a mean of 40 and variance of 100 (standard deviation of 10), what is the probability that a participant randomly selected from the population of study participants would be aged > 50 or < 30?

First we are interested in the probability that a randomly selected participant will be > 50 years of age. The transformed value for $X = 50$ is:

$$Z = \frac{50 - 40}{10} = 1.$$

As a result of this transformation, $P(X > 50)$ corresponds to $P(Z > 1)$. The probability, $P(Z > 1)$, can be obtained from Appendix 1, the look-up table for areas under the standard normal distribution curve. As seen in Appendix 1, the area under the standard normal distribution curve for $Z > 1$ is 0.159.

Then we would like to know what the probability is that a randomly selected participant will be < 30 years of age. The transformed value for $X = 30$ is:

$$Z = \frac{30 - 40}{10} = -1.$$

As above, $P(X < 30)$ is equal to $P(Z < -1)$. Using Appendix 1 as a reference, the area under the standard normal distribution curve for $Z < -1$ is 0.159.

The probability of interest is obtained by summing the two probabilities associated with $P(Z > 1)$ and $P(Z < -1)$ because the two events are mutually exclusive. That is, a participant cannot be both < 30 and > 50, so the probability of interest is 0.159 + 0.159 or 0.318.

At first glance it may seem that this transformation method is useful only in a few instances (when the random variable is known to have a normal distribution) and contrived ones at that. However, it is actually useful in many instances. Many random variables can be shown to have approximately normal distributions. The reason for this is given shortly. It turns out that, if a random variable has an approximate normal distribution, for which the mean and variance are known, a transformation results in a random variable that has an approximate standard normal distribution.

6.6 Classical probability and relative frequency probability

Before concluding the first part of this chapter on the fundamentals of probability it is important to point out that there are two ways to estimate a probability. To contrast these two types of probability we consider the question: "What is the probability of observing a 'head' when tossing a coin?"

The first type of probability, termed by some "classical" probability, is based on an assumption about the state of the experiment and some basic mathematical expressions. For example, we would begin answering this question by assuming that the coin was fair. Further, we would note to ourselves that a fair coin has two sides, the only two outcomes of a coin toss are "heads" and "tails," and only one of these two outcomes is the one of interest. The probability of observing a head from a single toss of a fair coin is therefore ½ or 0.5. The most straightforward way to solve classical probability problems is to write out all of the unique possible outcomes, the sample space, and then identify the number of times that the outcome of interest would occur. In this case the sample space is "heads" or "tails." The event of interest, observing heads, is represented by just one of these events, so the probability of interest is ½. The use of the binomial distribution to calculate the probability distribution of observing the number of assignments to active is another example. In that case we knew (by design) that the probability of assignment to active was exactly half.

Many Statistics students have suffered immensely over the years by having to solve classical probability problems. Marilyn vos Savant (1997) stumped many readers with the following classical probability problem:

> A woman and a man (unrelated) each have two children. At least one of the woman's children is a boy, and the man's older child is a boy. Do the chances that the woman has two boys equal the chances that the man has two boys?
>
> vos Savant (1997, p 15)

What is your answer? We leave it to you to conduct an online search to investigate the controversy surrounding this problem. We do not dwell any further on this method of estimating probabilities because we also dislike them, and the second type is more useful for us anyway.

The second type of probability, relative frequency probability, is calculated by repeating an experiment a large number of times (say n) and counting the number of times out of n that the outcome of interest (say m) occurred. The probability of the event is then calculated as:

$$P(\text{event}) = \frac{m}{n}.$$

The calculated probability is simply an estimate of the true probability (which remains unknown).

Using a relative frequency approach to estimating the probability of observing a head we would toss the coin a number of times (for example, 10), count the number of times a head landed face up (for example, 4 times), and then calculate the probability as 4/10 or 0.4. It is perhaps not surprising that the estimated probability here is not exactly 0.5. We were only one head shy of 5/10, so the relatively small number of coin tosses may have had an impact. You can imagine tossing the coin 100 times and observing 46 heads for a probability of 0.46. That would be much closer to the classical probability solution. Another possible reason that only four heads came up could be that the coin really was not fair at all. For the classical probability solution to this problem we would need to assume that the coin was fair or be told that it was. The relative frequency solution has the advantage of not requiring the assumption of a fair coin, but has the disadvantage of possibly being limited by the number of experiments.

Our initial probability estimate of 0.4 from 10 coin tosses does not seem to be that far off because we might reason that observing 4, 5, or 6 heads would be expected from 10 tosses of a fair coin. You may be intuitively thinking that, if we were to repeat the 10 coin tosses, we would probably count 4–6 heads again. Your intuition would be correct, and there is a statistical concept that explains how results from experiments vary from sample to sample. The magnitude of expected differences from sample to sample enables us to estimate a quantity that we can never really know, one that represents the truth. In the case of the coin-tossing experiment, our goal would be to infer whether or not the true probability of observing a head was 0.5.

6.7 The law of large numbers

In clinical trials we do not know what the probability of observing a particular serious adverse event is, but we observe a large number of outcomes (for example, participants exposed to a new treatment) to estimate it. As the sample size increases the estimate becomes more precise (that is, closer to the truth). An illustration of the "law of large numbers" is provided in Figure 6.5. Suppose that a relatively uncommon adverse event is represented by the chance event of two thrown dice landing with a total of two (or "snake eyes"). The classical probability solution to estimating the probability of this event is $(1/6)^2 = 0.02778$. The relative frequency solution can be obtained by rolling two dice a large number of times (n), counting the number of times "snake eyes" occurs (m), and estimating the probability as m/n. The most convenient means to conduct this experiment is using computer simulation. As seen in Figure 6.5, the estimates of the probability (denoted by the oscillating curve) vary quite a bit from the truth (represented by the horizontal reference line) until the sample size is around 10 000. The implication of the law of large numbers for clinical trials of new drugs is that the unknown quantities of interest are more precisely estimated with larger samples. These would include the mean change in SBP (systolic blood pressure) or the proportion of participants experiencing a serious adverse event. Given the limited size of most clinical development programs, the most precise estimates of risks of new therapies become evident only once a new drug has been marketed and used by many thousands of patients.

