Chapter 2
Statistical Background

The goal of this chapter is to lay the foundation for the rest of this book by presenting the statistical theory and methods that are utilized in the rest of this book. It is necessarily a mathematical chapter. The subsequent chapters are based largely on the general formulations and theory found here. It is possible to skip this chapter and simply apply the methodology found in the chapters that follow; however, the fundamentals of the underlying approaches are found here.

We begin this chapter with definitions of some statistical terms, primarily involving random variables. That is followed by introductory probability theory. We strive here for a level of comprehension that is at about the level of a senior undergraduate or master's student in a quantitative discipline. The assumption is that readers of this chapter are familiar with integration, basic notions of probability and limits, for example. Citations are given to more advanced works that underpin the mathematical and statistical structure here. In this chapter we present no examples, since our aim is not to explicate but rather to remind the reader of relevant results. The organization of this chapter begins with some background on probability and probability models. Having established these, we turn to statistical methods. The methods that we focus on here and throughout this text are hypothesis testing and confidence intervals. To that end, we next discuss resampling methods including bootstrap methods that will play a prominent role in later chapters. Finally, we discuss sample size and power calculations which are methods for determining the number of observations that need to be taken to achieve a particular set of statistical aims.

2.1 Some Preliminaries

Before moving to the mathematical materials, it is important to keep in mind the big picture of our goals. We would like to use statistical inference to study how well a given biometric authentication system performs. To that end, data is sampled from an ongoing process and we analyze that data in order to assess the system performance. As with any analysis, the quality of the resulting output is dependent upon the quality of the input. Statistical inference is not an exception to this dependence.

M.E. Schuckers, *Computational Methods in Biometric Authentication,* Information Science and Statistics,
DOI 10.1007/978-1-84996-202-5_2, © Springer-Verlag London Limited 2010

In particular, many statistical methods *assume* a random sample from a process to permit unbiased inference. The following quote is one that best captures the relevance of such an assumption. In a section entitled "The Central Role of Practical Assumptions Concerning 'Representative Data' " Hahn and Meeker [43] on pp. 5–6 of their book *Statistical Intervals: A Guide for Practitioners* include the following discussion:

> Departures from these implicit assumptions are common in practice and can invalidate the entire analyses. Ignoring such assumptions can lead to a false sense of security, which, in may applications, is the weakest link in the inference chain. . . .
> In the best of situations, one can rely on physical understanding, or information from outside the study, to justify the practical assumptions. Such assessment is, however, principally the responsibility of the engineer—or "subject-matter expert." Often, the assessment is far from clear-cut and an empirical evaluation may be impossible. In any case, one should keep in mind that the intervals described in this book reflect on the statistical uncertainty, and, thus, provide a *lower bound* on the true uncertainty. The generally nonquantifiable deviations of the practical assumptions from reality provide an added *unknown* element of uncertainty. If there were formal methods to reflect this further uncertainty (occasionally there are, but often there are not), the resulting interval, expressing the *total* uncertainty, would clearly be longer than the statistical interval alone. This observation does, however, lead to a rationale for calculating a statistical interval for situations where the basic assumptions are questionable. In such case, if it turns out that the calculated statistical interval is long, we then know that our estimates have much uncertainty—even *if* the assumptions were all correct. On the other hand, a narrow statistical interval would imply a small degree of uncertainty *only if* the required assumptions hold.

The emphasis here is from the original. Thus, we recommend the methods in this book based upon the utility of the statistical method, but we caution users that the assumptions of a particular approach should not be overlooked or ignored.

Another notion that is less common in statistical discussions needs to be addressed at this juncture. Throughout the rest of this text we will discuss inference in terms of a process rather than in terms of a population. The relevant difference was first articulated by W. Edwards Deming and is an important one for biometric systems, Deming [22]. Deming, in particular, differentiated between an enumerative study and an analytic study. An enumerative study is one that is concerned with understanding a fixed population while an analytic one is concerned with understanding a process and, hence, prediction. As Deming wrote on p. 249 of [22] "[i]n analytic problems the concern is not this one bowl alone but the sequence of bowls that will be produced in the future. . . " where samples have been taken from a single bowl. Since biometric systems are developed and assessed with an aim toward implementation and with interest in future outcomes of these systems, we view these studies as analytic rather than enumerative. As a consequence, the language that we use will be of an ongoing *process* rather than a fixed population.

2.2 Random Variables and Probability Theory

In this section, we define random variables and some of the properties of these variables. Because we aim to keep this discussion at a relatively low mathematical

level, we take a less formal approach to some of the definitions that follow. More formal (and complete) structure for these constructs can be found in, for example, Casella and Berger [14]. For those who are interested, a mathematically higher level introduction to these topics can be found in Billingsley [7] or Schervish [82]. More thorough background at approximately the same level as this book can be found in Wackerly et al. [99] or in DeGroot and Schervish [21].

Definition 2.1 An *experiment* is an event whose outcome is not completely known with certainty in advance.

Definition 2.2 A *sample space* is the collection of all possible outcomes for an experiment.

A sample space can be thought of as all of the possible results from an event whose outcome is uncertain. A classic example is the toss of a coin with the sample space being the set {*Heads*, *Tails*}. A more interesting example might be the high temperature in London, UK tomorrow. If we restrict ourselves to whole Celsius units, then the sample space will be some subset of the set of all integers, $\mathbb{Z}$. If, instead, we consider fractional Celsius values, e.g. 18.5 or 13.265 or 29.48733..., then the sample space becomes some subset of the real number line, $\mathbb{R}$.

Definition 2.3 A *parameter* is a real-valued numerical summary of a process or population.

Definition 2.4 A *random variable* (RV) is a real-valued function defined on a sample space.

We can think of a random variable as the outcome of some future experiment or of some as yet unobserved event. For mathematical integrity we say that the values that the random variable can take must be numerical.

Definition 2.5 A *discrete random variable* is a random variable that can only take a countable number of values.

Definition 2.6 A *continuous random variable* is a random variable that can take values on a subset of the real line $\mathbb{R} = (-\infty, \infty)$.

For the examples given above, the tossing of a coin is a discrete random variable—we can map heads and tails to the values 0 and 1—as is the measurement of the high temperature in London using whole Celsius units. If we use decimal units to make our measurements, then the random variable is (usually) treated as a continuous one. Most of the work in this book will focus on discrete random variables but some sections involve continuous random variables.

Definition 2.7 A *probability mass distribution* for a discrete random variable is a table, graph or formula that gives the probability, $P(v)$, of a particular outcome.

We have the following rules for probability mass distributions:

1. $0 \leq P(v) \leq 1$ for all values v, and
2. $\sum_v P(v) = 1$.

Definition 2.8 A *cumulative density function*, $F_X(x)$, for a random variable X is equal to the probability that the value of the random variable, X, is less than or equal to the specific value x.

$$F_X(x) = P(X \leq x). \tag{2.1}$$

Definition 2.9 A *probability density function*,

$$f_X(x) = \frac{dF_X(x)}{dx}, \tag{2.2}$$

is the derivative of the cumulative density function if the cumulative density function is differentiable.

Result 2.1 The probability that a random variable X takes values at or below x is written as

$$P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt. \tag{2.3}$$

It is important to remember that a random variable (or its cumulative distribution function (cdf) or probability density function (pdf)) or its probability mass function is a mathematical/ statistical model of reality. The choice of which random variable to use is one that highly depends on the system or the process we are trying to model.

Definition 2.10 We will denote an *indicator function* of a statement A as the function that is defined to be one if the statement A is true and zero otherwise. We will denote this function by I_A and formally define it as

$$I_A = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

Definition 2.11 The *expected value* for a random variable, V, written as $E[V]$ is the mean of that random variable. For a discrete random variable, V_d, the expected value is calculated as

$$E[V_d] = \sum_v vP(V = v). \tag{2.5}$$

For a continuous random variable, V_c, the expected value is calculated as

$$E[V_c] = \int_{-\infty}^{\infty} vf(v)dv. \tag{2.6}$$

For a given random variable, discrete or continuous, the expected value of that random variable is often known as the expectation of that random variable.

The expected value is a measure of central tendency meaning that is meant to be a measure of the 'center' or average of a distribution. The expected value is known as the first moment. We also note that for some collection of outcomes A, $E(I_A) = P(A)$.

Result 2.2 The expected value of a constant, a, is just the constant, i.e. $E[a] = a$.

Definition 2.12 The *expected value for a function* $g()$ *of a random variable*, $E[g(V)]$, is defined as

$$E[g(V)] = \sum_v g(v)P(V = v) \tag{2.7}$$

for a discrete random variable and as

$$E[g(V)] = \int_{-\infty}^{\infty} g(v)f(v)dv \tag{2.8}$$

for a continuous random variable.

