CHAPTER 1

Introduction & Basics

1.1 Statistics—the Field

In some ways, we are all born statisticians. Inferring general patterns from limited knowledge is nearly as automatic to the human consciousness as breathing. Yet, when inference is formalized through the science of mathematics to the field called **Statistics**, it often becomes clouded by preconceptions of abstruse theory. Let's see if we can provide some formalization to this natural process of rational inference without getting bogged down in theoretical details.

The purpose of the field of **Statistics** is to characterize a *population* based on the information contained in a *sample* taken from that population. The sample information is conveyed by functions of the observed data, which are called *statistics*. The field of **Statistics** is a discipline that endeavors to determine which functions are the most relevant in the characterization of various populations. (The concepts of 'populations', 'samples', and 'characterization' are discussed in this chapter.)

For example, the arithmetic mean might be the most appropriate statistic to help characterize certain populations, while the median might be more appropriate for others. Statisticians use statistical and probability theory to develop new methodology and apply the methods best suited for different types of data sets.

Applied Statistics can be viewed as a set of methodologies used to help carry out scientific experiments. In keeping with the *scientific method*, applied statistics consists of developing a hypothesis, determining the best experiment to test the hypothesis, conducting the experiment, observing the results, and making conclusions. The statistician's responsibilities include: study design, data collection, statistical analysis, and making appropriate inferences from the data. In doing so, the statistician seeks to limit bias, maximize objectivity, and obtain results that are scientifically valid.

Populations

A *population* is a universe of entities to be characterized but is too vast to study in its entirety. The population in a clinical trial would be defined by its limiting conditions, usually specified via study inclusion and exclusion criteria.

Examples of populations include:

- patients with mild-to-moderate hypertension
- obese teenagers
- adult, insulin-dependent, diabetic patients.

The first example has only one limiting factor defining the population, that is, mild-to-moderate hypertension. This population could be defined more precisely as patients with diastolic blood pressure within a specific range of values as an inclusion criterion for the clinical protocol. Additional criteria would further limit the population to be studied.

The second example uses both age and weight as limiting conditions, and the third example uses age, diagnosis, and treatment as criteria for defining the population.

It is important to identify the population of interest in a clinical study at the time of protocol development, because the population is the 'universe' to which statistical inferences might apply. Severely restricting the population by using many specific criteria for admission might ultimately limit the clinical indication to a restricted subset of the intended market.

Samples

You can describe a population by describing some representative entities in it. Measurements obtained from sample entities tend to characterize the entire population through inference.

The degree of representation of the entities in a sample that is taken from the population of interest depends on the sampling plan used. The simplest type of sampling plan is called a 'simple random sample'. It describes any method of selecting a sample of population entities such that each entity has the same chance of being selected as any other entity in the population. It's easy to see how random samples should represent the population, and the larger the sample, the greater the representation.

The method of obtaining a simple random sample from the population-of-interest is not always clear-cut. Simple random samples are rarely, if ever, used in clinical trials. Imagine the patients who comprise the populations in the three examples cited earlier, living all over the world. This would make the collection of a simple random sample an overwhelming task.

Although inferences can be biased if the sample is not random, adjustments can sometimes be used to control bias introduced by non-random sampling. An entire branch of **Statistics,** known as S*ampling Theory*, has been developed to provide alternative approaches to simple random sampling. Many of these approaches have the goal of minimizing bias. The techniques can become quite complex and are beyond the scope of this overview.

For logistical reasons, clinical studies are conducted at a convenient study center with the assumption that the patients enrolled at that center are typical of those that might be enrolled elsewhere. Multi-center studies are often used to blunt the effect of characteristics of the patient or of procedural anomalies that might be unique to any specific center.

Stratified sampling is another technique that is often used to obtain a better representation of patients. Stratified sampling uses random samples from each of several subgroups of a population, which are called 'strata'. Enrollment in a study is sometimes stratified by disease severity, age group, or some other characteristic of the patient.

Because inferences from non-random samples might not be as reliable as those made from random samples, the clinical statistician must specifically address the issue of selection bias in the analysis. Statistical methods can be applied to determine whether the treatment group assignment 'appears' random for certain response variables. For example, baseline values might be lower for Group A than Group B in a comparative clinical study. If Group A shows a greater response, part of that perceived response might be a regression-toward-the-mean effect, that is, a tendency to return to normal from an artificially low baseline level. Such effects should be investigated thoroughly to avoid making faulty conclusions due to selection bias.

Additional confirmatory studies in separate, independent samples from the same population can also be important in allaying concerns regarding possible sampling biases.

► *Characterization*

So how is the population characterized from a sample? Statistical methods used to characterize populations can be classified as descriptive or inferential.

Descriptive statistics are used to describe the distribution of population measurements by providing estimates of central tendency and measures of variability, or by using graphical techniques such as histograms. *Inferential* methods use probability to express the level of certainty about estimates and to test specific hypotheses.