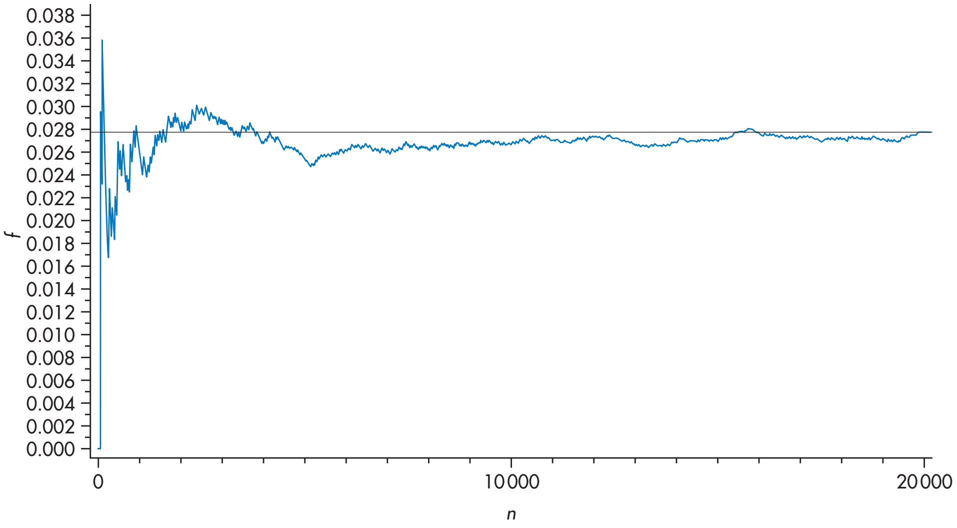

Figure 6.5 Illustration of the law of large numbers: Proportion of two dice coming up as "snake eyes" as a function of sample size (n)

6.8 Sample statistics and population parameters

The unknown quantities of interest described in the previous section are examples of parameters. A parameter is a numerical property of a population. One may be interested in measures of central tendency or dispersion in populations. Two parameters of interest for our purposes are the mean and standard deviation. The population mean and standard deviation are represented by μ and σ, respectively. The population mean, μ, could represent the average treatment effect in the population of individuals with a particular condition. The standard deviation, σ, could represent the typical variability of treatment responses about the population mean. The corresponding properties of a sample, the sample mean and the sample standard deviation, are typically represented by $\bar{x}$ and s, which were introduced in Chapter 5. Recall that the term "parameter" was encountered in Section 6.5 when describing the two quantities that define the normal distribution. In statistical applications, the values of the parameters of the normal distribution cannot be known, but are estimated by sample statistics. In this sense, the use of the word "parameter" is consistent between the earlier context and the present one. We have adhered to convention by using the term "parameter" in these two slightly different contexts.

An expression that defines how individual observations are used to derive a numerical estimate is called an estimator (much like a formula is used to calculate a number). The sample mean,

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

is considered an estimator for the population mean, μ. When individual observations are applied to the estimator, the result is a numeric value or estimate. When a single value is calculated, it represents a best guess of sorts, and is called a point estimate. No single estimate could be expected to be perfect so "interval estimates" are commonly used to reflect more accurately a range of plausible values.

Inferential statistics comprises two distinct, although closely related, procedures. In each case observations from a sample are used to:

- calculate an interval estimate that includes the unknown population parameter with some degree of confidence; in clinical trials, it is common practice to use a 95% confidence interval
- test whether or not a sample statistic is consistent with or contrary to a hypothesized value of the population parameter.

Inferences about a population are made on the basis of a sample taken from that population. The process of inferential statistics requires:

- identification of a representative sample of participants from a population of interest
- collection of individual observations
- calculation of sample statistics from the individual observations
- a statistical method to relate the sample statistic to the parameter of interest; this can be done in one of two ways:
  - – estimation of plausible values of the parameter
  - – testing a hypothesis of a proposed value of the parameter.

We discuss the former method, confidence intervals, first, after a necessary introduction to the concept of sampling variation. The latter method (hypothesis testing) is discussed later. First, however, it is useful to introduce a few other ideas.

6.9 Sampling variation

If we take a sample of 100 numbers from a population of 100 000 numbers, that sample's mean, which is precisely known, will provide an estimate of the population mean. The same is true for the standard deviation, that is:

- $\bar{x}$ is an estimate of μ
- s is an estimate of σ.

If we replaced the first sample of 100 numbers and then took another sample of 100 numbers, it is likely (effectively guaranteed) that the numbers would not be identical to those in the first sample, and that the calculated sample mean would be different from the first one. This logic applies to any number of means taken. Suppose that we were to repeat this process a number of times (in a simulated manner using computer software) and, at the end of each replication, tabulate the values of the sample mean or plot their relative frequencies. The shape of the resulting distribution of values would be recognizable. We would notice that a typical value would be apparent (the population mean μ), as would a symmetrical bell-shaped distribution. In short, the sample statistic from a sample of size n (in this case the sample mean) varies from sample to sample and its distribution has a mean and a standard deviation. Such a distribution is called a sampling distribution. An important general result for the sampling distribution of the sample mean is as follows:

- For any continuous random variable X which has a distribution with population mean, μ, and variance, σ^2, the sampling distribution of the mean for samples of size n has a distribution with population mean, μ, and variance, σ^2/n.

The square root of the variance,

$$\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}},$$

is the population standard error of the mean, which describes the typical variability of sample means around the population mean. If we know, or can assume, that the random variable X has a normal distribution with population mean, μ, and variance, σ^2, the sampling distribution of the mean of samples of size n will also have a normal distribution with population mean, μ, and variance, σ^2/n. Using the notation described earlier, this result can be summarized in this manner:

$$\text{If } X \sim N(\mu,\sigma^2) \text{ then } \bar{X}_n \sim N(\mu,\frac{\sigma^2}{n}).$$

6.10 Estimation: General considerations

It is not possible to *know* whether any single sample estimate, like the sample mean, is a good estimate of the population parameter that it is intended to estimate. However, it is possible to use the fact that most estimates of the sample statistic (for example, sample mean) are not too far removed from the population parameter, as specified by the shape of the sampling distribution, to define a range of values of the population parameter (for example, population mean) that are best supported by the sample data.

As exact knowledge of the population parameter is not possible, we must settle for a range of values that, with some specified probability or confidence, are most plausible. In other words, we would like to know the lower limit (LL) and upper limit (UL) of the most probable range of values of the true population parameter. In the case of the population mean, we seek two values, LL and UL, such that:

$$P(\text{LL} < \mu < \text{UL}) = 1 - \alpha.$$

The quantity α is the probability that the interval estimate does not include the value of the parameter of interest – that is, μ in this case. In most cases small values of α are desirable (for example, 0.10 or 0.05). Depending on the importance of the decision to be made on the basis of the interval estimate defined by LL and UL, very small values of α may be desirable (for example, 0.01 or 0.001).

When conducting a clinical trial, we do not know if our sample was representative of the population or not. We have only data from a sample and the statistics calculated from the sample data. Yet, our ultimate interest is not in the sample but in the population. In this chapter we consider the sample statistics for the mean and the standard deviation, $\bar{x}$ and s. A clinical trial represents a situation in which we can take only one sample from a population. Given that, what degree of certainty can we have that the mean of that sample represents the mean of the population? Before we answer this question fully we define a confidence interval for the sample mean in a special case. This special case will serve as our starting point for more realistic and common cases.