The expected value is also known as the mean.

Definition 2.13 The *variance* of a random variable, V, is the expected squared distance from the mean, $E[(V - E(V))^2]$, and is written as $Var[V] = E[(V - E(V))^2]$. For a discrete random variable, V_d, the variance is calculated as

$$Var[V_d] = \sum_v (v - E[V_d])^2 P(V = v). \tag{2.9}$$

For a continuous random variable, V_c, the variance is calculated as

$$Var[V_c] = \int_{-\infty}^{\infty} (v - E[V_c])^2 f(v)dv. \tag{2.10}$$

We will use the notation $Var[\ ]$ as well as $V[\ ]$ for the variance throughout this book.

Definition 2.14 The *standard deviation* for a random variable, V, is the square root of the variance, $\sqrt{Var[V]}$.

We will often use the Greek letter (μ) "mu" to represent the expected value of a random variable. The variance and standard deviation are measures of the variability, or spread, of a distribution and are often denoted by the Greek characters σ^2 and σ, respectively. In particular the variance is the average squared difference of a random variable from its mean. The variance is also known as the 2nd central moment.

We are often interested in combinations of two or more random variables or the relationship between these random variables. As such, we need to define a structure that accommodates all of the possible relationships that may exist between random variables. Below we focus on the case of the relationships between two random variables. Generalization to more than two random variables can be found in Casella and Berger [14], for example.

Definition 2.15 The *joint probability density function*, f, for the continuous random variables $V_1, \ldots, V_n$ is $f(V_1, \ldots, V_n)$, if the function $f(V_1, \ldots, V_n)$ satisfies the following properties:

1. $f(v_1, \ldots, v_n) \geq 0$ for all values of $v_1, \ldots, v_n$,
2. $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(v_1, \ldots, v_n) dv_1 \ldots dv_n = 1$.

Definition 2.16 The *joint probability distribution* for the discrete random variables $V_1, \ldots, V_n$ is $P(V_1 = v_1, \ldots, V_n = v_n)$ and must satisfy the following properties following properties:

1. $P(V_1 = v_1, \ldots, V_n = v_n) \geq 0$ for all $v_1, \ldots, v_n$,
2. $\sum_{v_1} \ldots \sum_{v_n} P(V_1 = v_1, \ldots, V_n = v_n) = 1$.

It is possible to have a joint distribution of a continuous random variable and a discrete random variable but this case is not relevant for the methods described in this book.

Definition 2.17 Let $f(V_1, V_2)$ be the joint probability density function for V_1 and V_2. Then the *marginal density function* for V_1 is given by

$$f_1(v_1) = \int_{-\infty}^{\infty} f(v_1, v_2) dv_2. \tag{2.11}$$

Definition 2.18 Two continuous random variables, V_1 and V_2, with joint probability density function f are *independent* if and only if

$$f(v_1, v_2) = f_1(v_1) f_2(v_2) \tag{2.12}$$

for all pairs of values v_1 and v_2. Two discrete random variables, V_3 and V_4, are *independent* if and only if

$$P(v_3, v_4) = P(v_3) P(v_4) \tag{2.13}$$

for all values of v_3 and v_4.

Definition 2.19 Let V_1 and V_2 be two continuous random variables with joint probability density function $f(v_1, v_2)$. Further, let V_3 and V_4 be two discrete random variables with joint probability distribution $P(V_3, V_4)$. The *expected value of a function of more than one random variable* $E[g(V_1, V_2)]$ or $E[g(V_3, V_4)]$ is defined to

be

$$E[g(V_1, V_2)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(v_1, v_2) f(v_1, v_2) dv_1 dv_2 \tag{2.14}$$

for the continuous random variables V_1 and V_2 and

$$E[g(V_3, V_4)] = \sum_{v_3} \sum_{v_4} g(v_3, v_4) P(v_3, v_4) \tag{2.15}$$

for the discrete random variables V_3 and V_4.

Definition 2.20 The *covariance* of two random variables, $Cov(V_1, V_2)$, is defined to be

$$Cov(V_1, V_2) = E(V_1 - E[V_1])(V_2 - E[V_2]). \tag{2.16}$$

We also note that

$$Cov(V_1, V_2) = Cov(V_2, V_1). \tag{2.17}$$

Result 2.3 We note here that $Cov(V, V) = Var[V]$. This is a direct result of the definition of these two quantities.

Definition 2.21 The *correlation* between two random variables, V_1 and V_2, is defined to be

$$Corr(V_1, V_2) = \frac{Cov(V_1, V_2)}{\sqrt{Var[V_1]}\sqrt{Var[V_2]}}. \tag{2.18}$$

As with the covariance, the correlation is symmetric, i.e.

$$Corr(V_1, V_2) = Corr(V_2, V_1). \tag{2.19}$$

Definition 2.22 We will say that two random variables V_1 and V_2 are *uncorrelated* if $Corr(V_1, V_2) = 0$.

Comment 2.1 We note specifically here that independence is a stronger condition than being uncorrelated and, hence, independence implies uncorrelated. That is, if we know that V_1 and V_2 are independent then we know they are uncorrelated. However, it is possible for two random variables to be uncorrelated and not independent.

The covariance and the correlation are measures of how two random variables vary together or how they *co-vary*. The correlation is a unitless measure that attempts to calibrate the way the two variables co-vary in a way that controls for the individual variability in each.

Here we have defined joint densities and joint expectations for two variables. There are equivalent multivariable extensions of these notions. Similarly there are multivariate notions of independence.

Definition 2.23 We will call a sequence of random variables $V_1, \ldots, V_n, \ldots$ *stationary* if the mean and the covariances of the V_i's do not depend on i. That is, we will call a process stationary if $E[V_i]$ and $Cov(V_i, V_j)$ do not depend on i and j.

Definition 2.23 is usually denoted as *weak stationarity*. It is the sense that we will use here. More detail on stationarity for processes can be found in Grimmett and Stirzaker [40] and Brockwell and Davis [10].

We offer the following results for linear combinations of random variables. Let $a_1, a_2, \ldots$ be fixed, known constants and let $E[V_i] = \mu_i$ and $Var[V_i] = \sigma_i^2$ for $i = 1, 2, 3, \ldots$.

Result 2.4

$$E[a_i V_i] = a_i \mu_i. \tag{2.20}$$

Result 2.5

$$Var[a_i V_i] = a_i^2 \sigma_i^2. \tag{2.21}$$

Result 2.6

$$E\left[\sum_i a_i V_i\right] = \sum_i a_i \mu_i. \tag{2.22}$$

Result 2.7

$$Var\left[\sum_i a_i V_i\right] = \sum_i a_i^2 \sigma_i^2 + \sum_i \sum_{j \neq i} a_i a_j \sigma_i \sigma_j Cov(V_i, V_j). \tag{2.23}$$

Comment 2.2 We note that for Result 2.7 the second term on the right hand side is eliminated if all of the variables are pairwise uncorrelated. If some correlation exists, that is, if there is at least one non-zero correlation, then that term cannot be eliminated.

Result 2.8 If $V_1, \ldots, V_n$ have the same mean, $E[V_i] = \mu$, and the same variance, $Var[V_i] = \sigma^2$, then for a given set of constants $a_1, \ldots, a_n$ Result 2.7 can be written as

$$Var\left[\sum_i a_i V_i\right] = \sum_i a_i^2 \sigma_i^2 + \sum_i \sum_{j \neq i} \sigma_i \sigma_j a_i a_j Corr(V_i, V_j) \tag{2.24}$$

which becomes

$$Var\left[\sum_i a_i V_i\right] = \sigma^2 \sum_i a_i^2 + \sigma^2 \sum_i \sum_{j \neq i} a_i a_j Corr(V_i, V_j). \tag{2.25}$$

Comment 2.3 Result 2.8 is *crucial* to the rest of this book because we will be dealing with a large number of random variables that we can treat as having a common mean and common variance but which are correlated. We will use the assumption of a stationary process to justify our assumption about a common mean and a common variance. Thus, we will see Result 2.8 many more times.

Result 2.9 We can use the above results to derive the variance for a mean of a process that is stationary. If $E[V_i] = \mu$ and $Var[V_i] = \sigma^2$, for all i, then with $a_i = \frac{1}{n}$ Result 2.7 becomes

$$Var\left[\frac{1}{n}\sum_i V_i\right] = \frac{\sigma^2}{n^2}\left[n + \sum_i \sum_{j \neq i} Corr(V_i, V_j)\right]. \tag{2.26}$$

Note that based upon Result 2.9 for a stationary process, we need only specify the mean, variance, and correlation of the random variables that encompass the process in order to be able to determine the variance of the mean. We further note that a proportion is simply a mean of random variables that take the value 0 or 1. So this result also applies to proportions.

