2 Probability

2.1 Uncertainty in observational science

Probability theory models uncertainty. Observational scientists often come across events whose outcome is uncertain. It may be physically impossible, too expensive or even counterproductive to observe all the inputs. The astronomer might want to measure the location and motions of all stars in a globular cluster to understand its dynamical state. But even with the best telescopes, only a fraction of the stars can be located in the two dimensions of sky coordinates with the third distance dimension unobtainable. Only one component (the radial velocity) of the three-dimensional velocity vector can be measured, and this may be accessible for only a few cluster members. Furthermore, limitations of the spectrograph and observing conditions lead to uncertainty in the measured radial velocities. Thus, our knowledge of the structure and dynamics of globular clusters is subject to considerable restrictions and uncertainty.

In developing the basic principles of uncertainty, we will consider both astronomical systems and simple familiar systems such as a tossed coin. The outcome of a toss, heads or tails, is completely determined by the forces on the coin and Newton's laws of motion. But we would need to measure too many parameters of the coin's trajectory and rotations to predict with acceptable reliability which face of the coin will be up. The outcomes of coin tosses are thus considered to be uncertain even though they are regulated by deterministic physical processes. Similarly, the observed properties of a quasar have considerable uncertainty, even though the physics of accretion disks and their radiation are based on deterministic physical processes.

The uncertainty in our knowledge could be due to the current level of understanding of the phenomenon, and might be reduced in the future. Consider, for example, the prediction of solar eclipses. In ancient societies, the motions of Solar System bodies were not understood and the occurrence of a solar eclipse would have been modeled as a random event (or attributed to divine intervention). However, an astronomer noticing that solar eclipses occur only on a new moon day could have revised the model with a monthly cycle of probabilities. Further quantitative prediction would follow from the Babylonian astronomers' discovery of the 18-year saros eclipse cycle. Finally, with Newtonian celestial mechanics, the phenomenon became essentially completely understood and the model changed from a random to a deterministic model subject to direct prediction with known accuracy.

The uncertainty of our knowledge could be due to future choices or events. We cannot predict with certainty the outcome of an election yet to be held, although polls of the voting public will constrain the prediction. We cannot accurately predict the radial velocity of a globular star prior to its measurement, although our prior knowledge of the cluster's velocity dispersion will constrain the prediction. But when the election results are tabulated, or the astronomical spectrum is analyzed, our level of uncertainty is suddenly reduced.

When the outcome of a situation is uncertain, why do we think that it is possible to model it mathematically? In many physical situations, the events that are uncertain at the micro-level appear to be deterministic at the macro-level. While the outcome of a single toss of a coin is uncertain, the proportion of heads in a large number of tosses is stable. While the radial velocity of a single globular cluster star is uncertain, we can make predictions with some confidence based on a prior measurement of the global cluster velocity and our knowledge of cluster dynamics from previous studies. Probability theory attempts to capture and quantify this phenomenon; the Law of Large Numbers directly addresses the relationship between micro-level uncertainty and macro-level deterministic behavior.

2.2 Outcome spaces and events

An **experiment** is any action that can have a set of possible results where the actually occurring result cannot be predicted with certainty prior to the action. Experiments such as tossing a coin, rolling a die, or counting of photons registered at a telescope, all result in sets of outcomes. Tossing a coin results in a set Ω of two outcomes $\Omega = \{H, T\}$; rolling a die results in a set of six outcomes $\Omega = \{1, 2, 3, 4, 5, 6\}$; while counting photons results in an infinite set of outcomes $\Omega = \{0, 1, 2, \dots\}$. The number of neutron stars within 1 kpc of the Sun is a discrete and finite sample space. The set of all outcomes Ω of an experiment is known as the **outcome space** or sample space.

An **event** is a subset of a sample space. For example, consider now the sample space Ω of all exoplanets, where the event E describes all exoplanets with eccentricity in the range 0.5–0.6, and the event F describes that the host star is a binary system. There are essentially two aspects to probability theory: first, assigning probabilities to simple outcomes; and second, manipulating probabilities or simple events to derive probabilities of complicated events.

In the simplest cases, such as a well-balanced coin toss or die roll, the inherent symmetries of the experiment lead to equally likely outcomes. For the coin toss, $\Omega = \{H, T\}$ with probabilities $P(H) = 0.5$ and $P(T) = 0.5$. For the die roll, $\Omega = \{1, 2, 3, 4, 5, 6\}$ with $P(i) = \frac{1}{6}$ for $i = 1, 2, \dots, 6$. Now consider the more complicated case where a quarter, a dime and a nickel are tossed together. The outcome space is

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}, \tag{2.1}$$

where the first letter is the outcome of the quarter, the second of the dime and the third of the nickel. Again, it is reasonable to model all the outcomes as equally likely with probabilities

$\frac{1}{8}$. Thus, when an experiment results in m equally likely outcomes, $\{e_1, e_2, \ldots, e_m\}$, then the probability of any event A is simply

$$P(A) = \frac{\#A}{m}, \tag{2.2}$$

where # is read "the number of". That is, $P(A)$ is the ratio of the number of outcomes favorable to A and the total number of outcomes.

Even when the outcomes are not equally likely, in some cases it is possible to identify the outcomes as combinations of equally likely outcomes of another experiment and thus obtain a model for the probabilities. Consider the three-coin toss where we only note the number of heads. The sample space is $\Omega = \{0, 1, 2, 3\}$. These outcomes cannot be modeled as equally likely. In fact, if we toss three coins 100 times, then we would observe that $\{1, 2\}$ occur far more frequently than $\{0, 3\}$. The following simple argument will lead to a logical assignment of probabilities. The outcome $\omega \in \Omega$ in this experiment is related to the outcomes in (2.1):

$\omega = 0$ when TTT occurs
$\omega = 1$ when HTT, THT or TTH occurs
$\omega = 2$ when HHT, HTH or THH occurs
$\omega = 3$ when HHH occurs.

Thus $P(0) = P(3) = 0.125$ and $P(1) = P(2) = 0.375$.

For finite (or countably infinite) sample spaces $\Omega = \{e_1, e_2, \ldots\}$, a probability model assigns a nonnegative weight p_i to the outcome e_i for every i in such a way that the p_i's add up to 1. A finite (or countably infinite) sample space is sometimes called a **discrete sample space**. For example, when exploring the number of exoplanets orbiting stars within 10 pc of the Sun, we consider a discrete sample space. In the case of countable sample spaces, we define the probability $P(A)$ of an event A as

$$P(A) = \sum_{i\,:\,e_i \in A} p_i. \tag{2.3}$$

In words, this says that the probability of an event A is equal to the sum of the individual probabilities of outcomes e_i belonging to A.

If the sample space Ω is uncountable, then not all subsets are allowed to be called events for mathematical and technical reasons. Astronomers deal with both countable spaces – such as the number of stars in the Galaxy, or the set of photons from a quasar arriving at a detector – and uncountable spaces – such as the variability characteristics of a quasar, or the background noise in an image constructed from interferometry observations.

2.3 Axioms of probability

A **probability space** consists of the triplet $(\Omega, \mathcal{F}, P)$, with sample space Ω, a class $\mathcal{F}$ of events, and a function P that assigns a probability to each event in $\mathcal{F}$ that obey three axioms of probability:

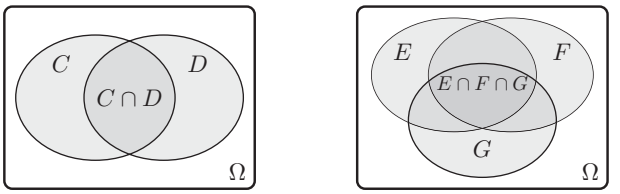

Fig. 2.1 Union and intersection of events.

Axiom 1 $0 \leq P(A) \leq 1$, for all events A
Axiom 2 $P(\Omega) = 1$
Axiom 3 For mutually exclusive (pairwise disjoint) events $A_1, A_2, \ldots,$

$$P(A_1 \cup A_2 \cup A_3 \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots,$$

that is, if for all $i \neq j$, $A_i \cap A_j = \emptyset$ ($\emptyset$ denotes the empty set or null event), then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Here, $\cup$ represents the union of sets while $\cap$ represents their intersection. Axiom 3 states that the probability that at least one of the mutually exclusive events A_i occurs is the same as the sum of the probabilities of the events A_i, and this should hold for infinitely many events. This is known as the **countable additivity** property. This axiom, in particular, implies that the finite additivity property holds; that is, for mutually exclusive (or disjoint) events A, B (i.e. $A \cap B = \emptyset$),

$$P(A \cup B) = P(A) + P(B). \tag{2.4}$$

This in particular implies that for any event A, the probability of its complement $A^c = \{\omega \in \Omega : \omega \notin A\}$, the set of points in the sample space that are not in A, is given by

$$P(A^c) = 1 - P(A). \tag{2.5}$$

(A technical comment can be made here: in the case of an uncountable sample space Ω, it is impossible to define a probability function P that assigns zero weight to singleton sets and satisfying these axioms for all subsets of Ω.)

Using the above axioms, it is easy to establish that for any two events $C,\ D$

$$P(C \cup D) = P(C) + P(D) - P(C \cap D); \tag{2.6}$$

that is, the probability of the union of the two events is equal to the sum of the event probabilities minus the probability of the intersection of the two events. This is illustrated in the left-hand panel of Figure 2.1.

For three events $E,\ F,\ G$,

$$\begin{aligned} P(E \cup F \cup G) \ = \ & P(E) + P(F) + P(G) - P(E \cap F) - P(F \cap G) \\ & - P(E \cap G) + P(E \cap F \cap G) \end{aligned} \tag{2.7}$$

as shown in the right-hand panel of Figure 2.1. The generalization to n events, $E_1, \ldots, E_n$ is called the **inclusion–exclusion formula**;

$$P(E_1 \cup E_2 \cup \cdots \cup E_n) = \sum_{i=1}^{\infty} P(E_i) - \sum_{i_1<i_2} P(E_{i_1} \cap E_{i_2}) + \cdots \\ \times (-1)^{r+1} \sum_{i_1<i_2<\cdots<i_r} P(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_r}) \\ + \cdots + (-1)^{n+1} P(E_1 \cap E_2 \cap \cdots \cap E_n), \tag{2.8}$$

where the summation

$$\sum_{i_1<i_2<\cdots<i_r} P(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_r}) \tag{2.9}$$

is taken over all of the $\binom{n}{r}$ possible subsets of size r of the set $\{1, 2, \ldots, n\}$.

2.4 Conditional probabilities

Conditional probability is one of the most important concepts in probability theory and can be tricky to understand. It often helps in computing desired probabilities, particularly when only partial information regarding a result of an experiment is available. Bayes' theorem at the foundation of Bayesian statistics uses conditional probabilities.

Consider the following simple example. When a die is rolled, the probability that it turns up one of the numbers $\{1, 2, 3\}$ is $1/2$, as each of the six outcomes is equally likely. Now consider that someone took a brief glimpse at the die and found that it turned up an even number. How does this additional information influence the assignment of probability to $A = \{1, 2, 3\}$? In this case, the weights assigned to the points are reassessed by giving equal weights, $1/3$, to each of the three even integers in $B = \{2, 4, 6\}$ and zero weights to the odd integers. Since it is already known that B occurred, it is intuitive to assign the scale and probability 1 to B and probability 0 to the complementary event B^c. Now as all the points in B are equally likely, it follows that the required probability is the ratio of the number of points of A that are in B to the total number of points in B. Since 2 is the only number from A in B, the required probability is $\#(A \cap B)/\#B = 1/3$.

Generalizing this, let us consider an experiment with m equally likely outcomes and let A and B be two events. If we are given the information that B has happened, what is the probability that A has happened in light of the new knowledge? Let $\#A = k$, $\#B = n$ and $\#(A \cap B) = i$. Then, as in the rolled die example above, given that B has happened, the new probability allocation assigns probability $1/n$ to all the outcomes in B. Out of these n, $\#(A \cap B) = i$ outcomes belong to A. Noting that $P(A \cap B) = i/m$ and $P(B) = n/m$, it leads to the conditional probability, $P(A \mid B) = i/n$, of A given B,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{2.10}$$

Equation (2.10) can be considered a formal definition of conditional probabilities providing $P(B) > 0$, even in the more general case where outcomes may not be equally likely. As a consequence, the multiplicative rule of probability for two events,

$$P(A \cap B) = P(A \mid B)P(B) \tag{2.11}$$

holds. The **multiplication rule** easily extends to n events:

$$\begin{aligned} P(A_1 \cap A_2 \cap \ldots A_n) = P(A_1)\, P(A_2 \mid A_1)\ \ldots\ P(A_{n-1} \mid A_1, \ldots A_{n-2}) \\ \times\ P(A_n \mid A_1, \ldots A_{n-1}). \end{aligned} \tag{2.12}$$

These concepts are very relevant to observational sciences such as astronomy. Except for the rare circumstance when an entirely new phenomenon is discovered, astronomers are measuring properties of celestial bodies or populations for which some distinctive properties are already available. Consider, for example, a subpopulation of galaxies found to exhibit Seyfert-like spectra in the optical band (property A) that have already been examined for nonthermal lobes in the radio band (property B). Then the conditional probability that a galaxy has a Seyfert nucleus given that it also has radio lobes is given by Equation (2.10), and this probability can be estimated from careful study of galaxy samples. The composition of a Solar System minor body can be predominately ices or rock. Icy bodies are more common at large orbital distances and show spectral signatures of water (or other) ice rather than the spectral signatures of silicates. The probability that a given asteroid, comet or Kuiper Belt Object is mostly icy is then conditioned on its semi-major axis and spectral characteristics.

2.4.1 Bayes' theorem

We are now ready to derive the famous Bayes' theorem, also known as Bayes' formula or Bayes' rule. It is named for the mid-eighteenth-century British mathematician and Presbyterian minister Thomas Bayes, although it was recognized earlier by James Bernoulli and Adrian de Moivre, and was later fully explicated by Pierre Simon Laplace. Let $B_1, \ldots, B_k$ be a partition of the sample space Ω. A partition of Ω is a collection of mutually exclusive (pairwise disjoint) sets whose union is Ω; that is, $B_i \cap B_j = 0$ for $i \neq j$. If A is any event in Ω, then to compute $P(A)$, one can use probabilities of pieces of A on each of the sets B_i and add them together to obtain

$$P(A) = P(A \mid B_1)P(B_1) + \cdots + P(A \mid B_k)P(B_k). \tag{2.13}$$

This is called the **law of total probability** and follows from the observation

$$P(A) = P(A \cap B_1) + \cdots + P(A \cap B_k), \tag{2.14}$$

and the multiplicative rule of probability, $P(A \cap B_i) = P(A \mid B_i)P(B_i)$.

Now consider the following example, a bit more complicated than those treated above. Suppose a box contains five quarters, of which one is a trick coin that has heads on both sides. A coin is picked at random and tossed three times. It was observed that all three tosses turned up heads.

If the type of the coin chosen is known, then one can easily compute the probability of the event H that all three tosses yield heads. If it is the two-headed coin, then the probability is 1, otherwise it is $1/8$. That is, $P(H \mid M) = 1$ and $P(H \mid M^c) = 1/8$, where M denotes the event that the two-headed coin is chosen and M^c is the complimentary event that a regular quarter is chosen. After observing three heads, what is the probability that the chosen coin has both sides heads? Bayes' theorem helps to answer this question. Here, by using the law of total probability and the multiplication rule, one obtains,

$$\begin{aligned} P(M \mid H) &= \frac{P(M \cap H)}{P(H)} = \frac{P(H \mid M)P(M)}{P(H \mid M)P(M) + P(H \mid M^c)P(M^c)} \\ &= \frac{1/5}{(1/5) + (1/8) \times (4/5)} = \frac{2}{3}. \end{aligned} \tag{2.15}$$

For a partition $B_1, \ldots, B_k$ of Ω, Bayes' theorem generalizes the above expression to obtain $P(B_i \mid A)$ in terms of $P(A \mid B_j)$ and $P(B_j)$, for $j = 1, \ldots, k$. The result is very easy to prove, and is the basis of Bayesian inference discussed in Section 3.8.

Theorem 2.1 (Bayes' theorem) *If $B_1, B_2, \ldots, B_k$ is a partition of the sample space, then for $i = 1, \ldots, k$,*

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A \mid B_1)P(B_1) + \cdots + P(A \mid B_k)P(B_k)}. \tag{2.16}$$

Bayes' theorem thus arises directly from logical inference based on the three axioms of probability. While it applies to any form of probabilities and events, modern Bayesian statistics adopts a particular interpretation of these probabilities, which we will present in Section 3.8.

2.4.2 Independent events

The examples above show that, for any two events A and B, the conditional probability of A given B, $P(A \mid B)$, is not necessarily equal to the unconditional probability of A, $P(A)$. Knowledge of B generally changes the probability of A. In the special situation where $P(A \mid B) = P(A)$ where the knowledge that B has occurred has not altered the probability of A, A and B are said to be **independent events**. As the conditional probability $P(A \mid B)$ is not defined when $P(B) = 0$, the multiplication rule $P(A \cap B) = P(A \mid B)P(B)$ will be used to formally define independence:

Definition 2.2 Two events A and B are defined to be independent if

$$P(A \cap B) = P(A)P(B).$$

This shows that if A is independent of B, then B is independent of A. It is not difficult to show that if A and B are independent, then A and B^c are independent, A^c and B are independent and also A^c and B^c are independent.

Note that three events E, F, G satisfying $P(E \cap F \cap G) = P(E)P(F)P(G)$ cannot be called independent, as it does not guarantee independence of E, F or independence of F, G or independence of E, G. This can be illustrated with a simple example of a sample

space $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8\}$, where all the points are equally likely. Consider the events $E = \{1, 2, 3, 4\}$, $F = G = \{4, 5, 6, 7\}$. Clearly $P(E \cap F \cap G) = P(E)P(F)P(G)$. But neither E and F, F and G, nor E and G are independent.

Similarly, we note that independence of A and B, B and C, and A and C together does not imply $P(A \cap B \cap C) = P(A)P(B)P(C)$. If we consider the events $A = \{1, 2, 3, 4\}$, $B = \{1, 2, 5, 6\}$, $C = \{1, 2, 7, 8\}$ then clearly A and B are independent, B and C are independent, and also A and C are independent, as A, B, C each contain exactly four numbers,

$$P(A) = P(B) = P(C) = \frac{4}{8} = \frac{1}{2}, \tag{2.17}$$

but $A \cap B = B \cap C = A \cap C = A \cap B \cap C = \{1, 2\}$ and

$$P(A \cap B) = P(B \cap C) = P(A \cap C) = P(\{1, 2\}) = \frac{2}{8} = \frac{1}{4}. \tag{2.18}$$

However,

$$P(A \cap B \cap C) = P(\{1, 2\}) = \frac{1}{4} \neq \frac{1}{8} = P(A)P(B)P(C). \tag{2.19}$$

Though A and B are independent, and A and C are independent, $P(A \mid B \cap C) = 1$. So A is not independent of $B \cap C$. This leads to the following definition:

Definition 2.3 (Independent events) A set of $A_1, \ldots, A_n$ events is said to be independent if, for every subcollection $A_{I_1}, \ldots, A_{i_r}$, $r \leq n$,

$$P(A_{I_1} \cap A_{I_2} \cdots \cap A_{I_r}) = P(A_{I_1})P(A_{I_2}) \cdots P(A_{I_r}). \tag{2.20}$$

An infinite set of events is defined to be independent if every finite subcollection of these events is independent. It is worth noting that for the case of three events, A, B, C are independent if all the following four conditions are satisfied:

$$\begin{aligned} P(A \cap B \cap C) &= P(A)P(B)P(C), \\ P(A \cap B) &= P(A)P(B), \\ P(B \cap C) &= P(B)P(C), \\ P(A \cap C) &= P(A)P(C). \end{aligned} \tag{2.21}$$

2.5 Random variables

Often, instead of focusing on the entire outcome space, it may be sufficient to concentrate on a summary of outcomes relevant to the problem at hand, say a function of the outcomes. In tossing a coin four times, it may be sufficient to look at the number of heads instead of the order in which they are obtained. In observing photons from an astronomical source, it may be sufficient to look at the mean number of photons in a spectral band over some time interval, or the ratio of photons in two spectral bands, rather than examining each photon individually.