6.10.1 Confidence interval for the population mean when the population variance is known

Assume that the random variable X has a normal distribution with an *unknown* population mean, μ, and with a *known* population variance, σ^2. For a sample size of n, the sampling distribution of the sample mean has a normal distribution with population mean, μ, and variance, σ^2/n. The implication of this result is that, for example:

- 90% of the sample mean values lie between $\mu \pm 1.645 \frac{\sigma}{\sqrt{n}}$
- 95% of the sample mean values lie between $\mu \pm 1.960 \frac{\sigma}{\sqrt{n}}$
- 99% of the sample mean values lie between $\mu \pm 2.576 \frac{\sigma}{\sqrt{n}}$.

In general, the following statement is true: For samples of size n, $(1 - \alpha)$% of sample means $\bar{x}$ lie in the range:

$$\mu \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where $z_{1-\alpha/2}$ is the value from the standard normal distribution that defines the upper and lower tail areas of size $\alpha/2$. Note that $z_{1-\alpha/2}$ is a particular example of a reliability factor. As the Z distribution is symmetric it is also true that the Z value on the negative side that cuts off an area of size $\alpha/2$ in the lower tail is equal to the Z value on the positive side (change in sign) that cuts off an area of size $\alpha/2$ in the upper tail. Equivalently, in mathematical terms, this means:

$$|z_{\alpha/2}| = z_{1-\alpha/2}.$$

We can therefore express a two-sided $(1 - \alpha)$% confidence interval for the population mean as:

$$P(\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha.$$

This expression for a two-sided $(1 - \alpha)$% confidence interval can be shortened in the following manner:

$$\bar{x} \pm z_{1-\alpha/2} (\sigma/\sqrt{n}).$$

The assumption of a normal distribution for the random variable X is somewhat restrictive. However, for any random variable, as the sample size increases, the sampling distribution of the sample mean becomes approximately normally distributed according to a mathematical result called the central limit theorem. For a random variable X that has a population mean, μ, and variance, σ^2, the sampling distribution of the mean of samples of size n (where n is large, that is, > 200) will have an approximately normal distribution with population mean, μ, and variance, σ^2/n. Using the notation described earlier, this result can be summarized as:

$$\bar{X}_n \to N(\mu, \frac{\sigma^2}{n})$$

when n is large. This is an important result, because it holds no matter the shape of the original distribution of the random variable, X. The reader is encouraged to search for online references that illustrate, through animation, this important theorem.

Therefore, the expression written above for the confidence interval for the population mean also applies to any continuous random variable as long as the sample size is large (as just noted, of the order of 200 or more). The other rather restrictive assumption required for this confidence interval is that the population variance be known. Such a scenario is neither common nor realistic.

We now apply the fundamental concept of the confidence interval as developed here to the case where the population variance is not known, but there is interest in defining a confidence interval for the population mean.

6.10.2 Confidence interval for the population mean when the population variance is unknown

A reasonable suggestion for devising a confidence interval for the population mean would be to substitute the sample estimate, s, for the corresponding population parameter, α and proceed as described earlier in Section 6.10. However, when the sample size is small (particularly < 30) the use of the Z distribution is less appropriate. William S Gossett, writing anonymously as "Student" while employed at Guinness Brewery, proposed the following statistic as an alternative. When X is a normally distributed variable and the sample size is small, the statistic

$$\mathrm{t} = \frac{\bar{x} - \mu}{\mathrm{s}/\sqrt{n}}$$

follows a t distribution ("Student's t"). The single parameter defining its shape is $(n - 1)$ degrees of freedom (df), the sufficient number of observations needed to estimate the sample mean. The t distribution is symmetric about its mean (zero) and looks like a normal distribution with, in cases of sample sizes > 200, heavier "tails."

Three density functions are plotted for t distributions with 5, 30, and 200 df in Figure 6.6. The greater the number of df, the "flatter" the tails. In the figure, the two curves that are closest together are associated with 30 and 200 df. It is interesting to note (and a convenient fact) that the area under the density curve between any two points for the case with 30 df is not appreciably different from the case with 200 df.

As was the case with the normal distribution, the shape of the t distribution can be used to find two values that define a central area under the density curve of size $(1 - \alpha)$. It can be shown that, once a value of t associated with an area of interest is determined, the difference between the sample mean $\bar{x}$ is within $t(s/\sqrt{n})$ of the population mean, μ. This enables us to calculate a confidence interval for the population mean when the sample size is small and the population variance unknown.

The interval estimate of the population mean, the two-sided $(1 - \alpha)\%$ confidence interval for the population mean, is:

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

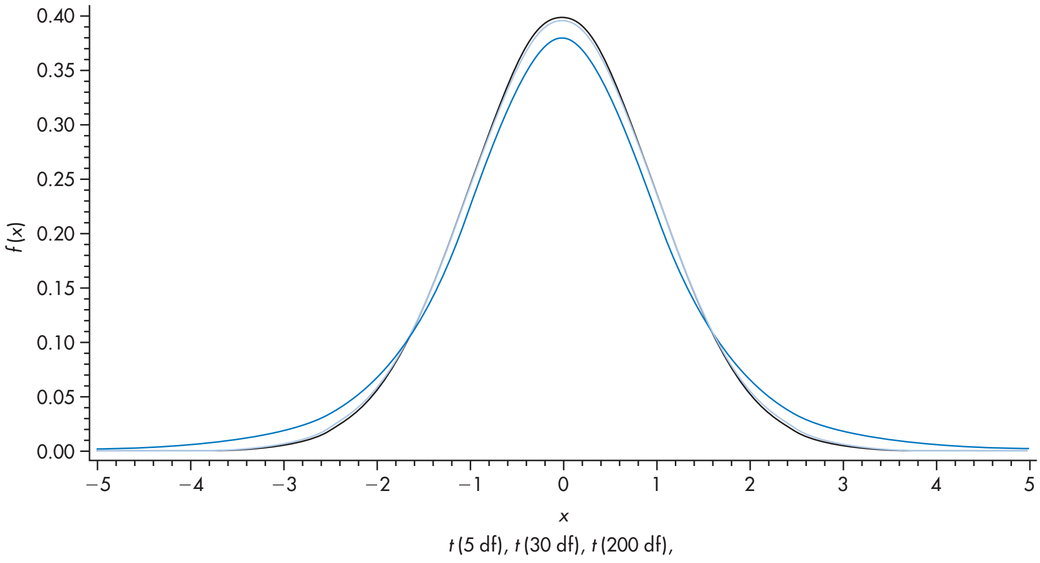

Figure 6.6 The t distribution with 5, 30, and 200 degrees of freedom

As in Section 6.10, the confidence interval has three components:

1. point estimate
2. standard error
3. reliability factor.

