

CHAPTER 1

INTRODUCTION & BASICS

WHAT IS STATISTICS?

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 the process of inference is formalized through mathematics to a field called **Statistics**, it often becomes clouded by preconceptions of abstruse theory. Let's see if we can provide some formalization by appealing to the natural process of rational inference without getting caught up in theoretical developments.

The purpose of **Statistics** is to characterize a *population* based on the information contained in a *sample* taken from the population. The concepts of 'populations', 'samples' and what we mean by 'characterizing' are discussed through this chapter.

The sample information is conveyed by functions of the observed data called *statistics*. The field of **Statistics** is a discipline which endeavors to determine which functions are the most relevant in the characterization of various populations.

The arithmetic mean, for example, may be the most appropriate statistic to help characterize certain populations, while the median may 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. The scientific method consists of developing an 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 drawing appropriate inferences from the data. In doing so, the applied statistician carries out the scientific method, attempting to limit bias, maximize objectivity and obtain results which are scientifically valid.

► Populations

By *population*, we mean a universe of entities which we would like to characterize but is too vast to study in its entirety. The population to be studied in a clinical trial is defined by its limiting conditions, usually specified by way of study inclusion and exclusion requirements.

Examples of populations include:

- 1) patients with mild to moderate hypertension
- 2) obese teenagers
- 3) adult insulin-dependent diabetic patients.

In the first example, there is only one limiting factor defining the population, that being mild to moderate hypertension. This might be defined more precisely as patients with diastolic blood pressure within a specific range as an inclusion criterion for the clinical protocol. Additional criteria would further limit the population.

Example 2 uses both age and weight as limiting conditions, while Example 3 uses age, diagnosis and treatment as criteria in defining the population.

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

► Samples

Intuitively, we can describe a population by describing some representative entities from it. Measurements obtained from the sample entities would tend to characterize the entire population through inference. **Statistics** provides a method of formalizing such intuition, as discussed later in this chapter.

The degree of representation of the entities in a sample taken from the population of interest depends on the sampling plan used. Conceptually, the simplest type of sampling plan is called a 'simple random sample'. This type of plan 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 is intuitively 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 clearcut. Simple random samples are rarely, if ever, used in clinical trials. One can envision the patients comprising the populations in the three examples cited being scattered all over the world, making 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 '*Sampling Theory*' has been developed to handle alternative approaches to simple random sampling which minimize 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 would be typical of those that might be enrolled elsewhere. Multi-center studies are often used to blunt the effect of patient characteristics or procedural anomalies which might be unique to any single center.

Stratified sampling is another technique which is often used to better target representative patients. This method uses

random samples from each of several subgroups of the population, called 'strata'. Study enrollment is sometimes stratified by disease severity, age group or some other patient characteristic.

Because inferences from non-random samples may 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 is found to show a larger response, part of that response may 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? Two types of statistical procedures used to characterize populations include *descriptive* and *inferential* procedures.

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 since hypotheses are not pre-planned. Given a large data set, it is very likely that at least one statistically significant result can be found using exploratory analyses. Such results are 'hypothesis-generating' and often lead to new studies prospectively designed to test these new hypotheses.

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

PROBABILITY DISTRIBUTIONS

An understanding of basic probability concepts is essential to grasp the fundamentals of statistical inference. Most introductory statistics texts discuss these basics, which are not repeated here. We do, however, review some elements of probability distributions.

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

As an example, consider 2 cancer patients, each with a 50-50 chance of surviving at least 3 months. Three months later, there are 4 possible outcomes, shown in the table below.

Outcome	Patient 1	Patient 2	X	$\Pr\{X\}$
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 a random variable, X , defined as the number of patients surviving at least 3 months. X can take values 0, 1 or 2 with probabilities 0.25, 0.50 and 0.25, respectively, since 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 above example is a *discrete* probability distribution, since 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 example just discussed involving two cancer patients is an example of a binomial distribution with 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 serum bilirubin are usually considered continuous random variables since 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 in the discrete example above. To do so 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 , say P_x . Such functions may have a form such as

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

or,

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

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 below.