Exploratory analyses represent a third type of statistical procedure used to characterize populations. Although exploratory methods use both descriptive and inferential techniques, conclusions cannot be drawn with the same level of certainty because hypotheses are not pre-planned. Given a large data set, it is very

likely that at least one statistically significant result can be found by using exploratory analyses. Such results are 'hypothesis-generating' and often lead to new studies prospectively designed to test these new hypotheses.

Two main inferential methods are confidence interval estimation and hypothesis testing, which are discussed in detail later in this chapter.

1.2 Probability Distributions

An understanding of basic probability concepts is essential to grasp the fundamentals of statistical inference. Most introductory statistics texts discuss these basics, therefore, only some brief concepts of probability distributions are reviewed here.

Each outcome of a statistical experiment can be mapped to a numeric-valued function called a 'random variable'. Some values of the random variable might be more likely to occur than others. The probability distribution associated with the random variable X describes the likelihood of obtaining certain values or ranges of values of the random variable.

For example, consider two cancer patients, each having a 50-50 chance of surviving at least 3 months. Three months later, there are 4 possible outcomes, which are shown in Table 1.1.

TABLE 1.1. Probability Distribution of Number of Survivors (n=2)

Outcome	Patient 1	Patient 2	X	Probability
1	Died	Died	0	0.25
2	Died	Survived	1	0.25
3	Survived	Died	1	0.25
4	Survived	Survived	2	0.25

Each outcome can be mapped to the random variable X, which is defined as the number of patients surviving at least 3 months. X can take the values 0, 1, or 2 with probabilities 0.25, 0.50, and 0.25, respectively, because each outcome is equally likely.

The probability distribution for X is given by P_x as follows:

X	P_x
0	0.25
1	0.50
2	0.25

▶ *Discrete Distributions*

The preceding example is a *discrete probability* distribution because the random variable X can only take discrete values, in this case, integers from 0 to 2.

The *binomial* distribution is, perhaps, the most commonly used discrete distribution in clinical biostatistics. This distribution is used to model experiments involving n independent trials, each with 2 possible outcomes, say, '*event*' or 'non-*event*', and the probability of '*event*', p, is the same for all n trials. The preceding example, which involves two cancer patients, is an example of a binomial distribution in which n = 2 (patients), p = 0.5, and '*event*' is survival of at least 3 months.

Other common discrete distributions include the *poisson* and the *hypergeometric* distributions.

▶ *Continuous Distributions*

If a random variable can take any value within an interval or continuum, it is called a *continuous* random variable. Height, weight, blood pressure, and cholesterol level are usually considered continuous random variables because they can take any value within certain intervals, even though the observed measurement is limited by the accuracy of the measuring device.

The probability distribution for a continuous random variable cannot be specified in a simple form as it is in the discrete example above. To do that would entail an infinite list of probabilities, one for each possible value within the interval. One way to specify the distribution for continuous random variables is to list the probabilities for ranges of X-values. However, such a specification can also be very cumbersome.

Continuous distributions are most conveniently approximated by functions of the random variable X, such as P_x. Examples of such functions are

$$P_x = 2x \quad \text{for } 0 < x < 1$$

or

$$P_x = a e^{-ax} \quad \text{for } 0 < x < \infty$$

The *normal* distribution is the most commonly used continuous distribution in clinical research statistics. Many naturally occurring phenomena follow the normal distribution, which can be explained by a powerful result from probability theory known as the *Central Limit Theorem*, discussed in the next section.

The normal probability distribution is given by the function

$$P_x = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{for } -\infty < x < \infty$$

where μ and σ are called 'parameters' of the distribution. For any values of μ and σ (>0), a plot of P_x versus x has a 'bell' shape (illustrated in Appendix B).

Other common continuous distributions are the *exponential* distribution, the *chi-square* distribution, the *F*-distribution and the Student *t*-distribution. Appendix B lists some analytic properties of common continuous distributions used in statistical inference (mentioned throughout this book). The *normal*, *chi-square*, *F*- and *t*-distributions are all interrelated, and some of these relationships are shown in Appendix B.

Whether discrete or continuous, every probability distribution has the property that the sum of the probabilities over all X-values equals 1.

▶ *The Central Limit Theorem*

The *Central Limit Theorem* states that, regardless of the distribution of measurements, sums and averages of a large number of like measurements tend to follow the normal distribution. Because many measurements related to growth, healing, or disease progression might be represented by a sum or an accumulation of incremental measurements over time, the normal distribution is often applicable to clinical data for large samples.

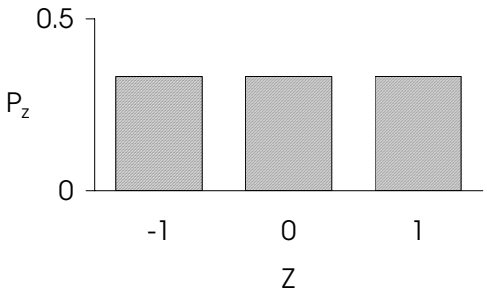
To illustrate the *Central Limit Theorem*, consider the following experiment. A placebo (inactive pill) is given to n patients, followed by an evaluation one hour later. Suppose that each patient's evaluation can result in 'improvement,' coded as +1, 'no change' (0), or 'deterioration' (–1), with each result equally probable. Let X_1, X_2, ..., X_n represent the measurements for the n patients, and define Z to be a random variable that represents the sum of these evaluation scores for all n patients,

$$Z = X_1 + X_2 + ... + X_n$$

For n = 1, the probability distribution of Z is the same as X, which is constant for all possible values of X. This is called a 'uniform' distribution. See Fig. 1.1.