The point estimate in this case is the sample mean, which represents the best estimate of the population mean.

The second component is the standard error of the mean, which quantifies the extent to which the process of sampling has mis-estimated the population mean. The standard error of the mean has the same meaning as in the case for normally distributed data – that is, the standard error describes the degree of uncertainty present in our assessment of the population mean on the basis of the sample mean. It is also the standard deviation of the sampling distribution of the mean for samples of size *n*. The smaller the standard error, the greater the certainty with which the sample mean estimates the population mean. When *n* is very large the standard error is very small, and therefore the sample mean is a very precise estimate of the population mean. As we know the standard deviation of the sample, *s*, we can make use of the following formula to determine the standard error of the mean, SE:

$$\text{SE} = \frac{s}{\sqrt{n}}.$$

At this point, it is worth emphasizing the difference between the terms "standard error" and "standard deviation," which, despite the same initial word, represent very different aspects of a data set. Standard error is a measure of how certain we are that the sample mean represents the population mean. Standard deviation is a measure of the dispersion of the original random variable. There is a standard error associated with any statistical estimator, including a sample proportion, the difference in two means, the difference in two proportions, and the ratio of two proportions. When presented with the term "standard error" in these applications the concept is the same. The standard error quantifies the extent to which an estimator varies over samples of the same size. As the sample size increases (for the same standard deviation) there is greater precision in the estimate of the population mean because the standard error becomes smaller as a result of the division of the square root of the sample size.

The third component of the interval estimate is a reliability factor, which represents the number of standard deviations required to enclose (1 − α)% of the sample means from the sampling distribution. It is used to quantify how close we would like our estimate to be to the real population mean or, in short, how reliable it is. The particular value of the reliability factor chosen above, $t_{1-\alpha/2,n-1}$, is the value of *t* with ($n - 1$) df that "cuts off" an area of α/2 in the upper tail. As the *t* distribution is symmetric it is also true that the *t* value on the negative side that cuts off an area of size α/2 in the lower tail is equal to the same *t* value but with a change in sign that cuts off α/2 in the upper tail. Equivalently, in mathematical terms, this means that $|t_{\alpha/2,n-1}| = t_{1-\alpha/2,n-1}$. Values of $t_{1-\alpha/2,n-1}$ are provided in Appendix 2 for various values of α.

For a sample size of 100 (99 df) the reliability factors for two-sided 90%, 95%, and 99% confidence intervals are 1.66, 1.98, and 2.63. The implication of these three values is that, all other things being equal (that is, $\bar{x}$, *s*, and *n*), requiring greater confidence in the interval estimate results in wider interval estimates. The more confidence that is required, the less reliable is the single sample estimate, and therefore greater numerical uncertainty is expressed in the interval estimate. This very important point is illustrated in the following example.

The following values of age ($n = 100$) were examined using a stem-and-leaf display in Chapter 5:

53, 69, 72, 48, 60, 61, 49, 71, 43, 31, 62, 51, 58, 61, 70, 66, 78, 39, 75, 63, 59, 53, 49, 61, 50, 88, 51, 80, 68, 75, 78, 81, 57, 70, 68, 66, 43, 60, 57, 35, 75, 61, 71, 45, 50, 82, 52, 65, 61, 77, 80, 58, 50, 59, 55, 59, 50, 39, 78, 72, 71, 79, 48, 55, 52, 55, 62, 59, 68, 63, 81, 69, 67, 67, 58, 57, 70, 73, 49, 43, 76, 73, 71, 77, 61, 62, 72, 73, 67, 62, 64, 40, 66, 74, 77, 67, 49, 83, 73, 59.

Assuming that these observations represent a simple random sample from the population of interest and that the age values in the population are normally distributed, the task is to calculate the two-sided 90%, 95%, and 99% confidence intervals for the population mean age.

The sample mean age is 62.6 and the standard deviation 12.01. The standard error of the mean is calculated as:

$$\text{SE} = 12.01/\sqrt{100} = 1.20.$$

With these numbers calculated, all that is left to compute the three confidence intervals are the reliability factors associated with each. For the 90% confidence interval, the value of the reliability factor will be the value of *t* that cuts off the upper 5% of the area (half the size of α) under the *t* distribution with 99 df. This value is 1.66 and can be verified from a table of values or from statistical software. Note that the *t* value of −1.66 is the value of *t* that cuts off the lower 5% of the area (half of the size of α) under the *t* distribution with 99 df. The reliability factors listed previously for the two-sided 95% and 99% confidence intervals can also be used to compute the following interval estimates:

- 90% CI = 62.6 ± (1.20)(1.66) = (60.6, 64.6)
- 95% CI = 62.6 ± (1.20)(1.98) = (60.2, 65.0)
- 99% CI = 62.6 ± (1.20)(2.63) = (59.4, 65.8).

Note that the two values that comprise the lower and upper limits of the confidence interval are typically placed in parentheses. The width of the confidence intervals (the difference between the upper and lower limits) increases because greater confidence (corresponding to smaller values of α) is required.

A statistical interpretation of these results is to say that we are 90% confident that the mean age of the population from which this sample was selected is enclosed in the interval 60.6–64.6 years. If greater confidence is required, we can say that with 99% confidence the mean age of the population is enclosed in the interval 59.4–65.8 years. Another interpretation of these confidence intervals is that they represent the most plausible values of the population mean. It is important to note that the lower and upper limits of the confidence interval are random variables. The population mean is considered to be an unknown fixed quantity for which the confidence interval serves as an estimate.

To summarize, the computational aspects of confidence intervals involve a point estimate of the population parameter, some error attributed to sampling, and the amount of confidence (or reliability) required for interpretation. We have illustrated the general framework of the computation of confidence intervals using the case of the population mean. It is important to emphasize that interval estimates for other parameters of interest will require different reliability factors because these depend on the sampling distribution of the estimator itself and different calculations of standard errors. The calculated confidence interval has a statistical interpretation based on a probability statement.

Another useful interpretation of confidence intervals is that the values that are enclosed within the confidence interval are those that are considered the most plausible values of the unknown population parameter. Values outside the interval are considered less plausible. All other things being equal, the need for greater confidence in the estimate results in wider confidence intervals, and confidence intervals become narrower (that is, more precise) as the sample size increases. This last fact is explored in greater detail in Chapter 12 because it is directly relevant to the estimation of the required sample size for a clinical trial. The methods to use for the calculation of confidence intervals for other population parameters of interest are provided in subsequent chapters.

6.11 Hypothesis testing: General considerations

As this book focuses on clinical trials our primary interest is in providing you with relevant examples of hypothesis testing in that arena. However, it is useful initially to lay some conceptual foundations with simpler examples. As for many other examples in statistics and probability, we illustrate these concepts first with flips of a coin.