The normal probability distribution is given by the function:

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

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 mound or 'bell' shape (illustrated in Appendix B).

Other common continuous distributions include 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 the *t*-distributions are all inter-related, 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.

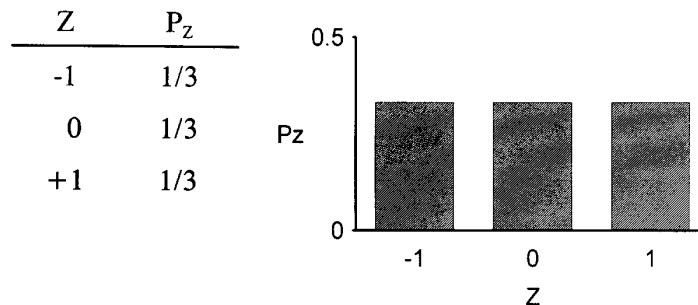
► *The Central Limit Theorem*

Briefly, the *Central Limit Theorem* states that, regardless of the distribution of measurements, sums and averages of a large number of measurements tend to follow the normal distribution. Since many measurements related to growth, healing, or disease progression might be represented by a

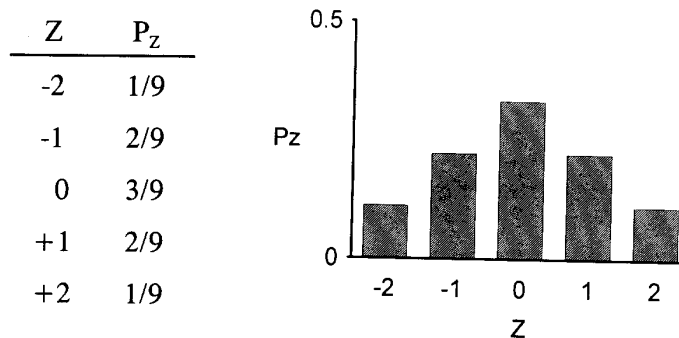
sum or accumulation of incremental measurements over time, the normal distribution is often applicable to clinical data for large samples.

To illustrate the *Central Limit Theorem*, we 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), each equally likely. Let X_1, X_2, \dots, X_n represent the measurements for the n patients, and define Z to be a random variable representing the sum of these evaluation scores for all n patients, $Z = X_1 + X_2 + \dots + X_n$.

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, shown as follows:



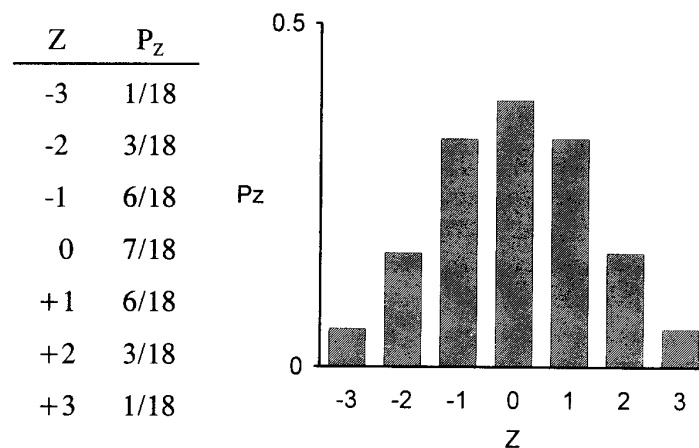
For $n = 2$, Z has the probability distribution shown as follows:



since, there are 9 equally likely outcomes resulting in 5 possible values for Z :