FIGURE 1.1. Probability Distribution for $Z = X_1$

Z	P_z
-1	1/3
0	1/3
+1	1/3



For n = 2, there are 9 equally probable outcomes resulting in 5 possible, distinct values for Z, as shown in Table 1.2.

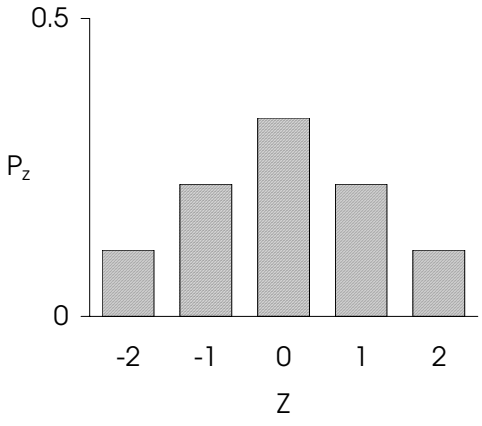
TABLE 1.2. All Possible Equally Probable Outcomes (n=2)

Patient 1	Patient 2	Z	Prob.
-1	-1	-2	1/9
-1	0	-1	1/9
0	-1	-1	1/9
-1	+1	0	1/9
0	0	0	1/9
+1	-1	0	1/9
0	+1	+1	1/9
+1	0	+1	1/9
+1	+1	+2	1/9

The resulting probability distribution for Z is shown in Figure 1.2.

FIGURE 1.2. Probability Distribution for $Z = X_1+X_2$

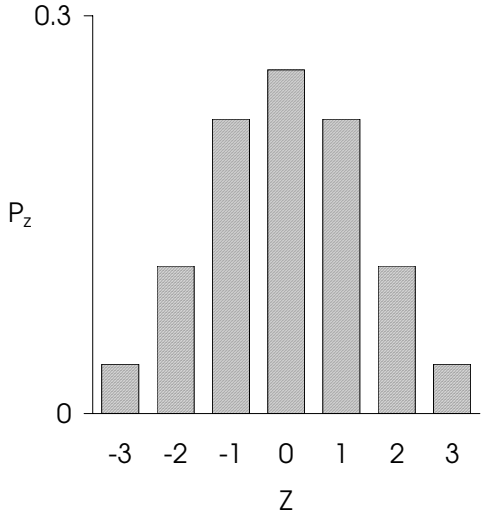
Z	P_z
-2	1/9
-1	2/9
0	3/9
+1	2/9
+2	1/9



For n = 3, Z can take values from -3 to +3. See Figure 1.3 for the distribution.

FIGURE 1.3. Probability Distribution for $Z = X_1+X_2+X_3$

Z	P_z
-3	1/27
-2	3/27
-1	6/27
0	7/27
+1	6/27
+2	3/27
+3	1/27

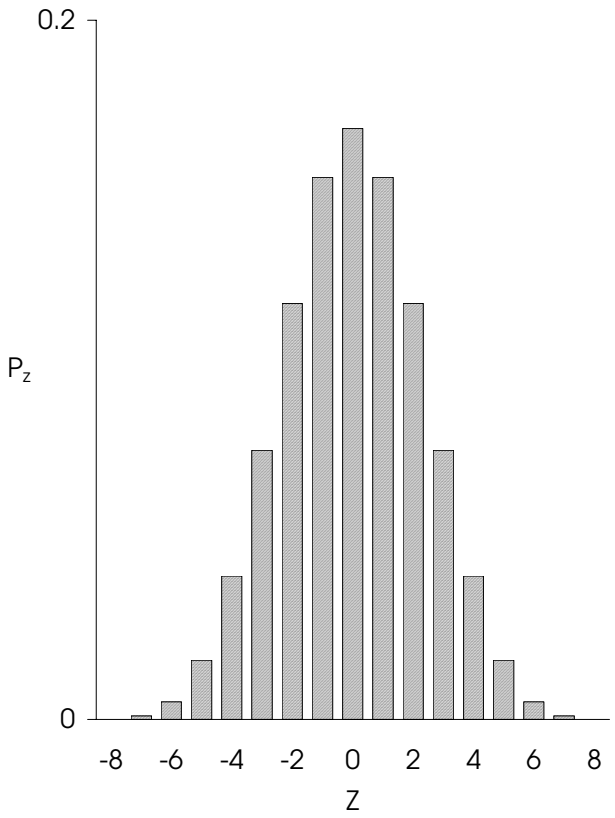


You can see from the histograms that, as n becomes larger, the distribution of Z takes on the bell-shaped characteristic of the normal distribution. The distribution of Z for 8 patients (n = 8) is shown in Figure 1.4.