Imagine the following scenario. You are holding a half-dollar coin. Our question to you is: Do you think this coin is fair or not? You examine it and hold it and with no other information you decide that you really cannot tell without more information. You propose to flip the coin twice. Flipping it twice, you get the following results: Heads (H) and heads (H). If we forced you to answer our question at this time, you may guess, based on these two observations, that the coin is not fair. After all, if the coin were really fair you would "expect" one head (H) and one tail (T). However, you are not at all confident with your answer because you note that the probability of observing two heads is not that small. It is (0.5)(0.5) = 0.25. This means that an outcome like this results 25% of the time that you conduct such an experiment. Accordingly, you wisely recognize that it would be better to have additional data before making your guess, because with just two heads observed out of two flips there is a non-trivial chance that you have guessed incorrectly.

Suppose you then revise the experiment and request that the results from 10 flips of the coin be recorded. You reason that, if the coin were fair, you would expect five heads and five tails. If you were to observe that only one or as many as nine heads came up out of ten tosses, you would conclude that the coin was not fair. Your logic is that, by chance alone, a fair coin would not very likely yield such a lopsided result. If you were to observe an event with even more extreme result, that is, 0 or 10 heads out of 10 tosses, you would also have concluded, perhaps with even more confidence, that the coin was not fair.

The rule that you intuitively arrived at was that if you observed as few as 0 or 1 or as many as 9 or 10 heads out of 10 coin flips, you would conclude that the coin was not fair. How likely is it that such a result would happen? In other words, suppose you repeated this experiment a number of times with a truly fair coin. What proportion of experiments conducted in the same manner would result in an erroneous conclusion on your part because you followed the evidence in this way? This is the point where the rules of probability come into play. You can find the probability of making the wrong conclusion (calling the fair coin biased) by following such a decision rule using the binomial distribution.

Using the binomial distribution, the probability of observing 9 heads out of 10 when the probability of observing a head with each trial is ½ is 0.00977. Likewise, the probability of observing 10 heads out of 10 is 0.00098. So the probability of observing either 9 or 10 heads is the sum of these two (we sum them because these are mutually exclusive outcomes). That probability is 0.01075 (around 1%). We note that the probability of observing 0 or 1 heads is the same as for observing 9 or 10. Therefore the probability of observing a result as extreme as 1 or fewer or 9 or more heads is around 0.02. If after 10 coin flips, we have observed 1 or fewer heads or 9 or more heads, we would conclude that the coin was biased because a fair coin would yield such a result only with probability around 2% (not very often). Put another way, if we conducted this experiment many times and used such a rule when we have observed such a result, we would be incorrect in 2% of the experiments. That seems like an acceptable risk to take. Besides, in this scenario, there seems to be no adverse consequence to being wrong except for a bit of damaged pride.

This rather simple example is an illustration of the conceptual components of hypothesis testing. The basis of hypothesis testing is "proof by contradiction." We use the word "proof" rather liberally here because the scientific standard for establishing proof is more rigorous than a single trial or set of trials could possibly provide. Hypothesis testing is a statistical method in which we use data (evidence) to choose between two decisions, each with their own course of action and related implications. The real world implication of making either decision depends on the field of study. In the world of new drug development, these decisions could be to decide that a drug is not efficacious at any dose studied, and is therefore not worth studying further. Another decision could be to select one particular dose (among many studied) for further development in confirmatory trials.

The process of testing a hypothesis usually begins with the statement of the hypothesis that we would like to conclude as a result of the research (we refer to this as the alternate hypothesis). There is another hypothesis that we need to define and it is referred to as the null hypothesis. The null hypothesis can be viewed as a "straw man" hypothesis, one that we would like to knock over by collecting evidence that contradicts it in favor of the alternate hypothesis. One important consideration in the statement of the two hypotheses is that they should represent all possible outcomes. In the context of our coin-tossing illustration, the alternate hypothesis would be that the coin is biased and the null hypothesis would be that the coin is fair. In that experiment, we counted the number of heads and were looking for evidence that would contradict the null hypothesis and compel us to conclude that the alternate hypothesis was true. Evidence that contradicted the null hypothesis would be a very high or very low proportion of heads because a fair coin would yield approximately the same number of heads as tails. Importantly, these two hypotheses cover the only two possible outcomes: The coin is either fair or biased.

The next part of the hypothesis-testing process is to decide on a numerical result (a test statistic) that, if observed, would sufficiently contradict the null hypothesis such that the null hypothesis would be rejected in favor of the alternate hypothesis. As we discovered with our coin-tossing example, some results would not be all that rare by chance alone. Therefore, our decision rule should be defined such that erroneous conclusions are not made more often than we are willing to tolerate.

You will recall that we might have chosen other results before we concluded that the coin was biased, but we chose results that would rarely be expected by chance alone. In fact, the decision rule is based on our chosen probability of rejecting the null hypothesis when it is really true. For the coin example, this is the probability of claiming that the coin is biased when it is really fair. When asked to take part in this experiment the fairness of the coin remains unknown to us, but we choose a decision rule that is consistent with results that would not be expected by chance very often.

6.11.1 Type I errors and type II errors

Rejecting the null hypothesis when it is true is called a type I error. The probability of making a type I error is called alpha (α). There is another kind of error that we might commit by using data from our sample (in this case, 10 coin tosses) to make an inference about the state of nature. This second kind of error is called a type II error and results from failing to reject the null hypothesis (suppose we observed seven heads) when, in fact, the alternate hypothesis is true (that is, the coin was biased). We would then act as if the coin were fair – perhaps taking part in a new challenge that involved wagering a lot of money.

When making decisions of any type, whether they are as inconsequential as our coin-tossing experiment or as important and costly as developing a new drug, we would like to minimize the chances that we make the wrong decision. In planning a new study or experiment, such as a clinical trial, it is worthwhile to consider minimizing the probability of committing each of these errors. The two types of errors are presented in Table 6.3. In clinical trials a type I error is committed when we claim that the new antihypertensive is superior to placebo but it really is similar. A type II error is committed when we fail to claim the new antihypertensive is superior to placebo but it really is. In reality we cannot know the truth, but the study design, including the sample size and the statistical analyses used to evaluate the trial, will enable us to limit the probability of committing each of these errors.

6.11.2 Probability of type I and II errors

An important aspect of study design is defining the probabilities of committing each of these two kinds of errors. A type I error could mean that a new drug is approved for marketing but really does not provide a benefit. Ideally, the probability of committing a type I error of this type would be fairly small. Committing a

Table 6.3 Two possible errors in hypothesis testing

	Truth about null hypothesis	
Decision based on test statistic	**True**	**False**
Fail to reject null hypothesis	Correct	Type II error
Reject null hypothesis	Type I error	Correct

type II error in a superiority trial of a new antihypertensive is not appealing for a study sponsor because it could lead to discontinuation of a development program for a new treatment that is actually efficacious. Therefore, it is desirable to limit the probability of committing a type II error as well. We have more to say about these two probabilities in subsequent chapters, but for now it is sufficient to identify them formally.