These real-valued functions on the outcome space or sample space are called **random variables**. Data are realizations of random variables. Typically a random variable X is a function on the sample space Ω. In the case of countable sample spaces Ω, this definition always works. But in the case of uncountable Ω, one should be careful. As mentioned earlier, not all subsets of an uncountable space can be called an event, or a probability assigned to them. A random variable is a function such that $\{\omega \in \Omega : X(\omega) \leq a\}$, is an event for all real numbers a. In practical situations, the collection of events can be defined to be inclusive enough that the set of events follows certain mathematical conditions (closure under complementation, countable unions and intersections). So in practice, the technical aspects can be ignored.

Note that in casual usage, some people label a phenomenon as "random" to mean that the events have equal chances of possible outcomes. This concept is correctly called **uniformity**. The concept of **randomness** does not require uniformity. Indeed, the following sections and Chapter 4 are largely devoted to phenomena that follow nonuniform distributions.

2.5.1 Density and distribution functions

A random variable is called a **discrete** random variable if it maps a sample space to a countable set (e.g. the integers) with each value in the range having probability greater than or equal to zero.

Definition 2.4 (Cumulative distribution function) The **cumulative distribution function (c.d.f.)** or simply the **distribution function** F of a random variable X is defined as

$$F(x) = P(X \leq x) = P(\omega \in \Omega : X(\omega) \leq x), \tag{2.22}$$

for all real numbers x. In the discrete case when X takes values $x_1, x_2, \ldots$, then

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i). \tag{2.23}$$

The c.d.f. F is a nondecreasing, right-continuous function satisfying

$$\lim_{x \to -\infty} F(x) = 0 \text{ and } \lim_{x \to \infty} F(x) = 1. \tag{2.24}$$

The c.d.f. of a discrete random variable is called a **discrete distribution**. A random variable with a continuous distribution function is referred to as a **continuous** random variable. A continuous random variable maps the sample space to an uncountable set (e.g. the real numbers). While the probability that a continuous random variable takes any specific value is zero, the probability that it belongs to an infinite set of values such as an interval may be positive. It should be understood clearly that the requirement that X is a continuous random variable does not mean that $X(\omega)$ is a continuous function; in fact, continuity does not make sense in the case of an arbitrary sample space, Ω.

Often some continuous distributions are described through the **probability density function** (p.d.f.). A nonnegative function f is called the probability density function of a distribution function if for all x

$$F(x) = \int_{-\infty}^{x} f(y)dy. \tag{2.25}$$

We warn astronomers that the statistician's term "density" has no relationship to "density" in physics where it measures mass per unit volume. If the probability density exists, then the distribution function F is continuous. The converse of this statement is not always true, as there exist continuous distributions that do not have corresponding density functions.

Instead of a single function on a sample space, we might consider several functions at the same time. For example, consider the sample space (or outcome space) where Ω consists of exoplanets within 50 pc of the Sun, and the random variables $\{X_1, X_2, X_3, X_4\}$ denote the exoplanet masses, radii, surface temperatures, and orbital semi-major axes. There may be important relationships among these variables of scientific interest, so it is crucial to study these variables together rather than individually. Such a vector $(X_1, \ldots, X_k)$ of random variables is called a **random vector**. The distribution of random vector F is given by

$$\begin{aligned} F(x_1, \ldots, x_k) &= P(X_1 \leq x_1, \ldots, X_k \leq x_k) \\ &= P(\omega \in \Omega : X_1(\omega) \leq x_1, \ldots, X_k(\omega) \leq x_k), \end{aligned} \tag{2.26}$$

where $x_1, \ldots, x_k$ are real numbers. Note that F is a nondecreasing right-continuous function in each coordinate. The one-dimensional distributions F_i of individual variables X_i are called **marginal distributions** given by

$$F(x_i) = P(X_i \leq x_i), \quad i = 1, \ldots, k. \tag{2.27}$$

Definition 2.5 (Independent random variables) The random variables $X_1, \ldots, X_n$ are said to be independent if the joint distribution is the product of the marginal distributions. That is

$$P(X_1 \leq a_1, \ldots, X_n \leq a_n) = P(X_1 \leq a_1) \ldots P(X_n \leq a_n),$$

for all real numbers $a_1, \ldots, a_n$.

An infinite collection $\{X_i\}$ of random variables are said to be independent if every finite subcollection of random variables is independent.

An important indicator of the central location of a random variable's distribution is the first moment or **mean** of the random variable. The mean of a random variable is defined as the weighted average where the weight is obtained from the associated probabilities. The terms **mathematical expectation** or the **expected value** of a random variable are often used interchangeably with the more familiar term "mean". For a random variable X taking values $x_1, x_2, \ldots, x_n$, the expected value of X is denoted by $E[X]$ defined by

$$E[X] = \sum_i x_i P(X = x_i). \tag{2.28}$$

If h is a real-valued function, then $Y = h(X)$ is again a random variable. The expectation of this Y can be computed without deriving the distribution of Y using only the distribution of X:

$$E[Y] = E[h(X)] = \sum_i h(x_i) P(X = x_i). \tag{2.29}$$

The notation $E[X]$ is equivalent to the notations $\langle X \rangle$ or $\bar{X}$ familiar to physical scientists.

The same definition of expectation as in (2.28) and (2.29) can be used for any discrete random variable X taking infinitely many nonnegative values. However, difficulties may be encountered in defining the expectation of a random variable taking infinitely many positive and negative values. Consider the case where W is a random variable satisfying

$$P(W = 2^j) = P(W = -2^j) = 2^{-j-1}, \text{ for } j = 1, 2, \ldots \tag{2.30}$$

In this case, the expectation $E[W]$ cannot be defined, as both the positive part $W^+ = \max(0, W)$ and the negative part $W^- = \max(0, -W)$ have infinite expectations. This would make $E[X]$ to be $\infty - \infty$, which is meaningless. However, for a general discrete random variable, $E[X]$ can be defined as in (2.28) provided

$$\sum_i |x_i| \, P(X = x_i) < \infty. \tag{2.31}$$

In case the distribution F of a random variable X has density f as in (2.25), then the **expectation** is defined as

$$E[X] = \int_{-\infty}^{\infty} y f(y) dy, \text{ provided } \int_{-\infty}^{\infty} |y| f(y) dy < \infty. \tag{2.32}$$

The expectation of a function h of a random variable X can be defined similarly as in (2.29), provided $\sum_i |h(x_i)| P(X = x_i) < \infty$ in the discrete case, and

$$E[h(X)] = \int h(y) f(y) dy \quad \text{provided} \quad \int |h(y)| f(y) dy < \infty \tag{2.33}$$

in case the distribution of X has density f.

Another important and commonly used function of a distribution function that quantifies the spread is the second moment centered on the mean, known as the **variance** and often denoted by σ^2. The variance is defined by

$$\sigma^2 = Var(X) = E\left[(X - \mu)^2\right] = E[X^2] - \mu^2, \tag{2.34}$$

where $\mu = E[X]$.

The mean and variance need not be closely related, as seen in the following simple example. Let X be a random variable taking values 1 and -1 with probability 0.5 each, and let Y be a random variable taking values 1000 and -1000 with probability 0.5. Both X and Y have the same mean ($= 0$), but $Var(X) = 1$ and $Var(Y) = 10^6$.

It is helpful to derive the variance of the sum of random variables. If $X_1, X_2, \ldots, X_n$ are n random variables, we find that

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] \tag{2.35}$$

and the variance of the sum $\sum_{i=1}^{n} X_i$ can be expressed as

$$\begin{aligned} Var\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} Var(X_i) + \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} Cov(X_i, X_j), \quad \text{where} \\ Cov(X, Y) &= E[(X - E[X])(Y - E[Y])]. \end{aligned} \tag{2.36}$$

The Cov quantity is the **covariance** measuring the relation between the scatter in two random variables. If X and Y are independent random variables, then $Cov(X, Y) = 0$ and $Var(\sum_{i-1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i)$, while the converse is not true; some dependent variables may have zero covariance.

If all the X_i variables have the same variance σ^2, the situation is called **homoscedastic**. If $X_1, X_2, \ldots, X_n$ are independent, the variance of the sample mean $\bar{X} = (1/n) \sum_{i=1}^{n} X_i$ is given by

$$Var(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}. \tag{2.37}$$

The variance essentially measures the mean square deviation from the mean of the distribution. The square root of the variance, σ, is called the **standard deviation**. The mean μ and the standard deviation σ of a random variable X are often used to convert X to a **standardized form**

$$X_{std} = \frac{X - \mu}{\sigma} \tag{2.38}$$

with mean zero and variance unity. This important transformation also removes the units of the original variable. Other transformations also reduce scale and render a variable free from units, such as the logarithmic transformation often used by astronomers. It should be recognized that the logarithmic transformation is only one of many optional variable transformations. Standardization is often preferred by statisticians with mathematical properties useful in statistical inference.

The third central moment $E[(X - E[X])^3]$ provides information about the **skewness** of the distribution of a random variable; that is, whether the distribution of X leans more towards right or left. Higher order moments like the k-th order moment $E[X^k]$ also provide some additional information about the distribution of the random variable.

2.5.2 Independent and identically distributed random variables

When repeated observations are made, or when an experiment is repeated several times, the successive observations lead to independent random variables. If the data are generated from the same population, then the resultant values can be considered as random variables with a common distribution. These are a sequence of **independent and identically distributed** or i.i.d. random variables. In the i.i.d. case, the random variables have a common mean and variance (if these moments exist).

Some observational studies in astronomy produce i.i.d. random variables. The redshifts of galaxies in an Abell cluster, the equivalent widths of absorption lines in a quasar spectrum, the ultraviolet photometry of a cataclysmic variable accretion disk, and the proper motions of a sample of Kuiper Belt bodies will all be i.i.d. if the observational conditions are unchanged. But the i.i.d. conditions are often violated. The sample may be heterogeneous with objects drawn from different underlying distributions. The observations may have been taken under different conditions such that the measurement errors differ. This leads to

a condition called **heteroscedasticity** that violates the i.i.d. assumption. Heteroscedasticity means that different data points have different variances.

Since a great many methods of statistics, both classical and modern, depend on the i.i.d. assumption, it is crucial that astronomers understand the concept and its relationship to the datasets under study. Incorrect use of statistics that require i.i.d. will lead to incorrect quantitative results, and thereby increase the risk of incorrect or unsupported scientific inferences.

2.6 Quantile function

The cumulative distribution function $F(x)$ estimates the value of the population distribution function at a chosen value of x. But the astronomer often asks the inverse question: "What value of x corresponds to a specified value of $F(x)$?" This answers questions like "What fraction of galaxies have luminosities above L^*?" or "At what age have 95% of stars lost their protoplanetary disks?" This requires estimation of the **quantile function** of a random variable X, the inverse of F, defined as

$$Q(u) = F^{-1}(u) = \inf\{y : F(y) \geq u\} \tag{2.39}$$

where $0 < u < 1$. Here inf (infimum) refers to the smallest value of y with the property specified in the brackets.

When large samples are considered, the quantile function is often convenient for scientific analysis as the large number of data points are reduced to a smaller controlled number of interesting quantiles such as the 5%, 25%, 50%, 75% and 95% quantiles. A quantile function for an astronomical dataset is compared to the more familiar histogram in Figure 6.1 of Chapter 6. Quantile-quantile (Q-Q) plots are often used in visualization to compare two samples or one sample with a probability distribution. Q-Q plots are illustrated in Figures 5.4, 7.2, 8.2 and 8.6.

But when small samples are considered, the quantile function can be quite unstable. This is readily understood: for a sample of $n = 8$ points, the 25% and 75% quartiles are simply the values of the second and sixth data points, but for $n = 9$ interpolation is needed based on very little information about the underlying distribution of $Q(u)$. Also, the asymptotic normality of the quantile function breaks down when the distribution function has regions with low density.

For many datasets, computation of the quantile function is straightforward: the data are sorted in order of increasing X, and the values of chosen quantiles are estimated by local interpolation. But for extremely large datasets, sorting the entire vector is computationally infeasible and the total number of data points may be ill-defined. In these cases, quantiles can be estimated from a continuous data stream if the sampling is random using the **data skeleton algorithm** (McDermott *et al*. 2007).

2.7 Discrete distributions

We have encountered the discrete uniform distribution in some examples above, such as the probability $P(i) = 1/6$ for $i = 1, 2, \ldots, 6$ for a single roll of a balanced die. This is only the simplest of a range of probability distributions that take on discrete values which are frequently encountered. We outline other discrete probability distributions here; more mathematical properties of some with particular importance in astronomy are presented in Chapter 4.

Bernoulli distribution: Suppose that an experiment results in a success or failure (True or False) and we define $X = 1$ if the outcome is a success and $X = 0$ if the outcome is a failure. X is called a **binary variable** or **dichotomous variable**, and the distribution $P(X = 1) = p = 1 - P(X = 0)$ is called the **Bernoulli distribution**, where $0 < p < 1$.

Binomial distribution: Suppose this experiment is repeated n times independently and the number of successes are denoted by X. Then the distribution of X is given by

$$P(X = i) = \binom{n}{i} p^i (1-p)^{n-i}, \text{ where } \binom{n}{i} = \frac{n!}{i!(n-i)!}, \tag{2.40}$$

for $i = 0, 1, \ldots, n$. In this case, the random variable X is said to have the **binomial distribution** and is denoted by $X \sim Bin(n, p)$. The mean and variance of X are given by

$$E[X] = np, \text{ and } Var[X] = np(1-p). \tag{2.41}$$

The binomial probabilities $P(X = i)$ first increase monotonically and then decrease. Its highest value is reached when i is the largest integer less than or equal to $(n + 1)p$.

If the number of trials n in a binomial distribution is large, it is practically impossible to compute the probabilities. A good approximation is needed. In 1773, De Moivre established the special case for a binomial random variable $X \sim Bin(n, p)$,

$$P\left(a \leq \frac{X - np}{\sqrt{np(1-p)}} \leq b\right) \to \Phi(b) - \Phi(a), \tag{2.42}$$

as $n \to \infty$, where Φ is the cumulative normal distribution function. This is a special case of the Central Limit Theorem (CLT) (Section 2.10). Note that the CLT applies when the underlying probability p is any fixed number between 0 and 1 and n approaches infinity.

Poisson distribution: A random variable X is said to have a **Poisson distribution**, denoted $X \sim Poi(\lambda)$, with rate $\lambda > 0$ if

$$P(X = i) = \frac{\lambda^i}{i!} e^{-\lambda}, \text{ for } i = 0, 1, \ldots \tag{2.43}$$

In this case

$$E[X] = Var(X) = \lambda. \tag{2.44}$$

The binomial and Poisson distributions are closely related. Suppose $X \sim Bin(n, p_n)$. If n is large and i is close to $n/2$, it is extremely difficult to compute the binomial coefficients $\binom{n}{i}$ in Equation (2.40). However, the binomial probabilities can be approximated by the Poisson probabilities when p_n is small, n is large, and $\lambda = np_n$ is not too large or too small. That is,

$$P(X = i) = \binom{n}{i} p_n^i (1 - p_n)^{n-i} \approx e^{-\lambda} \frac{\lambda^i}{i!}, \tag{2.45}$$

for $i = 0, 1, \ldots$. Poisson distributed random variables appear in many astronomical studies and it is thus very important to understand them well.

Negative binomial distribution: Let the random variable X_r denote the number of trials until a total of r successes is accumulated. Then

$$\begin{aligned} P(X_r = i) &= \binom{n-1}{r-1} (1-p)^{n-r} p^r, \text{ for } n = r, r+1, \ldots, \\ E[X_r] &= \frac{r}{p}, Var(X_r) = \frac{qr}{p^2}, \end{aligned} \tag{2.46}$$

and X_r is said to have a **negative binomial distribution** with parameters (r, p). When $r = 1$ the distribution is known as the **geometric distribution**.

2.8 Continuous distributions

We introduce here four important continuous distributions: uniform, exponential, normal or Gaussian, and lognormal distributions. Figure 2.2 at the end of this chapter plots some of these distributions for a few typical parameter values, showing both the p.d.f. f and the c.d.f. F defined in Equations (2.23) and (2.25). Further properties of these and other continuous distributions are discussed in Chapter 4.

Uniform distribution: The distribution function F is called **uniform** if its p.d.f. f is given by

$$f(x) = \frac{1}{b-a} \quad \text{for} \quad a < x < b, \tag{2.47}$$

and zero otherwise. A random variable X is uniformly distributed on (a, b) if its distribution is uniform on (a, b). In this case it is denoted by $X \sim U(a, b)$. For such a random variable, $E[X] = (a + b)/2$ and $Var(X) = \frac{1}{12}(b - a)^2$. If Y has a continuous distribution F, then $F(Y) \sim U(0, 1)$.

Uniformly distributed random variables play an important role in simulations. To simulate a uniform random variable, flip a coin repeatedly, and define $X_n = 1$ or $X_n = 0$ according

as head or tail turned up on the n-th flip. Then it can be shown that

$$V = \sum_{n=1}^{\infty} \frac{X_n}{2^n} \sim U(0, 1). \tag{2.48}$$

If F is any c.d.f., continuous or otherwise, the random variable

$$W = \inf\{x : F(x) \geq V\}, \tag{2.49}$$

the smallest x such that $V \leq F(x)$, has distribution F. This provides a method for generating realizations of a random variable with any given distribution. A random variable with any distribution can be generated this way by flipping a coin.

Exponential distribution: A random variable with the distribution function

$$F(x) = 1 - e^{-\lambda x} \tag{2.50}$$

for $x > 0$ and $F(x) = 0$ for $x \leq 0$ is called **exponential with rate** $\lambda > 0$. F has p.d.f. f given by

$$f(x) = \lambda \exp(-\lambda x), \quad x \geq 0, \tag{2.51}$$

and zero otherwise. A random variable with the exponential distribution has the so-called **memoryless property**

$$P(X > t + s \mid X > s) = P(X > t) \quad \text{for all } s, t \geq 0. \tag{2.52}$$

This property is essential in modeling the waiting times in Poisson processes. The mean and variance of an exponential random variable X are given by

$$E[X] = \frac{1}{\lambda} \quad \text{and} \quad Var(X) = \frac{1}{\lambda^2}. \tag{2.53}$$

The top panels of Figure 2.2 show the exponential density and distribution for three different values for λ.