While the probability distribution of the measurements (X) is 'uniform', the sum of these measurements (Z) is a random variable that tends toward a normal distribution as n increases. The *Central Limit Theorem* states that this will be the case regardless of the distribution of the X measurements. Because the sample mean, $\bar{x}$, is the sum of measurements (multiplied by a constant, 1/n), the *Central Limit Theorem* implies that $\bar{x}$ has an approximate normal distribution for large values of n regardless of the probability distribution of the measurements that comprise $\bar{x}$.

FIGURE 1.4. Probability Distribution for $Z = X_1+X_2+X_3+X_4+X_5+X_6+X_7+X_8$

Z	P_z
-8	0.000
-7	0.001
-6	0.005
-5	0.017
-4	0.041
-3	0.077
-2	0.119
-1	0.155
0	0.169
+1	0.155
+2	0.119
+3	0.077
+4	0.041
+5	0.017
+6	0.005
+7	0.001
+8	0.000



1.3 Study Design Features

Sound statistical results can be valid only if the study plan is well thought out and accompanied by appropriate data collection techniques. Even the most sophisticated statistical tests might not lead to valid inferences or appropriate characterizations of the population if the study itself is flawed. Therefore, it is imperative that statistical design considerations be addressed in clinical studies during protocol development.

There are many statistical design considerations that go into the planning stage of a new study. The probability distribution of the primary response variables will help predict how the measurements will vary. Because greater variability of the measurements requires a larger sample size, distributional assumptions enable the computation of sample-size requirements to distinguish a real trend from statistical variation. Determining the sample size is discussed in Chapter 2.

Methods to help reduce response variability can also be incorporated into the study design. Features of controlled clinical trials such as *randomization* and *blinding*, and statistical 'noise-reducing' techniques (such as the use of covariates, stratification or blocking factors, and the use of within-patient controls) are ways to help control extraneous variability and focus on the primary response measurements.

► *Controlled Studies*

A controlled study uses a known treatment, which is called a 'control', along with the test treatments. A control may be inactive, such as a placebo or sham, or it may be another active treatment, perhaps a currently marketed product.

A study that uses a separate, independent group of patients in a control group is called a *parallel-group* study. A study that gives both the test treatment and the control to the same patients is called a *within-patient control* study.

A controlled study has the advantage of being able to estimate the pure therapeutic effect of the test treatment by comparing its perceived benefit relative to the benefit of the control. Because the perceived benefit might be due to numerous study factors other than the treatment itself, a conclusion of therapeutic benefit cannot be made without first removing those other factors from consideration. Because the controls are subject to the same study factors, treatment effect *relative to* control, instead of absolute perceived benefit, is more relevant in estimating actual therapeutic effect.

► *Randomization*

Randomization is a means of objectively assigning experimental units or patients to treatment groups. In clinical trials, this is done by means of a randomization schedule generated prior to starting the enrollment of patients.

The randomization scheme should have the property that any randomly selected patient has the same chance as any other patient of being included in any treatment group. Randomization is used in controlled clinical trials to eliminate systematic treatment group assignment, which might lead to bias. In a non-randomized setting, patients with the most severe condition might be assigned to a group based on the treatment's anticipated benefit. Whether this assignment is intentional or not, this creates bias because the treatment groups would represent samples from different populations, some of whom might have more severe conditions than others. Randomization filters out such selection bias and helps establish baseline comparability among the treatment groups.

Randomization provides a basis for unbiased comparisons of the treatment groups. Omitting specific responses from the analysis is a form of tampering with this randomization and will probably bias the results if the exclusions are made in a non-randomized fashion. For this reason, the primary analysis of a clinical trial is often based on the 'intent-to-treat' principle, which includes all randomized patients in the analysis even though some might not comply with protocol requirements.

► *Blinded Randomization*

Blinded (or masked) randomization is one of the most important features of a controlled study. Single-blind, double-blind, and even triple-blind studies are common among clinical trials.

A *single-blind* study is one in which the patients are not aware of which treatment they receive. Many patients actually show a clinical response with medical care even if they are not treated. Some patients might respond when treated with a placebo but are unaware that their medication is inactive. These are examples of the well-known *placebo effect*, which might have a psychological component dependent on the patient's belief that he is receiving appropriate care. A 20% placebo response is not uncommon in many clinical indications.

Suppose that a response, Y, can be represented by a true therapeutic response component, TR, and a placebo effect, PE. Letting subscripts A and P denote 'active' and 'placebo' treatments, respectively, the estimated therapeutic benefit of the active compound might be measured by the difference

$$Y_A - Y_P = (TR_A + PE_A) - (TR_P + PE_P)$$

Because a placebo has no therapeutic benefit, $TR_P = 0$. With $PE_\Delta = PE_A - PE_P$, you obtain

$$Y_A - Y_P = TR_A + PE_\Delta$$

When patients are unaware of their treatment, the placebo effect (PE) should be the same for both groups, making $PE_\Delta = 0$. Therefore, the difference in response values estimates the true therapeutic benefit of the active compound.