Result 2.10 For two sets of random variables $V_1, \ldots, V_n$ and $R_1, \ldots, R_m$, we have that

$$Cov\left(\sum_{i=1}^{n} V_i, \sum_{j=1}^{m} R_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} Cov(V_i, R_j). \tag{2.27}$$

It is often the case that it is easier to write a linear combination of random variables by using vector notation. Consequently, let $\mathbf{V} = (V_1, V_2, \ldots, V_n)^T$ be a random (column) vector (where ^{T} represents the transpose of the vector). For example, the linear combination $\sum a_i V_i$ can be written as $\mathbf{a}^T\mathbf{V}$ where $\mathbf{a} = (a_1, a_2, \ldots, a_n)^T$. The definitions below further develop these ideas.

Definition 2.24 The *expected value of a random vector* is the expected value of the elements in that vector. That is,

$$E[\mathbf{V}] = (E[V_1], E[V_2], \ldots, E[V_n])^T = (\mu_1, \mu_2, \ldots, \mu_n)^T. \tag{2.28}$$

Definition 2.25 The *covariance of a random vector* is a matrix whose individual terms, σ_{ij} are the individual covariances, $\sigma_{ij} = Cov(V_i, V_j)$. We will use the notation $\boldsymbol{\Sigma}$ to represent covariance matrices. $\boldsymbol{\Sigma}$ is a symmetric matrix defined as

$$\boldsymbol{\Sigma} = Cov(\mathbf{V}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_{nn} \end{pmatrix}. \tag{2.29}$$

If we let $\mathbf{a}$ be a (column) vector of constants that is the same length as $\mathbf{V}$, then the expected value and variance of $\mathbf{a}^T\mathbf{V}$ is:

Result 2.11

$$E[\mathbf{a}^T\mathbf{V}] = \mathbf{a}^T E[\mathbf{V}] = \mathbf{a}^T\boldsymbol{\mu} \tag{2.30}$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^T$,

and

Result 2.12

$$Var[\mathbf{a}^T\mathbf{V}] = \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a}. \tag{2.31}$$

2.2.1 Specific Random Variables

Definition 2.26 We will call a random variable, V, *a Bernoulli random variable* if it has the following characteristics:

1. V only takes two values, say zero (failure) and one (success),
2. $P(V = 1) = p$ and $P(V = 0) = 1 - p$.

We will denote a random variable, V, as being a Bernoulli random variable with probability of success p as $V \sim Bern(p)$.

Result 2.13 If V is a Bernoulli random variable ($V \sim Bern(p)$), then $E[V] = p$ and $Var[V] = p(1 - p)$.

Definition 2.27 We will refer to a random variable as a *binomial random variable* if it is the sum of n independent Bernoulli random variables, each with the same probability of a success, p. Let $V_i \sim Bern(p)$ for $i = 1, \ldots, n$ and assume that the V_i are independent. We will denote a random variable $V = \sum_{i=1}^{n} V_i$ as being a binomial random variable with parameters n and p as $V \sim Bin(n, p)$.

Result 2.14 If V is a Binomial random variable ($V \sim Bin(n, p)$), then $E[V] = np$ and $Var[V] = np(1 - p)$.

Comment 2.4 The variance of the sum of n uncorrelated Bernoulli random variables is the same as that for a binomial random variable.

Definition 2.28 We will call V *a Gaussian or normal random variable* with mean μ and variance σ^2 if it has the following density function:

$$f_V(v) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(v-\mu)^2}. \tag{2.32}$$

We will denote a Gaussian random variable, V, with mean μ and variance σ^2 as $V \sim N(\mu, \sigma^2)$. The terms Gaussian and normal are interchangeable for this distribution but we will use Gaussian in most cases in this book.

Result 2.15 Let $V \sim N(\mu, \sigma^2)$. then $Z = \frac{V-\mu}{\sigma} \sim N(0, 1)$. The distribution $N(0, 1)$ will be referred to as a standard Gaussian distribution.

Tables 9.1 and 9.2 give percentiles for a standard Gaussian distribution and will be used extensively throughout this text. A Gaussian random variable is the one that is typically associated with a bell-shaped curve.

2.2.2 Estimation and Large Sample Theory

In this section, we review some of the results from large sample probability theory. Large sample probability theory investigates functions of sequences of random variables, $f(V_1, V_2, \ldots)$.

Definition 2.29 An *estimator*, $\hat{\theta}$, of a parameter θ is a function of random variables, $\hat{\theta}(V_1, \ldots, V_n)$. We will use a $\hat{}$ to denote an estimator of a parameter.

Definition 2.30 One estimator of importance is the *sample mean*

$$\bar{V} = \frac{1}{n} \sum_{i=1}^{n} V_i. \tag{2.33}$$

Though, perhaps obvious, the sample mean is an estimator of the mean of a process.

Definition 2.31 We will call an estimator, $\hat{\theta}$, of θ *unbiased* for θ if $E[\hat{\theta}] = \theta$.

Definition 2.32 For an estimator $\hat{\theta} = \hat{\theta}_n(V_1, V_2, \ldots, V_n)$ where $V_1, V_2, \ldots, V_n$ are random variables, the probability distribution of θ is called the *sampling distribution* of θ.

Definition 2.33 If random variables $V_1, V_2, \ldots, V_n$ each have the same probability distribution and they are all independent of each other, then we will say that these random variables are *independently and identically distributed* (*iid*).

We will use the notation *iid* to represent a collection of random variables that are independently and identically distributed.

Theorem 2.1 *Let* $V_1, \ldots, V_n$ *be a sequence of* iid *random variables with expected value* μ *and variance* σ^2. *Then*,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} V_i = \lim_{n\to\infty} \bar{V}_n = \mu. \tag{2.34}$$

Theorem 2.2 *Let* $V_1, V_2, \ldots, V_n$ *be* iid *Gaussian random variables with expected value (mean)* μ *and variance* $\sigma^2 < \infty$. *That is,* $V_i \sim N(\mu, \sigma^2)$. *Then*

$$\lim_{n\to\infty} P\left(\frac{\bar{V}_n - \mu}{\sigma} \leq x\right) \to P(Z \leq x) \tag{2.35}$$

where $Z \sim N(0, 1)$.

The following result relaxes the assumption of Gaussianity necessary to achieve a mean which follows a Gaussian distribution.

Theorem 2.3 *Let* $V_1, V_2, \ldots, V_n$ *be* iid *random variables with expected value (mean)* μ *and variance* $\sigma^2 < \infty$. *Then*

$$\lim_{n\to\infty} f_{\frac{\bar{V}_n - \mu}{\sigma/\sqrt{n}}} \to f_Z \tag{2.36}$$

where Z *is a standard Gaussian random variable*.

Comment 2.5 Theorem 2.3 is known as the Central Limit Theorem (CLT). In more general language it states that if we average a large number of random variables, then the distribution of the sample mean will be a Gaussian distribution. This version of the central limit theorem is based upon *iid* data. There are numerous other central limit theorems that relax these conditions. We will use those as necessary. The interested reader is directed to Jacod and Shiryaev [50], for example. Below we offer an extension to Result 2.3. Recently, Dietz and Schuckers [23] proposed some biometric specific extensions to the CLT given above.

Theorem 2.4 *Let* $V_1, V_2, \ldots,$ *be a sequence of independent random variables that are each the sum of* m_i *correlated binary random variables such that* $E[V_i] = m_i\pi$ *and* $\sigma^2 = Var[V_i] = m_i\pi(1-\pi)(1+\rho(m_i-1))$ *then by Moore* [69] *we have the following result. Let*

$$\hat{\pi} = \frac{\sum_{i=1}^{n} V_i}{\sum_{i=1}^{n} m_i}, \tag{2.37}$$

then

$$\lim_{n\to\infty} f_{\frac{\hat{\pi}_n - \pi}{\sigma_{\hat{\pi}}}} \to f_Z \tag{2.38}$$

where $Z \sim N(0, 1)$,

$$\sigma_{\hat{\pi}} = \sqrt{\frac{\pi(1-\pi)(n\bar{m} + \rho\sum_{i=1}^{n}(m_i - 1))}{n^2\bar{m}^2}} \tag{2.39}$$

and

$$\bar{m} = \frac{\sum_{i=1}^{n} m_i}{n}. \tag{2.40}$$

The mean and variance that we use in Theorem 2.4 are the result of a particular correlation structure that we will see in Chaps. 3 and 7. This result is another form of a central limit theorem.