Normal or Gaussian distribution: In 1733, French mathematician Abraham De Moivre introduced the probability density function (p.d.f.)

$$\phi(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \quad \text{where} \quad -\infty < \mu < \infty, \ \sigma > 0, \tag{2.54}$$

to approximate the binomial distribution $Bin(n, p)$ when n is large. He named Equation (2.54) the "exponential bell-shaped curve". Its full potential was realized only in 1809, when the German mathematician Karl Friedrich Gauss used it in his astronomical studies of celestial mechanics. During the following century, statisticians found that many datasets representing many types of random variables have histograms closely resembling the Gaussian density. This is explained by the Central Limit Theorem (Section 2.10). The curve (2.54) has come to be known both as the Gaussian and the normal probability density. The

probability distribution function obtained from the bell-shaped curve (2.54) is called the **Gaussian** or **normal distribution function**,

$$\Phi(x;\mu,\sigma^2)=\int_{-\infty}^{x}\phi(y;\mu,\sigma^2)dy=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}dy. \tag{2.55}$$

A random variable with a p.d.f. (2.54) is called a **normal random variable**. A normal random variable X with parameters μ and σ^2 is denoted by $X\sim N(\mu,\sigma^2)$. Its mean and variance are given by

$$E[X]=\mu \quad \text{and} \quad Var(X)=\sigma^2. \tag{2.56}$$

The middle panels of Figure 2.2 show the normal density and distribution for four different combinations of mean μ and variance σ^2.

Lognormal distribution: For a random variable $X\sim N(\mu,\sigma^2)$, $Y=e^X$ has a **lognormal distribution** with p.d.f. f given by

$$f(x)=\frac{1}{\sigma x\sqrt{2\pi}}\exp\left\{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right\},\ \text{for } x>0, \tag{2.57}$$

and $f(x)=0$ for $x\le 0$. A lognormal random variable Y with parameters μ and σ^2 is denoted by $Y\sim \ln N(\mu,\sigma^2)$. The mean and variance of Y are given by

$$E[Y]=e^{\mu+(1/2)\sigma^2} \quad \text{and} \quad Var(Y)=\left(e^{\sigma^2}-1\right)e^{2\mu+\sigma^2}. \tag{2.58}$$

The bottom panels of Figure 2.2 show the lognormal density and distribution for $\mu=0$ and four values of variance σ^2.

2.9 Distributions that are neither discrete nor continuous

Of course, there are distributions that are neither discrete nor continuous. Suppose X is uniformly distributed on $(-1,1)$, that is

$$P(a<X<b)=\frac{1}{2}(b-a) \text{ for all } -1<a<b<1. \tag{2.59}$$

If $Y=\max(0,X)$, then clearly $P(Y=0)=P(X\le 0)=1/2$ and

$$P(a<X<b)=\frac{1}{2}(b-a) \text{ for all } 0<a<b<1. \tag{2.60}$$

So the distribution F of Y is continuous except at 0. In the most general case, where the random variable X is neither discrete nor has a density, as in this example, $E[X]$ is defined as

$$E[X]=\int_{-\infty}^{\infty}y\,dF(y),\ \text{provided}\ \int_{-\infty}^{\infty}|y|dF(y)<\infty, \tag{2.61}$$

where the integral is the Riemann–Steiljes integral with respect to the distribution function F, where F is the c.d.f. of X.

2.10 Limit theorems

Probability theory has many powerful mathematical results which establish or constrain properties of random variables. One profound phenomenon, mentioned in Section 2.1, is that uncertainty at the micro-level (e.g. a single measurement or event) leads to deterministic behavior at the macro-level. This is a consequence of the Law of Large Numbers:

Theorem 2.6 (Law of Large Numbers) *Let $X_1, X_2, \ldots$ be a sequence of independent random variables with a common distribution and $E[|X_1|] < \infty$. Then*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu = E[X_1], \tag{2.62}$$

as $n \to \infty$, i.e. *for all $\epsilon > 0$, $P(|\bar{X}_n - \mu| > \epsilon) \to 0$, as $n \to \infty$.*

The theorem states that the **sample mean** $\bar{X}_n$ gives a good approximation to the **population mean** $\mu = E[X_1]$ when n is large. There is a crucial distinction between the sample mean, which is a random quantity, and the population mean, which is a population parameter and is a fixed number. It is important to note that this result is valid for discrete random variables, continuous random variables, and for general random variables.

An even more powerful result is the Central Limit Theorem (CLT). This states that, for a sequence of independent random variables $X_1, X_2, \ldots$ with a common distribution, the distribution of $\bar{X}_n$ can be approximated by a Gaussian (normal) distribution provided $E[|\ X_1\ |^2] < \infty$. A formal statement of the theorem follows.

Theorem 2.7 (Central Limit Theorem) *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean $\mu = E[X_1]$ and finite variance $\sigma^2 = E[(X_1 - \mu)^2] > 0$. Then*

$$P\left(\sqrt{n}(\bar{X}_n - \mu) \leq x\sigma\right) \to \Phi(x), \tag{2.63}$$

for all x, where

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)dt \quad \textit{and} \quad \phi(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}.$$

The CLT is an extremely useful result for approximating probability distributions of sums of large numbers of independent random variables. In later chapters, we will repeatedly refer to a statistic or estimator exhibiting **asymptotic normality**. This arises directly from the CLT where the distribution approaches the Gaussian for large sample sizes.

2.11 Recommended reading

Ross, S. (2010) *Introduction to Probability Models* 10th ed., Academic Press, New York

An excellent undergraduate textbook laying the foundations of probability theory. Coverage includes conditional probability, Bayes' theorem, random variables, elementary

probability distributions, renewal theory, queueing theory, reliability, stationary stochastic processes and simulation techniques.

2.12 R applications

We give here the first **R** scripts of the volume with associated graphics. The reader is encouraged to first read the introduction to **R** in Appendix B.

R provides automatic computation of about 20 common probability distribution functions, and many more are available in **CRAN** packages. In the script below, we make the upper left panel of Figure 2.2 showing the p.d.f. of the exponential distribution for three values of the rate λ in Equation (2.51). The script starts with the construction of a closely spaced vector of x-axis values using **R**'s *seq* function. Here we make a vector of 250 elements ranging from 0 to 5. The function *dexp* gives the density (p.d.f.) of the exponential distribution to be evaluated at these values.

A bivariate graphics frame is created by the generic *plot* function. We first calculate it for the rate $\lambda = 0.5$ and then add curves for $\lambda = 1.0$ and 1.5. Note how the function *dexp* can be compactly embedded within the *plot* function call. Several commonly used parameters of *plot* are used: plot type (points, lines, histogram, steps), axis limits, axis labels, font size scaling, line width, and line type (solid, dashed, dotted). After the first curve is plotted, we superpose other curves using the *lines* command using the 'add=TRUE' option to place new curves on the same plot. These and other options are used very often, and the **R** practitioner is encouraged to learn them from the documentation given by *help*(par). The *legend* function permits annotations inside the plotting window. Another generic option for annotating graphs is the *text* function.

```
# Set up 6 panel figure

par(mfrow=c(3,2))

# Plot upper left panel with three illustrative exponential p.d.f. distributions

xdens <- seq(0,5,0.02)
plot(xdens,dexp(xdens,rate=0.5), type='l', ylim=c(0,1.5), xlab='',
   ylab='Exponential p.d.f.',lty=1)

lines(xdens,dexp(xdens,rate=1), type='l', lty=2)
lines(xdens,dexp(xdens,rate=1.5), type='l', lty=3)
legend(2, 1.45, lty=1, substitute(lambda==0.5), box.lty=0)
legend(2, 1.30, lty=2, substitute(lambda==1.0), box.lty=0)
legend(2, 1.15, lty=3, substitute(lambda==1.5), box.lty=0)

# Help files to learn these functions

help(seq) ; help(plot) ; help(par) ; help(lines) ; help(legend)
```

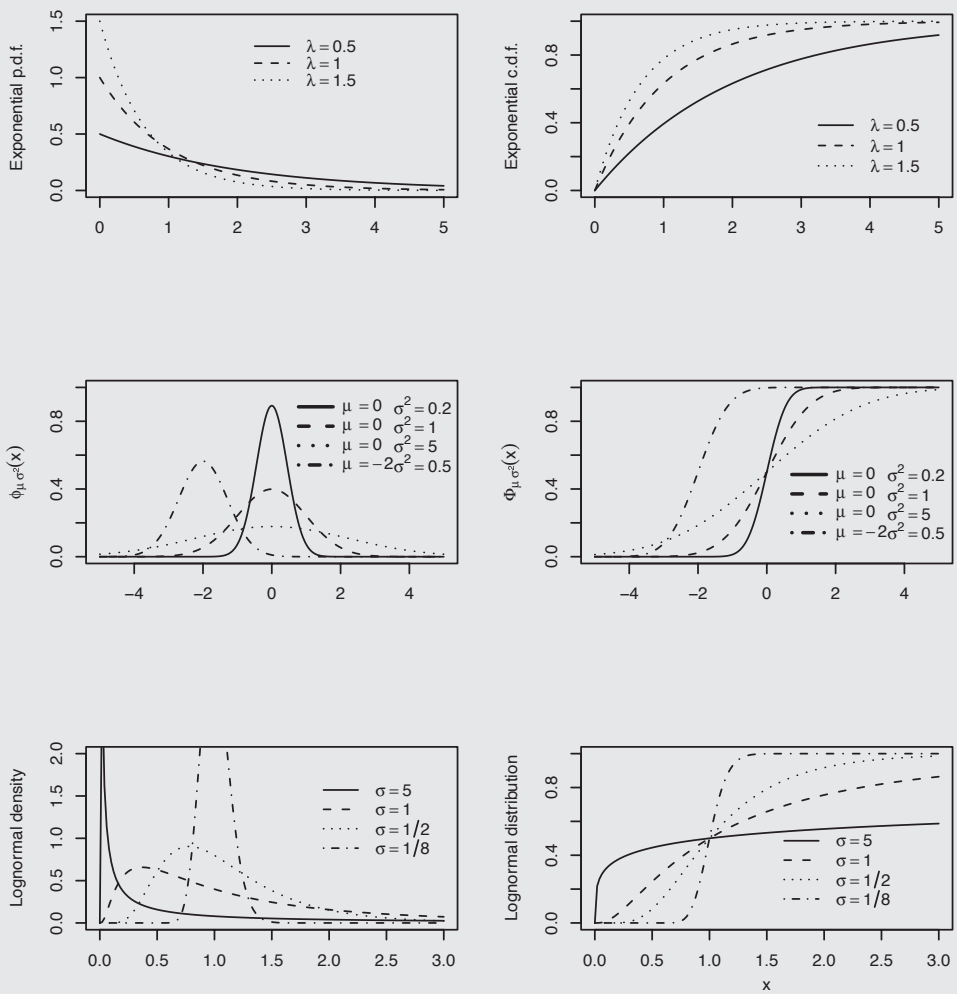

Fig. 2.2 Examples of three continuous statistical distributions: exponential (*top*), normal (*middle*), and lognormal (*bottom*). Left panels are (differential) probability density functions (p.d.f.) and right panels are cumulative distribution functions (c.d.f).

The other five panels of Figure 2.2 illustrating other distribution shapes are made with very similar scripts shown below. The top right panel substitutes *dexp* giving the exponential p.d.f. with the *pexp* function giving the exponential c.d.f. The middle panels similarly use *dnorm* and *pnorm* for the p.d.f. and c.d.f. of the normal (Gaussian) distribution, and the bottom panels use the *dlnorm* and *plnorm* functions for the lognormal distribution p.d.f. and c.d.f.

The labels and annotations of these graphs have options commonly used by astronomers such as Greek letters, superscripts and subscripts. One option shown above uses the *substitute* function to show Greek letters, while another more flexible option below uses the

expression function. The "==" operator is used to assign specific values to parameters within functions.

```
# Plot upper right panel with three illustrative exponential c.d.f. distributions

plot(xdens, pexp(xdens,rate=0.5), type='l', ylim=c(0,1.0), xlab='',
   ylab='Exponential c.d.f.', lty=1)
lines(xdens, pexp(xdens,rate=1), type='l', lty=2)
lines(xdens, pexp(xdens,rate=1.5), type='l', lty=3)
legend(3, 0.50, lty=1, substitute(lambda==0.5), box.lty=0)
legend(3, 0.38, lty=2, substitute(lambda==1.0), box.lty=0)
legend(3, 0.26, lty=3, substitute(lambda==1.5), box.lty=0)

# Plot middle panels with illustrative normal p.d.f. and c.d.f.

xdens <- seq(-5, 5, 0.02)
ylabdnorm <- expression(phi[mu~sigma^2] (x))
plot(xdens, dnorm(xdens, sd=sqrt(0.2)), type='l', ylim=c(0,1.0), xlab='',
   ylab=ylabdnorm, lty=1)
lines(xdens, dnorm(xdens, sd=sqrt(1.0)), type='l', lty=2)
lines(xdens, dnorm(xdens, sd=sqrt(5.0)), type='l', lty=3)
lines(xdens, dnorm(xdens, mean=-2.0, sd=sqrt(0.5)), type='l', lty=4)
leg1 <- expression(mu^' '==0, mu^' '==0, mu^' '==0, mu^' '==-2)
leg2 <- expression(sigma^2==0.2, sigma^2==1.0, sigma^2==5.0, sigma^2==0.5)
legend(0.5, 1.0, lty=1:4, leg1, lwd=2, box.lty=0)
legend(3.0, 1.01, leg2, box.lty=0)

ylabpnorm <- expression(Phi[mu~sigma^2] (x))
plot(xdens, pnorm(xdens, sd=sqrt(0.2)), type='l', ylim=c(0,1.0), xlab='',
   ylab=ylabpnorm, lty=1)
lines(xdens, pnorm(xdens, sd=sqrt(1.0)), type='l', lty=2)
lines(xdens, pnorm(xdens, sd=sqrt(5.0)), type='l', lty=3)
lines(xdens, pnorm(xdens, mean=-2.0, sd=sqrt(0.5)), type='l', lty=4)
leg1 <- expression(mu^' '==0, mu^' '==0, mu^' '==0, mu^' '==-2)
leg2 <- expression(sigma^2==0.2, sigma^2==1.0, sigma^2==5.0, sigma^2==0.5)
legend(0.5, 0.6, lty=1:4, leg1, lwd=2, box.lty=0)
legend(3.0, 0.61, leg2, box.lty=0)

# Plot bottom panels with illustrative lognormal p.d.f. and c.d.f.

xdens <- seq(0,3, 0.02)
plot(xdens, dlnorm(xdens, meanlog=0, sdlog=5), type='l', ylim=c(0,2), xlab='',
   ylab='Lognormal density', lty=1)
```

```
lines(xdens, dlnorm(xdens, meanlog=0, sdlog=1), type='l', lty=2)
lines(xdens, dlnorm(xdens, meanlog=0, sdlog=1/2), type='l', lty=3)
lines(xdens, dlnorm(xdens, meanlog=0, sdlog=1/8), type='l', lty=4)
leg1 <- expression(sigma==5, sigma==1, sigma==1/2, sigma==1/8)
legend(1.8,1.8,lty=1:4,leg1,box.lty=0)

plot(xdens, plnorm(xdens, meanlog=0, sdlog=5), type='l', ylim=c(0,1), xlab='x',
    ylab='Lognormal distribution',lty=1)
lines(xdens, plnorm(xdens, meanlog=0, sdlog=1), type='l', lty=2)
lines(xdens, plnorm(xdens, meanlog=0, sdlog=1/2), type='l', lty=3)
lines(xdens, plnorm(xdens, meanlog=0, sdlog=1/8), type='l', lty=4)
leg1 <- expression(sigma==5, sigma==1, sigma==1/2, sigma==1/8)
legend(1.5, 0.6, lty=1:4, leg1, box.lty=0)

# Return plot to single-panel format

par(mfrow=c(1,1))
```

The **CRAN** package *prob* provides capabilities to define random variables based on experiments and to examine their behaviors. The package has built-in simple experiments such as coin tosses and urn draws, but allows the user to define new calculated associated probabilities, including marginal and conditional distributions. This could be useful to astronomers considering the statistical outcomes of hierarchical experiments; for instance, the selection of a sample from an underlying population using a sequence of observational criteria. The code requires that the full sample space be specified.

3 Statistical inference

3.1 The astronomical context

Statistical inference helps the scientist to reach conclusions that extend beyond the obvious and immediate characterization of individual datasets. In some cases, the astronomer measures the properties of a limited sample of objects (often chosen to be brighter or closer than others) in order to learn about the properties of the vast underlying population of similar objects in the Universe. Inference is often based on a **statistic**, a function of random variables. At the early stages of an investigation, the astronomer might seek simple statistics of the data such as the average value or the slope of a heuristic linear relation. At later stages, the astronomer might measure in great detail the properties of one or a few objects to test the applicability, or to estimate the parameters, of an astrophysical theory thought to underly the observed phenomenon.

Statistical inference is so pervasive throughout these astronomical and astrophysical investigations that we are hardly aware of its ubiquitous role. It arises when the astronomer:

- smooths over discrete observations to understand the underlying continuous phenomenon
- seeks to quantify relationships between observed properties
- tests whether an observation agrees with an assumed astrophysical theory
- divides a sample into subsamples with distinct properties
- tries to compensate for flux limits and nondetections
- investigates the temporal behavior of variable sources
- infers the evolution of cosmic bodies from studies of objects at different stages
- characterizes and models patterns in wavelength, images or space

and many other situations. These problems are discussed in later chapters of this volume: nonparametric statistics (Chapter 5), density estimation (Chapter 6), regression (Chapter 7), multivariate analysis (Chapter 8) and classification (Chapter 9), censoring and truncation (Chapter 10), time series analysis (Chapter 11) and spatial analysis (Chapter 12). In this chapter, we lay some of the foundations of statistical inference.

Consider the effort to understand the dynamical state and evolution of a globular cluster of stars. Images give two-dimensional locations of perhaps 1% of the stars, with individual stellar masses estimated from color–magnitude diagrams. Radial velocity measurements from spectra are available for even fewer cluster members. Thus, only three of the six dimensions of phase space (three spatial and three velocity dimensions) are accessible. From this limited information, we have gleaned significant insights with the use of astrophysical models. The three-dimensional structure was originally modeled as a truncated isothermal

sphere with two-body dynamical interactions causing the more massive stars to settle towards the core. Modern models involve computationally intensive N-body dynamical simulations to study the effects of binary star systems and stellar collisions on cluster evolution. Statistical inference allows quantitative evaluation of parameters within the context of astrophysical models, giving insights into the structure and dynamics of the globular star cluster. Statistical modeling is thus a crucial component of the scientific inferential process by which astrophysical understanding is developed from incomplete observational information.