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

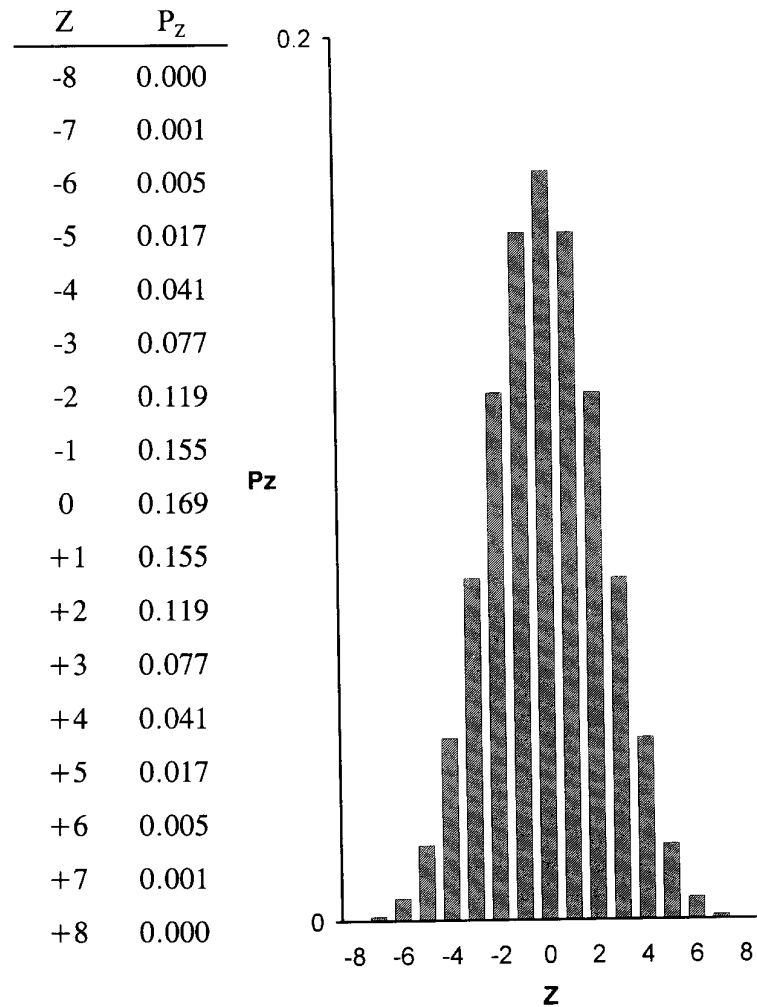
For $n = 3$, Z can take values from -3 to +3 with the distribution:



We 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 on the next page.

While the probability distribution of the measurements (X) is 'uniform', the sum of these measurements (Z) is a random variable which 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. Since the sample mean, \bar{x} , is the sum of measurements (times a constant, $1/n$), the *Central Limit Theorem* implies that \bar{x} has an approximate normal

distribution for large n regardless of the probability distribution of the measurements comprising \bar{x} .



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 may not lead to valid inferences or appropriate characterizations of the population if the study itself is flawed. It is imperative, therefore, that statistical design considerations be addressed in clinical studies during protocol development.

There are a number of 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. Since 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. Sample size determination is discussed further in Chapter 2.

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

► *Controlled Studies*

A controlled trial is one in which a known treatment, called a 'control', is used in the same study as the test treatments. Controls may be inactive, such as a placebo or sham, or another active treatment, perhaps a currently marketed product. A study which uses a separate, independent group of patients in a control group is called a *parallel-group* study. Studies which give both the test treatments and control to the same patients are called *within-patient control* studies.

► *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 commencement of patient enrollment. 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 may lead to bias. In a non-randomized setting, patients with the

most severe condition may be assigned to a group based on the treatment's perceived benefit, whether intentional or not. This creates bias since the treatment groups would represent samples from different populations, some more severe than others. Randomization filters out such selection bias and helps establish baseline comparability among the treatment groups.

► *Blinding*

Blinded 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. Others may also respond when treated with a placebo, but are unaware that their medication is inactive. These are examples of the well-known 'placebo effect', which may 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:

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

Since placebo has no therapeutic benefit, $TR_P = 0$, and with $PE_\Delta = PE_A - PE_P$, we 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 $PE_\Delta = 0$. Thus, the difference in response values estimates the true therapeutic benefit of the active compound.

If, however, patients know which treatment they have been assigned, the 'placebo effect' of the active group may 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} .

Bias may affect the investigator's evaluations as well. Evaluation of study measurements such as global assessments and decisions regarding dosing changes, visit timing, use of concomitant medications and degree of followup on adverse events or abnormal labs, may be affected by the investigator's knowledge of the patient's treatment, whether conscious or not. Blinding the investigator and the patient will help eliminate these biases. Such studies are known as *double-blind* studies.