Theorem 2.5 *Let $V_1, V_2, \ldots$ be a sequence of independent random variables with $E[V_i] = \mu_i$, $Var[V_i] = \sigma_i^2$ and $r_i^3 = E[\mid V_i - \mu_i \mid^3] < \infty$. Also let $S_n = \sum_{i=1}^n V_i$. If*

$$\lim_{n\to\infty} \frac{(\sum_{i=1}^n r_i^3)^{1/3}}{(\sum_{i=1}^n \sigma_i^2)^{1/2}} = 0, \tag{2.41}$$

then

$$f_{S_n^*} \to N(0, 1) \tag{2.42}$$

where

$$S_n^* = \frac{S_n - \sum_{i=1}^n \mu_i}{(\sum_{i=1}^n \sigma_i^2)^{1/2}}. \tag{2.43}$$

Equation (2.41) is known as the *Lyapunov Condition* and provides a central limit theorem for *non-iid* random variables.

Definition 2.34 The *effective sample size* is the equivalent number of observations that would result in the same variance of the estimator if the data were *iid*. We will generally denote an effective sample size with a superscript dagger, i.e. †. The effective sample size is calculated by taking the actual sample size say, n, and multiplying it by the ratio of the variance if the data were *iid* to the observed variance. The equation is

$$n^\dagger = n\left[\frac{V_{iid}[\hat{\theta}]}{V[\hat{\theta}]}\right] \tag{2.44}$$

where V_{iid} is the variance assuming the data is collected via a simple random sample or assuming the data is *iid*.

The effective sample size is used to measure the amount of independent information in a sample. It is something that we will consistently use later as we discuss methods for evaluating the performance of a classification or a biometric system.

2.3 Statistical Inference

In this section, we discuss statistical inference. Statistical inference draws conclusions or makes statements about a process parameter based upon a sample. We treat a sample as composed of n random variables. For statistical inference throughout this book, we will focus on confidence interval estimation and hypothesis testing. We assume that the reader has been exposed previously to confidence intervals and

hypothesis tests. A brief refresher of these topics is given. An introduction is then given to some less well known statistical methods. These include randomization tests, bootstrap , jackknife methods, sample size calculations, power calculations and prediction intervals. Bootstrap, jackknife and randomization methods all fall under the category of non-parametric or distribution-free approaches. Prediction intervals are inferential intervals for some function of future observations from a process. This is in contrast to confidence intervals which are generally for a single parameter of that process.

Non-parametric here will mean that no attempt is made to model the distribution or shape of the observations. Our methods will attempt to maintain the covariances and correlations between observations. The bootstrap and jackknife approaches are methods for estimating the sampling distribution by resampling the collected data. Randomization tests follow a similar idea to resampling methods in that they are non-parametric methods. The basic approach of a randomization test is to combine the data from two or more groups and then randomly re-assign those observations back to those groups and recalculate a given test statistic. This process is then repeated multiple times creating a distribution for the given test statistic assuming a null hypothesis of equality of some parameter among the groups. These methods are an alternative to the large sample methodology which *does* depend upon a specific distribution.

All of the methods in this section are dependent upon having a quality sample from a population or from a process. The following comment addresses some concerns about the general use of statistical methods.

Comment 2.6 Most statistical methods start with an assumption that the sample upon which inference is based is a random sample. We note here that there are statistically appropriate methods for collection observations beyond a simple random sample. (See Definition 2.35.) Frequently, in biometric systems testing the sample that is used is a convenience sample, i.e. one that is readily available. Some in the biometrics community have stated that this disqualifies the use of statistical tools. Obviously, a book on statistical methods will argue against such a broad statement. Here we note that this does not completely preclude the use of statistical methods. We quote now from p. 17 of Hahn and Meeker's influential text *Statistical Intervals: A Guide for Practitioners* [43]:

> Because one is not sampling randomly, statistical intervals strictly speaking are not applicable for convenience sampling. In practice, however, one uses experience and understanding of the subject matter to decide on the applicability of applying statistical inferences to the results of convenience sampling. Frequently, one might conclude that the convenience sample will provide data that, for all practical purposes, are as "random" as those that one would obtain by a simple random sample. ... Our point is that, treating a convenience sample as if it were a random sample *may sometimes* be reasonable from a practical point of view. However, the fact that this assumption is being made needs to be recognized, and the validity of using statistical intervals, as if a random sample had been selected needs to be critically assessed based upon the specific circumstances.

The emphasis in the above paragraph is from the original text.

The larger point that we wish to make here is that there are limitations to the inference that can be made from statistical methods. These limitations should not be

ignored. These limitations need to be thoughtfully considered as part of any conclusions that one wants to draw from a data analysis.

Definition 2.35 A *simple random sample* or SRS is a sample of size n taken such that each possible combination of n units is equally likely.

Definition 2.36 For a sample of n random variables, $V_1, \ldots, V_n$, we will call s the *sample standard deviation* where

$$s = \sqrt{\frac{\sum_{i=1}^{n}(V_i - \bar{V})^2}{n-1}} \tag{2.45}$$

and

$$\bar{V} = n^{-1}\sum_{i=1}^{n} V_i. \tag{2.46}$$

Result 2.16 If $V_1, V_2, \ldots, V_n$ are *iid* random variables with $Var(V_i) = \sigma^2$ for all i, then s^2 is unbiased for σ^2 where s is given by (2.45). We refer to s^2 as the sample variance.

Definition 2.37 If we have n samples from a distribution, $V_1, \ldots, V_n$ and we estimate the population standard deviation, σ, using the sample standard deviation, s, then

$$\frac{\bar{V} - \mu}{s/\sqrt{n}} \sim t(n-1) \tag{2.47}$$

where $t(n-1)$ represents a *t-distribution* with $n-1$ degrees of freedom (df).

A table of percentile of the *t-distribution* can be found in Table 9.3 for a variety of degrees of freedom. We use *degrees of freedom* here in the statistical sense, not the biometrics sense that is used by Daugman [19].

Definition 2.38 The *Student's t-distribution or t-distribution probability density function* with ν degrees of freedom is

$$f_T(t) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}(1 + t^2/\nu)^{-(\nu+1)/2} \tag{2.48}$$

for a random variable T.

Definition 2.39 The *Chi-squared distribution* with ν degrees of freedom is

$$f_X(x) = \frac{x^{(\nu/2)-1}\mathrm{e}^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)} \tag{2.49}$$

for a random variable X when $x > 0$ and 0 otherwise.

Definition 2.40 The *standard error* for an estimator $\hat{\theta}$ is the estimated standard deviation of the sampling distribution of that estimator.

Result 2.17 If we take a random sample of size n from a process or population with mean μ and finite variance σ^2 and n is large, then the sampling distribution for the normalized sample mean

$$\frac{\bar{V}_n - \mu}{s/\sqrt{n}} \tag{2.50}$$

is a *t-distribution*.

Result 2.18 If we take a random sample of size n from a process or population with proportion p and large n, then the sampling distribution for the sample proportion is a Gaussian distribution.

Result 2.19 The standard error for the mean is $\frac{s}{\sqrt{n}}$ for a simple random sample and is denoted by $s_{\bar{X}}$.

Result 2.20 The standard error for a proportion, $\hat{p} = \frac{X}{n}$ taken from a simple random sample is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ and is denoted by $s_{\hat{p}}$.

Results 2.17 and 2.18 will also hold if we do not have correlated samples. The sample size needed for convergence to the reference distributions, t and Gaussian, respectively, increases as the amount of dependence increases. If we have independent samples, then the general guidance that is given is that the sample size for the mean needs to be at least $n \geq 30$ and the sample size for a proportion needs to be large enough so that $np \geq 10$ and $n(1-p) \geq 10$, see, for example, Deveaux et al. [96].

2.3.1 Confidence Intervals

Statistical methods are predicated on the idea that information that we have collected from a population or process is incomplete. Consequently, the estimated values for a population or process parameter are unlikely to be equal to the parameters themselves. To reflect this uncertainty, intervals or ranges of values are created. In the case of multivariate parameters, confidence regions or bands are created. Below we discuss in some detail confidence intervals for parameter estimation. In Sect. 2.3.2, we introduce the idea of prediction intervals for future observations. A complete list of statistical intervals along with a thorough description of the issues involved in each can be found in Hahn and Meeker [43].

