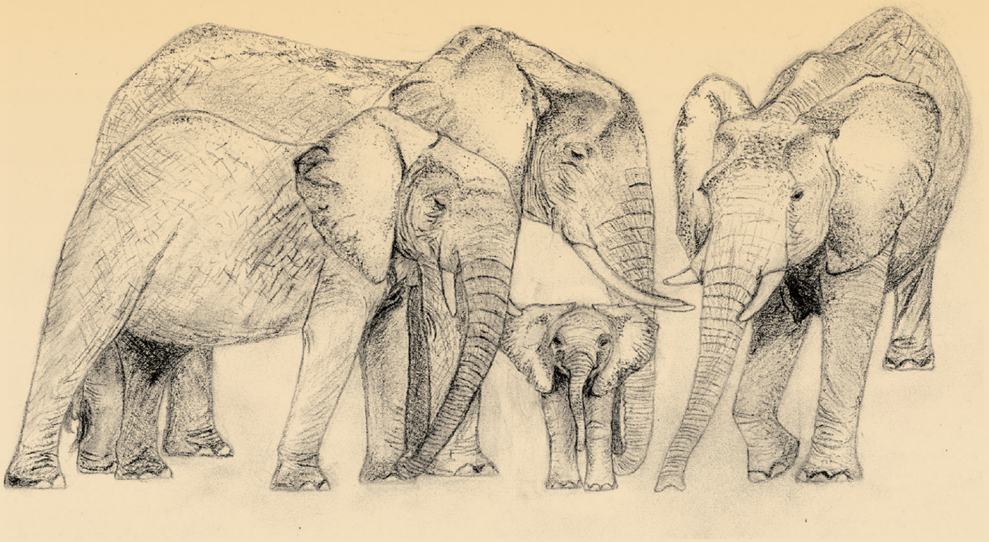
CHAPMAN & HALL/CRC
APPLIED ENVIRONMENTAL STATISTICS

STATISTICAL METHODS FOR FIELD AND LABORATORY STUDIES IN BEHAVIORAL ECOLOGY



Scott Pardo
Michael Pardo

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Statistical Methods for Field and Laboratory Studies in Behavioral Ecology

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Statistical Methods for Field and Laboratory Studies in Behavioral Ecology

Scott A. Pardo
Michael A. Pardo

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper

International Standard Book Number-13: 978-1-138-74336-6 (Hardback)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Preface

Behavioral ecology is a field that largely depends upon empirical investigation and observation, and as such leans heavily on statistical methods. Thus, behavioral ecologists require some instruction into those statistical concepts and methods that will be of use to their work. For example, in the September 2017 issue of the journal *Behavioral Ecology*, we counted more than 50 different statistical techniques. Although all the methods mentioned in that work could be found in various statistical texts, it would be difficult to find all of them in one place. This text was written with the behavioral ecologist in mind. Not only does it contain information on methods that have been widely used by behavioral ecologists, it also provides a little depth into the theory under which those methods were derived. Without getting overly mathematical, the theoretical aspects are described in order to elucidate the assumptions and limitations of the methods. In that way, the scientist will have a better view as to when these methods are applicable, and the appropriate level of skepticism required when interpreting results. Sometimes there may be more than one technique for analyzing the same data and providing the same type of conclusion. This text will also compare such methods, describe their assumptions, and hopefully provide some insight into which technique the researcher might choose. In particular, methods that require few assumptions about the underlying probability distributions of populations or data-generating processes will be described, together with associated computer programs. The computer programs provided are written in the R language, which has gained much popularity in the scientific world. Datasets provided are mostly based, at least to some degree, on real studies, but the data themselves are simulated, and the examples are simplified for pedagogical purposes. Those studies providing the inspiration for the simulated data are cited in the text.

It is assumed that the reader has had exposure to statistics through a first introductory course at least, and also has sufficient knowledge of R. This is not a primer for R or for statistics. However, some introductory material is included to aid the less initiated reader. The first five chapters largely consist of material covered in many first courses on statistics for biologists. However, there is mention of some intermediate notions, such as rank-based methods, permutation tests, and bootstrapping. In most chapters, at least two different methods are presented, together with their primary assumptions, for analyzing the exact same data. As such, this is not a book about parametric, nonparametric, frequentist, or Bayesian statistics. Rather, with no sword to grind, statistical methods are presented to the researcher in order to familiarize him or her with techniques described in scholarly literature.