However, if patients know which treatment they have been assigned, the placebo effect in the active group might differ from that of the control group, perhaps due to better compliance or expectation of benefit. In this case, the estimate of therapeutic benefit is contaminated by a non-zero PE_Δ.

In addition, bias, whether conscious or not, might arise if the investigator does not evaluate all patients uniformly. Evaluation of study measurements (such as global assessments and decisions regarding dosing changes, visit timing, use of concomitant medications, and degree of follow-up relating to adverse events or abnormal labs) might be affected by the investigator's knowledge of the patient's treatment. Such bias can be controlled by *double-blinding* the study, which means that information regarding treatment group assignment is withheld from the investigator as well as the patient.

Double-blinding is a common and important feature of a controlled clinical trial, especially when evaluations are open to some degree of subjectivity. However, double-blinding is not always possible or practical. For example, test and control treatments might not be available in the same formulation. In such cases, treatment can sometimes be administered by one investigator and the evaluations performed

by a co-investigator at the same center in an attempt to maintain some sort of masking of the investigator.

Studies can also be *triple-blind*, wherein the patient, investigator, and clinical project team (including the statistician) are unaware of the treatment administered until the statistical analysis is complete. This reduces a third level of potential bias -- that of the interpretation of the results.

Selection of appropriate statistical methods for data analysis in confirmatory studies should be done in a blinded manner whenever possible. Usually, this is accomplished through the development of a statistical analysis plan prior to completing data collection. Such a plan helps remove the potential for biases associated with data-driven methodology. It also eliminates the ability to select a method for the purpose of producing a result closest to the outcome that is being sought.

► *Selection of Statistical Methods*

Features of controlled clinical trials, such as randomization and blinding, help to limit bias when making statistical inferences. The statistical methods themselves might also introduce bias if they are 'data-driven', that is the method is selected based on the study outcomes. In most cases, the study design and objectives will point to the most appropriate statistical methods for the primary analysis. These methods are usually detailed in a formal analysis plan prepared prior to data collection and, therefore, represent the best 'theoretical' methodology not influenced by the data.

Often, sufficient knowledge of the variability and distribution of the response in Phase 3 or in pivotal trials is obtained from previous studies. If necessary, there are ways to confirm distributional assumptions based on preliminary blinded data in order to fully pre-specify the methodology. Because different statistical methods might lead to different conclusions, failure to pre-specify the methods might lead to the appearance of selecting a method that results in the most desirable conclusion.

Methodology bias is one concern addressed by an analysis plan. More importantly, pre-specifying methodology helps to ensure that the study objectives are appropriately addressed. The statistical method selected will depend very strongly on the actual objective of the study. Consider a trial that includes three doses of an active compound and an inactive placebo. Possible study objectives include determining if

- there is any difference among the four groups being studied.
- any of the active doses is better than the placebo.
- the highest dose is superior to the lower doses.
- there is a dose-response.

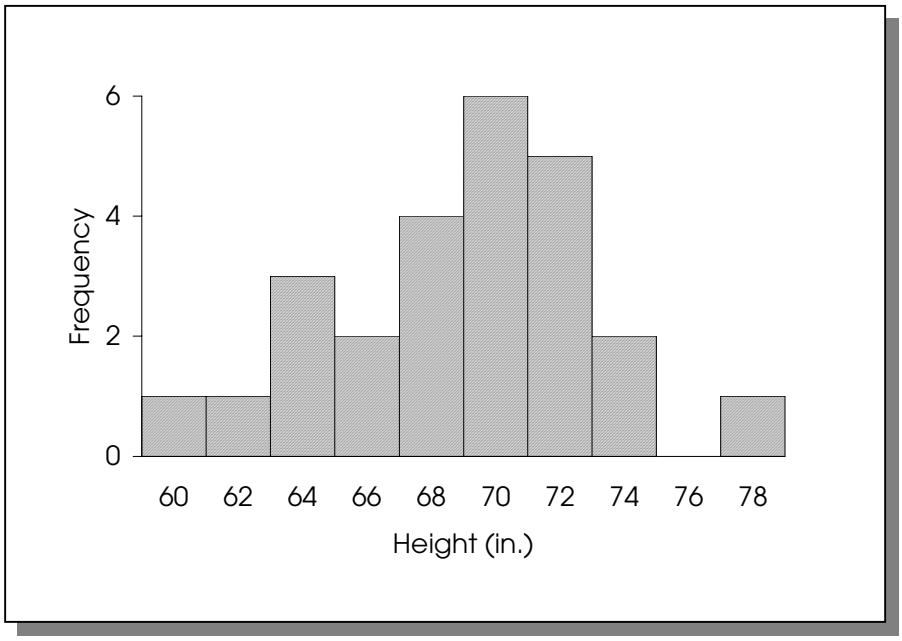
A different statistical method might be required for each of these objectives. The study objective must be clear before the statistical method can be selected.

1.4 Descriptive Statistics

Descriptive statistics describe the probability distribution of the population. This is done by using histograms to depict the shape of the distribution, by estimating distributional parameters, and by computing various measures of central tendency and dispersion.