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 may 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 investigator blind.

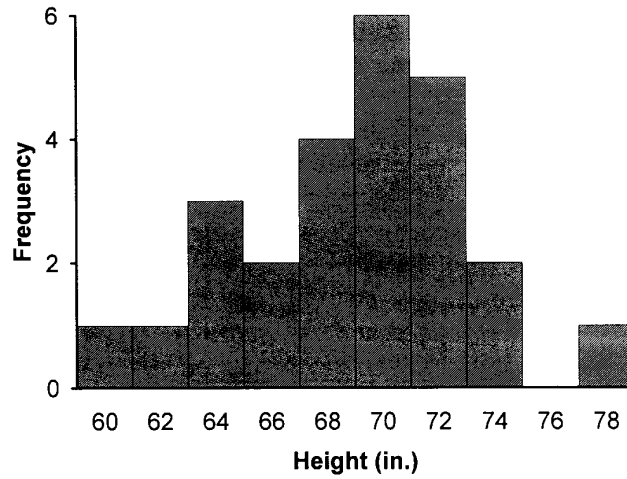
Studies can also be *triple-blind*, wherein the patient, investigator and clinical project team (including the statistician), are all masked as to the treatment administered until the statistical analysis is complete. This reduces a third level of potential bias -- that of the interpretation of the results.

DESCRIPTIVE STATISTICS

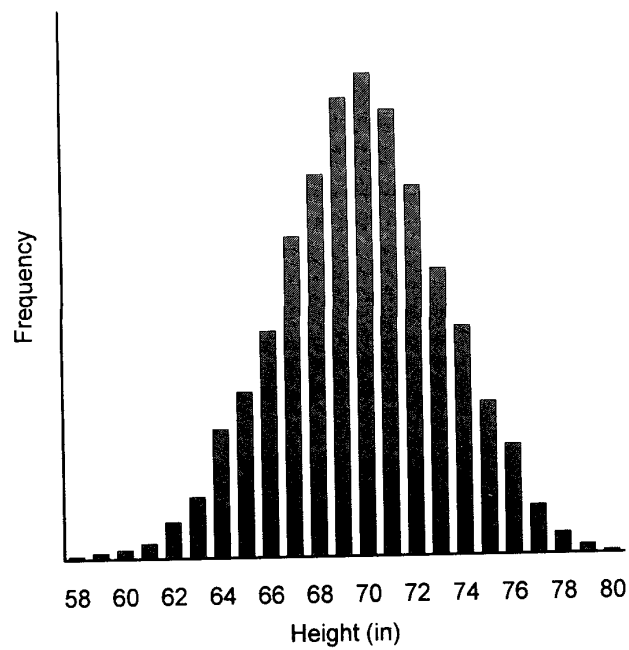
Descriptive statistics are used to describe the probability distribution of the population. This is done by using histograms to depict the shape of the distribution, by estimating distributional parameters and by computing various measures of central tendency and dispersion.

A histogram is a plot of the measured values of a random variable by their frequency. Height measurements for 16-

year olds, for example, can be described by a sample histogram as follows, based on 25 students:



If more and more measurements are taken, the histogram may begin looking like a 'bell-shaped' curve, which is characteristic of the normal distribution, as follows.



If we assume the population distribution can be modelled with a known distribution, such as the normal, we 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, \dots, x_n is obtained. A number of 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 2-1 on the following page.

In addition to distributional parameters, we sometimes want to estimate parameters associated with a statistical 'model'. If an unknown response can be modelled as a function of known or controlled variables, we 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, and are estimated in a manner which results in the greatest consistency between the model and the observed data.

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 omission of other features of a controlled trial, statistical inference may not be possible in such studies. However, trends or patterns observed in the data using descriptive or exploratory methods will often help in building hypotheses to be tested in a more controlled manner in subsequent studies. Inferential statistical methods are used in such situations.