Hopefully, the text will remove the perception of the magical aura that statistical methods often evoke.

The remaining chapters cover methods that each have multiple books written on them. As such, this can only be viewed as an introduction, and an introduction to some more fundamental but not elementary methods. Nevertheless, the material presented should at least get the reader started on the path.

Something should be said about the organization of material within a chapter. Except for Chapters 1, 15, and 16, each chapter is divided into five sections:

General Ideas
Examples with R Code
Theoretical Aspects
Key Points
Exercises and Questions

Hopefully, the first two sections, General Ideas and Examples with R Code, can get the reader started in the process of analyzing data. The Theoretical Aspects section will help provide some explanation of how the methods actually work, why they work, and what assumptions are necessary for them to work correctly. We strongly recommend that the student reads the Theoretical Aspects sections in order to gain a better understanding of the methods, their strengths, and their limitations.

As in the case of all texts, some very important topics have been omitted. In particular, the uses of statistical methods for phylogenetic analyses and spatial modeling have not been discussed. These, and other advanced methods, are beyond the scope of this book.

Acknowledgments

The authors would like to acknowledge and thank Yehudah A. and Jeremy D. Pardo for their unwitting contributions to this work. The authors would also like to thank Maria Modanu and Sarah Bluher for their review, explanations, and suggestions about decision and game theory. Scott would like to acknowledge and thank Dr. Rezi Zawadzki for her encouragement and suggestions. Finally, both authors, Scott and Michael, owe a great debt to their wife and mother, respectively, for first of all giving them the idea and suggesting they write this book, and for continually making suggestions during the writing. They especially owe her for conceiving of and suggesting the last chapter, which would not exist if she hadn't thought of it, and if she hadn't persisted in encouraging its writing.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

About the Authors

Scott Pardo has been a professional statistician for 37 years. He has worked in a diverse set of environments and applications, including the U.S. Army, satellite communications, cardiac pacemakers, pharmaceuticals, and blood glucose meters. He has a PhD in systems and industrial engineering from the University of Southern California, is a Six Sigma Master Blackbelt, and an accredited professional statistician, PStat®.

Michael Pardo is a PhD candidate in behavioral ecology at Cornell University, and has been conducting field-based research in animal behavior for over 10 years. He holds a BS in environmental biology from the State University of New York (SUNY) College of Environmental Science and Forestry. His primary research interests are in vocal communication and social cognition, particularly with mammals and birds. He has studied eastern gray squirrels, Asian elephants, and acorn woodpeckers.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

1

Statistical Foundations

Statistics has its foundation in probability. The basic building block is known as the random variable. Without being overly mathematical, random variables are those things that can be expressed in some sort of quantitative fashion, and whose values cannot be perfectly predicted. Random variables will take the form of observations or measurements made on experimental units. Experimental units are very often individual animals, but could be a collective, such as a flock, herd, hive, family, or other collection of individuals. The observations and measurements to be discussed in this text will be things that can be quantified. For example, a variable might have only two possible values: Say, one, if a particular, predefined behavior is observed under particular conditions; and zero if it is not. Another variable could be the distance traveled by an individual in some fixed period of time. The random nature of these variables implies that they have a probability distribution associated with their respective values. The analyses of data will be all about features of these distributions, such as means, standard deviations, and percentiles.

By way of a taxonomy for observations or measurements, we will refer to those whose values can be expressed as an integer as *discrete*, and those whose values can be expressed as a decimal number or fraction as *continuous*. Analyses for these types of variables are different in details, but have similar aims.

Statistical analyses involve three basic procedures:

1. Estimation
2. Inference and decision making
3. Model building: Discrimination and prediction

In all cases, statistics is the science of applying the laws and rules of probability to samples, which are collections of values of a random variable or in fact a collection of random variables. The type of sample upon which we will most heavily rely is called the *random sample*. A random sample can be defined as a subset of individual values of a random variable where the individuals selected for the subset all had an equal opportunity for selection. This does not mean that in any given data-gathering exercise there could not be more than one group or class of individuals, but that within a class the individuals chosen should not have been chosen with any particular bias.