3.2 Concepts of statistical inference

Statistical inference helps in making judgments regarding the likelihood that a hypothesized effect in data arises by chance or represents a real effect. It is particularly designed to draw conclusions about the underlying population when the observed samples are subject to uncertainties.

The term **statistical inference** is very broad. Two main aspects of inference are **estimation** and the **testing of hypotheses**. Regression, goodness-of-fit, classification and many other statistical procedures fall under its framework. Statistical inference can be parametric, nonparametric and semi-parametric. Parametric inference requires that the scientist makes some assumptions regarding the mathematical structure of the underlying population, and this structure has parameters to be estimated from the data at hand. Linear regression is an example of parametric inference. Nonparametric procedures make no assumption about the model structure or the distribution of the population. The Kolmogorov–Smirnov hypothesis test and the rank-based Kendall's τ correlation coefficient are examples of nonparametric procedures. Semi-parametric methods combine nonparametric and parametric procedures; local regression models are examples of semi-parametric procedures.

A classic development of the theory of statistical inference is presented by Erich Lehmann in his volumes *Theory of Point Estimation* (Lehmann & Casella 1998) and *Testing Statistical Hypotheses* (Lehmann & Romano 2005). The undergraduate-level texts by Rice (1995) and Hogg *et al.* (2005), and the graduate-level text by Wasserman (2004), are recommended for modern treatments. We outline here several central concepts that these volumes cover in more detail.

Point estimation If the shape of the probability distribution, or relationship between variables, of the underlying population is well-understood, then it remains to find the parameters of the distribution or relationship. For a dataset drawn from a Poisson distribution (Sections 2.7 and 4.2), for example, we seek the value of the rate λ, while for a normal distribution, we want to estimate the mean μ and variance σ^2. Typically a probability distribution or relationship is characterized by a p-dimensional vector of model parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$. For example, the model of a planet in a Keplerian orbit around a star has a vector of six parameters: semi-major axis, eccentricity, inclination, ascending node longitude, argument of periastron and true anomaly. Analogous parameter lists can be

made for a tidally truncated isothermal sphere of stars (King model), a turbulent viscous accretion disk (Shakura–Sunyaev α-disk model), the consensus model of cosmology with dark matter and dark energy (ΛCDM model), and many other well-developed astrophysical theories. The goal of estimating plausible or "best" values of θ based on observations is called **point estimation** (Section 3.3).

Likelihood methods One of the most popular methods of point estimation is **maximum likelihood estimation** (MLE). Likelihood is the hypothetical probability that a past event would yield a specific outcome. The concept differs from that of a probability in that a probability refers to the occurrence of future events, while a likelihood refers to past events with known outcomes. MLE is an enormously popular statistical method for fitting a mathematical model to data. Modeling real-world data by maximizing the likelihood offers a way to tune the free parameters of the model to provide a good fit. Developed in the 1920s by R. A. Fisher, MLE is a conceptual alternative to the least-squares method of the early nineteenth century (Section 1.2.3), but is equivalent to least squares under Gaussian assumptions. The Cramér–Rao inequality, which sets a lower bound on the variance of a parameter, is an important mathematical result of MLE theory (Section 3.4.4). MLE can be used for nontrivial problems such as mixture models for multimodal distributions or nonlinear models arising from astrophysical theory, and is readily applied to multivariate problems. While the likelihood function can be maximized using a variety of computational procedures, the EM algorithm developed in the 1970s is particularly effective (Section 3.4.5).

Confidence intervals Point estimates cannot be perfectly accurate as different datasets drawn from the same population will give rise to different inferred parameter values. To account for this, we estimate a range of values for the unknown parameter that is usually consistent with the data. This is the parameter's **confidence interval** or confidence set around the best-fit value (Section 3.4.4). The confidence intervals will vary from sample to sample. A confidence interval associated with a particular confidence level, say 95%, is a random interval that is likely to contain a one-dimensional parameter with probability 95%. Confidence intervals can be estimated for different methods of point estimation including least squares and MLE.

Resampling methods Understanding the variability of a point estimation is essential to obtaining a confidence interval, or to assess the accuracy of an estimator. In many situations encountered in the physical sciences, the variance may not have a closed-form expression. Resampling methods developed in the 1970s and 1980s come to the rescue in such cases. Powerful theorems demonstrated that they provide inference on a wide range of statistics under very general conditions. Methods such as the "bootstrap" involved constructing hypothetical populations from the observations, each of which can be analyzed in the same way to see how the statistics of interest depend on plausible random variations in the observations. Resampling the original data preserves whatever structures are truly present in the underlying population, including non-Gaussianity and multimodality. Although they typically involve random numbers, resampling is not an arbitrary Monte Carlo simulation; it is simulation from the observed data.

Testing hypotheses As the name implies, the goal here is not to estimate parameters of a function based on the data, but to test whether a dataset is consistent with a stated hypothesis. The scientist formulates a null hypothesis and an alternative hypothesis. The result of the test is to either reject or not reject the null hypothesis at a chosen significance level. Note that failure to reject the null hypothesis does not mean that the null hypothesis is correct. Statistical testing of a hypothesis leads to two types of error: wrongly rejecting the null hypothesis (**Type 1 errors** or **false positives**) and wrongly failing to reject the null hypothesis (**Type 2 errors** or **false negatives**). It is impossible to bring these two errors simultaneously to negligible values. Classical hypothesis testing is not symmetric; interchanging the null and alternative hypotheses gives different results. An important astronomical application is the detection of weak signals in noise. Hypothesis tests can also address questions like: Is the mean, or other statistic, of the data equal to a preset value (perhaps zero)? Is the current dataset consistent with the same underlying population as a previous dataset (two-sample test)? Are a group of datasets consistent with each other (k-sample test)? Is the dataset consistent with an underlying population that follows a specified functional relationship (goodness-of-fit test)?

Bayesian inference A growing approach to statistical inference is based on an interpretation of Bayes' theorem (Section 2.4.1) where observational evidence is used to infer (or update) inferences. As evidence accumulates, the degree of belief in a hypothesis or model ought to change. Bayesian inference is based not only on the likelihood that the data follow the model but also on a **prior** distribution. The quality of a Bayesian analysis depends on how one can best convert the prior information into a mathematical prior probability. Various Bayesian methods for parameter estimation, model assessment and other inference problems are outlined in Section 3.8.

3.3 Principles of point estimation

In parametric point estimation, the astronomer must be very careful in setting up the statistical calculation. Two decisions must be made. First, the functional model and its parameters must be specified. If the model is not well-matched to the astronomical population or astrophysical process under study, then the best fit obtained by the inferential process may be meaningless. This problem is called **model misspecification**. Statistical procedures are available to assist the scientist in **model validation** (or **goodness-of-fit**) and **model selection**.

Second, the method by which best-fit parameters are estimated must be chosen. The method of moments, least squares (LS) and maximum likelihood estimation (MLE) are important and commonly used procedures for constructing estimates of the parameters. The choice of estimation method is not obvious, but can be guided by the scientific goal. Statistical measures are available to assist the scientist in choosing a method: some may give best-fit parameters closest to the true value, with the greatest accuracy (smallest variance),

or with the highest probability (maximum likelihood). Fortunately, for many situations we can find single best-fit parameter values that are simultaneously unbiased, have minimum variance and have maximum likelihood.

In classical parametric estimation, the observations are assumed to be independently and identically distributed (i.i.d.) random variables with known probability distributions. The dataset $x_1, x_2, \ldots, x_n$ is assumed to be a realization of independent random variables $X_1, X_2, \ldots, X_n$ having a common probability distribution function (p.d.f.) f. We now consider distribution functions characterized by a small number of parameters, $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$. The point estimator of the vector of true parameter values θ is designated $\hat{\theta}$, pronounced "theta-hat". The estimator $\hat{\theta}$ of θ is a function of the random variables $(X_1, X_2, \ldots, X_n)$ under study,

$$\hat{\theta} = g(X_1, X_2, \ldots, X_n). \tag{3.1}$$

The point estimator is thus a function of random variables of the underlying population that is computed from a realization of the population in a particular data sample.

Providing one validates that the data are consistent with the model, the result of parameter estimation can be a great simplification of a collection of data into a few easily interpretable parameters. Often the astronomer believes the data instantiate a deterministic astrophysical theory where the functional relationship is established by physical processes such as gravity, quantum mechanics and thermodynamics. Here the functions can be relatively simple, such as an elliptical orbit of two orbiting bodies, or extremely complex, such as the spectrum from a stellar atmosphere with different atomic species in different excited states subject to radiative transfer.

A great deal of mathematics and discussion lies behind the simple goal of obtaining the "best" estimates of the parameters θ. One aspect relates to the method of obtaining estimators. During the nineteenth century, the method of moments and method of least squares were developed. Astronomy played a crucial role in least-squares theory, as outlined in Section 1.2.2. As a young man in the 1910s and 1920s, R. A. Fisher formulated the "likelihood" that a dataset fits a model, and inaugurated the powerful methods of maximum likelihood estimation (MLE). Minimum variance unbiased estimators (MVUEs) later rose to prominence. As computers became more capable, numerically intensive methods with fewer limitations than previous methods became feasible. The most important, developed in the 1970s and 1980s, is the bootstrap method. With the advent of numerical methods like Markov chain Monte Carlo simulations, nontrivial Bayesian inferential computations became feasible during the 1990s. Bayesian computational methods are being actively developed today.

It is thus common that different point estimators, often derived with different methods, are available to achieve the same data-analytic or scientific goal. A great deal of effort has been exerted to interpret possible meanings of the word "best" in "best-fit parameter" because statistical point estimators have several important properties that often cannot be simultaneously optimized. Statisticians take into consideration several important criteria of a point estimator:

Unbiasedness The bias of an estimator $\hat{\theta}$ is defined to be the difference between the mean of estimated parameter and its true value,

$$B(\hat{\theta}) = E[\hat{\theta}] - \theta. \tag{3.2}$$

This is not the error of a particular instantiation of $\hat{\theta}$ from a particular dataset. This is an intrinsic offset in the estimator.

An estimator $\hat{\theta}$ of a parameter θ is called **unbiased** if $B(\hat{\theta}) = 0$; that is, the expected value of $\hat{\theta}$ is $E[\hat{\theta}] = \theta$. For some biased estimators $\hat{\theta}$, $B(\hat{\theta})$ approaches zero as the data size approaches infinity. In such cases, $\hat{\theta}$ is called **asymptotically unbiased**. Heuristically, $\hat{\theta}$ is an unbiased if its long-term average value is equal to θ. If $\hat{\theta}$ is an unbiased estimator of θ, then the variance of the estimator $\hat{\theta}$ is given by $E[(\hat{\theta} - \theta)^2]$. The smaller the variance of the estimator, the better the estimation procedure. However, if the estimator $\hat{\theta}$ is biased, then $E[(\hat{\theta} - \theta)^2]$ is not the variance of $\hat{\theta}$. In this case,

$$E[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + (E[\hat{\theta} - \theta])^2 \tag{3.3}$$

$$MSE = \text{Variance of } \hat{\theta} + (\text{Bias})^2. \tag{3.4}$$

This quantity, the sum of the variance and the square of the bias, is called the **mean square error** (MSE) and is very important in evaluating estimated parameters.

Minimum variance unbiased estimator (MVUE) Among a collection of unbiased estimators, the most desirable one has the smallest variance, $Var(\hat{\theta})$.

Consistency This criterion states that a **consistent estimator** will approach the true population parameter value as the sample size increases. More precisely, an estimator $\hat{\theta}$ for a parameter θ is **weakly consistent** if for any small $\epsilon > 0$

$$P[|\hat{\theta} - \theta| \geq \epsilon] \longrightarrow 0 \tag{3.5}$$

as $n \to \infty$. The estimator is **strongly consistent** if

$$P[\hat{\theta} \longrightarrow \theta \text{ as } n \to \infty] = 1. \tag{3.6}$$

Asymptotic normality This criterion requires that an ensemble of consistent estimators $\hat{\theta}(n)$ has a distribution around the true population value θ that approaches a normal (Gaussian) distribution with variance decreasing as $1/n$.

3.4 Techniques of point estimation

Parameter estimation is motivated by the problem of fitting models from probability distributions or astrophysical theory to data. Many commonly used probability distributions (such as Gaussian, Poisson, Pareto or power-law) or astrophysical models (such as the temperature and pressure of a uniform gas, or masses and eccentricity in a planetary orbit) depend only on a few parameters. Once these parameters are known, the shape and scale of the curve, and the corresponding properties of the underlying population, are completely determined. For example, the one-dimensional Gaussian distribution depends only on two

parameters, the mean μ and the standard deviation σ, as the probability density function ϕ of the Gaussian distribution is given by

$$\phi(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{3.7}$$

Point estimates $\hat{\mu}$ and $\hat{\sigma}$ can be obtained using a variety of techniques. The most common methods for constructing estimates are: the method of moments, least squares and maximum likelihood estimation (MLE). The methods are assessed by the scientific goal of the estimation effort, and by criteria that are important to the scientist.

We will illustrate the common methods of estimation using the two parameters of a population that satisfies a normal (Gaussian) density, the mean μ and standard deviation σ. In this simple and familiar case, the different methods often – but not always – give the same estimators. But in more complicated situations, as we discuss in later chapters, the estimators will often differ.

In the Gaussian case (Equation 3.7), the sample mean and sample variance

$$\begin{aligned} \hat{\mu} &= \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \\ \hat{\sigma}^2 &= S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \end{aligned} \tag{3.8}$$

are estimators of μ and σ^2, respectively. The factor $n-1$ instead of n in the denominator of the estimator of σ^2 is required for unbiasedness. S_X is not an unbiased estimator of the standard deviation σ and $E[S_X] < \sigma$. In the Gaussian case, $\bar{X}$ and S_X^2 are unbiased estimators of μ and σ^2. This is because $E[X_i] = \mu$ and $E[(X_i - \mu)^2] = \sigma^2$ for each i. A simple calculation indicates that the estimator S_X^2 is not an unbiased estimator of σ^2 if $n-1$ in the denominator is replaced by n.

3.4.1 Method of moments

The **method of moments** for parameter estimation dates to the nineteenth century. The moments are quantitative measures of the parameters of a distribution: the first moment describes its central location; the second moment its width; and the third and higher moments describe asymmetries. As presented in Section 2.5.1, the distribution function F of a random variable is defined as $F(x) = P(X \leq a)$ for all a. The k-th moment of a random variable X with distribution function F is given by

$$\mu_k(X) = E[X^k] = \int x^k dF(x). \tag{3.9}$$

For the random sample X_i, the k-th sample moment is

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k. \tag{3.10}$$

Various parameters of a distribution can be estimated by the method of moments if one can first express the parameters as simple functions of the first few moments. Replacing

the population moments in the functions with the corresponding sample moments gives the estimator.

To illustrate some moment estimators, first consider the exponential distribution with p.d.f. $f(x) = \lambda \exp(-\lambda x)$ introduced in Equation (2.51). The moment estimator for the rate λ is $\hat{\lambda} = 1/\bar{X}$ as $E[X] = 1/\lambda$. This result is not immediately obvious from a casual examination of the distribution.

Second, consider the normal distribution with the mean μ and variance σ^2. The population moments and their sample estimators are

$$\mu = E[X] \quad \text{and} \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$\sigma^2 = E[X^2] - \mu^2 \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \hat{\mu}^2. \tag{3.11}$$

Note that the variance is the central second moment. The moment-based variance estimator is not unbiased, as the unbiased variance has a factor $1/(n-1)$ rather than $1/n$ before the summation.

3.4.2 Method of least squares

As discussed in our historical overview in Section 1.2.2, parameter estimation using least squares was developed in the early nineteenth century to solve problems in celestial mechanics, and has since been very widely used in astronomy and other fields. We discuss least-squares estimation extensively for regression problems in Section 7.3, and give only a brief introduction here.

Consider estimation of the population mean μ. The least-squares estimator $\hat{\mu}$ is obtained by minimizing the sum of the squares of the differences $(X_i - \mu)$,

$$\hat{\mu}_{LS} = \arg\min_{\mu} \sum_{i=1}^{n} (X_i - \mu)^2; \tag{3.12}$$

that is, $\hat{\mu}_{LS}$ is the value of μ that minimizes $\sum_{i=1}^{n}(X_i - \mu)^2$. We can derive this as follows:

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \mu)^2 &= \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2 + 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \bar{X}) \\ &= \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2\end{aligned}$$

To show the unbiasedness of S_x^2, note that

$$\begin{aligned}E\left[\sum_{i=1}^{n}(X_i - \mu)^2\right] &= \sum_{i=1}^{n} E[(X_i - \mu)^2] = n\sigma^2 \\ E[\bar{X} - \mu]^2 &= Var(\bar{X}) = \frac{1}{n}\sigma^2 \\ E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] &= n\sigma^2 - \sigma^2 = (n-1)\sigma^2.\end{aligned} \tag{3.13}$$

We thus find that $S_X^2 = 1/(n-1)\sum_{i=1}^n (X_i - \bar{X})^2$ satisfies $E[S_X^2] = \sigma^2$, demonstrating that the sample variance S_X^2 is an unbiased estimator of the population variance σ^2. If we use instead $T^2 = 1/n\sum_{i=1}^n (X_i - \bar{X})^2$ as an estimator of σ^2, then $E[T^2] = [(n-1)/n]\sigma^2 \neq \sigma^2$ showing that T^2 is a biased estimator of the variance.

In this simple case, $\hat{\mu} = \bar{X}$ which is the intuitive solution. But in more complex estimation problems, particularly in the context of regression with a functional relationship between two or more variables, this method provides solutions that are not intuitively obvious. Consider the linear regression $Y_i = \xi_i + \epsilon_i$ where ϵ_i are random variables with mean zero and $\xi_i = \sum_{j=1}^k a_{ij}\beta_j$ is a known linear combination of parameters $(\beta_1, \beta_2, \ldots, \beta_k)$. The estimators of β_j can be obtained by minimizing the sum of squares of $(Y_i - \xi_i)$ provided all the ϵ_i variables have the same variance (homoscedastic errors). If the error variances σ_i^2 of X_i are also different (heteroscadastic), then one can minimize the weighted sum of squares

$$\sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left(X_i - \sum_{j=1}^{k} a_{ij}\beta_j \right)^2, \tag{3.14}$$

over $\beta_1, \ldots, \beta_k$. This is called the weighted least-squares method and is related to some procedures astronomers call **minimum χ^2 regression**. Least-squares regression weighted by measurement error is discussed in detail in Section 7.4.

3.4.3 Maximum likelihood method

British mathematicians were actively discussing theoretical and practical approaches to estimation during the early twentieth century. The most brilliant was R. A. Fisher, starting with his critique of least squares as an undergraduate in 1912. He advocated that one should instead calculate the "chance of a given set of observations occurring" as the product of individual probabilities given the data and the model, stating that "the most probable set of values for the θ's [model parameters] will make P [the product of probabilities, later called the likelihood] a maximum". During the next decade, he actively criticized Pearson's χ^2 procedure, as well as the classical method of moments and least squares, promoting instead methods that give the most probable outcome.