Confidence intervals are statistical tools for making inference about a parameter. We defined a parameter above in Definition 2.3. A point estimate is a single

point that is an estimate of a parameter. Because there is variability in sampling, our estimate rarely has the same value as the *estimand*, the quantity we are trying to estimate. Therefore, we recognize that a range of values is necessary to summarize our knowledge of the parameter based upon the information in our sample. Note that the width of a particular interval is indicative of how much information we have about the parameter. That is, if we believe that a rate or proportion falls in the interval $(0.20, 0.40)$ we have far less information than if we believe it is in the interval $(0.29, 0.32)$. Thus, the width of an interval is a measure of how much information we have about the parameter. Wider intervals represent less information; narrower intervals represent more information about the process under consideration.

Definition 2.41 A *confidence level*, $1 - \alpha$, is defined to be the proportion of times that a confidence interval 'captures' the parameter in repeated sampling from a process or a population. Here by capture we mean that the parameter value in contained in the confidence interval.

Typical values for confidence levels are 90%, 95% and 99% which correspond to $\alpha = 0.10, 0.05$ and 0.01, respectively. As mentioned above, there are multidimensional equivalents to a confidence interval. These are commonly referred to as confidence regions.

Definition 2.42 The $100 \times k$th *percentile* is the point, v_k, in the distribution of a random variable V where $P(V \leq v_k) \leq k$ and $P(V \geq v_k) \geq 1 - k$.

Definition 2.43 A $100 \times (1 - \alpha)\%$ *confidence interval* for a parameter θ is defined to be an interval formed by two real-valued functions of the data $U(V_1, \ldots, V_n)$ and $L(V_1, \ldots, V_n)$ such that

$$P(L(V_1, \ldots, V_n) \leq \theta \leq U(V_1, \ldots, V_n)) \geq 1 - \alpha. \tag{2.51}$$

U and L are meant to denote the upper and lower endpoints of the interval and are themselves random variables. Note that it is possible for U or L to be defined as constants.

The typical methodology for making a confidence interval is to take the $\alpha/2$th percentile and the $1 - \alpha/2$th percentile of the sampling distribution for an estimator and use those to make a confidence interval. This is the typical *two-sided* confidence interval. It is the most commonly used because for estimation of means and proportions and it yields the narrowest interval for a given confidence level. Other choices are possible (and sometimes advisable). One other common choice is to create an interval that is one-sided, i.e. that has either a set upper or a set lower bound but not both. This is done by setting $U(V_1, \ldots, V_n) \stackrel{set}{=} \infty$ or $L(V_1, \ldots, V_n) \stackrel{set}{=} -\infty$. This is appropriate if we are only interested in estimation of one tail for the parameter. For example, we might only be interested in the worst case false accept rate that is credible based on a sample from the process. In that case we are uninterested in the

lower bound and we would set $L(V_1, \ldots, V_n) \stackrel{set}{=} -\infty$. (For a proportion we could equivalently, and without loss of information, set $L(V_1, \ldots, V_n) \stackrel{set}{=} 0$.)

Below we present methodology for two confidence intervals assuming random samples from an *iid* process. It is not always the case in classification performance that one has independent (or uncorrelated) samples. We'll spend a good deal of the rest of this text on cases that are not *iid* but here as a heuristic foundation we present *iid* cases. Thus, we provide these tools as a basis for discussion and as a summary of the background that is assumed for the rest of this text. We will provide generalizations for these methods, as appropriate, in later chapters.

Result 2.21 Let $R_1, \ldots, R_n$ be *iid* samples from a population with fixed mean, say μ_R. If n is large (generally $n \geq 30$) then we can use the following to make a $(1-\alpha)100\%$ *confidence interval for* μ_R:

$$\bar{R} \pm t_{\alpha/2;n-1} \frac{s_R}{\sqrt{n}} \tag{2.52}$$

where

$$\bar{R} = \frac{1}{n} \sum_{i=1}^{n} R_i \tag{2.53}$$

and

$$s_R = \sqrt{\frac{\sum_{i=1}^{n} (R_i - \bar{R})^2}{n-1}} \tag{2.54}$$

and $t_{\alpha/2;n-1}$ is the $(1-\alpha/2) \times 100$th percentile of a t-distribution with $n-1$ degrees of freedom. Table 9.3 gives percentiles for some t-distributions.

Some comments on the use of the interval above are important. We have two possible sources of a Gaussian distribution for the sampling distribution of the sample mean. The first happens if we are sampling from a Gaussian distribution and then we get a Gaussian sampling distribution for any sample size. The second happens if we are sampling from a non-Gaussian distribution and in that case we need our sample to be sufficiently large. As a consequence, it is often possible to use the methods described in Result 2.21 when the sample size is less than 30. This will be true if the observed values in the sample are roughly bell-shaped and symmetric.

Result 2.22 Let R be a Binomial random variable with probability of success p and number of trials n. $R \sim Bin(n, p)$. If n is large (generally $n\hat{p} \geq 10$ *and* $n(1-\hat{p}) \geq 10$) then we can use the following to make a $(1-\alpha)100\%$ *confidence interval for* p*, the success rate or probability for* R:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{2.55}$$

where

$$\hat{p} = \frac{r}{n} \tag{2.56}$$

which is the proportion observed in the sample and $z_{\alpha/2}$ is the $(1-\alpha/2) \times 100$th percentile from a Gaussian distribution. r is the observed value of the RV R.

Note that a proportion is just an average of zero's and one's. Thus Theorem 2.3 holds for proportions. Confidence intervals can be created for a wide range of process quantities, including medians, variances and percentiles. We will not cover these confidence intervals in this book.

Comment 2.7 In order to use the confidence interval given in Result 2.22, we must have a large sample size, that is, n needs to be sufficiently large. Here we say that n is large if $n\hat{p} \geq 10$ *and* $n(1-\hat{p}) \geq 10$. We follow De Veaux et al. [96] among other introductory statistics texts in using the value 10 as a cutoff. We do note that this is an area that has seen some recent attention in the statistics literature. The interested reader should consult, for example, Agresti and Coull [4], Brown et al. [11], or Brown et al. [12] for more details.

2.3.2 Prediction Intervals

Above we have defined the basics of confidence intervals. Again, see Hahn and Meeker [43] for a thorough treatment of these ideas. Here we point out that a confidence interval is meant to take information collected on a particular process and give us information (in the form of a range of values) about a parameter or parameters measured on that process. It is not meant to predict future performance of that process, although it can be a useful tool for this prediction. Prediction intervals, on the other hand, are meant to do exactly that—provide an inferential interval for a function of future observations.

Definition 2.44 A $100 \times (1-\alpha)\%$ *prediction interval* for a real-valued function of future observations, $g(V_{n+1}, \ldots, V_{n+m})$, is defined to be an interval formed by two real-valued functions of n observed observations $U_p(V_1, \ldots, V_n)$ and $L_p(V_1, \ldots, V_n)$ such that

$$P(L_p(V_1, \ldots, V_n) \leq g(V_{n+1}, \ldots, V_{n+m}) \leq U_p(V_1, \ldots, V_n) \mid V_1, \ldots, V_n) \geq 1-\alpha \tag{2.57}$$

where $V_1, \ldots, V_n$ are n observed values and $V_{n+1}, \ldots, V_{n+m}$ are m random variables from the same process representing future unknown values. U_p and L_p are meant to denote the upper and lower endpoints of the interval and are themselves random variables until $V_1, \ldots, V_n$ are observed. As above, it is possible for U or L to be defined as constants.

Prediction intervals are confidence intervals but they are confidence intervals for future values of a process. Hence they are called *prediction* intervals. Below we present prediction intervals for a mean and for a proportion both from *iid* processes.

Result 2.23 Let $R_1, \ldots, R_n$ be an *iid* sample from a population with fixed mean, say μ_R, and let $\bar{R}_n^\diamond$ be the mean of the $n^\diamond$ unobserved future values, $R_{n+1}, \ldots, R_{n+n^\diamond}$. If n is large (generally ≥ 30) then we can use the following to make a $(1-\alpha)100\%$ *prediction interval for* $\bar{R}_n^\diamond$:

$$\bar{R} \pm t_{\alpha/2;n-1} s_R \sqrt{\frac{1}{n^\diamond} + \frac{1}{n}} \tag{2.58}$$

where

$$\bar{R} = \frac{1}{n} \sum_{i=1}^{n} R_i \tag{2.59}$$

is the sample mean and

$$s_R = \sqrt{\frac{\sum_{i=1}^{n} (R_i - \bar{R})^2}{n-1}} \tag{2.60}$$

is the sample standard deviation following Definition 2.36 calculated on $R_1, \ldots, R_n$.

Result 2.24 Let R be a Binomial random variable with probability of success p and n observed trials and let R_2 be a Binomial random variable with probability of success p from $n^\diamond$ unobserved trials. We assume that R_2 takes observations independently of R, but with the same probability of success p.