The nature of all three types of procedures can be subdivided into two basic classes:

1. Parametric
2. Nonparametric

By parametric, we mean that there is some underlying "model" that describes the data-generating process (e.g., the normal, or Gaussian, distribution), and that model can be described by a few (usually one to three) numerical parameters. By nonparametric, we mean analyses that are not dependent on specifying a particular form of model for the data-generating process. Both paradigms for statistical analyses are useful and have a place in the data analyst's toolbox. As such, both classes of analyses will be discussed throughout the text.

Some Probability Concepts

Parametric distributions are described by mathematical functions. The fundamental function is called the *probability density function for continuous variables*, or in the case of discrete variables, it is often called the *probability mass function*. The idea is to describe the probability that the random variable, call it X, could take on a particular value, or have values falling within some specified range. In the case of continuous variables, the probability that X is exactly equal to a particular value is always zero. This rather curious fact is based on a set of mathematical ideas called *measure theory*. Intuitively, the probability of finding an individual with exactly some specific characteristic (say, a weight of 2.073192648 kg) is, well, zero. This is not to say that once you find such an individual, you must be hallucinating. The notion of zero probability (and in fact any probability) relates to a priori determination, that is, before any observation. Once an observation is made, the probability of observing whatever it is you observed is in fact 1, or 100 percent.

In general, capital letters, like X, will refer to a random variable, whereas lower case letters, like x, will refer to a specific value of the random variable, X. Often, in order to avoid confusing discrete and continuous variables, the symbol $f_X(x)$ will refer to the density function for variable X, evaluated at the value x, and $p_X(x_k)$ to a probability mass function for a discrete variable X evaluated at the value x_k. The notation $Pr\{\}$ will refer to the probability that whatever is inside the curly brackets will happen, or be observed. If the symbol *"dx"* means a very small range of values for X, and x_k represents a particular value of a discrete random variable, then

$$f_X(x)dx = Pr\{x - dx \leq X \leq x + dx\}$$

and

$$p_X(x_k) = \Pr\{X = x_k\}$$

There is a particularly important function called the *cumulative distribution function* (CDF) that is the probability $Pr\{X \le x\}$, which is usually defined in terms of density or mass functions, namely

$$F_X(x) = \int_{-\infty}^{x} f_X(\xi)\, d\xi$$

for continuous variables, and

$$F_X(x) = \sum_{x_k \le x} p_X(x_k)$$

for discrete variables.

As mentioned earlier, the functions $f_X(.)$ and $p_X(.)$ generally have parameters, or constants, that dictate something about the particular nature of the shape of the density curve. Table 1.1 shows the parameter lists, density or mass functions for several common distributions.

In the case of the binomial and beta distributions, the symbol p was used to denote a parameter (binomial), or as a value of a random variable (beta), and not the mass function itself. The function $\Gamma(x)$ is called the gamma function (oddly enough) and has a definition in terms of an integral:

$$\Gamma(x) = \int_{0}^{\infty} \xi^{x-1} e^{-x}\, d\xi$$

Aside from the CDF, there are some other important functions of $f_X(x)$ and $p_X(x_k)$. In particular, there is the expected value, or mean:

$$E[X] = \mu = \begin{cases} \sum_{k} x_k p_X(x_k) \\ \int_{-\infty}^{+\infty} \xi f_X(\xi)\, d\xi \end{cases}$$

and the variance:

TABLE 1.1

Some Probability Density and Mass Functions

Name	Parameters	Density or Mass Function	Range of Values
Normal	μ, σ	$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$	$-\infty < x < +\infty$
Gamma	n, λ	$\frac{\lambda^n}{\Gamma(n)}x^{n-1}\exp(-\lambda x)$	$x > 0$
Chi-Squared	ν	$\frac{(1/2)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}x^{\frac{\nu}{2}-1}\exp\left(-\frac{1}{2}x\right)$	$x > 0$
Student's t	ν	$\frac{\Gamma\left(\frac{1}{2}(\nu+1)\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{1}{2}\nu\right)}\left[1+\frac{x^2}{\nu}\right]^{-\frac{(\nu+1)}{2}}$	$-\infty < x < +\infty$
F	ν_1, ν_2	$\frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{1}{2}\nu_1\right)\Gamma\left(\frac{1}{2}\nu_2\right)}\frac{x^{\left(\frac{\nu}{2}\right)-1}}{(1+x)^{(\nu_1+\nu_2)/2}}$	$x > 0$
Poisson	λ	$\frac{\lambda^k e^{-\lambda}}{k!}$	$k = 0, 1, 2, \ldots$
Binomial	n, p	$\binom{n}{k}p^k(1-p)^{n-k}$	$k = 0, 1, 2, 3, \ldots n$
Beta	α, β	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}$	$0 < p \le 1$

$$V[X] = E[X-\mu^2] = \sigma^2 = \begin{cases} \sum_k (x_k-\mu)^2 p_X(x_k) \\ \int_{-\infty}^{+\infty} (\xi-\mu)^2 f_X(\xi)\,d\xi \end{cases}$$

Commonly the Greek letter μ is used to symbolize the expected value, and σ^2 is used to represent the variance. The variance is never negative (it is a sum of squared values). The square root of the variance is called the standard deviation, and has its most important role in random variables having a normal distribution. The expected value has units that are the same as

individual measurements or observations. The variance has squared units, so that the standard deviation has the same units as the measurements.

Often we must deal with more than one random variable simultaneously. The density or mass function of one variable might depend on the value of some other variable. Such dependency is referred to as *conditioning*. We symbolize the conditional density of X, given another variable, say Y, is equal to a particular value, say y, using the notation:

$$f_{X|Y}\left(x|Y=y\right)$$

Typically, the fact that $Y = y$ will affect the particular values of parameters. Also, we will usually drop the subscript $X|Y$, since the conditional nature of the density is made obvious by the "|" notation.

It is possible that the value of one random variable, say Y, has no effect on the probability distribution of another, X. It turns out that any two random variables have what is called a joint density function. The joint density of X and Y could be defined as

$$f_{XY}(x,y)dxdy = Pr\{x-dx \le X \le x+dx,\ AND\ y-dy \le Y \le y+dy\}$$

The joint density quantifies the probability that random variable X falls in a given range and at the same time random variable Y falls in some other given range.

It turns out that this joint density can be expressed in terms of conditional densities:

$$f_{XY}(x,y) = f_{X|Y}(x|y)f_Y(y)$$

The marginal density of one variable (say X) is the density of X without the effect of Y, and is computed as

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dy$$

When X and Y are independent of each other, then

$$f_{X|Y}\left(x|y\right) = f_X(x)$$

So that

$$f_{XY}(x,y) = f_X(x)f_Y(y)$$

In other words, when X and Y are independent, their joint density is the product of their marginal densities.

In addition to joint distributions, the expected values and variances of sums and differences of random variables find themselves in many applications. So, if X and Y are two random variables:

$$E[X \pm Y] = E[X] \pm E[Y]$$

If X and Y are independent, then

$$V[X \pm Y] = V[X] + V[Y]$$

While the sign of the operator (±) follows along with the expected values, the variance of the difference is the sum of the variances.

Another set of facts we will use relating to conditional densities or mass functions is based on something called Bayes' theorem. Briefly, Bayes' theorem states that if X is a random variable with density f, and Y is a random variable with density g, then

$$g\left(x|Y=y\right) = \frac{g(y)f\left(x|Y=y\right)}{\int_{-\infty}^{+\infty} f\left(x|Y=\xi\right)g(\xi)d\xi}$$

As long as Y is continuous, this particular formula holds even if X is discrete, and f is the mass function of X. If, however, Y is discrete, and g is its mass function, then the integral is replaced with a summation:

$$g\left(x|Y=y\right) = \frac{g(y)f\left(x|Y=y\right)}{\sum_k f\left(x|Y=\xi_k\right)g(\xi_k)}$$

It should be noted that it is possible for a random variable to not actually have a density function associated with it. However, that situation probably never exists in nature, so we will assume the density always exists.