In a crucial paper, Fisher (1922) clearly formulated the principle underlying maximum likelihood: "The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observation should be that observed." This paper introduced the concepts of consistency, sufficiency, efficiency and information. Aldrich (1997) gives an interesting historical account of this decade in the history of statistics. Even today, Fisher's 1922 arguments play a fundamental role in much of the conceptual and operational methodology for statistical inference.

The method is based on the "likelihood", the probability density (or for discrete distributions, the probability mass) function viewed as a function of the data given particular values of the model parameters. Here we use the notation $f(\cdot;\theta)$ for a probability density with parameter θ. For example, for an exponential random variable, $f(\cdot;\theta)$ is given by $f(x;\theta) = \theta\exp(-\theta x)$ for $x > 0$, and $f(x;\theta) = 0$ for $x \leq 0$. For i.i.d. random variables

$X_1, X_2, \ldots, X_n$ with a common density function $f(\cdot;\theta)$, the likelihood L and loglikelihood ℓ are given by

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(X_i;\theta) \\ \ell(\theta) &= \ln L(\theta) = \sum_{i=1}^{n} \ln f(X_i;\theta) \end{aligned} \tag{3.15}$$

where "ln" represents the natural logarithm. The likelihood at parameter θ is thus the product of densities evaluated at all the random variables. The sample realization of the likelihood is

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) \tag{3.16}$$

when $(x_1, x_2, \ldots, x_n)$ are the observed data. It is usually computationally easier to work with ℓ than L.

To illustrate maximum likelihood estimation, let us consider a situation where the data follow the geometric distribution. This is a discrete analog of the exponential distribution. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sequence of random variables with probability mass function

$$f(x;p) = (1-p)^{x-1}p, \quad \text{where } x = 1, 2, 3, \ldots \tag{3.17}$$

and $0 < p < 1$. The likelihood function is given by

$$\begin{aligned} L(p) &= (1-p)^{X_1-1}p \ldots (1-p)^{X_n-1}p \\ &= p^n(1-p)^{\sum X_i - n}. \end{aligned} \tag{3.18}$$

Defining $\ell(p) = \ln L(p)$, the maximum likelihood estimator for the parameter p is obtained by equating the derivative of $\ell(p)$ to zero,

$$\frac{d\ell(p)}{dp} = \frac{n}{p} - \frac{(\sum_{i=1}^{n} X_i) - n}{1-p} = 0. \tag{3.19}$$

This leads to the solution

$$\hat{p}_{MLE} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}}. \tag{3.20}$$

If the actual data were $3, 4, 7, 2, 9$ with $n = 5$, then the MLE of p would be $p_{MLE} = 5/(3+4+7+2+9) = 1/5 = 0.2$. If all of the observed values are equal to 1, then there is no MLE.

MLEs have many strong mathematical properties. For most probability structures considered in astronomy, the MLE exists and is unique. As we saw for the normal variance, the MLE $\hat{\theta}$ is usually consistent but may not be unbiased. But this can often be overcome by multiplying $\hat{\theta}$ by a constant. Another property is that, for many nice functions g of the parameter, $g(\hat{\theta})$ is the MLE of $g(\theta)$, whenever $\hat{\theta}$ is the MLE of θ. A crucial property is that, for many commonly occurring situations, maximum likelihood parameter estimators $\hat{\theta}$ have an approximate normal distribution when n is large. This asymptotic normality is

very useful for estimating confidence intervals for MLE parameters. In most cases, the MLE estimator satisfies

$$\begin{aligned} \hat{\theta} &\doteq \theta \\ Var(\hat{\theta}) &\doteq \frac{1}{I(\theta)} \quad \text{where} \\ I(\theta) &= nE\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right)^2, \end{aligned} \tag{3.21}$$

where the symbol $\doteq$ means "approximately". $I(\theta)$ is called the **Fisher information**. When θ is a vector of parameters, this is the **Fisher information matrix** with off-diagonal terms of the form $\partial^2 f/\partial\theta_i\partial\theta_j$.

3.4.4 Confidence intervals

The confidence interval of a parameter θ, a statistic derived from a dataset X, is defined by the range of lower and upper values $[l(X), u(X)]$ that depend on the variable(s) X defined such that

$$P[l(X) < \theta < u(X)] = 1 - \alpha \tag{3.22}$$

where $0 < \alpha < 1$ is usually a small value like $\alpha = 0.05$ or 0.01. That is, if θ is the true parameter, then the **coverage probability** that the interval $[l(X), u(X)]$ contains θ is at least $1 - \alpha$. The quality of confidence intervals is judged using criteria including validity of the coverage probability, optimality (the smallest interval possible for the sample size), and invariance with respect to variable transformations. For $\alpha = 0.05$, the estimated 95% confidence interval of an estimator of some parameter θ is an interval (l, μ) such that $P(l < \hat{\theta} < u) = 0.95$. If the experiment were repeated 100 times, an average of 95 intervals obtained will contain the parameter value θ.

The idea is illustrated with a simple example for the random variables $X_1, X_2, \ldots, X_n$ drawn from a normal distribution with mean μ and variance 1. Then $\bar{X}$ is a good estimator of μ, its variance is $1/n$ and $\sqrt{n}(\bar{X} - \mu)$ has exactly the normal distribution with mean zero and unit variance. This can be expressed in two ways,

$$\begin{aligned} P(-1.96 < \sqrt{n}(\bar{X} - \mu) < 1.96) &= 0.95, \\ P(\bar{X} - 1.96/\sqrt{n} < \mu < \bar{X} + 1.96/\sqrt{n}) &= 0.95 \end{aligned} \tag{3.23}$$

for all values of μ. Thus, $(\bar{X} - 1.96/\sqrt{n}, \bar{X} + 1.96/\sqrt{n})$ is the 95% confidence interval for μ. A confidence interval for the variance can be similarly derived based on the χ^2 distribution which applies to the variables that are squares of normally distributed variables.

If there are two or more unbiased estimators, the one with the smaller variance is often preferred. Under some regularity conditions, the **Cramér–Rao inequality** gives a lower bound on the minimum possible variance for an unbiased estimator. It states that if $\hat{\theta}$ is an unbiased estimator based on i.i.d. random variables $X_1, X_2, \ldots, X_n$ with a common density function $f(\cdot;\theta)$ where θ is a parameter, then the smallest possible value that $Var(\hat{\theta})$ can attain is $1/I(\theta)$ where I is the Fisher information in Equation (3.21).

Consider again the situation where X is exponentially distributed with density f that is,

$$f(x;\theta) = \theta^{-1}\exp(-x/\theta), \quad x > 0. \tag{3.24}$$

A simple application of the MLE shows that $\hat{\theta} = \bar{\mathbf{X}}$ and the Fisher information is $I(\theta) = n\theta^{-2}$. From the Cramér–Rao inequality, the smallest possible value of the variance of an estimator of θ is θ^2/n. This is attained by $\bar{X}$; hence, $\bar{X}$ is the best possible unbiased estimator of θ. It is the minimum variance unbiased estimator (MVUE).

A subtlety, often missed by astronomers, is that the resulting MLE confidence intervals on the mean depend on a precise statement of what is known in advance about other parameters of the problem, the variance σ^2 and the sample size n. The $100(1-\alpha)\%$ confidence interval for μ is

$$\left[\bar{X} - \frac{1}{\sqrt{n}}\, c_{\alpha/2}, \bar{X} + \frac{1}{\sqrt{n}}\, c_{\alpha/2}\right],$$

where

$$c_b = \begin{cases} z_b\,\sigma & \text{if } \sigma \text{ is known} \\ z_b\,S_X & \text{if } \sigma \text{ is unknown and } n \text{ is large} \\ t_b(n-1)\,S_X & \text{if } \sigma \text{ is unknown and } n \text{ is small.} \end{cases} \tag{3.25}$$

Here $S_X = 1/(n-1)\sum_{i=1}^{n}(X_i - \bar{X})^2$, and z_b, t_b denote respectively the numbers such that $P(Z > z_b) = P(T_m > t_b(m)) = \alpha$, where Z has a standard normal distribution and T_m has a t-distribution with m degrees of freedom. The middle of the three solutions for c_b is most commonly used, but is not always the appropriate solution.

Similarly, the confidence intervals for differences of two means are not always the simple case of quadratic error propagation commonly used by astronomers (Bevington 1969). Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be two independent samples from two normal populations with means μ and ν and variances σ^2 and τ^2, respectively. Let $\bar{Y}$ denote the sample mean of the Y's, and define

$$S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2, \quad S_p^2 = \frac{1}{n+m-2}\left[\{(n-1)S_X^2 + (m-1)S_Y^2\}\right].$$

The maximum likelihood estimator for the difference between the two means $\mu - \nu$ is $\bar{X} - \bar{Y}$, as expected, but the $100(1-\alpha)\%$ confidence interval for the difference depends on the situation with the other parameters:

$$\left[\bar{X} - \bar{Y} - \frac{1}{\sqrt{n}}\, d_b(\alpha/2, m, n), \bar{X} - \bar{Y} + \frac{1}{\sqrt{n}}\, d_b(\alpha/2, m, n)\right],$$

where

$$d_b(m,n) = \begin{cases} z_b\sqrt{\frac{\sigma^2}{n} + \frac{\tau^2}{m}} & \text{if } \sigma, \tau \text{ are known} \\ z_b\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}} & \text{if } \sigma, \tau \text{ are unknown and } n, m \text{ are large} \\ t_b(n+m-2)\,S_p\sqrt{\frac{1}{n} + \frac{1}{m}} & \text{if } \sigma = \tau \text{ but the common value is} \\ & \quad \text{unknown and } n, m \text{ are small.} \end{cases} \tag{3.26}$$

The confidence limits above are **two-sided** where the scientific question permits values either higher or lower than the best-fit parameter value. But sometimes the scientific question

involves an asymmetrical **one-sided** confidence limit. The one-sided lower $100(1-\alpha)\%$ confidence interval for μ is given by $[\bar{X} - \frac{1}{\sqrt{n}}\, c_\alpha, \infty)$. Similarly, the one-sided lower $100(1-\alpha)\%$ confidence interval for $\mu - \nu$ is $[\bar{X} - \bar{Y} - d_\alpha(m,n), \infty)$.

Finally, we consider the problem of estimating a proportion, or ratio, in a binary trial experiment. While typically phrased in terms of the proportion of heads and tails from flipping a coin, the issue arises in astronomy when we seek the fraction of quasars that are radio-loud, the ratio of brown dwarfs to stars, or the "hardness" ratio of photons above and below some critical energy. Let y be the number of successes in n Bernoulli trials with probability $0 < p < 1$ of success on each trial. The best-fit value (e.g. by MLE) for the unknown fraction p is easily found to be $p = y/n$, the intuitive value.

But it is less easy to obtain the confidence intervals for such a proportion. When n is large and asymptotic normality applies, the $100(1-\alpha)\%$ approximate confidence interval for p is

$$\left[\frac{y}{n} - z_{\alpha/2}\sqrt{\frac{1}{n}\left(\frac{y}{n}\left(1-\frac{y}{n}\right)\right)},\ \frac{y}{n} + z_{\alpha/2}\sqrt{\frac{1}{n}\left(\frac{y}{n}\left(1-\frac{y}{n}\right)\right)}\right]. \tag{3.27}$$

The case of the difference, rather than the ratio, between two Bernoulli trial experiments can also be considered. Let y_1 and y_2 be two independent binomial random variables with parameters p_1 and p_2, and the corresponding number of trials n_1 and n_2. Then the estimate of the difference is

$$\widehat{p_1 - p_2} = y_1/n_1 - y_2/n_2 \tag{3.28}$$

and the $100(1-\alpha)\%$ approximate confidence interval is given by

$$\left[\frac{y_1}{n_1} - \frac{y_2}{n_2} - z_{\alpha/2}\, a(n_1,n_2,y_1,y_2),\ \frac{y_1}{n_1} - \frac{y_2}{n_2} + z_{\alpha/2}\, a(n_1,n_2,y_1,y_2)\right],$$

where

$$a(n_1,n_2,y_1,y_2) = \sqrt{\frac{1}{n_1}\left(\frac{y_1}{n_1}\left(1-\frac{y_1}{n_1}\right)\right) + \frac{1}{n_2}\left(\frac{y_2}{n_2}\left(1-\frac{y_2}{n_2}\right)\right)}. \tag{3.29}$$

We will return to this and similar problems involving ratios of discrete distributions in Sections 3.8 and 4.1.1 to illustrate the subtleties and options that can arise in estimation even for easily stated problems that often appear in astronomy.

3.4.5 Calculating MLEs with the EM algorithm

Likelihoods can be maximized by any numerical optimization method. During the mid-twentieth century before computers, simple analytical statistical models were emphasized where the maximum of the likelihood function could be obtained by differential calculus. For a model with p parameters $\theta_1, \theta_2, \ldots, \theta_p$, the equations

$$\frac{\partial L(\theta_i)}{\partial \theta_i} = 0 \tag{3.30}$$

often gave a system of p equations in p unknowns that could be solved using algebraic techniques.

Alternatively, the maximum of the likelihood could be found numerically using iterative numerical techniques like the Newton–Raphson and gradient descent methods and their modern variants (e.g. Levenberg–Marquardt, Davidon–Fletcher–Powell, and Broyden–Fletcher–Goldfarb–Shanno methods; Nocedal & Wright 2006). These techniques may converge slowly, encountering problems when the derivative is locally unstable and when the likelihood function has multiple maxima. Other techniques, such as simulated annealing and genetic algorithms, are designed to assist in finding the global maximum in likelihood functions with complex structure.

One particular numerical procedure, the **EM algorithm**, has been enormously influential in promoting maximum likelihood estimation since the seminal papers of Dempster *et al.* (1977) and Wu (1983). In accord with the statisticians' concern with modeling uncertain and incomplete data to describe a hidden phenomenon, the EM algorithm considers the mapping of a set of datasets to an unknown complete dataset. The method was independently developed in astronomy for image deconvolution by Richardson (1972) and Lucy (1974). Here the data are the observed image of the sky blurred by the telescope's optics, the model is the telescope point spread function, and the missing dataset is the true sky image.

One begins the EM algorithm with initial values of the model parameter values θ and the dataset. These might be estimated by least squares, or represent guesses by the scientist. The algorithm proceeds by iteration of two steps. The **expectation step** (E) calculates the likelihood for the current values of the parameter vector θ. The **maximization step** (M) updates the missing dataset values with the criterion that the likelihood of the values with respect to the current model is maximized. This updated dataset then takes the place of the original dataset, and the algorithm is iterated until convergence. In many situations, maximization of the complete data is easier than the original incomplete data.

The algorithm is successful for many MLE problems because each iteration is guaranteed to increase the likelihood over the previous iteration. Local minima are ignored and convergence is usually rapid. However, there is still no guarantee that the achieved maximum is global over the full parameter space. Research on accelerating the EM algorithm and improving its convergence for maximum likelihood estimation is still actively pursued (McLachlan & Krishnan 2008).

The EM algorithm has been fruitful for many types of MLE calculations even when missing data are not obviously present. These include and linear and nonlinear regression (Chapter 7), normal (Gaussian) mixture models, pattern recognition and image deconvolution, modeling binned or categorical data, multivariate analysis and classification (Chapters 8 and 9), modeling censored and truncated data (Chapter 10) and time series anlaysis (Chapter 11).

3.5 Hypothesis testing techniques

We now turn to statistical testing of hypotheses. This is formulated in terms of deciding between two competing statements, H_0 and H_a, called the **null hypothesis** and **alternative**

hypothesis, based on the data. There are two possible errors in adjudicating between these hypotheses:

Type 1 error Here one wrongly rejects the null hypothesis H_0 giving a **false positive** decision. For example, when searching for a faint signal in noise, this occurs when we incorrectly infer that a signal is present when it truly is not.

Type 2 error Here one fails to reject the null hypothesis when the alternative is true, giving a **false negative** decision. In our example, we would incorrectly infer that a signal is absent when it truly is present.

Ideally, one likes to minimize these two errors to negligible levels, but it is not possible to achieve this. So the scientist must decide what errors are more important for the goals of the test.

A traditional choice is to construct the critical regions to keep Type 1 errors under control at the 5% level, allowing Type 2 errors to be uncontrolled. This choice of 5% is called the **significance level** of the hypothesis test, and represents the probability of generating false positives; that is, incorrectly rejecting the null hypothesis. The **power** of a test is the probability of correctly rejecting the null when the alternative hypothesis is true. The power is $1 - \beta$ where β here is the Type 2 error or false negative rate. The **uniformly most powerful** (UMP) test is the test statistic that give the highest power for all parameter values for a chosen significance level. This is often the preferred test statistic.

A result of a hypothesis test is called **statistically significant** if it is unlikely to have occurred by chance. That is, the hypothesis is significant at level α if the test rejects the null hypothesis at the prescribed significance level α. Typical significance levels used in many fields are $\alpha = 0.05$ or 0.01. Note that the common standard in astronomy of 3σ, where σ is the standard deviation corresponding to $\alpha = 0.003$ for the normal distribution.

Along with the binary "Yes/No" results of a statistical test, the so-called ***p*-value** is often reported. The p-value of a hypothesis test is the probability, assuming the null hypothesis is true, of observing a result at least as extreme as the value of the test statistic. It is important to note that the null hypothesis and the alternative hypothesis are not treated symmetrically. We can only reject the null hypothesis at a given level of significance; we can never accept a null hypothesis. In the case of signal detection, rejecting the null hypothesis leads to detecting a signal. So it is often the case that the alternative hypothesis is chosen as the statement for which there is likely to be supporting evidence.

A common difficulty is that significance levels must be adjusted when many hypothesis tests are conducted on the same dataset. This situation often occurs in astronomical image analysis. A large image or data cube may be searched at millions of locations for faint sources, so that one must seek a balance between many false positives and sensitivity. A new procedure for combining multiple hypothesis tests called the **false detection rate** provides a valuable way to control for false positives (Benjamini & Hochberg 1995).

For a two-sided hypothesis test $H_0 : \theta = \theta_0$ vs. $H_a : \theta \neq \theta_0$, the set theoretic complement of the rejection region at level of significance α serves as the $(1 - \alpha)100\%$ confidence region for θ. Recall that z_α and $t_\alpha(m)$ denote, respectively, the numbers such that $P(Z > z_\alpha) = P(T_m > t_\alpha(m)) = \alpha$, where Z has the standard normal distribution and T_m has the t distribution with m degrees of freedom.

Table 3.1 Hypotheses for one proportion ($H_0: p = p_0$).

	H_1	Critical region
$x = \dfrac{(y/n) - p_0}{\sqrt{p_0(1-p_0)/n}}$,	$p > p_0$	$x \geq z_\alpha$
	$p < p_0$	$x \leq -z_\alpha$
	$p \neq p_0$	$\lvert x \rvert \geq z_{\alpha/2}$

Table 3.2 Hypotheses for two proportions ($H_0: p_1 = p_2$). $\hat{p}_1 = \frac{y_1}{n_1}$, $\hat{p}_2 = \frac{y_2}{n_2}$, $\hat{p} = \frac{y_1+y_2}{n_1+n_2}$

	H_1	Critical region
$x = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$,	$p_1 > p_2$	$x \geq z_\alpha$
	$p_1 < p_2$	$x \leq -z_\alpha$
	$p_1 \neq p_2$	$\lvert x \rvert \geq z_{\alpha/2}$

Table 3.3 Hypotheses for one mean for normal data ($H_0: \mu = \mu_0$).