The probability of committing a type I error is the probability of rejecting the null hypothesis when it is true (for example, claiming that the new treatment is superior to placebo when they are equivalent in terms of the outcome). The probability of committing a type I error is called α, which is sometimes referred to as the size of the test. The probability of committing a type II error is the probability of failing to reject the null hypothesis when it is false. This probability is also called beta (β). The quantity (1 − β) is referred to as the power of the statistical test. It is the probability of rejecting the null hypothesis (in favor of the alternate) when the alternate is true. As stated earlier it is desirable to have low error probabilities associated with a test. As we would like α and β to be as low as possible the quantities (1 − α) and (1 − β) are typically fairly large. These probabilities are provided in Table 6.4.

6.11.3 Hypothesis testing and research questions

Statistical hypothesis testing represents a means to formulate and answer the research question in a quantitative manner. The null hypothesis is the hypothesis that is tested. If quantitative data are produced that are not consistent with the null hypothesis, it is rejected.

Before proceeding with this statistical approach, a research question must be posed, which will then prompt the design of a study that will lead to the collection of data and an appropriate statistical analysis. A simple research question from a drug development program, as stated in Chapter 3, is “Does the investigational drug lower blood pressure?” A way to answer this research question is to design a study to estimate the mean change from baseline in SBP. If the mean change from baseline is negative, the answer to the research question would be that the investigational drug does lower blood pressure. This example will be used to illustrate the concept of hypothesis testing.

Table 6.4 Probabilities of outcomes (conditional on the null hypothesis) in hypothesis testing

	Truth about null hypothesis	
Decision based on test statistic	**True**	**False**
Fail to reject null hypothesis	1 − α	β
Reject null hypothesis	α	1 − β (i.e., power)

6.12 Hypothesis test of a single population mean

Suppose our interest is in testing whether the population mean was equal to a particular hypothesized value, μ_0. A hypothesis testing process typically starts with a statement of the null and alternate hypotheses. The null hypothesis can be stated in the following manner:

$$H_0: \mu = \mu_0.$$

If data are found to contradict the null hypothesis, it will be rejected in favor of the alternate hypothesis:

$$H_A: \mu \neq \mu_0.$$

The alternate hypothesis is two sided in the sense that values clearly less than μ_0 would be consistent with it as would values that were clearly greater than μ_0. Rejection of the null hypothesis because $\mu_0 << \mu$ (μ is much greater than the hypothesized value μ_0) may lead to one decision (for example, continue with the development of the new drug with a larger study) whereas rejection of the null hypothesis because $\mu_0 >> \mu$ (μ is much less than the hypothesized value μ_0) may lead to a completely different decision (for example, to stop development of the new drug because it has no effect on SBP or actually increases SBP). What is important is that, *a priori*, either outcome is possible.

The next step of the hypothesis testing process is to identify a numeric criterion by which the plausibility of the null hypothesis is tested. This numeric criterion is called the test statistic, and we use it to decide if the value that resulted from the study contradicts the null hypothesis or not. The test statistic to be used in this case is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

If the null hypothesis is true – that is, the population mean is the hypothesized value, μ_0 – the value of the test statistic will be close to 0. The further the test statistic value is from 0 (either negative or positive) the less plausible is the hypothesized value, μ_0 – that is, the null hypothesis should be rejected in favor of the alternate. The next step of hypothesis testing is to determine those values of the test statistic that would lead to rejection of the null hypothesis, that is, to determine the critical region.

Assuming that the random variable is normally distributed (or approximately so if the sample size is > 30) and if the null hypothesis is true, the test statistic just defined has a *t* distribution with $(n - 1)$ df. Referring to Figure 6.6 you will see that most values of a random variable that follow a *t* distribution fall in the range -1 to $+1$. A value in this range would be expected just by chance alone. However, values < -2 or $> +2$ occur much less frequently, that is, there is less area to the left of -2 and to the right of $+2$. We would like to define a critical region that is associated with small tail areas because values in the tail do not occur frequently, whereas values in the center of the distribution are very common. In other words, we would like to define the test so that we do not reject the null hypothesis very often when in fact it is true – that is, we would like to define a critical region so that the probability of committing a type I error, α, is small.

In most scientific endeavors the choice of α is 0.05. We are willing to accept a 1 in 20 chance that, at the end of the study, it is concluded that the population mean is not the hypothesized value when in fact it really is. It is important to remember that the choice of α is part of the study design, and not a result of a study. Also, it is important to note here that there is nothing special about the value of 0.05. Depending on the stage of development or the severity or importance of the disease for which we wish to develop the drug, we may choose a value of α that is higher or lower than 0.05. What is important in the choice of α are the implications (for sponsors, regulatory bodies, clinicians, and patients) of committing a type I error. Having alerted you to the possibility of choosing other values for α, and the fact that this choice has various implications, we adopt the conventional value of α of 0.05 in subsequent discussions.

Knowledge of the distribution of the test statistic enables us to define a critical region that would erroneously lead to rejection with probability of 0.05. In the case of the current test, the critical region will be any value of the test statistic such that:

$t < t_{\alpha/2,n-1}$ or $t > t_{1-\alpha/2,n-1}$.

Similarly to the case of the standard normal distribution, the critical values can be obtained from a series of tabulated values or from statistical software. A number of percentiles of various *t* distributions are provided in Appendix 2. It is important to note that there is not just one *t* distribution; there are many of them, and their shapes are determined by the number of degrees of freedom. As either low or high values of the test statistic could lead to rejection, the hypothesis test is considered a two-sided test. The probability of committing a type I error is 0.05, but, because the critical region is evenly split between low values and high values, the probability of committing a type I error in favor of one direction (for example, large values of *t*) is α/2.

Once the critical region of the test has been defined, the next step of hypothesis testing is to calculate the value of the test statistic from the sample data. The test statistic is calculated as:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where $\bar{x}$ is the sample mean, μ_0 the hypothesized value of the population mean, *s* the sample standard deviation, and *n* the sample size.

If the value of the test statistic is in the critical region the null hypothesis is rejected and the conclusion is made that the population mean is not equal to μ_0. When the null hypothesis is rejected, such a result is considered "statistically significant" at the α level, meaning that the result was unlikely (with probability no greater than α) to have been observed by chance alone. If the value of the test statistic is not in the critical region we fail to reject the null hypothesis. It is important to emphasize the fact that we cannot claim that the population mean is equal to μ_0, but simply that the data were not sufficient to conclude that they were different.