If n and $n^\diamond$ are large (generally $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$) then we can use the following to make a $(1-\alpha)100\%$ prediction interval for the future proportion of successes $\frac{R_2}{n^\diamond}$:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p}) \left(\frac{1}{n} + \frac{1}{n^\diamond} \right)} \tag{2.61}$$

where $\hat{p} = \frac{R_1}{n}$ which is the proportion of observed successes in the original sample.

Prediction intervals are wider than the equivalent confidence intervals because there are two sources of variability: one for estimation of the relevant parameter—which is the appropriate variability for a confidence interval—and one for the sampling variability of $g(V_{n+1}, \ldots, V_{n+n^\diamond})$ a function of the future observations.

2.3.3 Hypothesis Testing

Hypothesis tests are a statistical counterpart to confidence intervals. Hypothesis testing is a methodology for determining whether a particular parameter value is sup-

ported by the data. This methodology is in opposition to confidence intervals which start with no particular value in mind. Hypothesis tests are particularly useful in biometrics for testing whether an error rate is below a particular value. For example, hypothesis testing would be used to determine if a false accept rate is less than 1.5% or the failure to enrol rate is less than 4%. The statistical theory also allows for the possibility that the hypothesis we want to test is simply that the parameter is not equal to a particular value. We outline the fundamental pieces of a hypothesis test below:

1. The *null hypothesis* H_0: is written first and represents the value of the parameter that we would like to test against.
2. The *alternative hypothesis* H_1 is written next and it represents the hypothesis regarding the parameter we would like to suggest is true.
3. The *test statistic* is what follows next and it is generally a method of measuring how large the discrepancy is between the parameter value specified by the null hypothesis and the observed estimator of the parameter calculated from the data.
4. The *p-value* is the probability that an observed test statistic will be more extreme (relative to the alternative hypothesis) than the test statistic that was observed, if H_0 were true.
5. The *decision rule* is what determines whether the null hypothesis is rejected or not rejected. *The decision rule is to reject the null hypothesis if the p-value is small*. If we reject the null hypothesis, we conclude that the parameter estimate based upon data collected represented a significant difference from the hypothesized parameter value given in the null hypothesis. The alternative conclusion that we can make is to 'fail to reject' the null hypothesis.

The logic of hypothesis testing is related to the idea of falsifiability, Popper [77]. The null hypothesis is usually a value for a parameter that is specified *a priori* data collection. The null hypothesis is the statement that it is possible to falsify. It is rarely possible to show that a particular hypothesis is true—that would require a census of our process outcomes—but rather we can find enough empirical evidence to conclude that it is not. Thus the language of our decision rule is to *reject* the null hypothesis or to *fail to reject* the null hypothesis. How much evidence is necessary is set by how small the p-value needs to be in order to reject. The common choices for the significance level are 0.10, 0.05 and 0.01. These values—significance levels usually denoted by the Greek letter α, see Definition 2.46 below—are instituted structurally by some disciplines and are backed by historical precedence. The exact value to be used in practice need not be any of these; however, if a value for α is chosen, it needs to be established before data is analyzed. A selected α should be clearly stated as part of any reports on the performance of a classification or matching system. Further, it is possible and increasingly the case in some fields, such as medicine, to report only the *p-value* without specifying α and letting readers determine the significance of a given result particularly for those results where different choices of α could lead to different conclusions.

In any decision making process that involves uncertainty or incomplete information, errors can result. One source of these errors is the incomplete nature of any

data collection to obtain information about a process. Statisticians distinguish between two types of errors: Type I and Type II errors. We define these below as well as the notation for the probabilities of each.

Definition 2.45 A *type I error* occurs when an experimenter rejects the null hypothesis and, in fact, the null hypothesis is true.

Definition 2.46 The *significance level*, denoted by the Greek letter α, is the probability that we reject the null hypothesis when indeed the null hypothesis is true. Formally,

$$\alpha = P(\text{Reject } H_0 \mid H_0 \text{ is true}). \tag{2.62}$$

The significance level is the probability of making a type I error.

Definition 2.47 A *type II error* occurs when an experimenter fails to reject the null hypothesis if the null hypothesis is false. Thus,

$$P(\text{Fail to Reject } H_0 \mid H_0 \text{ is not true}). \tag{2.63}$$

Definition 2.48 Statistical *power* is the probability of rejecting the null hypothesis when the null hypothesis is false. This is one minus the probability of a type II error, and is usually denoted by $Power = 1 - P(\text{type II error}) = 1 - \beta$ where $\beta = P(\text{type II error})$.

The evaluation of classification and matching performance is often saddled by a difficulty of notation and definitions. This confusion occurs because we are trying to estimate and make inference about classification errors. Since these classification rates are themselves similar to the type I and type II error rates, confusion can arise. This misunderstanding becomes acute if we are testing the classification error rates. In this book, we aim to be as explicit as possible in identifying both the quantity of interest and the relevant statistical error rate.

Below we present two example hypothesis tests, one for a process mean and one for a process proportion. Both of these are large sample tests meaning that they use a central limit theorem to allow an approximation to a specific sampling distribution. Most of the measures that we will describe in this text do not assume an *iid* data collection but are general enough to include such a structure if it is appropriate. Additionally, we will present hypothesis tests that do not depend upon a particular limiting distribution from a large sample.

Result 2.25 Here we present a large sample hypothesis test for a process mean. Let $R_1, \ldots, R_n$ be an *iid* sample from a process with finite variance. If $n \geq 30$ or the observations $R_1, \ldots, R_n$ are sampled from a Gaussian distribution, then we can use the following hypothesis test to test whether the process mean is less than a

hypothesized value μ_0.

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu < \mu_0.$$

Test Statistic:

$$t = \frac{\bar{R} - \mu_0}{s_R/\sqrt{n}}. \tag{2.64}$$

p-value: $p = P(T < t)$ where T is a t-distribution random variable with df $n - 1$.

We will reject the null hypothesis, H_0, if *p-value* is small. If a significance level, α, has been prespecified then we will reject the null hypothesis if $p < \alpha$. Table 9.3 has percentiles for certain t-distributions.

To test whether the population mean is greater than μ_0 we would change our alternative hypothesis to $H_1 : \mu > \mu_0$ and change our *p-value* calculation to be $p = P(T > t)$.

Result 2.26 This is a large sample hypothesis test for a process proportion. Let B_n be a binomial random variable such that $B_n \sim Bin(n, p)$. If $B_n \geq 10$ and $n - B_n \geq 10$, then we can use the following hypothesis test to determine whether the probability of interest for this process is less than p_0.

$$H_0 : p = p_0$$
$$H_1 : p < p_0.$$

Test Statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}. \tag{2.65}$$

p-value: $p = P(Z < z)$ where Z is a standard Gaussian random variable and we can use Table 9.1 to determine the *p-value*.

We will reject the null hypothesis, H_0, if $p < \alpha$ for a particular significance level. If a significance level has not been determined *a priori*, then we will reject the null hypothesis if the *p-value* is small.

For the hypothesis tests above, we used a one-sided alternative hypothesis of $H_1 : \mu < \mu_0$ and $H_1 : p < p_0$. The reason for focusing on these particular tests is that testing to determine if an error rate is below a particular boundary is a common goal for biometric authentication. We will focus on one-sided tests with an alternative that is 'less than' in the rest of this book; however, other alternative hypotheses are possible. The other one-sided alternative hypotheses would be $H_1 : \mu > \mu_0$ and $H_1 : p > p_0$. Respectively, the *p-values* for those tests are $p = P(T > t)$ and $p = P(Z > z)$. If we are unsure about the direction of the alternative hypothesis (either higher or lower), we can use a two-sided alternative hypothesis. For the two

examples given above, we would use $H_1 : \mu \neq \mu_0$ for testing a process mean and $H_1 : p \neq p_0$ for testing a process proportion. The *p-values* that are associated with these two sided tests are $2P(T > |t|)$ and $2P(Z > |z|)$, respectively.

2.3.4 Sample Size and Power Calculations

Researchers often want to know how many observations they will have to collect for a particular sample. These type of determinations generally fall under a group of statistical methods known as sample size calculations. Some attempts have been made in biometrics to provide sample size calculations for false match and false non-match rates, e.g. Schuckers [87]. It is possible to take the confidence intervals and hypothesis testing methods that we have outlined above and solve them for the number of individuals that we want to test. We do this by establishing a criterion and then determining the number of samples needed for that criterion. In the case of a confidence interval, we are often interested in specifying the total width of a confidence interval for a given confidence level, while for hypothesis testing we often want to specify the power, or the probability of not making a type II error. In general, determining the amount of data to be collected for a confidence interval is considered a sample size calculation, while a similar determination for the power of a hypothesis test is considered a power calculation.