Some Statistical Concepts

Earlier we mentioned that statistical problems could be classified into the categories:

1. Estimation
2. Inference and decision making
3. Model building: Discrimination and prediction

Estimation is the process of using data to guess at the value of parameters or some feature of a probability distribution assumed to be governing the data-generating process. Probably the most common is estimating the expected value of a distribution. The expected value of the random variable's distribution is

$$E[X] = \mu = \begin{cases} \sum_k x_k p_X(x_k) \\ \int_{-\infty}^{+\infty} \xi f_X(\xi) d\xi \end{cases}$$

One of the useful mathematical properties of expected value is that it is a linear operator, namely:

$$E[X_1 + X_2 + \cdots + X_n] = E[X_1] + E[X_2] + \cdots + E[X_n]$$

and

$$E[aX] = aE[X]$$

when a is a non-random constant. An estimate based on a sample of observations from the data-generating process is

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

We use the notation $\hat{\mu}$ instead of a perhaps more well-recognized symbol $\bar{x}$, to emphasize the fact that we are using the data to estimate the expected value. There are many such estimation formulae (called estimators), and many are used in different contexts for different reasons. The main point is that data can be used to estimate parameters or other features of probability distributions. The other point is that, since estimators use data, they themselves are random variables. Thus, if two researchers studying the same population of finches each make independent observations on either two sets (samples) of birds or even on the same sample, but at two different times, and each researcher calculates an average, the two averages most likely won't be exactly the same.

There are different methods used to derive estimator formulas for various parameters. Perhaps the best known is called the method of maximum likelihood. The idea is that if you have a random sample of measurements (X),

you can find values of parameters that maximize something called the likelihood function, which generally depends on assuming the form of the distribution for the data-generating process. Suppose that the values x_1, x_2, …, x_n represent n values sampled from a normally distributed data-generating process, with unknown expected value and variance the density function evaluated at x_i, say, would be given by

$$f(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

The likelihood function for the sample would be the product of all the valuations of the density function:

$$L(x_1, x_n, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$$

Of course, this likelihood function cannot be computed without knowing μ and σ. The idea of maximum likelihood is to find values $\hat{\mu}$ and $\hat{\sigma}$ that maximize L. Usually the log of the likelihood function is taken before attempting to solve the maximization. Maximizing the log of L is equivalent to maximizing L, since the log is a monotonic increasing function. The log of a product is the sum of the logs of the factors:

$$\log L = \sum_{i=1}^{n} \log\left(f(x_i)\right)$$

Maximizing the sum is easier mathematically than maximizing the product.

What is important to note is that first we had to pick a parametric form for the density function of the random variable from which we were sampling, the parameter values are unknown, and our guess for the parameter values is based on a criterion that gives us the best guess. It turns out that for the normal model, the maximum likelihood estimators for μ and σ^2 are

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\mu}\right)^2$$

Some may notice that the maximum likelihood estimator for σ^2 differs from the formula used in most elementary texts, in that it divides by n and not $n - 1$. Dividing the sum by $n - 1$ to estimate σ^2 gives the formula a property known as *unbiasedness*. While this is important, in the case of this estimator the effects are fairly small. Another estimation method is called *least squares*. Rather than maximize a likelihood function, least squares chooses estimators that minimize an "error" function. A common context for least squares estimation is linear regression. More will be said about least squares. For now, just recognize it as a method for estimating parameters.

Statistical estimates, since they are based on a finite sample of observations or measurements made on individuals taken from some population or data-generating process, have themselves a random variation component. Inasmuch as a statistical estimate is attempting to help make a guess about a parameter, it would be good to know that the formula used to compute the estimate has a reasonable chance of getting close to the actual value of the parameter. One such property has already been described, namely, maximum likelihood. Another property that is desirable is unbiasedness, which was also mentioned earlier. An estimation formula is said to be unbiased if its expected value is equal to the parameter to be estimated. For example, assuming a random sample, $x_1, x_2, \ldots, x_n$, then the expected value of each x_i is the population mean, μ, and

$$E[\hat{\mu}] = \frac{1}{n}\sum_{i=1}^{n}E[x_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

Thus, our arithmetic mean estimator for μ is in fact unbiased. Conversely, the maximum likelihood estimate of σ^2 is not unbiased (or, in other words, biased). It turns out that

$$E[\hat{\sigma}^2] = \frac{1}{n}\sum_{i=1}^{n}E\left[(x_i - \hat{\mu})^2\right] = \frac{n-1}{n}\sigma^2$$

Thus, the maximum likelihood estimator of σ^2 slightly underestimates the variance. The point of the discussion about unbiasedness is that estimation formulae are themselves random variables, and as such we will need to consider their probabilistic characteristics.