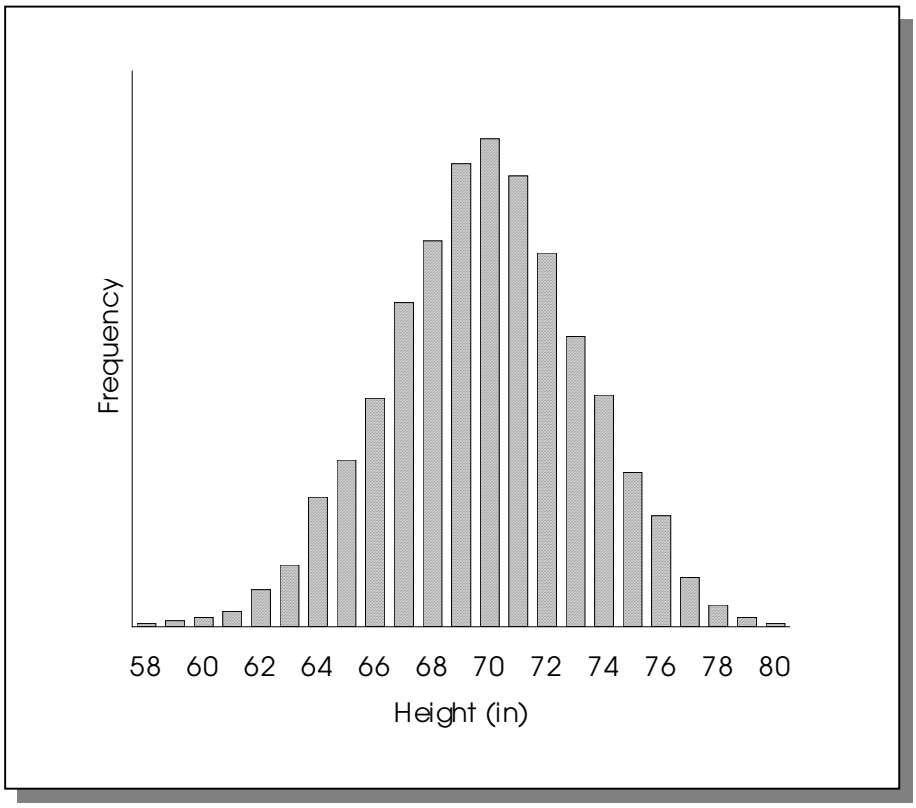
A *histogram* is a plot of the measured values of a random variable by their frequency. For example, height measurements for 16-year-old male students can be described by a sample histogram based on 25 students. See Figure 1.5.

Figure 1.5. Histogram of Height Measurements (n=25)



If more-and-more measurements are taken, the histogram might begin looking like a 'bell-shaped' curve, which is characteristic of a normal distribution. See Figure 1.6.

Figure 1.6. Histogram of Height Measurements (n=300)



If you assume the population distribution can be modeled with a known distribution (such as the normal), you need only estimate the parameters associated with that distribution in order to fully describe it. The binomial distribution has only one parameter, p, which can be directly estimated from the observed data. The normal distribution has two parameters, μ and σ^2, representing the mean and variance, respectively.

Suppose a sample of n measurements, denoted by $x_1, x_2, ..., x_n$ is obtained. Various descriptive statistics can be computed from these measurements to help describe the population. These include measures of *central tendency*, which describe the center of the distribution, and measures of *dispersion,* which describe the variation of the data. Common examples of each are shown in Table 1.3.

In addition to distributional parameters, you sometimes want to estimate parameters associated with a statistical model. If an unknown response can be modeled as a function of known or controlled variables, you can often obtain valuable information regarding the response by estimating the weights or coefficients of each of these known variables. These coefficients are called *model parameters*. They are estimated in a way that results in the greatest consistency between the model and the observed data.

TABLE 1.3. Common Descriptive Statistics

Measures of 'Central Tendency'	
Arithmetic Mean	$\bar{x} = (\sum x_i) / n = (x_1 + x_2 + ... + x_n) / n$
Median	the middle value, if n is odd; the average of the two middle values if n is even (50^{th} percentile)
Mode	the most frequently occurring value
Geometric Mean	$(\prod x_i)^{1/n} = (x_1 \cdot x_2 \cdot ... \cdot x_n)^{1/n}$
Harmonic Mean	$n / \sum(x_i)^{-1} = n\{(1/x_1) + (1/x_2) + ... + (1/x_n)\}^{-1}$
Weighted Mean	$\bar{x}_w = (\sum w_i x_i) / W$, where $W = \sum w_i$
Trimmed Mean	Arithmetic mean omitting the largest and smallest observations
Winsorized Mean	Arithmetic mean after replacing outliers with the closest non-outlier values

Measures of 'Dispersion'	
Variance	$s^2 = \sum(x_i - \bar{x})^2 / (n - 1)$
Standard Deviation	s = square root of the variance
Standard Error (of the mean)	$(s^2 / n)^{1/2}$ = Standard deviation of $\bar{x}$
Range	Largest value - Smallest value
Mean Absolute Deviation	$(\sum \lvert x_i - \bar{x} \rvert) / n$
Inter-Quartile Range	75^{th} percentile – 25^{th} percentile
Coefficient of Variation	$s / \bar{x}$

Descriptive statistical methods are often the only approach that can be used for analyzing the results of pilot studies or Phase I clinical trials. Due to small sample sizes, the lack of blinding, or the omission of other features of a controlled trial, statistical inference might not be possible. However, trends or patterns observed in the data by using descriptive or exploratory methods will often help in building hypotheses and identifying important cofactors. These new hypotheses can then be tested in a more controlled manner in subsequent studies, wherein inferential statistical methods would be more appropriate.