It is important to note that if the process changes, then any inferences about the process based upon pre-change observations would not be appropriate. Further, the calculations below as well as similar ones in the rest of the text will depend upon some estimation of the process parameters, especially those involving the variability of the process. Without those, it is not possible to derive methods for the amount of data needed. Care and thought need to be used in consideration of which estimates to use for these parameters. Schuckers [87] provides some insight in the context of biometrics for the ways to undertake this process. We provide a brief discussion of these strategies below.

2.3.4.1 Sample Size Calculations

We start with a determination of the number of observations necessary for creating a $(1-\alpha)100\%$ confidence interval for a mean to have a margin of error of B. The margin of error for a traditional confidence interval is the portion after the $\pm$. We note that the notation 6 ± 3 here means the interval $(6-3, 6+3) = (3, 9)$.

Result 2.27 It is possible to 'invert' a CI for mean to determine the number of observations needed to get a CI with a margin of error of B.

$$B = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{2.66}$$

$$B^2 = z_{\alpha/2}^2 \frac{\sigma^2}{n}. \tag{2.67}$$

Solving for n,

$$n = \left\lceil z_{\alpha/2}^2 \frac{\sigma^2}{B^2} \right\rceil, \tag{2.68}$$

where $\lceil\ \rceil$ is the ceiling function or the 'next largest integer' function.

Comment 2.8 Note that here we have used the $1-\alpha/2$th percentile for a Gaussian or normal distribution here, $z_{\alpha/2}$. The formula for a confidence interval for the mean uses a t-distribution. Here we use the $z_{\alpha/2}$ as an approximation to the percentile of the t-distribution since the latter depends upon the degrees of freedom of the estimated variance, here $n-1$ which is one less than the observed sample size. It is possible to recursively determine the appropriate sample size by iteratively using the following

$$n_{m+1} = \left\lceil (t_{\alpha/s;n_m-1})^2 \frac{\sigma^2}{B^2} \right\rceil \tag{2.69}$$

until n_m converges. For most applications the sample size using the Gaussian approximation is sufficient. Here we need to specify *a priori* an estimate for the variance σ^2, as well as the confidence level, $1-\alpha$, in order to be able to calculate n, the number of observations to collect.

Result 2.28 Here we derive the sample size for a confidence interval for a proportion from a process. As with the mean case above, we can solve the equation

$$B = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \tag{2.70}$$

$$B^2 = z_{\alpha/2}^2 \frac{p(1-p)}{n} \tag{2.71}$$

for n to get a sample size for estimation of a process proportion to within a given width. We then get,

$$n = \left\lceil z_{\alpha/2}^2 \frac{p(1-p)}{B^2} \right\rceil. \tag{2.72}$$

Comment 2.9 For both of the sample size calculations given here—Result 2.27 and Result 2.28—it is necessary to specify *a priori* some quantities. In the case of the mean it is necessary to specify the confidence level $(1-\alpha)$, the margin of error (B) and the standard deviation, (σ), or equivalently the variance, (σ^2). How to specify these quantities is often a nominal hurdle for some practitioners. In particular, the requirement to specify σ can be vexing since this is a process parameter and

gaining knowledge about the process of interest is often the goal of a data collection. Thus, an investigator may not have a good sense of how to specify σ. Statisticians have developed a series of suggestions and guidelines for choosing these values.

For process parameters, there are three basic methods available to obtain useable values: a pilot study, a previous study on a similar process, or an educated guess. A pilot study is a small-scale data collection to ascertain information about both the process of interest and the data collection tools. To that end, it can be used to gain valuable information. In particular, the standard deviation from observations collected on a pilot study from the process of interest is a valid point estimate for σ. Similarly, a standard deviation reported by another study from a similar process would likewise be appropriate. Lastly, an approximation based upon available information is possible to use. One useful tool in this endeavor is to imagine the range of values (*maximum minus minimum*) that the process of interest can take and divide that range by 6 to get an estimate of σ. This procedure comes from approximating the total range by 6σ which is based upon the fact that over 99% of a Gaussian distribution will fall within three standard deviations of the mean. It is important to note that the sample size, n, varies with the square of the standard deviation (or with the variance), σ^2. Thus, if a practitioner wants to be conservative (in a statistical sense), then it is reasonable to inflate an observed or estimated standard deviation, σ, to compensate for the fact that most values for σ that are used in sample size calculations are estimates.

Comment 2.10 For determining the sample size for a proportion—Result 2.28—it is likewise necessary to specify a process parameter, the process proportion (p), in order to obtain a sample size calculation. This is tricky since the data collection is aimed at learning about p, but it is necessary to specify the estimand in order to determine the sample size. As with the sample size calculation for the mean, we can use information from a pilot study or from a previous study on a similar process to glean a plausible value for p. Additionally, in the case of inference about a process proportion, we can also use the value $p = 0.5$ since that value is the most conservative possible choice meaning that it will produce the largest possible n.

Comment 2.11 In addition to estimation of process parameter(s) for sample size calculations, it is also important to specify the confidence level $(1 - \alpha)$ and the margin of error (B). The selection of these is often guided by external constraints such as those required by regulating agencies. We point out here that by decreasing the margin of error, we increase the sample size. Intuitively, more observations (and hence more information) allows for the construction of intervals with greater precision. Similarly, we note that for a given margin of error B, having a higher confidence level means a larger sample size. Thus, decreasing α (and, hence, increasing the confidence level) while simultaneously maintaining B, means that relatively more observations will be required.

2.3.4.2 Power Calculations

Power calculations are to hypothesis testing as sample size calculations are to confidence intervals. We use power calculations to determine the number of observations that are needed for a hypothesis test. For the sample size calculations above, our criterion was the width of the confidence interval. For power calculations we determine the sample size needed to achieve a given power, $1-\beta = 1-P$ (type II error), for a specific significance level of a test, α. As with sample size calculations, it is necessary to provide estimates about the process of interest in order to be able to utilize these calculations.

Power calculations start with a hypothesis test that we wish to carry out at the end of the data collection. With the desired test in mind, we specify the desired or required significance level of the test, α, and the desired power $= 1-\beta$ for the test. We must also specify a value for our estimand in the alternative hypothesis that will give us the desired power. This is done so that we can calculate the power explicitly. For example, if we might want to test the hypotheses: $H_0 : \theta = 0.05$ against $H_a : \theta < 0.05$, we would have to specify a value for $\theta < 0.05$, say $\theta = 0.03$. This value would then allow the calculation of power and the determination of the number of samples needed to achieve that power.

Recall that for a single proportion from *iid* observations that are Bernoulli, i.e. zero's or one's, we have the following hypothesis test:

$$H_0 : p = p_0$$
$$H_1 : p < p_0.$$

Suppose that we want to test at the $\alpha = 0.05$ significance level whether a particular process has a failure rate of less than $p = 0.10$ and we want the power to be 80% if $p = 0.05$. This means that $0.80 = 1-\beta$ and so $\beta = 0.20$. (Note that there are several proportions in this case. This is one example where it is easy to confuse performance rates, p, p_0 and p_a, with hypothesis testing error rates, α and β.) Here p_0 is 0.10 and we will call the value that we want to detect p_a which is 0.05 in this case. Ideally, we would like the power to be 100% but, as with all estimation, 100% accuracy is very—read infinitely—costly in terms of the number of observations that are needed. Consequently, we must trade off the desired power with its cost. Generally, a subscript of 'zero' will indicate a value to be tested and a subscript of 'a' will indicate a value inside the range of the alternative hypothesis. Having specified the desired power, the level of the test, and an alternative value for the failure rate we can turn to calculation of the appropriate sample size.

To determine the number of observations that would be required to achieve the above conditions, we start by noting that these conditions translate to the following probability statement

$$P\left(\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} < -z_\alpha \middle| p = p_a\right) = 1-\beta, \tag{2.73}$$

where z_α is the $1-\alpha$th $\times$ 100 percentile of a $N(0,1)$ which assumes a large enough sample size so that we can use a Gaussian approximation to the sampling distribution of $\hat{p}$. We can then rewrite (2.73) as

$$P\left(\hat{p} < p_0 - z_\alpha\sqrt{p_0(1-p_0)/n} \mid p = p_a\right) = 1-\beta \tag{2.74}$$

which becomes

$$P\left(\frac{\hat{p}-p_a}{\sqrt{p_a(1-p_a)/n}} < \frac{p_0 - z_\alpha\sqrt{p_0(1-p_0)/n} - p_a}{\sqrt{p_a(1-p_a)/n}}\right) = 1-\beta. \tag{2.75}$$

If our sample size is sufficiently large, then the quantity on the left side of the inequality in (2.75) is a standard Gaussian random variable and, thus, we can equate the right hand side of the inequality to the appropriate percentile, z_β of a standard Gaussian distribution. We then have

$$z_\beta = \frac{p_0 - z_\alpha\sqrt{p_0(1-p_0)/n} - p_a}{\sqrt{p_a(1-p_a)/n}}. \tag{2.76}$$

We can rearrange (2.76) to get the following

$$z_\alpha\sqrt{p_0(1-p_0)/n} + z_\beta\sqrt{p_a(1-p_a)/n} = p_a - p_0. \tag{2.77}$$

We can then square both sides and solve for n. We then get the following result.