Inference is about making a priori guesses about parameter values and then using data to decide if you were correct. Suppose, for example, you guessed that the average duration of a courtship display was 30 seconds. How would you decide whether to believe your guess, or not? First you would gather data, by timing courtship displays of several individuals, say n. Then you would probably compute the maximum likelihood estimates of mean and variance. Suppose the estimate of the mean was 31.5 seconds, and the standard deviation (square root of variance) estimate was three seconds. OK, so it wasn't 30. Were you wrong? The question becomes one of how much variation there might be if the experiment were repeated. The idea of statistical inference is to make a decision about what to believe, and not what actually is the truth. Our decision has risk associated with it, namely the risk (or probability) of saying our guess is wrong when in fact it is correct, and the risk of saying our guess is correct when in fact it is not. There is a formalism for expressing the notions of inference. There are two competing hypotheses, or guesses, about the parameter or parameters of interest. One is called the "null" hypothesis, symbolized as H_0. The logical negation of the null hypothesis is called, not surprisingly, the alternate hypothesis, and is often symbolized as H_1. So, in the example of the courting display question, we might have

$$H_0: \mu = 30$$
$$H_1: \mu \neq 30$$

The error of deciding that H_0 is false when in fact it was true is called a Type I error. The error of believing H_0 is true when it is not is called a Type II error. The next thing required is a rule, based on data, that lets the decisionmaker decide whether to believe H_0 or H_1. Since data are subject to variation, the rule is necessarily probabilistic. It turns out that, conveniently, the calculation

$$t = \frac{\hat{\mu} - 30}{\hat{\sigma}/\sqrt{n}}$$

has a known probability distribution, the familiar Student's t, provided that the null hypothesis is actually correct (i.e., that $\mu = 30$). This formula is known as a *test statistic*, because it is the quantity we will use to decide whether to believe (accept) the null hypothesis, or disbelieve it (reject). In fact, a common feature of all inference is determining the distribution of the test statistic if H_0 were actually true. The probability of making a Type I error is symbolized with the letter α. The probability of a Type II error is traditionally symbolized with the letter β. We can find a range of values that t would fall in between with probability $1 - \alpha$, given that H_0 is true, even before we

gathered any data. In fact, the range of possible values only depends on the sample size, n, and the desired probability content of the range. If, for example, the sample size was $n = 10$, and we wanted the probability content to be $100(1 - \alpha)\% = 95$ percent, then the range of values for t we would expect if the null hypothesis was correct would be approximately ±2.228. The range ($t \leq -2.228$, $t \geq +2.228$) is called the *critical region* of size α. If the value of the test statistic falls in the critical region, we say that the test statistics is significant, and we REJECT the null hypothesis is favor of the alternative. The particular region for this example is partially based on the presumption that we computed the maximum likelihood estimate for standard deviation. If after getting data we computed the value of t using the formula above, and its value fell within the range ±2.228, we would continue to believe the null hypothesis, because there is a fairly "high" (95 percent) chance of t falling inside this range if H_0 is correct. Conversely, there is a relatively "low" chance that t would fall outside the "critical" range if H_0 was correct. Unfortunately, we cannot make the same statement about the alternative hypothesis, H_1, since there are an infinite number of possible values (anything other than 30) that would make it correct. Thus, it is easier to fix the chance of making the mistake of deciding that H_0 is false when in fact it is true. Once this risk is decided upon, the decision rule for either believing H_0 or not believing it is fairly easy to compute, provided we know something about the distribution of the test statistic, given that the null hypothesis is true.