1.5 Inferential Statistics

The two primary statistical methods for making inferences are confidence interval estimation and hypothesis testing.

► *Confidence Intervals*

Population parameters, such as the mean (μ) or the standard deviation (σ), can be estimated by using a point estimate, such as the sample mean ($\bar{x}$) or the sample standard deviation (s). A *confidence interval* is an interval around the point estimate that contains the parameter with a specific high probability or confidence level. A 95% confidence interval for the mean (μ) can be constructed from the sample data with the following interpretation: If the same experiment were conducted a large number of times and confidence intervals were constructed for each, approximately 95% of those intervals would contain the population mean (μ).

The general form of a confidence interval is $[\theta_L - \theta_U]$, where θ_L represents the lower limit and θ_U is the upper limit of the interval. If the probability distribution of the point estimate is symmetric (such as the normal distribution), the interval can be found by

$$\hat{\theta} \pm C \cdot \sigma_{\hat{\theta}}$$

where $\hat{\theta}$ is the point estimate of the population parameter θ, $\sigma_{\hat{\theta}}$ is the standard error of the estimate, and C represents a value determined by the probability distribution of the estimate and the significance level that you want. When $\sigma_{\hat{\theta}}$ is unknown, the estimate $\hat{\sigma}_{\hat{\theta}}$ may be used.

For example, for α between 0 and 1, a $100(1-\alpha)\%$ confidence interval for a normal population mean (μ) is

$$\bar{x} \pm Z_{\alpha/2} \cdot \sigma / \sqrt{n}$$

where the point estimate of μ is $\bar{x}$, the standard error of $\bar{x}$ is $\sigma/\sqrt{n}$, and the value of $Z_{\alpha/2}$ is found in the normal probability tables (See Appendix A.1). Some commonly used values of α and the corresponding critical Z-values are

α	$Z_{\alpha/2}$
0.10	1.645
0.05	1.96
0.02	2.33
0.01	2.575

In most cases, the standard deviation (σ) will not be known. If it can be estimated using the sample standard deviation (s), a $100(1-\alpha)\%$ confidence interval for the mean (μ) can be formed as

$$\bar{x} \pm t_{\alpha/2} \cdot s / \sqrt{n}$$

where $t_{\alpha/2}$ is found from the Student-t probability tables (see Appendix A.2) based on the number of degrees of freedom, in this case, n–1. For example, a value of $t_{\alpha/2} = 2.093$ would be used for a 95% confidence interval when n = 20.

Many SAS procedures will print point estimates of parameters with their standard errors. These point estimates can be used to form confidence intervals using the general form for $\hat{\theta}$ that is given above. Some of the most commonly used confidence intervals are for population means (μ), differences in means between two populations ($\mu_1-\mu_2$), population proportions (p), and differences in proportions between two populations ($p_1 - p_2$). For each of these, the form for $\hat{\theta}$ and its standard error are shown in Table 1.4.

TABLE 1.4. Confidence Interval Components Associated with Means and Proportions

θ	$\hat{\theta}$	$\sigma^2_{\hat{\theta}}$	$\hat{\sigma}^2_{\hat{\theta}}$	C
μ	$\bar{x}$	σ^2 / n	s^2 / n	$Z_{\alpha/2}$ if σ is known; $t_{\alpha/2}$ if σ is unknown
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\sigma_1^2/n_1 + \sigma_2^2/n_2$	$s^2 (1/n_1 + 1/n_2)$	$Z_{\alpha/2}$ if σ_1 and σ_2 are known; $t_{\alpha/2}$ if σ_1 or σ_2 is unknown. If unknown, assume equal variances and use $s^2 = (n_1-1)s_1^2 + (n_2-1)s_1^2/ (n_1 + n_2 - 2)$
p	$\hat{p} = x/n$	$p (1- p) / n$	$\hat{p} (1- \hat{p}) / n$	$Z_{\alpha/2}$ (x 'events' in n binomial trials)*
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$p_1(1- p_1) / n_1 + p_2(1- p_2) / n_2$	$\hat{p}_1 (1- \hat{p}_1) / n_1 + \hat{p}_2 (1- \hat{p}_2) / n_2$	$Z_{\alpha/2}$ ($\hat{p}_i = x_i/n_i$ for i = 1,2)*

* applies to large samples

► *Hypothesis Testing*

Hypothesis testing is a means of formalizing the inferential process for decision-making purposes. It is a statistical approach for testing hypothesized statements about population parameters based on logical argument.