The use of this method, the one-sample *t* test, is appropriate when:

- the observations represent a simple random sample from the population of interest
- the random variable is continuous
- the random variable is normally distributed or approximately normally distributed (mound shaped) with a sample size of at least 30.

This hypothesis test is illustrated with the following simple example.

Imagine that, having identified a promising new investigational antihypertensive drug, a pharmaceutical company would like to administer it to a group of 10 hypertensive individuals to see if the drug has the desired effect. For simplicity we assume that there is no control group. The first study of the new antihypertensive will be a single-dose, nonrandomized, uncontrolled trial in 10 participants. SBP was recorded at the start of the study before initiation of treatment (baseline) and at the end of 4 weeks (end of study). The research question of interest is: Does the new drug lower SBP? The scientists designing the trial would like to maintain a type I error of 0.05, that is, $\alpha = 0.05$. As the test conducted is two sided, the probability of making a type I error in favor of the drug having a beneficial effect (one side of the critical region) is 0.025.

The null hypothesis is:

H_0: $\mu = 0$.

And the alternate hypothesis is:

H_A: $\mu \neq 0$.

The one-sample *t* test will be used to test the null hypothesis. As there are 10 observations and assuming the change scores (the random variable of interest) are normally distributed, the test statistic will follow a *t* distribution with 9 df. A table of critical values for the *t* distribution (Appendix 2) will inform us that the two-sided critical region is defined as $t < -2.26$ and $t > 2.26$ – that is, under the null hypothesis, the probability of observing a *t* value < -2.26 is 0.025 and the probability of observing a *t* value > 2.26 is 0.025.

Baseline and end-of-study values of SBP are presented for the 10 participants in Table 6.5, along with their respective change scores.

The mean change score is −7 and the standard deviation is 7.1. (We leave it to you to verify this.) The test statistic is therefore calculated as:

$$t = \frac{-7 - 0}{7.1/\sqrt{10}} = -3.10.$$

Table 6.5 Systolic blood pressure (SBP) values and change scores

Study participant	Baseline SBP (mmHg)	End-of-study SBP (mmHg)	Change in SBP (mmHg)
1	143	147	4
2	152	144	−8
3	162	159	−3
4	158	157	−1
5	147	131	−16
6	149	133	−16
7	150	145	−5
8	148	144	−4
9	154	150	−4
10	149	132	−17

As this calculated test statistic is in the critical region ($t = -3.10 < -2.26$) the null hypothesis is rejected. The result is considered statistically significant at the $\alpha = 0.05$ level because there was less than a 5% chance of such a result being observed by chance alone. The conclusion from the study is that the new drug did lower SBP by a mean of 7 mmHg. Scientists from the sponsor company may use this information as sufficient preliminary evidence to continue with the development of the new drug.

6.13 The *p* value

One shortcoming of the hypothesis testing approach is the arbitrary choice of a value for α. Depending upon our risk tolerance for committing a type I error, the conventional value of 0.05 may not be acceptable. Another way to convey the "extremeness" of the resulting test statistic is to report a *p* value.

A *p* value is the probability that the result obtained or one more extreme (in favor of the alternate) would be observed by chance alone. We know from the definition of the critical region that a value of the test statistic $t < -2.26$ or $t > 2.26$ would have occurred with probability ≤ 0.05 by chance alone. In fact, the test statistic value was −3.10 which lies to the left of −2.26. A value of −3.10 led to rejection, as would values < -3.10 or > 3.10. The *p* value in this case is the area under the *t*-distribution density curve with 9 df associated with values of $t < -3.10$ or > 3.10 and is equal to 0.01. This means that there is only a 1% chance of observing a value of the test statistic as large as 3.10 (in absolute magnitude) or larger by chance alone. The difference between α (a design parameter) and the *p* value (a study result) can be seen in Figure 6.7, where the areas to the left and right of the dashed lines represent α and the areas to the left and right of the solid line represent the *p* value.

The *p* values can be estimated from a table of values from the appropriate *t* distribution (for example, by finding the tail areas associated with a particular value of the test statistic). More commonly, however, statistical software is used for all statistical analyses and *p* values are included in the results. The following is a helpful way to interpret *p* values:

- Hypothesis tests are rejected if the calculated *p* value $\leq \alpha$.
- Hypothesis tests are not rejected if the calculated *p* value $> \alpha$.

It is not uncommon for results of hypothesis tests to be represented simply by the *p* value. However, it is not a wise practice to rely solely on them. Recall that increasing the sample size reduces the standard error, which increases the size of the test statistic and therefore reduces the *p* value. This serves as a reminder that it is not just the statistical significance of the result (that is, the *p* value) that counts. The clinical relevance of the size of the effect (for example, the

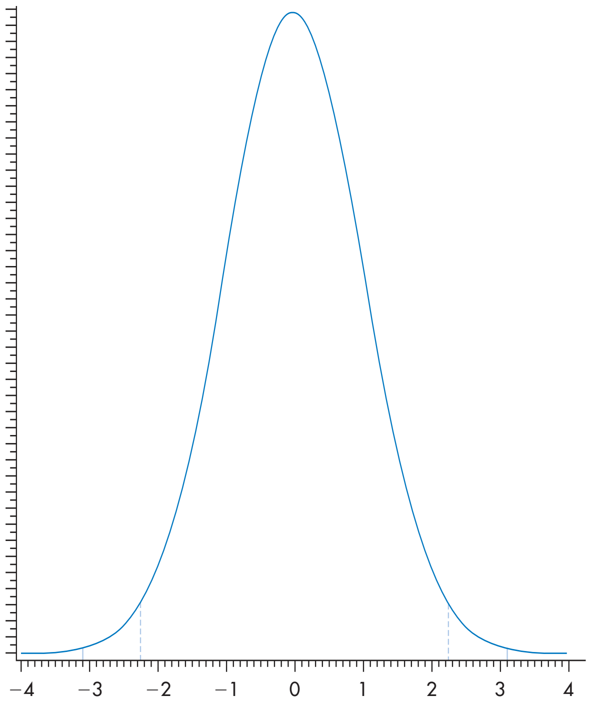

Figure 6.7 The *t* distribution with 9 degrees of freedom, critical region (dashed line) and *p* value (solid line). Note that the critical region is represented by the tail areas to the left and right of the dashed lines; the *p* value is represented by the tail areas to the left and right of the solid lines

confidence interval for the parameter of interest) is probably more important than the *p* value, as we argue in the following section.

6.14 Relationship between confidence intervals and hypothesis tests

Confidence intervals can be used to test a number of hypotheses. This is illustrated using the study data from the previous example in Section 6.12.