Another way of determining a rule for rejecting or accepting the null hypothesis is to *compute a probability* of observing the data you got IF the null hypothesis was actually correct. This probability is usually referred to as a *p-value*. Thus, in our example, if in fact $\mu = 30$, then the test statistic

$$t = \frac{\hat{\mu} - 30}{\hat{\sigma}/\sqrt{n}}$$

has a Student's t distribution with degrees-of-freedom parameter equal to n (since we used the maximum likelihood estimate of σ). Suppose we had data that yielded a sample estimate of μ, say,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = 31.5$$

and an estimate of σ^2:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2 = 9$$

If $n = 25$, then the sample test statistic would be

$$t = \frac{31.5 - 30}{3/\sqrt{25}} \approx 2.50$$

Since the alternative hypothesis is $\mu \neq 30$, we compute the probability that the test statistic would be outside the range (–2.50, +2.50). To compute this, we can use the R function *pt*():

$$\text{pt}(\text{q} = -2.5,\ \text{df} = 25,\ \text{lower.trail} = \text{TRUE}) \approx 0.01934$$

and

$$\text{pt}(\text{q} = +2.5,\ \text{df} = 25,\ \text{lower.trail} = \text{FALSE}) \approx 0.01934$$

The "two-sided" *p*-value is 0.01934 + 0.01934 = 0.03868.

Since Student's *t* distribution is symmetric about zero, the probability for the "lower tail" of $-t$ is equal to the probability for the "upper tail" of $+t$.

If our threshold of *p*-values is $\alpha = 0.05$, then since 0.03868 is less than 0.05, we will no longer believe that the null hypothesis is correct, and reject it. With a sample size of $n = 25$, and $1 - \alpha = 0.95$, then the critical region is $t \leq -2.06$, $t \geq +2.06$. Since $t = 2.50 > +2.06$, we would reject the null hypothesis. Regardless of whether you determine a critical region of size α, or choose α to be a threshold for *p*-values, the conclusions would be identical.

Another methodology that is somewhere between estimation and inference is called confidence interval building. The confidence interval again employs that risk level, α, but in a slightly different manner. Suppose we wanted to know that value of the parameters that would correspond to the limits of the critical range for the test statistic. Using the previous example, let

$$t_{low} = -2.06 = \frac{\hat{\mu} - \mu_{low}}{\hat{\sigma}/\sqrt{n}}$$

and

$$t_{high} = -2.06 = \frac{\hat{\mu} - \mu_{high}}{\hat{\sigma}/\sqrt{n}}$$

Solving for μ_{low} and μ_{high} gives:

$$\mu_{low} = \hat{\mu} - 2.06\frac{\hat{\sigma}}{\sqrt{n}}$$

and

$$\mu_{high} = \hat{\mu} + 2.06\frac{\hat{\sigma}}{\sqrt{n}}$$

The range of values (μ_{low}, μ_{high}) is called the 100(1 – α)% "confidence interval" for parameter μ. It can be thought of as a feasible range for the unknown values of μ. That is, we are not certain about the actual value of μ, but we are nearly certain (100(1 – α)% certain) that it lies somewhere in the interval (μ_{low}, μ_{high}). So, in our example with $\hat{\mu} = 31.5$, $\hat{\sigma} = 3$, and $n = 25$, the 95-percent confidence interval would be

$$\mu_{low} = 31.5 - 2.06\frac{3}{\sqrt{25}} \approx 30.26$$

$$\mu_{high} = 31.5 + 2.06\frac{3}{\sqrt{25}} \approx 32.74$$

Since the hypothetical value for μ, namely 30, is not contained in the confidence interval (30.26, 32.74), we do not believe that 30 is a feasible value for μ.

When the null hypothesis is rejected, we say that the difference between our estimate of the parameter and the null value is *statistically significant at the 100α% level*. Another way of stating the same thing is that if we reject the null hypothesis, we would believe that the results of our analyses are repeatable.

Model building is a special application of estimation, but it usually has some inference associated with it. The idea is to postulate some mathematical relationship between some variables, some random and some without any random component. Then we estimate the values of model parameters. Finally, we test to see if we should believe that the form of the model we postulated was reasonable. Models can be predictive or discriminatory/classificatory. A simple example of a predictive model would be a simple linear regression. Suppose there is a continuously valued random variable, *Y*, and another continuously valued nonrandom variable, *X*. *Y* could be things such as response time, elapsed time, distance traveled, or other random variables that can be expressed as a decimal number. In this simple case, we

are assuming the X variable is not random. In other words, X is something whose value would have no random variation, and whose value is known perfectly without error. Y is referred to as the response variable, and X is the predictor or regressor variable. The general form of the linear model is

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

The coefficients β_0 and β_1 are unknown, and need to be estimated. The variable ε represents random "noise," indicating that the value of Y is on the average a linear function of X, but the actual observed values may have some perturbations, or noise, or sometimes-called errors associated with them.