To understand the concept behind the hypothesis test, let's examine a form of deductive argument from logic, using the following example:

If you have an apple, you do not have an orange. You have an orange. Therefore, you do not have an apple.

The first two statements of the argument are premises and the third is the conclusion. The conclusion is logically deduced from the two premises, and its truth depends on the truth of the premises.

If **P** represents the first premise and **Q** represents the second premise, the argument may be formulated as

if **P** then not **Q**	(conditional premise)
Q	(premise)
_____	
therefore, not **P**	(conclusion)

This is a deductively valid argument of logic that applies to any two statements, **P** and **Q**, whether true or false. Note that if you have both an apple and an orange, the conditional premise would be false, which makes the conclusion false because the argument is still valid.

Statistical arguments take the same form as this logical argument, but statistical arguments must account for random variations in statements that might not be known to be completely true. A statistical argument might be paraphrased from the logical argument above as

if **P** then *probably* not **Q**	(conditional premise)
Q	(premise)
_____	
therefore, *probably* not **P**	(conclusion)

The following examples illustrate such 'statistical arguments'.

Example 1

Statements:
P = the coin is fair
Q = you observe 10 tails in a row

Argument:
If the coin is fair, you would probably not observe 10 tails in a row. You observe 10 tails in a row. Therefore, the coin is probably not fair.

Example 2

Statements:
P = Drug A has no effect on arthritis
Q = from a sample of 25 patients, 23 showed improvement in their arthritis after taking Drug A

Argument:
If Drug A has no effect on arthritis, you would probably not see improvement in 23 or more of the sample of 25 arthritic patients treated with Drug A. You observe improvement in 23 of the sample of 25 arthritic patients treated with Drug A. Therefore, Drug A is probably effective for arthritis.

In the first example, you might initially suspect the coin of being biased in favor of tails. To test this hypothesis, assume the null case, which is that the coin is fair. Then, design an experiment that consists of tossing the coin 10 times and recording the outcome of each toss. You decide to reject the hypothesis concluding that the coin is biased in favor of tails if the experiment results in 10 consecutive tails.

Formally, the study is set out by identifying the hypothesis, developing a test criterion, and formulating a decision rule. For Example 1,

- **Null hypothesis**: the coin is fair
- **Alternative**: the coin is biased in favor of tails
- **Test criterion**: the number of tails in 10 consecutive tosses of the coin
- **Decision rule**: reject the null hypothesis if all 10 tosses result in 'tails'

First, establish the hypothesis **P**. The hypothesis is tested by observing the results of the study outcome **Q**. If you can determine that the probability of observing **Q** is very small when **P** is true and you do observe **Q**, you can conclude that **P** is probably not true. The degree of certainty of the conclusion is related to the probability associated with **Q**, assuming **P** is true.

Hypothesis testing can be set forth in an algorithm with 5 parts:

- the null hypothesis (abbreviated H_0)
- the alternative hypothesis (abbreviated H_A)
- the test criterion
- the decision rule
- the conclusion.

The null hypothesis is the statement **P** translated into terms involving the population parameters. In Example 1, 'the coin is fair' is equivalent to 'the probability of tails on any toss is ½'. Parametrically, this is stated in terms of the binomial parameter p, which represents the probability of tails.

$$H_0: \ p \leq 0.5$$

The alternative hypothesis is 'not **P**', or

$$H_A: \ p > 0.5$$

Usually, you take 'not **P**' as the hypothesis to be demonstrated based on an acceptable risk for defining '*probably*' as used in Examples 1 and 2.

The test criterion or 'test statistic' is some function of the observed data. This is statement **Q** of the statistical argument. Statement **Q** might be the number of tails in 10 tosses of a coin or the number of improved arthritic patients, as used in Examples 1 and 2, or you might use a more complex function of the data. Often the test statistic is a function of the sample mean and variance or some other summary statistics.

The decision rule results in the rejection of the null hypothesis if unlikely values of the test statistic are observed when assuming the test statistic is true. To determine a decision rule, the degree of such 'unlikeliness' needs to be specified. This is referred to as the *significance level* of the test (denoted α) and, in clinical trials, is often (but not always) set to 0.05. By knowing the probability distribution of the test statistic when the null hypothesis is true, you can identify the most extreme $100\alpha\%$ of the values as a rejection region. The decision rule is simply, reject H_0 when the test statistic falls in the rejection region.

See Chapter 2 for more information about significance levels.

1.6 Summary

This introductory chapter provides some of the basic concepts of statistics, gives an overview of statistics as a scientific discipline, and shows that the results of a statistical analysis can be no better than the data collected. You've seen that the researcher must be vigilant about biases that can enter into a data set from a multitude of sources. With this in mind, it is important to emphasize the correct application of statistical techniques in study design and data collection as well as at the analysis stage.

Statistical methods used to characterize populations from sample data can be classified as descriptive or inferential, most notably, parameter estimates by confidence intervals and hypothesis testing. These techniques are the focus of the methods presented in this book, Chapters 4 through 22.