H_1	Critical region, σ known	Critical region, variance unknown
$\mu > \mu_0$	$\bar{X} \geq \mu_0 + z_\alpha \sigma/\sqrt{n}$	$\bar{X} \geq \mu_0 + t_\alpha(n-1) S_X/\sqrt{n}$
$\mu < \mu_0$	$\bar{X} \leq \mu_0 - z_\alpha \sigma/\sqrt{n}$	$\bar{X} \leq \mu_0 - t_\alpha(n-1) S_X/\sqrt{n}$
$\mu \neq \mu_0$	$\lvert \bar{X} - \mu_0 \rvert \geq z_{\alpha/2}\sigma/\sqrt{n}$	$\lvert \bar{X} - \mu_0 \rvert \geq t_{\alpha/2}(n-1) S_X/\sqrt{n}$

Table 3.4 Hypotheses for one mean for nonnormal data when n is large.

H_0	H_1	Critical region
$\mu = \mu_0$	$\mu > \mu_0$	$\bar{X} \geq \mu_0 + z_\alpha S_X/\sqrt{n}$
$\mu = \mu_0$	$\mu < \mu_0$	$\bar{X} \leq \mu_0 - z_\alpha S_X/\sqrt{n}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert \bar{X} - \mu_0 \rvert \geq z_{\alpha/2} S_X/\sqrt{n}$

Critical regions (also called **rejection regions**) for some commonly used tests of hypotheses involving proportions, means and variances, are given in Tables 3.1–3.8. The tables provide inequalities involving values of statistics computed from the data. Some of the critical regions are based on the Central Limit Theorem. In these tables, $\bar{X}$ is the sample mean of n i.i.d random variables from a population with mean μ_X and variance σ^2, and $\bar{Y}$ is the sample mean of m i.i.d random variables, independent of the X's, from a population with mean μ_Y and variance τ^2. S_X^2 denotes the sample variance.

In the testing for a single proportion (Table 3.1), the denominator $\sqrt{p_0(1-p_0)/n}$ can be replaced by $\sqrt{(y/n)(1-(y/n))/n}$. Similarly in testing for equality of two proportions (Table 3.2), the denominator of z can be replaced by $\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$. Here the y, y_1 and y_2 values are observations from the binomial distribution with population proportions p, p_1 and p_2 with n, n_1 and n_2 trials, respectively.

If both the X's and Y's are from normal populations and m and n are not large, then asymptotic theory does not apply and different critical regions of the hypothesis tests are

Table 3.5 Hypotheses for the equality of two means when n and m are large.

H_0	H_1	Critical region
$\mu_X = \mu_Y$	$\mu_X > \mu_Y$	$\bar{X} - \bar{Y} \geq z_\alpha \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$
$\mu_X = \mu_Y$	$\mu_X < \mu_Y$	$\bar{X} - \bar{Y} \leq -z_\alpha \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$
$\mu_X = \mu_Y$	$\mu_X \neq \mu_Y$	$\lvert\bar{X} - \bar{Y}\rvert \geq z_{\alpha/2} \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$

Table 3.6 Confidence regions for common small-sample hypothesis tests. Hypotheses for the equality of two means for normal populations.

H_0	H_1	Critical region
$\mu_X = \mu_Y$	$\mu_X > \mu_Y$	$\bar{x} - \bar{y} \geq d_\alpha(m, n)$
$\mu_X = \mu_Y$	$\mu_X < \mu_Y$	$\bar{x} - \bar{y} \leq -d_\alpha(m, n)$
$\mu_X = \mu_Y$	$\mu_X \neq \mu_Y$	$\lvert\bar{x} - \bar{y}\rvert \geq d_{\alpha/2}(m, n)$

Table 3.7 Hypotheses for one variance (H_0: $\sigma^2 = \sigma_0^2$).

H_1	Critical region
$\sigma^2 > \sigma_0^2$	$(n-1)S_X^2 \geq \sigma_0^2 \chi_\alpha^2(n-1)$
$\sigma^2 < \sigma_0^2$	$(n-1)S_X^2 \leq \sigma_0^2 \chi_{1-\alpha}^2(n-1)$
$\sigma^2 \neq \sigma_0^2$	$(n-1)S_x^2 \geq \sigma_0^2 \chi_{\alpha/2}^2(n-1)$
	or $(n-1)S_x^2 \leq \sigma_0^2 \chi_{1-\alpha/2}^2(n-1)$

Table 3.8 Hypotheses for the equality of two variances (H_0: $\sigma_X^2 = \sigma_Y^2$).

H_1	Critical region
$\sigma_X^2 > \sigma_Y^2$	$S_X^2/S_Y^2 \geq F_\alpha(n-1, m-1)$
$\sigma_X^2 < \sigma_Y^2$	$S_Y^2/S_X^2 \geq F_\alpha(m-1, n-1)$
$\sigma_X^2 \neq \sigma_Y^2$	$S_X^2/S_Y^2 \geq F_{\alpha/2}(n-1, m-1)$
	or $S_Y^2/S_X^2 \geq F_{\alpha/2}(m-1, n-1)$

used as given in Table 3.6. Here the quantity $d_b(m, n)$ is defined in Equation (3.26). Recall that if $X_1, X_2, \ldots, X_n$ are i.i.d. normal random variables with mean μ and variance σ^2, then $(n-1)S_X^2/\sigma^2$ has a χ^2 distribution with $n-1$ degrees of freedom. This fact is used to construct tests concerning variances (Tables 3.7–3.8). Here we let $\chi_\beta^2(m)$ denote the number such that the probability that a χ^2 random variable with m degrees of freedom exceeds $\chi_\beta^2(m)$ is β. Finally, for the test of equality of variances of two populations, we rely on the property that the ratio of two independent χ^2 random variables, normalized by

their degrees of freedom, follows an F distribution. Thus, if $\sigma_X^2 = \sigma_Y^2$, then S_X^2/S_Y^2 has an F distribution with $(n-1, m-1)$ degrees of freedom. Let $F_\beta(n-1, m-1)$ be such that $P(F \geq F_\beta(n-1, m-1)) = \beta$.

3.6 Resampling methods

The classical statistical methods of earlier sections concentrated mainly on the statistical properties of the estimators that have a simple closed form and which can be analyzed mathematically. Except for a few important but simple statistics, these methods often involve unrealistic model assumptions. While it is often relatively easy to devise a statistic that measures a property of scientific interest, it is almost always difficult or impossible to determine the distribution of that statistic.

These limitations have been overcome in the last two decades of the twentieth century with a class of computationally intensive procedures known as **resampling methods** that provide inferences on a wide range of statistics under very general conditions. Resampling methods involve constructing hypothetical populations derived from the observations, each of which can be analyzed in the same way to see how the statistics depend on plausible random variations in the observations. Resampling the original data preserves whatever distributions are truly present, including selection effects such as truncation and censoring.

The **half-sample method** may be the oldest resampling method, where one repeatedly chooses at random half of the data points, and estimates the statistic for each resample. The inference on the parameter can be based on the histogram of the resampled statistics. It was used by P. C. Mahalanobis in 1946 under the name **interpenetrating samples**. An important variant is the Quenouille–Tukey **jackknife method**. For a dataset with n data points, one constructs exactly n hypothetical datasets each with $n-1$ points, each one omitting a different point. The most important of the resampling methods proved to be the **bootstrap** or resampling with replacement. B. Efron introduced the bootstrap method in 1979. Here one generates a large number of datasets, each with n data points randomly drawn from the original data. The constraint is that each drawing is made from the entire dataset, so a simulated dataset will miss some points and have duplicates or triplicates of others. Thus, the bootstrap can be viewed as a Monte Carlo method to simulate from existing data, without any assumption about the underlying population.

3.6.1 Jackknife

The jackknife method was introduced by M. Quenouille (1949) to estimate the bias of an estimator. The method was later shown to be useful in reducing the bias as well as in estimating the variance of an estimator. Let $\hat{\theta}_n$ be an estimator of θ based on n i.i.d. random vectors $X_1, \ldots, X_n$. That is, $\hat{\theta}_n = f_n(X_1, \ldots, X_n)$, for some function f_n. Let

$$\hat{\theta}_{n,-i} = f_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \tag{3.31}$$

be the corresponding recomputed statistic based on all but the i-th observation. The jackknife estimator of bias $E(\hat{\theta}_n) - \theta$ is given by

$$bias_J = \frac{(n-1)}{n} \sum_{i=1}^{n} \left(\hat{\theta}_{n,-i} - \hat{\theta}_n\right). \tag{3.32}$$

The jackknife estimator θ_J of θ is given by

$$\theta_J = \hat{\theta}_n - bias_J = \frac{1}{n} \sum_{i=1}^{n} \left(n\hat{\theta}_n - (n-1)\hat{\theta}_{n,-i}\right). \tag{3.33}$$

Such a bias-corrected estimator hopefully reduces the overall bias of the estimator. The summands above

$$\theta_{n,i} = n\hat{\theta}_n - (n-1)\hat{\theta}_{n,-i},\ i = 1, \ldots, n \tag{3.34}$$

are called **pseudo-values**.

In the case of the sample mean $\hat{\theta}_n = \bar{X}_n$, it is easy to check that the pseudo-values are simply

$$\theta_{n,i} = n\hat{\theta}_n - (n-1)\hat{\theta}_{n,-i} = X_i,\ i = 1, \ldots, n. \tag{3.35}$$

This provides motivation for the jackknife estimator of the variance of $\hat{\theta}_n$,

$$\begin{aligned} var_J(\hat{\theta}_n) &= \frac{1}{n(n-1)} \sum_{i=1}^{n} (\theta_{n,i} - \theta_J)(\theta_{n,i} - \theta_J)' \\ &= \frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}_{n,-i} - \bar{\theta}_n)(\hat{\theta}_{n,-i} - \bar{\theta}_n)', \end{aligned} \tag{3.36}$$

where $\bar{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{n,-i}$. For most statistics, the jackknife estimator of the variance is consistent; that is

$$Var_J(\hat{\theta}_n)/Var(\hat{\theta}_n) \to 1, \tag{3.37}$$

as $n \to \infty$ almost surely. In particular, this holds for a **smooth functional model**. To describe this, let the statistic of interest, $\hat{\theta}_n$, based on n data points be defined by $H(\bar{Z}_n)$, where $\bar{Z}_n$ is the sample mean of the random vectors $Z_1, \ldots, Z_n$ and H is continuously differentiable in a neighborhood of $E(\bar{Z}_n)$. Many commonly occurring statistics fall under this model, including: sample means, sample variances, central and noncentral t statistics (with possibly nonnormal populations), sample coefficient of variation, least-squares estimators, maximum likelihood estimators, correlation coefficients, regression coefficients and smooth transforms of these statistics.

However, consistency does not always hold; for example, the jackknife method fails for nonsmooth statistics such as the sample median. If $\hat{\theta}_n$ denotes the sample median in the univariate case, then in general,

$$Var_J(\hat{\theta}_n)/Var(\hat{\theta}_n) \to \left(\frac{1}{2}\chi_2^2\right)^2 \tag{3.38}$$

in distributions where χ^2_2 denotes a *chi-square* random variable with two degrees of freedom (Efron 1982, Section 3.4). So in this case, the jackknife method does not lead to a consistent estimator of the variance. However, the bootstrap resampling method does lead to a consistent estimator for this case.

3.6.2 Bootstrap

Bootstrap resampling constructs datasets with n points, rather than $n-1$ for the jackknife, where each point was selected from the full dataset; that is, resampling with replacement. The importance of the bootstrap emerged during the 1980s when mathematical study demonstrated that it gives a nearly optimal estimate of the distribution of many statistics under a wide range of circumstances, including the smooth function models listed above for the jackknife. In several cases, the method yields better results than those obtained by the classical normal approximation theory. However, one should caution that bootstrap is not the solution for all problems. The theory developed in the 1980s and 1990s shows that bootstrap fails in some nonsmooth situations. Hence, caution should be used and one should resist the temptation to use the method inappropriately.

We first describe an application of the bootstrap to estimate the variance of the sample mean. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be data drawn from an unknown population distribution F. Suppose $\hat{\theta}_n$, based on the data $\mathbf{X}$, is a good estimator of θ, a parameter of interest. The interest lies in assessing its accuracy in estimation. Determining the confidence intervals for θ requires knowledge of the sampling distribution G_n of $\hat{\theta}_n - \theta$; that is, $G_n(x) = \mathrm{P}(\hat{\theta}_n - \theta \leq x)$, for all x. The sample mean $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ is a good estimator of the population mean μ. To get the confidence interval for μ, we must find the sampling distribution of $\bar{X}_n - \mu$ which depends on the shape and other characteristics of the unknown distribution F.

Classical statistical theory applies the normal approximation obtained from the Central Limit Theorem to the sampling distribution. The problem is that, even if the sampling distribution is not symmetric, the CLT approximates using a normal distribution, which is symmetric. This can be seen from the following example. If $(X_1, Y_1), \ldots, (X_n, Y_n)$ denote observations from a bivariate normal population, then the maximum likelihood estimator of the correlation coefficient ρ is given by **Pearson's linear correlation coefficient**,

$$\hat{\rho}_n = \frac{\sum_{i=1}^n (X_i Y_i - \bar{X}_n \bar{Y}_n)}{\sqrt{\left(\sum_{i=1}^n (X_i - \bar{X}_n)^2\right)\left(\sum_{i=1}^n (Y_i - \bar{Y}_n)^2\right)}}. \tag{3.39}$$

For statistics with asymmetrical distributions, such as that of $\hat{\rho}_n$, the classical theory suggests variable transformations. In this case, **Fisher's Z transformation Z** given by

$$Z = \frac{\sqrt{(n-3)}}{2}\left(\ln\left(\frac{1+\hat{\rho}_n}{1-\hat{\rho}_n}\right) - \ln\left(\frac{1+\rho}{1-\rho}\right)\right) \tag{3.40}$$

gives a better normal approximation. This approximation corrects skewness and is better than the normal approximation of $\sqrt{n}(\hat{\rho}_n - \rho)$. The bootstrap method, when properly used,

avoids such individual transformations by taking into account the skewness of the sampling distribution. It automatically corrects for skewness.

The bootstrap method avoids such clumsy transformations for statistics with asymmetrical (or unknown) distributions. The bootstrap presumes that if $\hat{F}_n$ is a good approximation to the unknown population distribution F, then the behavior of the samples from $\hat{F}_n$ closely resemble that of the original data. Here $\hat{F}_n$ can be the empirical distribution function (e.d.f.) or a smoothed e.d.f. of the data $X_1, \ldots, X_n$, or a parametric estimator of the function F. The e.d.f. will be further discussed in Section 5.3.1. Once $\hat{F}_n$ is provided, datasets $\mathbf{X}^* = (X_1^*, \ldots, X_n^*)$ are resampled from $\hat{F}_n$ and the statistic θ^* based on $\mathbf{X}^*$ is computed for each resample. Under very general conditions Babu & Singh (1984) have shown that the difference between the sampling distribution G_n of $\hat{\theta}_n - \theta$ and the **bootstrap distribution** G_b (that is, the distribution of $\theta^* - \hat{\theta}_n$) is negligible. G_b can thus be used to draw inferences about the parameter θ in place of the unknown G_n.

In principle, the bootstrap distribution G_b, which is a histogram, is completely known, as it is constructed entirely from the original data. However, to get the complete bootstrap distribution, one needs to compute the statistics for nearly all of the $M = n^n$ possible bootstrap samples. For the simple example of the sample mean, presumably one needs to compute

$$\begin{aligned}
&X_1^{*(1)}, \ldots, X_n^{*(1)}, \quad r_1 = \bar{X}^{*(1)} - \bar{X} \\
&X_1^{*(2)}, \ldots, X_n^{*(2)}, \quad r_2 = \bar{X}^{*(2)} - \bar{X} \\
&\qquad \ddots \qquad\quad \ddots \; \ddots \qquad \ddots \\
&X_1^{*(M)}, \ldots, X_n^{*(M)}, \; r_M = \bar{X}^{*(M)} - \bar{X}.
\end{aligned}$$

The bootstrap distribution is given by the histogram of $r_1, \ldots, r_M$. Even for $n = 10$ data points, M turns out to be ten billion.

In practice, the statistic of interest, $\theta^* - \hat{\theta}_n$, can be accurately estimated from the histogram of a much smaller number of resamples. Asymptotic theory shows that the sampling distribution of $\theta^* - \hat{\theta}_n$ can be well-approximated by generating $N \simeq n(\ln n)^2$ bootstrap resamples (Babu & Singh 1983). Thus, only $N \sim 50$ simulations are needed for $n = 10$ and $N \sim 50{,}000$ for $n = 1000$. Thus, we can estimate the distribution of the statistic of interest for the original dataset from the histogram of the statistic obtained from the bootstrapped samples.

So far, we have described the most popular and simple bootstrap – the **nonparametric bootstrap** where the resampling with replacement is based on the e.d.f. of the original data. This gives equal weights to each of the original data points. Table 3.9 gives bootstrap versions of some common statistics. In the case of the ratio estimator and the correlation coefficient, the data pairs are resampled from the original data pairs (X_i, Y_i).

Bootstrap resampling is also widely used for deriving confidence intervals for parameters. However, one can only invert the limiting distribution to get confidence intervals when the limiting distribution of the point estimator is free from the unknown parameters. Such

Table 3.9 Statistics and their bootstrap versions.

Statistic	Bootstrap version
Mean, $\bar{X}_n$	$\bar{X}_n^*$
Ratio estimator, $\bar{X}_n/\bar{Y}_n$	$\bar{X}_n^*/\bar{Y}_n^*$
Variance, $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$	$\frac{1}{n}\sum_{i=1}^n (X_i^* - \bar{X}_n^*)^2$
Correlation coefficient, $\frac{\sum_{i=1}^n (X_i Y_i - \bar{X}_n \bar{Y}_n)}{\sqrt{(\sum_{i=1}^n (X_i - \bar{X}_n)^2)(\sum_{i=1}^n (Y_i - \bar{Y}_n)^2)}}$	$\frac{\sum_{i=1}^n (X_i^* Y_i^* - \bar{X}_n^* \bar{Y}_n^*)}{\sqrt{(\sum_{i=1}^n (X_i^* - \bar{X}_n^*)^2)(\sum_{i=1}^n (Y_i^* - \bar{Y}_n^*)^2)}}$

quantities are called **pivotal statistics**. It is thus important to focus on pivotal or approximately pivotal quantities in order to get reliable confidence intervals for the parameter of interest.