Scientists from the pharmaceutical company believe that reporting a 95% confidence interval for the population mean change in SBP may prove helpful. Following the confidence interval defined in Section 6.10, a 95% confidence interval for the population mean is:

$$-7 \pm 2.26(7.1/\sqrt{10}) = (-12.1, -1.9).$$

The scientists can report from this study that they are 95% confident that the true population mean change in SBP is within the interval (−12.1, −1.9). One interpretation of this interval is that the scientists are 95% confident that the drug works by reducing SBP, as evidenced by an upper limit of the confidence interval that is less than 0. Another less favorable interpretation is that the drug does not work all that well – after all, the confidence interval does not rule out some very minor reductions in SBP (upper limit of −1.9 mmHg). It is true that, had the scientists hypothesized a value of the population mean outside of the values of this 95% confidence interval, the null hypothesis would have been rejected at the $\alpha = 0.05$ level. For example, the following null hypotheses would have been rejected:

H_0: $\mu = 2$
H_0: $\mu = -15$.

Conversely, the following null hypotheses would not have been rejected:

H_0: $\mu = -8$
H_0: $\mu = -2$.

This relationship can be stated more generally as:

- All values outside the $(1 - \alpha)\%$ confidence interval for a parameter of interest would be rejected by a hypothesis test (of size α) of the parameter.
- Values within the $(1 - \alpha)\%$ confidence interval for a parameter of interest would fail to be rejected by a hypothesis test (of size α) of the parameter.

In this example, μ represents the population mean change from baseline SBP. If the upper limit of the 95% confidence interval excludes 0, negative values of population mean are most plausible, implying that the drug lowered SBP. If the lower limit of the 95% confidence interval excludes 0, positive values of the population mean are most plausible, implying that the drug actually increased SBP. If 0 is enclosed in the 95% confidence interval, negative and positive values of the mean are most plausible, implying that we cannot rule out the possibility that the drug had no effect. These three scenarios are displayed in Figure 6.8.

As confidence intervals can be used to test a number of hypotheses simultaneously, they convey much more information than a single *p* value resulting from a hypothesis test. In addition to being able to test various hypotheses (the null hypothesis of zero change from baseline was rejected) the confidence interval allows regulatory agencies and physicians who review the data to interpret the clinical relevance of the magnitude of the values within the confidence interval.

6.15 Brief review of estimation and hypothesis testing

This chapter started with an introduction to the concepts of probability and random variable distributions. The role of probability is to assist in our ability to make statistical inferences. Test statistics are the numeric results of an experiment or study. The yardstick by which a test statistic is measured is how extreme it is. The term “extreme” in Statistics is used in relation to a value that would have been expected if there was no effect, that is, the value that would be expected by random chance alone. Confidence intervals provide an interval estimate for a population parameter of interest. Confidence intervals of $(1 - \alpha)\%$ can also be used to test hypotheses, as seen in Chapter 8.

The process of hypothesis testing is carried out using the following steps, which will be highlighted in subsequent chapters:

- State the null and alternate hypotheses. It is sometimes easier to state the alternate hypothesis first because that is what we would like to conclude at the end of the study. The null hypothesis then covers the remainder of values of the population parameter. The specific statements of the null and alternate hypotheses depend on the type of study and the analysis approach used. We cover many different examples in later chapters.
- Determine the test statistic appropriate for the method used. Choosing the appropriate test statistic depends on the analysis method and the assumptions that we must make.
- Select a value of α (as noted earlier, our standard for this book is 0.05).

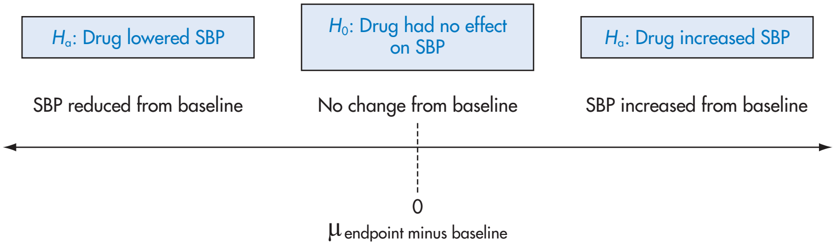
H_a: Drug lowered SBP
H_0: Drug had no effect on SBP
H_a: Drug increased SBP
SBP reduced from baseline
No change from baseline
SBP increased from baseline
0
$\mu_{\text{endpoint minus baseline}}$

Figure 6.8 Conclusions to be drawn from the population mean change from baseline

- Calculate the value of the test statistic under the null hypothesis and the corresponding *p* value. Compare the *p* value with the value of α.
- State the statistical decision either to reject or to fail to reject the null hypothesis.

Statistical inference is one way to use data to make a decision in the presence of uncertainty. The resulting decisions are not perfect. The commission of either a type I or a type II error can have significant impacts on drug companies, study participants, patients, and public health. Therefore, minimizing the probability that each might occur is an important part of the study design, including the manner in which data are analyzed and interpreted.

6.16 Review

1. Using Table 6.1, calculate the probability of selecting a participant who is:

   (a) female
   (b) female and ≥ 65 years of age
   (c) ≥ 65 years of age
   (d) female given that the participant is ≥ 65 years of age.

2. Show that the true negative rate of a diagnostic test is a function of the sensitivity and specificity of the test and the prevalence of the disease.

3. Assume that SBP among all adults aged 30 years and older in the UK has a normal distribution with mean 120 mmHg and variance 100 mmHg. What proportion of participants in this population has:

   (a) SBP $<$ 90 mmHg?
   (b) SBP $<$ 120 mmHg?
   (c) SBP $<$ 100 mmHg or SBP $>$ 140 mmHg?
   (d) SBP $>$ 160 mmHg?

4. What is the difference between standard deviation and standard error?

5. What is α? How does a researcher decide on a value for α?

6. What is β? How does a researcher decide on a value for β?

7. What are the three components of a confidence interval?

8. What is a two-sided hypothesis test?

9. The one-sample *t* test is being used for a two-sided test of the null hypothesis, H_0: $\mu = 0$. For each of the following scenarios, define the rejection region for the test:

   (a) $n = 10$; $\alpha = 0.10$
   (b) $n = 10$; $\alpha = 0.01$
   (c) $n = 30$; $\alpha = 0.05$
   (d) $n = 30$; $\alpha = 0.001$.

10. For each of the following 95% confidence intervals for the population mean, would a two-sided test of the null hypothesis, H_0: $\mu = 0$, be rejected or not rejected?

    (a) $(-4.0, 4.0)$
    (b) $(-2.0, -1.0)$
    (c) $(22.3, 44.6)$
    (d) $(-12.7, 0.01)$.

6.17 References

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

vos Savant M (1997). Ask Marilyn. *Parade Magazine* 30 March, p. 15.