Result 2.29 The number of samples needed from a binomial process to achieve an α level hypothesis test of $H_0 : p = p_0$ against $H_a : p < p_0$ with a power of $1-\beta$ for an alternative of p_a is

$$n = \left\lceil \frac{(z_\alpha\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p_a(1-p_a)})^2}{(p_0-p_a)^2} \right\rceil. \tag{2.78}$$

Note that the difference $p_0 - p_a$ is called the 'effect size.'

Using a similar process we can get the following result for a power calculation of a process mean.

Result 2.30 For a hypothesis test of $H_0 : \mu = \mu_0$ against $H_a : \mu < \mu_0$ with a significance level of α and a desired power of $1-\beta$ against an alternative of $\mu = \mu_a$, the number of samples needed from a *iid* process is

$$n = \left\lceil \sigma^2 \frac{(z_\alpha - z_\beta)^2}{(\mu_0-\mu_a)^2} \right\rceil. \tag{2.79}$$

Comment 2.12 Result 2.30 is based upon an assumption that the standard deviation (or variance) is the same regardless of whether the population mean is μ_0 or μ_a as well as a Gaussian approximation of the sampling distribution of the sample mean. The latter assumption can be alleviated by following an iterative procedure similar to the one outlined in Comment 2.8.

Comment 2.13 An admonition on sample size and power calculations is important in this context. Often in a real world data collection, the number of individuals that agree to start the data collection process is not the same number that will complete the process. There is a certain amount of attrition that may occur. This is especially true when testing involves human beings. Consequently, in planning an evaluation, it is worthwhile to take the statistically required sample sizes or power calculations and inflate them to adjust for this attrition.

2.3.5 Resampling Methods

Resampling methods are a class of statistical methods for approximating the sampling distribution of an estimator or a group of estimators. These methods have become increasingly utilized due to the ease of implementation, the advances in computation and the wide range of disciplines to which they can be applied. The advantage of these methods is that they make few assumptions about the shape of the sampling distribution. They are especially useful when sample sizes are small and, therefore, large sample methods are not appropriate. More details on, and examples of, resampling methods can be found in Manly [62], Edgington [28], Efron and Tibshirani [29] and Lahiri [54], for example. We start this section with a discussion of randomization tests, then move to jackknife and bootstrap methodology.

Definition 2.49 A *randomization test* is one where a test statistic is computed based upon the observed data and then the data is repeatedly permuted and the same test statistic is calculated for each permuted data set that results. This is useful, for example, when comparing a statistic on two different groups. A reference distribution is then calculated based upon the calculated test statistic from the permuted data. A *p-value* is then calculated by comparing the observed test statistic from the original data to the reference distribution of the test statistic based upon the permuted data. This definition follows from Edgington [28].

We note below that there is a distinction between randomization tests and permutation tests. The basics of each test are that we want to look at the distribution of a statistic had the observations been randomly reshuffled among the groups. This permuting of observations is appropriate if the null hypothesis holds. *Note that randomization tests, as presented here, are singularly for comparing two or more groups.* A randomization test is called a permutation test if all possible permutations of the observations are obtained. Thus, the randomization test *p-value* is an empirical estimate of the *p-value* that would be obtained under a permutation test. If the number of possible permutations is small, then it is often reasonable to do all possible combinations. However, the number of permutations grows exponentially with the number of observations, in which case it is sufficient to use a randomization test to obtain an approximate *p-value* assuming that the number of permutations obtained is reasonably large, say more than 1000.

Below we outline the jackknife and bootstrap methodology. Here we will provide a short definition. The goal of both methods is to provide approximations to the sampling distributions for an estimator or statistic, say $\hat{\phi}$. Here we outline the basic approaches. Later in the book as we use these techniques, we expand out the exact methods as appropriate.

Definition 2.50 A *jackknife procedure* is one where the sampling distribution for an estimator, say $\hat{\phi} = \hat{\phi}(V_1, \ldots, V_n)$, is estimated by recalculating $\hat{\phi}n$ times, each time leaving out one of the observed values. That is, the ith element of the sampling distribution for $\hat{\phi}$ is $\hat{\phi}(V_1, \ldots, V_{i-1}, V_{i+1}, \ldots, V_n)$ which is calculated based upon $n - 1$ observations.

The above is the definition for the traditional jackknife, sometimes known as the 'leave-one-out' jackknife, but there are other jackknife procedures that 'leave out' more than one observation. See Shao and Wu [90] or Friedl and Stampfer [34] for expositions of the properties of these other procedures.

Definition 2.51 A *bootstrap procedure* is one where the sampling distribution for an estimator, $\hat{\phi} = \hat{\phi}(V_1, \ldots, V_n)$ is estimated by repeatedly sampling n observations *with replacement* from the original sample of size n and calculating $\hat{\phi}$ upon each of these replicate sets. The resulting collection of estimates forms a distribution that approximates the sampling distribution of $\hat{\phi}$.

Result 2.31 We present here an algorithm for a $100 \times (1 - \alpha)\%$ bootstrap confidence interval for an *iid* process mean. Denote the mean of the process by μ. The steps are as follows:

1. Calculate
$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{2.80}$$
based upon the observed values of the sample $X_1, \ldots, X_n$.
2. Sample with replacement n integers from the list $\{1, 2, \ldots, n\}$ and call these values $b_1, b_2, \ldots, b_n$.
3. Calculate and store $e = \hat{\mu}_b - \hat{\mu}$ where
$$\hat{\mu}_b = \frac{1}{n} \sum_{i=1}^{n} X_{b_i}. \tag{2.81}$$
4. Repeat steps 2 and 3 above some large number of times, say M times.
5. Determine the $\alpha/2$th and the $1 - \alpha/2$th percentiles of e and represent those by e_L and e_U, respectively.
6. Then a $100 \times (1 - \alpha)\%$ confidence interval for μ can be made using
$$(\hat{\mu} - e_U, \hat{\mu} - e_L). \tag{2.82}$$

Comment 2.14 The basic bootstrap confidence interval approach we just delineated is one that we utilize in multiple ways throughout this text. The distribution of e is the distribution of differences from the statistic $\hat{\mu}$. The methodology given above differs slightly from the bootstrap confidence interval that is typically presented in biometric authentication. See, for example, Poh et al. [76] or Bolle et al. [9]. Here we made use of the so-called *Hall adjustment* which accounts for potential asymmetry in the sampling distribution of the statistic, Hall [44].

Comment 2.15 In Result 2.31, we presented a bootstrap confidence interval. In a similar manner, it is possible to create a hypothesis test for a particular process parameter. In order to ensure that the bootstrap approximation to the sampling distribution for a statistic is what would be expected when the null hypothesis is true, we have to adjust the sampling distribution so that it is centered at the value specified in the null hypothesis, for example, μ_0. To do this, we take $e = \hat{\mu}_b - \hat{\mu} + \mu_0$. In this way, we can approximate the distribution of the statistic, $\hat{\mu}$, assuming that the null hypothesis is true.

It is worthwhile mentioning that the methods above apply to a variety of data collection methodologies. That is, they can be applied to data that is not collected in a manner that would allow modeling using *iid* random variables. However, special care is needed in these circumstances to craft appropriate methodology. All of the methods in this section are quite powerful and have been underutilized in the performance evaluation of classification systems. Throughout this book, we will use these methods in a variety of circumstances.

There is an extensive amount of work that has been done on resampling methods and this continues to be an active area of research in statistics. An excellent overview of this work can be found in Manly [62]. Much of the recent interest in the statistics literature is on non-*iid* data. The interested (and *motivated*) reader is directed to the mathematically sophisticated work by Lahiri [54] for results in this area.

Springer

http://www.springer.com/978-1-84996-201-8

Computational Methods in Biometric Authentication
Statistical Methods for Performance Evaluation
Schuckers, M.E.
2010, XXV, 317 p., Hardcover
ISBN: 978-1-84996-201-8