Consider the confidence interval of the sample mean. If the data are normally distributed, $X_i \sim N(\mu, \sigma^2)$, then $\sqrt{n}(\bar{X} - \mu)/S_n$ has a t distribution with $n - 1$ degrees of freedom, and hence it is pivotal. In the nonnormal case, it is approximately pivotal where $S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$. To obtain the bootstrap confidence interval for μ, we compute $\sqrt{n}(\bar{X}^{*(j)} - \bar{X})/S_n$ for N bootstrap samples, and arrange the values in increasing order, $h_1 < h_2 < \cdots < h_N$. One can then read from the histogram (say) the 90% confidence interval of the parameter. That is, the 90% confidence interval for μ is given by

$$\bar{X} - h_m \frac{S_n}{\sqrt{n}} \leq \mu < \bar{X} - h_k \frac{S_n}{\sqrt{n}}, \tag{3.41}$$

where $k = [0.05N]$ and $m = [0.95N]$.

It is important to note that even when σ is known, the bootstrap version of $\sqrt{n}(\bar{X} - \mu)/\sigma$ is $\sqrt{n}(\bar{X}^* - \bar{X})/S_n$. One should not replace $\sqrt{n}(\bar{X}^* - \bar{X})/S_n$ by $\sqrt{n}(\bar{X}^* - \bar{X})/\sigma$.

For datasets of realistic size, the sampling distributions of several commonly occurring statistics are closer to the corresponding bootstrap distribution than the normal distribution given by the CLT. If pivotal statistics are used, then the confidence intervals are similarly reliable under very general conditions. The discussion here is applicable to a very wide range of functions that includes sample means and variances, least squares and maximum likelihood estimators, correlation and regression coefficients, and smooth transforms of these statistics.

A good overview of the bootstrap is presented by Efron and Tibshirani (1993), and Zoubir & Iskander (2004) provide a practical handbook for scientists and engineers. Bootstrap methodology and limiting theory are reviewed by Babu & Rao (1993). The bootstrap method has found many applications in business, engineering, biometrics, environmental statistics, image and signal processing, and other fields. Astronomers have started using the bootstrap also. In spite of their many capabilities, one should recognize that bootstrap methods fail under certain circumstances. These include nonsmooth statistics (such as the maximum value of a dataset), heavy-tailed distributions, distributions with infinite variances, and some nonlinear statistics.

3.7 Model selection and goodness-of-fit

The aim of model fitting and parameter estimation is to provide the most parsimonious "best fit" of a mathematical model to data. The model might be a simple, heuristic function to approximate phenomenological relationships between observed properties in a sample; for example, a linear or power-law (Pareto) function. The same procedures of mathematical statistics can be used to link data to complicated astrophysical models. The relevant methods fall under the rubrics of regression, goodness-of-fit, and model selection.

The common procedure has four steps:

1. choose a model family based on astrophysical knowledge or a heuristic model based on exploratory data analysis;
2. obtain best-fit parameters for the model using the methods described in Section 3.3;
3. apply a goodness-of-fit hypothesis test to see whether the best-fit model agrees with the data at a selected significance level; and
4. repeat with alternative models to select the best model according to some quantitative criterion. In this last step, a dataset may be found to be compatible with many models, or with no model at all.

The coupled problems of goodness-of-fit and model selection, steps 3 and 4 above, are among the most common in modern statistical inference. The principles are clear. After a model has been specified and best-fit parameters estimated, a reliable and broadly applicable test for the fit's validity given the dataset is needed. Once a satisfactory model has been found, exploration of alternative models is needed to find the optimal model. A final principle is parsimony: a good statistical model should be among the simplest consistent with the data. This idea dates to the Middle Ages, when the fourteenth-century English philosopher William of Ockham proposed that the simplest solution is usually the correct one. In statistical modeling, **Ockham's Razor** suggests that we leave off extraneous ideas better to reveal the truth. We thus seek a model that neither underfits, excluding key variables or features, nor overfits, incorporating unnecessary complexity. As we will discuss further in Section 6.4.1 on data smoothing, underfitting induces bias and overfitting induces high variability. A model selection criterion should balance the competing objectives of conformity to the data and parsimony.

The model selection problem is more tractable when **nested** models are compared, where the simpler model is a subset of the more elaborated model. In interpreting a continuum astronomical spectrum, for example, modeling might start with a single-temperature blackbody and proceed to multiple temperatures as needed to fit the spectrum. Perhaps the underlying physics is synchrotron rather than blackbody, and the spectrum should be fitted with synchrotron models. Comparison of **nonnested** models, such as thermal and nonthermal, is more difficult than comparison of nested models.

The most common goodness-of-fit procedure in astronomy involves the **reduced χ^2 statistic** (Bevington 1969; Press *et al.* 1997; see Chapter 7). Here the model is first compared to grouped (binned) data and best-fit parameters are obtained by weighted least squares where the weights are obtained from the dataset (e.g. $\sigma_i = \sqrt{N_i}$ where N_i is the number

of objects placed into the i-th bin) or from ancillary measurements (e.g. the noise level in featureless regions of the image, spectrum or time series). Mathematically, this procedure can be expressed as

$$\hat{\theta} = \arg\min_{\theta} X^2(\theta) = \arg\min_{\theta} \sum_{i=1}^{N} \left(\frac{y_i - M_i(\theta)}{\sigma_i} \right)^2 . \tag{3.42}$$

After the **minimum χ^2 parameters** are found, goodness-of-fit is evaluated using a criterion $X^2_\nu \simeq 1$ where ν represents the degrees of freedom in the problem. If X^2_ν is acceptably close to unity based on a chosen significance level and the assumption of an asymptotic χ^2 distribution, then the model is deemed acceptable. Model selection can proceed informally, with the $\chi^2_\nu \simeq 1$ test applied for each model examined. Note that we use the designation X^2 rather than χ^2 in Equation (3.42) because, under many conditions, it may not asymptotically follow the χ^2 distribution. Indeed, as discussed in Section 7.4, the astronomers' common use of minimum χ^2 procedures has a number of problems and is not recommended as a regression procedure under many circumstances. Similar difficulties arise when it is used as a general statistic for goodness-of-fit or model selection.

Both goodness-of-fit and model selection involve statistical hypothesis testing (Section 3.5), first to test model validity assuming a specified significance level, and second to compare two models as the null (H_0) and alternative (H_a) hypotheses. Classical hypothesis testing methods are generally used for nested models. However, they do not treat models symmetrically.

Informative reviews of the discussions on goodness-of-fit methods and model validation can be found in the multi-authored volume edited by Huber-Carol *et al*. (2002). Model selection techniques are discussed in the monograph by Burnham & Anderson (2002) and the review articles in the volume edited by Lahiri (2001). Information criteria are discussed in detail by Konishi & Kitagawa (2008).

3.7.1 Nonparametric methods for goodness-of-fit

The three well-known nonparametric hypothesis tests — based on the Kolmogorov–Smirnov (K-S), Cramér–von Mises, and Anderson–Darling statistics — can be used to compare the empirical distribution function of univariate data to the hypothesized distribution function of the model. These statistics and their corresponding tests are described in Section 5.3.1. They measure different distances between the data and the model, and have powerful theory establishing that their distribution functions are independent of the underlying distributions. We note that e.d.f.-based goodness-of-fit tests are not applicable when the problem has two or more dimensions.

However, for the purpose of the goodness-of-fit, it is important to recognize the distinction between the K-S-type **statistic** and the K-S-type **test** with tabulated probabilities. These widely available probabilities are usually not correct when applied to model-fitting situations where some or all of the parameters of the model are estimated from the same dataset. The standard probabilities of the K-S and related tests are only valid if the parameters are estimated independently of the dataset at hand, perhaps from some previous

datasets or prior astrophysical considerations. The inapplicability of these probabilities was demonstrated by Lilliefors (1969) and was developed further by others (e.g. Babu & Rao 2004). Astronomers are often not aware of this limitation, and are obtaining unreliable goodness-of-fit probabilities from the e.d.f.-based tests.

Fortunately, bootstrap resampling of the data, and repeated calculation of the K-S-type statistic with respect to a single model, gives a histogram of the statistic from which a valid goodness-of-fit probability can be evaluated (Babu & Rao 1993). We outline the methodology underlying these goodness-of-fit bootstrap calculations. Let $\{F(.;\theta) : \theta \in \Theta\}$ be a family of continuous distributions parametrized by θ. We want to test whether the dataset $X_1, \ldots, X_n$ comes from $F = F(.;\theta)$ for some $\theta = \theta_0$. The K-S-type statistics (and a few other goodness-of-fit tests) are continuous functionals of the process,

$$Y_n(x;\hat{\theta}_n) = \sqrt{n}\big(F_n(x) - F(x;\hat{\theta}_n)\big). \tag{3.43}$$

Here F_n denotes the e.d.f. of $X_1, \ldots, X_n$, $\hat{\theta}_n = \theta_n(X_1, \ldots, X_n)$ is an estimator of θ derived from the dataset, and $F(x;\hat{\theta}_n)$ is the model being tested. For a simple example, if $\{F(.;\theta) : \theta \in \Theta\}$ denotes the Gaussian family with $\theta = (\mu, \sigma^2)$, then $\hat{\theta}_n$ can be taken as $(\bar{X}_n, S_n^2)$ where $\bar{X}_n$ is the sample mean and S_n^2 is the sample variance based on the data $X_1, \ldots, X_n$. In modeling an astronomical spectrum, F may be the family of blackbody or synchrotron models.

The bootstrap can be computed in two different ways. The **parametric bootstrap** consists of simulated datasets obtained from the best-fit model $F(x;\hat{\theta}_n)$. Techniques for obtaining Monte Carlo realizations of a specified function are well-known (Press *et al*. 1997). The **nonparametric bootstrap**, discussed in Section 3.6.2, gives Monte Carlo realizations of the observed e.d.f. using a random-selection-with-replacement procedure.

In the parametric bootstrap, $\hat{F}_n = F(.;\hat{\theta}_n)$; that is, we generate data $X_1^*, \ldots, X_n^*$ from the model assuming the estimated parameter values $\hat{\theta}_n$. The process based on the bootstrap simulations,

$$Y_n^P(x) = \sqrt{n}\big(F_n^*(x) - F(x;\hat{\theta}_n^*)\big), \tag{3.44}$$

and the sample process,

$$Y_n(x) = \sqrt{n}\big(F_n(x) - F(x;\hat{\theta}_n)\big), \tag{3.45}$$

converge to the same Gaussian process Y. Consequently, for the K-S test,

$$\begin{aligned} L_n &= \sqrt{n}\,\sup_x |F_n(x) - F(x;\hat{\theta}_n)| \ \text{ and} \\ L_n^* &= \sqrt{n}\,\sup_x |F_n^*(x) - F(x;\hat{\theta}_n^*)| \end{aligned} \tag{3.46}$$

have the same limiting distribution. The critical values of L_n for the K-S statistic can be derived by constructing B resamples based on the parametric model ($B \sim 1000$ usually suffices), and arrange the resulting L_n^* values in increasing order to obtain 90 or 99 percentile points for getting 90% or 99% critical values. This procedure replaces the often incorrect use of the standard probability tabulation of the K-S test.

The nonparametric bootstrap involving resamples from the e.d.f.,

$$\begin{aligned} Y_n^N(x) &= \sqrt{n}\big(F_n^*(x) - F(x;\hat{\theta}_n^*)\big) - B_n(x) \\ &= \sqrt{n}\big(F_n^*(x) - F_n(x) + F(x;\hat{\theta}_n) - F(x;\hat{\theta}_n^*)\big), \end{aligned} \tag{3.47}$$

is operationally easy to perform but requires an additional step of bias correction

$$B_n(x) = \sqrt{n}(F_n(x) - F(x;\hat{\theta}_n)). \tag{3.48}$$

The sample process Y_n and the bias-corrected nonparametric process Y_n^N converge to the same Gaussian process Y. That is,

$$\begin{aligned} L_n &= \sqrt{n}\sup_x |F_n(x) - F(x;\hat{\theta}_n)| \quad \text{and} \\ J_n^* &= \sup_x |\sqrt{n}\left(F_n^*(x) - F(x;\hat{\theta}_n^*)\right) - B_n(x)| \end{aligned} \tag{3.49}$$

have the same limiting distribution. The critical values of the distribution of L_n can then be derived as in the case of the parametric bootstrap. The regularity conditions under which these results hold are detailed by Babu & Rao (2004).

3.7.2 Likelihood-based methods for model selection

To set up a general framework for model selection, let D denote the observed data and let $M_1, \ldots, M_k$ denote the models for D under consideration. For each model M_j, let $L(D|\theta_j; M_j)$ and $\ell(\theta_j) = \ln f(D|\theta_j; M_j)$ denote the likelihood and loglikelihood respectively, where θ_j is a p_j-dimensional parameter vector. Here $L(D|\theta_j; M_j)$ denotes the probability density function (in the continuous case) or the probability mass function (in the discrete case) evaluated at the data D. Most of the methodology can be framed as a comparison between two models, M_1 and M_2.

The model M_1 is said to be nested in M_2 if some elements of the parameter vector θ_1 are fixed (and possibly set to zero). That is, $\theta_2 = (\alpha, \gamma)$ and $\theta_1 = (\alpha, \gamma_0)$, where γ_0 is some known fixed constant vector. Comparison of M_1 and M_2 can then be considered as a classical hypothesis testing problem where $H_0 : \gamma = \gamma_0$. Nested models of this type occur frequently in astronomical modeling. In astrophysical modeling, stellar photometry might be modeled as a blackbody (M_1) with absorption (M_2), the structure of a dwarf elliptical galaxy might be modeled as an isothermal sphere (M_1) with a tidal cutoff (M_2), or a hot plasma might be modeled as an isothermal gas (M_1) with nonsolar elemental abundances (M_2).

A simple example in the statistics of nested models might be a normal model with mean μ and variance σ^2 (M_2) compared to a normal with mean 0 and variance σ^2 (M_1). This model selection problem can be framed as a hypothesis test of $H_0 : \mu = 0$ with free parameter σ. There are some objections to using hypothesis testing to decide between the two models M_1 and M_2, as they are not treated symmetrically by the test in which the null hypothesis is M_1. We cannot accept H_0 and show it is true. We can only reject, or fail to reject, H_0. Note

that with very large samples, even very small discrepancies lead to rejection of the null hypothesis, while for small samples, even large discrepancies may not lead to rejection.

We now look at three classical methods for testing H_0 based on maximum likelihood estimators that were developed during the 1940s: the Wald test, the likelihood ratio test, and Rao's score test. The three tests are equivalent to each other to the first order of asymptotics, but differ in the second-order properties. No single test among these three is uniformly better than the others. Here we restrict consideration to a one-dimensional vector of model parameters θ where θ_0 is a preselected value of the parameter.

To test the null hypothesis $H_0 : \theta = \theta_0$, the **Wald test** uses the statistic

$$W_n = \frac{(\hat{\theta}_n - \theta_0)^2}{Var(\hat{\theta}_n)}, \tag{3.50}$$

the standardized distance between θ_0 and the maximum likelihood estimator $\hat{\theta}_n$ based on data of size n. A. Wald showed that the distribution of W_n is approximately the χ^2 distribution with one degree of freedom. We saw in Section 3.4.4 that, although in general the variance of $\hat{\theta}_n$ is not known, a close approximation is $Var(\hat{\theta}_n) = 1/I(\hat{\theta}_n)$ where $I(\theta)$ is Fisher's information. Thus $I(\hat{\theta}_n)(\hat{\theta}_n - \theta_0)^2$ has a χ^2 distribution in the limit, and the Wald test rejects the null hypothesis H_0, when $I(\hat{\theta}_n)(\hat{\theta}_n - \theta_0)^2$ is large.

The **likelihood ratio test**, as its name implies, uses the ratio of likelihoods as a model selection statistic, or in logarithmic form,

$$LRT = \ell(\hat{\theta}_n) - \ell(\theta_0), \tag{3.51}$$

where $\ell(\theta)$ denotes the loglikelihood at θ. The likelihood ratio test is widely used in astronomy but not always correctly, as explained by Protassov *et al*. (2002).

The **score test** developed by C. R. Rao, also known as the Lagrangian multiplier test, uses the statistic

$$S(\theta_0) = \frac{\ell'(\theta_0))^2}{nI(\theta_0)}, \tag{3.52}$$

where ℓ' denotes the derivative of ℓ, and I again is Fisher's information.

3.7.3 Information criteria for model selection

While the classical Wald, likelihood ratio, and score tests are still widely used, an alternative approach based on **penalized likelihoods** has dominated model selection since the 1980s. If the model M_1 is nested within model M_2, then the largest likelihood achievable by M_2 will always be larger than that achievable by M_1 simply because there are more parameters that can adjust to more detailed variations. If a **penalty** is applied to compensate for the obligatory difference in likelihoods due to the different number of parameters in M_1 and M_2, the desired balance between overfitting and underfitting might be found. Many model selection procedures based on information criteria use penalty terms.

The traditional maximum likelihood paradigm, as applied to statistical modeling, provides a mechanism for estimating the unknown parameters of a model having a specified dimension and structure. H. Akaike (1973) extended this paradigm by considering a framework in which the model dimension is also unknown. He proposed a framework where both

model estimation and selection could be simultaneously accomplished. Grounding in the concept of entropy, Akaike proposed **an information criterion** now popularly known as the **Akaike information criterion** (AIC) defined for model M_j, as

$$AIC = 2\ell(\hat{\theta}_j) - 2p_j. \tag{3.53}$$

where p_j is the number of parameters in the j-th model. The term $2\ell(\hat{\theta}_j)$ is the goodness-of-fit term, and $2p_j$ is the penalty term. The penalty term increases as the complexity of the model grows and thus compensates for the necessary increase in the likelihood. The AIC selects the model M_i if $i = \arg\max_j 2\ell(\hat{\theta}_j) - 2p_j$. That is, the maximum AIC attempts to find the model that best explains the data with a minimum of free parameters.

Unlike the classical Wald, likelihood ratio, and score tests, the AIC treats all the models symmetrically, not requiring an assumption that one of the candidate models is the "correct" model. The AIC can be used to compare nonnested as well as nested models. The AIC can also be used to compare models based on different families of probability distributions. One of the disadvantages of AIC is the requirement of large samples, especially in complex modeling frameworks. In addition, it is not a consistent statistic: if p_0 is the correct number of parameters, and $\hat{p} = p_i$ ($i = \arg\max_j [2\ell(\hat{\theta}_j) - 2p_j]$), then $\lim_{n\to\infty} P(\hat{p} > p_0) > 0$. That is, even if we have a very large number of observations, $\hat{p}$ does not approach the true value.

The Schwarz information criterion, more commonly called the **Bayesian information criterion** (BIC), is the other widely used choice for penalized likelihood model selection. The BIC defined as

$$BIC = 2\ell(\hat{\theta}_j) - p_j \ln n, \tag{3.54}$$

where n is the number of data points, is consistent. It is derived by giving all the models under consideration equal weights; that is, equal prior probabilities to all the models under consideration. The BIC selects the model with highest marginal likelihood that, expressed as an integral, is approximated using Laplace's method. This in turn leads to the expression (3.54). Another model evaluation criterion based on the concept of **minimum description length** (MDL) is used in transmitting a set of data by coding using a family of probability models. The MDL model selection criterion is $-(1/2)BIC$. Like AIC, the models compared by the BIC need not be nested.