TABLE 2-1: Common Descriptive Statistics

Measures of 'Central Tendency'	
Arithmetic Mean	$\bar{x} = (\sum x_i)/n$ $= (x_1 + x_2 + \dots + x_n) / n$
Median	the middle value, if n is odd; the average of the 2 middle values if n is even
Mode	the most frequently occurring value
Geometric Mean	$(\prod x_i)^{1/n} = (x_1 \cdot x_2 \cdot \dots \cdot x_n)^{1/n}$
Harmonic Mean	$n/\sum(x_i)^{-1} =$ $n\{(1/x_1) + (1/x_2) + \dots + (1/x_n)\}^{-1}$
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	$(\sum x_i - \bar{x}) / n$
Inter-Quartile Range	75th percentile - 25th percentile
Coefficient of Variation	s/\bar{x}

INFERENCE 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 standard deviation, σ , can be estimated using a point estimate, such as the sample mean, \bar{x} , or sample standard deviation, s . A confidence interval is an interval around the point estimate which contains the parameter with a certain 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, 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 desired significance level.

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

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

\bar{x} is the point estimate of μ ,

$\frac{\sigma}{\sqrt{n}}$ is the standard error of \bar{x} , and

$Z_{\alpha/2}$ is found from the normal probability tables (e.g., Appendix A.1). Some commonly used values of α and the corresponding critical Z-values are shown below:

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

► 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, we first review the form of a certain deductive argument from logic.

Consider the following argument: "If you have an apple, then 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 only on the truth of the premises.

If we let **P** represent the statement 'you have an apple', and **Q** represent the statement 'you have an orange', 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 which 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 making the conclusion false since the argument is still valid.

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

if P then <i>probably</i> not Q	(conditional premise)
Q	(premise)

therefore, <i>probably</i> not P	(conclusion)

The following examples illustrate such 'statistical arguments':

Example 1

Statements:

P = 'the coin is fair'

Q = 'we observe 10 tails in a row'

Argument:

If the coin is fair, we would probably not observe 10 tails in a row. We 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** = '23 of a sample of 25 patients showed improvement in their arthritis after taking Drug A'

Argument:

If Drug A has no effect on arthritis, we would probably not see improvement in 23 or more of our sample of 25 arthritic patients with Drug A. We observe improvement in 23 of our sample of 25 arthritic patients on Drug A. Therefore, Drug A is probably effective for arthritis.

In the first example, we might initially suspect the coin of being biased in favor of tails. To test this hypothesis, we assume the null case, that is, that the coin is fair. We then design an experiment consisting of tossing the coin ten times and recording the outcomes of each toss. We 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 this example, we have:

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'

We establish the hypothesis, **P**. The hypothesis is tested by observing the results of a study whose outcome is **Q**. If we can determine that the probability of observing **Q** is very small when **P** is true, and we do observe **Q**, we 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:

- a. the null hypothesis, abbreviated H_0
 - b. the alternative hypothesis, abbreviated H_A
 - c. the test criterion
 - d. the decision rule
- and, e. the conclusion.

The null hypothesis is the statement **P** translated into terms involving the population parameters. In the first example, 'the coin is fair' is equivalent to 'the probability of tails on any toss is 1/2'. Parametrically, this is stated in terms of the binomial parameter, p , representing the probability of tails:

$$H_0: p \leq 0.5.$$

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

$$H_A: p > 0.5.$$

We generally take 'not **P**' as the hypothesis to be demonstrated based on an acceptable risk for defining 'probably' as used in the examples.

The test criterion or 'test statistic' is some function of the observed data. This is statement **Q** of our statistical argument, and may be the number of tails in 10 tosses of a coin or the number of improved arthritic patients, as in the examples, or we may 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 it 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, we 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".

Significance levels are discussed further in Chapter 2.

SUMMARY

This introduction discusses some of the basic concepts of statistics, provides 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. We have seen that the researcher must be vigilante of biases which can enter into a data set from a multitude of sources. With this in mind, it is important to emphasize the proper application of statistical techniques in study design and data collection as well as at the analysis stage.

Methods of characterizing populations from sample data include descriptive and inferential procedures, most notably parameter estimates by confidence intervals and hypothesis testing. These techniques are the focus of the methods presented in this book, Chapters 3-20.