Once the values of the parameters are estimated, a predicted value of Y can be computed for a given value of X. We would not usually consider X to have random noise associated with it. That is, when we get a value for X, we are (mostly) certain that the value would not vary if we measured or observed it a second time under exactly the same conditions. Rather, we suppose that given the value of X, we can predict on the average what Y would be, with the understanding that Y might vary from this average.

Another closely related type of model is also linear, but is classificatory or discriminatory. The X variables are not continuous, but are discrete categories. The goal is to determine if particular groupings of individuals actually discriminate between individuals. In other words, we want to know if individuals in different groups actually differ from each other with respect to Y. Perhaps the simplest example is the one-way analysis of variance (ANOVA). In this case, the single X variable is a set of discrete categories, and Y is the continuous random variable response. The question is not to find a prediction of Y for a given value of X, per se. Rather, the question is to estimate the difference in the average Y between the different categories. In the case of ANOVA, often the inferential part of modeling is of greater interest, namely, whether the difference in average values of Y between the different groups of X categories is in fact repeatable.

There are certainly more types of both predictive and classificatory modeling. The key notion here is that data can be used to create these sorts of models, through a combination of estimation and inference.

This is the classical parametric methodology for statistical inference. There is another set of methods, sometimes called nonparametric or distribution-free, of which neither term is strictly true. The idea is that the distribution of test statistics should not depend on the distribution of the data-generating process. The basic idea is still the same; you formulate a test statistic, you determine the "critical range" or "critical value" based on α, you get some data, and then you compute the test statistic to decide if you should accept or reject the null hypothesis.

A special set of nonparametric techniques is sometimes referred to as *resampling methods*. This book will in fact emphasize resampling methods where appropriate. The resampling techniques will generally fall into the bootstrap estimation process or the permutation hypothesis testing process. Both of these methods are computer-based, but given modern computing software such as R, they are fairly easy to perform.

Bayesian statistics is an alternate view of parameters, not as particular values to estimate or about which to make a guess about their true values, but treating them as if they themselves are random variables. Like the classic "frequentist" approach, Bayesian methods employ a likelihood function. However, these methods incorporate prior information about the parameters of interest. "Prior" to making observations, the analyst posits a distribution of the parameters of interest. The "prior" distribution expresses the knowledge about the parameter prior to performing the "next" experiment. So, for example, perhaps the mean response time to a stimulus is guessed to be most likely 10 seconds, but it could be as fast as 5 seconds and as delayed as 15 seconds. Rather than simply hypothesizing that the mean is exactly 10 seconds, the Bayesian method is to postulate a distribution that expresses the current level of knowledge and uncertainty in the parameter. Then, once data are gathered, Bayes' theorem is used to combine the prior distribution with the likelihood function, to update the prior knowledge. The updated distribution for the parameter is called the posterior distribution. So, if $f_{old}(\tilde{\mu})$ represents the prior density function for the parameter $\tilde{\mu}$, and $L[x_1, x_2, \ldots, x_n|\tilde{\mu}]$ the likelihood function for the sample, given a particular value of $\tilde{\mu}$, then the updated density function (called the posterior density) is

$$f_{new}(\tilde{\mu}|x_1, x_2, \ldots, x_n) = \frac{f_{old}(\tilde{\mu})L[x_1, x_2, \ldots, x_n|\tilde{\mu}]}{\int_{-\infty}^{+\infty} f_{old}(\xi)L[x_1, x_2, \ldots, x_n|\xi]d\xi}$$

Key Points for Chapter 1

- The primary concept for probability in our context is the random variable; it is generally defined in terms of measurements or observations made, where the values of those measurements or observations cannot be deduced exactly before they are made.
- Random variables come in two flavors: discrete and continuous.
- While the actual value of a random variable cannot be known a priori, statements can be made about the probability that a random