Conditions under which these two criteria are mathematically justified are often ignored in practice. AIC penalizes free parameters less strongly than does the BIC. The BIC has a greater penalty for larger datasets, whereas the penalty term of the AIC is independent of the sample size. Different user communities prefer one over the other. The user should also beware that sometimes these criteria are expressed with a minus sign so the goal changes to minimizing rather than maximizing the information criterion.

3.7.4 Comparing different model families

We briefly address the more difficult problem of comparing different model family fits to the dataset. Rather than asking "Which model within a chosen model family best fit the observed dataset?" we need to address "How far away is the unknown distribution

underlying the observed dataset from the hypothesized family of models?". For example, an astronomer might ask whether the continuum emission of a quasar is better fit by a power-law model associated with a nonthermal relativistic jet or by a multiple-temperature blackbody model associated with a thermal accretion disk. The functional form and parameters of the nonthermal and thermal models are completely unrelated to each other. If we assume a nonthermal model family and the quasar is actually emitting due to thermal processes, then we have **misspecified** the model. Least squares, maximum likelihood estimation and other inferential methods assume that the probability model is "correctly specified" for the problem under study. When misspecification is present, the classical model selection tests (Wald, likelihood ratio, and score tests) and the widely used Akaike information criterion are not valid for comparing fits from different model families. The treatment of this problem has been addressed by a number of statisticians (e.g. Huber 1967, White 1982, Babu & Rao 2003) and the findings rely on somewhat advanced mathematics.

We outline here one approach to the problem (Babu & Rao 2003). Let the original dataset $X_1, \ldots, X_n$ come from an unknown distribution H that may or may not belong to the family $\{F(.;\theta) : \theta \in \Theta\}$. Let $F(.,\theta_0)$ be the specific model in the family that is "closest" to H where proximity is based on the **Kullback–Leibler information**,

$$\int \ln\left(\frac{h(x)}{f(x;\theta)}\right) dH(x) \geq 0, \tag{3.55}$$

which arises naturally due to maximum likelihood arguments and has advantageous properties. Here h and f are the densities (i.e. derivatives) of H and F. If the maximum likelihood estimator $\hat{\theta}_n \to \theta_0$, then for any $0 \leq \alpha \leq 1$,

$$P\left(\sqrt{n}\sup_x |F_n(x) - F(x;\hat{\theta}_n) - (H(x) - F(x;\theta_0))| \leq C^*_\alpha\right) - \alpha \to 0 \tag{3.56}$$

where C^*_α is the α-th quantile of $\sup_x |\sqrt{n}\left(F^*_n(x) - F(x;\hat{\theta}^*_n)\right) - \sqrt{n}\left(F_n(x) - F(x;\hat{\theta}_n)\right)|$. This provides an estimate of the distance between the true distribution and the family of distributions under consideration.

3.8 Bayesian statistical inference

In Section 2.4.1, we showed how Bayes' theorem is a logical consequence of the axioms of probability. Bayesian inference is founded on a particular interpretation of Bayes' theorem. Starting with Equation (2.16), we let A represent an observable X, and B represent the space of models M that depend on a vector of parameters θ. Bayes' theorem can then be rewritten as

$$P(M_i(\theta) \mid X) = \frac{P(X \mid M_i(\theta))P(M_i(\theta))}{P(X \mid M_1(\theta))P(M_1(\theta)) + \cdots + P(X \mid M_k(\theta))P(M_k(\theta))}. \tag{3.57}$$

In this context, the factors in Bayes' theorem have the following interpretations and designations:

$P(M_i(\theta) \mid X)$ is the conditional probability of the i-th model given the data X. This is called the **posterior probability**.

$P(X \mid M_i(\theta))$ is the conditional probability of the observable for the i-th model. This is called the **likelihood function**.

$P(M_i(\theta))$ is the marginal probability of the i-th model. The observable has no input here. This is called the **prior information**.

In this framework, Bayes' theorem states that the posterior distribution of a chosen set of models given the observable is equal to the product of two factors — the likelihood of the data for those models and the prior probability on the models — with a normalization constant.

More generally, we can write Bayes' theorem in terms of the parameters θ of the models M and the likelihood $L(X \mid \theta)$,

$$P(\theta \mid X) = \begin{cases} \dfrac{P(\theta)L(X|\theta)}{\sum_j P(\theta_j)L(X|\theta_j)} & \text{if } \theta \text{ is discrete;} \\ \dfrac{P(\theta)L(X|\theta)}{\int P(u)L(X|u)\,du} & \text{if } \theta \text{ is continuous.} \end{cases} \tag{3.58}$$

Bayesian inference applies Bayes' theorem to assess the degree to which a given dataset is consistent with a given set of hypotheses. As evidence accumulates, the scientists' belief in a hypothesis ought to change; with enough evidence, it should become very high or very low. Thus, proponents of Bayesian inference say that it can be used to discriminate between conflicting hypotheses: hypotheses with very high support should be accepted as true and those with very low support should be rejected as false. However, detractors say that this inference method may be biased because the results depend on the prior distribution based on the initial beliefs that one holds before any evidence is ever collected. This is a form of inductive bias; it is generally larger for small datasets than large datasets, and larger for problems without much prior study than well-developed problems with extensive prior knowledge.

3.8.1 Inference for the binomial proportion

We examine here the Bayesian approach to statistical inference through applications to a simple problem: the estimation of the success ratio of a simple binary ("Yes/No") variable using Equation (3.57) with $k = 2$. This is called the binomial proportion problem.

Example 3.1 Consider the astronomer who has in hand a sample of optically selected active galactic nuclei (AGN such as Seyfert galaxies and quasi-stellar objects) and surveys them with a radio telescope to estimate the fraction of AGN that produce nonthermal radio emission from a relativistic jet. Let X denote the random variable giving the result of the radio-loudness test, where $X = 1$ indicates a positive and $X = 0$ indicates a negative result. Let θ_1 denote that radio-loudness is present and θ_2 denote that the AGN is radio-quiet. $P(\theta = \theta_1)$ denotes the probability that a randomly chosen AGN is radio-loud; this is the prevalence of radio-loudness in the optically selected AGN population.

From previous surveys, the astronomer expects the prevalence is around 0.1, but this is not based on the observational data under study. This is **prior** information that can be incorporated in the Bayesian analysis. Our goal is to use the new radio data to estimate posterior information in the form of $P(\theta|X = x)$, where as described above $x = 1$ denotes the existence of radio-loudness and $x = 0$ denotes radio-quietness. We start by adopting a prior of $P(\theta = \theta_1) = 0.1$. Based on our knowledge of the sensitivity of the radio telescope and the AGN redshift distribution, we establish that our radio survey is sufficiently sensitive to measure the radio emission 80% of the time in radio-loud AGN, $P(X = 1|\theta_1) = 0.8$. The failures are due to AGN at very high redshifts. However, the sample also includes some AGN at small redshifts where the telescope may detect thermal radio emission from star formation in the host galaxy. Let us say that this irrelevant positive detection of radio emission occurs 30% of the time, $P(X = 1|\theta_2) = 0.3$.

We are now ready to calculate the chances that an AGN is truly radio-loud when a positive radio detection is obtained, $P(\theta = \theta_1|X = 1)$. Bayes' theorem gives

$$P(\theta = \theta_1 \mid X = 1) = \frac{0.8 \times 0.1}{0.8 \times 0.1 + 0.3 \times 0.9} = \frac{0.08}{0.35} = 0.23. \tag{3.59}$$

This shows that the observational situation outlined here does not answer the scientific question very effectively: only 23% of the true radio-loud AGN are clearly identified, and 77% are either false negatives or false positives. The result is moderately sensitive to the assumed prior; for example, the fraction of true radio-loud AGN discoveries rises from 23% to 40% if the prior fraction is assumed to be 20% rather than 10%. But the major benefit would arise if we had additional data that could reduce the detection of irrelevant thermal radio emission. This might arise from ancillary data in the radio band (such as the radio polarization and spectral index) or data from other bands (such as optical spectral line-widths and ratios). If this fraction of erroneous detection is reduced from 30% to 5%, then the discovery fraction of radio-loud AGN increases from 23% to a very successful 95%. Significant astronomical efforts are devoted to addressing this problem (e.g. Ivecić *et al*. 2002).

This example shows that the measured conditional probabilities $P(X = x|\theta)$ are "inverted" to estimate the conditional probabilities of interest, $P(\theta|X = x)$. Inference based on Bayes' theorem is thus often called the theory of **inverse probability**. It originated with Thomas Bayes and Pierre Simon Laplace in the eighteenth and nineteenth centuries, and was revived by Sir Harold Jeffreys (who held the Plumian Professorship of Astronomy and Experimental Philosophy at the University of Cambridge), E. Jaynes, and other scholars in the mid-twentieth century.

3.8.2 Prior distributions

Bayesian inference is best performed when prior information is available to guide the choice of the prior distribution. This information can arise either from earlier empirical studies or from astrophysical theory thought to determine the observed phenomena. Even when prior knowledge is available, it may not be obvious how to convert that knowledge into a workable prior probability distribution. The choice of prior distribution influences the final

result. The numerator of Bayes' theorem can be viewed as an average of the likelihood function weighted by the prior, whereas the maximum likelihood estimator gives the mode of the likelihood function without any weighting. There is often a relationship between the type of prior distribution and the type of posterior distribution.

Consider, for example, that we have some prior knowledge about θ that can be summarized in the form of a **beta distribution** p.d.f.

$$\begin{aligned} P(x;\alpha,\gamma) &= \frac{x^{\alpha-1}(1-x)^{\gamma-1}}{B(\alpha,\gamma)}, \quad 0<x<1, \quad \text{where} \\ B(\alpha,\gamma) &= \frac{\Gamma(\alpha)\Gamma(\gamma)}{\Gamma(\alpha+\gamma)} \quad \text{and} \quad \Gamma(\alpha)=\int_0^\infty t^{\alpha-1}e^{-t}dt \end{aligned} \tag{3.60}$$

with two shape parameters, $\theta = (\alpha, \gamma)$. Note that special cases of the beta include the uniform distribution. In the case when the likelihood is binomial, the posterior will be distributed as beta with parameters $x+\alpha$ and $n-x+\gamma$ where n is the sample size. Such priors which result in posteriors from the same family are called **natural conjugate priors**.

Example 3.2 Consider the problem in Example 3.1 but restricted to the situation where we have no prior information available on the probability of radio-loudness in AGN; that is, the distribution of θ is unconstrained. For simplicity, we might assume that θ is uniformly distributed on the interval $(0, 1)$. Its prior density is then $p(\theta) = 1,\ 0 < \theta < 1$. This choice is sometimes called the **non-informative prior**. Often, Bayesian inference from such a prior coincides with classical inference. The likelihood for a binary variable is given by the binomial distributions discussed in Section 4.1. The posterior density of θ given X is now

$$\begin{aligned} P(\theta|X) &= \frac{p(\theta)L(X|\theta)}{\int P(u)L(X|u)\,du} \\ &= \frac{(n+1)!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}, \quad 0<\theta<1. \end{aligned}$$

In this example, the resulting function of θ is the same as the likelihood function $L(x|\theta) \propto \theta^X(1-\theta)^{n-X}$. Thus, maximizing the posterior probability density will give the same estimate as the maximum likelihood estimate,

$$\hat{\theta}_B = \hat{\theta}_{MLE} = \frac{X}{n}, \tag{3.61}$$

where the subscript B denotes Bayesian. However, the interpretation of $\hat{\theta}_B$ as an estimate of θ is quite different from the interpretation of $\hat{\theta}_{MLE}$. The Bayesian estimator is the most probable value of the unknown parameter θ conditional on the sample data x. This is called the **maximum a posteriori** (MAP) estimate or the **highest posterior density** (HPD) estimate; it is the maximum likelihood estimator of the posterior distribution. Note that if we chose a different prior distribution, the Bayesian maximum posterior result would differ from classical results.

The uncertainty in the most probable posterior value of θ can also be estimated because the $P(\theta|X)$ is now described in terms of a genuine probability distribution concentrated

around $\hat{\theta}_B = X/n$ to quantify our post-experimental knowledge about θ. The classical Bayes estimate $\hat{\theta}_B$ minimizes the posterior mean square error,

$$E[(\theta - \hat{\theta}_B)^2|X] = \min_a E[(\theta - a)^2|X]. \tag{3.62}$$

Hence $\hat{\theta}_B = E(\theta|X)$ is the mean of the posterior distribution. If $\hat{\theta}_B$ is chosen as the estimate of θ, the natural measure of variability of this estimate is obtained in the form of the posterior variance, $E[(\theta - E(\theta|X))^2|X]$, with the posterior standard deviation serving as a natural measure of the estimation of error. The resulting estimated confidence interval of the form

$$\hat{\theta}_B \pm c\sqrt{E[(\theta - E(\theta|x))^2|X]}. \tag{3.63}$$

We can compute the posterior probability of any interval containing the parameter θ. A statement such as $P(\hat{\theta}_B - k_1 \leq \theta \leq \hat{\theta}_B + k_2|X) = 0.95$ is perfectly meaningful, conditioned on the given dataset.

In Example 3.2, if the prior is a beta distribution with parameters $\theta = (\alpha, \gamma)$, then the posterior distribution of $\theta|X$ will be a beta distribution with parameters $X + \alpha$, $n - X + \gamma$. So the Bayes estimate of θ will be

$$\hat{\theta}_B = \frac{(X + \alpha)}{(n + \alpha + \gamma)} = \frac{n}{n + \alpha + \gamma}\frac{X}{n} + \frac{\alpha + \gamma}{n + \alpha + \gamma}\frac{\alpha}{\alpha + \gamma}. \tag{3.64}$$

This is a convex combination of the sample mean and prior mean with weights depending upon the sample size and the strength of the prior information as measured by the values of α and γ.

Bayesian inference relies on the concept of conditional probability to revise one's knowledge. In the above example, prior to the collection of sample data, the astronomer had some (perhaps vague) information constraining the distribution of θ. Then measurements of the new sample were made. Combining the model density of the new data with the prior density gives the posterior density, the conditional density of θ given the data.

3.8.3 Inference for Gaussian distributions

We now consider Bayesian analysis of a problem of univariate signal estimation where the data are thought to follow a normal distribution due either to intrinsic scatter in the population or to error in the measurement process. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be n independent and identically distributed (i.i.d.) random variables drawn from a normal distribution with mean μ and variance σ^2. We can express the data as

$$X_i = \mu + \sigma\eta_i \tag{3.65}$$

where μ is the signal strength (the parameter of scientific interest) and η_i are i.i.d. standard normal random variables. The likelihood is given by

$$L(\mathbf{X}; \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + n\left(\bar{X}_n - \mu\right)^2\right)\right\} \tag{3.66}$$

where $\bar{X}_n$ is the mean of $\mathbf{X}$. This likelihood is combined with an assumed prior in the numerator of Bayes' theorem. We consider two cases representing different levels of prior knowledge.

Case 1: σ is known Here the likelihood simplifies to

$$L(\mathbf{X}; \mu, \sigma^2) \propto \exp\left[-\frac{n}{2\sigma^2}(\bar{X}_n - \mu)^2\right]. \tag{3.67}$$

If we assume an informative prior π on μ and that it is normally distributed with mean μ_0 and variance τ^2, then the posterior is $P(\mu|\mathbf{X}) \propto L(\mathbf{X}; \mu, \sigma^2)\Pi(\mu)$. Here Π is the normal density with mean μ_0 and variance τ^2 evaluated at the parameter μ. After some algebra, this leads to the result that

$$\begin{aligned} P(\mu|\mathbf{X}) &\propto \exp\left\{-\frac{1}{\gamma^2}(\mu - \hat{\mu})^2\right\} \quad \text{where} \\ \gamma^2 &= \frac{\sigma^2}{n}\frac{\tau^2}{\tau^2 + \sigma^2/n} \quad \text{and} \\ \hat{\mu} &= \gamma^2\left(\frac{\mu_0}{\tau^2} + \frac{n\bar{X}_n}{\sigma^2}\right) = \left(\frac{\gamma^2}{\tau^2}\right)\mu_0 + \left(1 - \frac{\gamma^2}{\tau^2}\right)\bar{X}_n. \end{aligned} \tag{3.68}$$

That is, the posterior distribution of μ, given the data, is normally distributed with mean $\hat{\mu}$ and variance γ^2. Consequently the 95% HPD "credible interval" for μ is given by $\hat{\mu} \pm 1.96\gamma$. Bayesian versions of confidence intervals are referred to as **credible intervals**.

Note that the HPD Bayesian estimator for the signal strength μ_B is not the simple MLE value $\bar{X}_n$, but rather is a convex combination of the sample mean $\bar{X}_n$ and the prior mean μ_0. If $\tau \to \infty$, the prior becomes nearly flat and $\hat{\mu}_B \to \hat{\mu}_{MLE}$ and $\gamma^2 \to \sigma^2/n$, producing the frequentist results. The Bayesian estimator of the signal strength could be either larger or smaller than the sample mean, depending on the chosen values of μ_0 and τ^2 of the assumed prior distribution.

Case 2: σ is unknown In this case, the variance σ^2 is a nuisance parameter that should be removed, if possible. In the Bayesian context, we **marginalize** over the nuisance parameters by integrating σ^2 out of the joint posterior distribution. That is, if $P(\mu, \sigma^2)$ denotes the prior, then the corresponding joint posterior is

$$\begin{aligned} P(\mu, \sigma^2|\mathbf{X}) &\propto L(\mathbf{X}; \mu, \sigma^2)P(\mu, \sigma^2) \quad \text{and} \\ P(\mu|\mathbf{X}) &= \int_0^\infty P(\mu, \sigma^2|\mathbf{X})d\sigma^2. \end{aligned} \tag{3.69}$$

If we use the uninformative Jeffreys prior, then $P(\mu, \sigma^2) \propto 1/\sigma^2$. For the Jeffreys prior, algebraic manipulation shows that the posterior estimator μ_B follows the Student's t distribution with $n - 1$ degrees of freedom. Thus, the Jeffreys prior reproduces frequentist results.

While our examples here are restricted to rather simple parameter estimation problems, Bayes' theorem can be applied to any problem where the likelihood and the prior can be clearly specified. Complex real-life modeling problems often require a **hierarchical Bayes**

approach, where the statistical model requires two or more interrelated likelihoods. In astronomy, this arises in regression models incorporating measurement errors (Kelly 2007; Mandel *et al*. 2009; Hogg *et al*. 2010), instrumental responses (van Dyk *et al.* 2001), systematic errors (Shkedy *et al.* 2007) and many other problems. Time series and other parameter estimation situations with fixed variables can be formulated in a Bayesian framework. Bayesian applications have grown into many branches of applied statistics and, in some cases, have become the preferred approach.

3.8.4 Hypotheses testing and the Bayes factor

Bayesian principles can be effectively applied to estimate probabilities for hypothesis tests where one seeks probabilities associated with binary (e.g. Yes vs. No, H_o vs. H_1) questions as discussed in Section 3.5. These include two-sample tests, treatments of categorical variables, composite and multiple hypothesis testing.

We start testing alternative hypotheses based on the random vector $\mathbf{X}$,

$$H_0 : \theta \in \Theta_0 \quad \text{against} \quad H_1 : \theta \in \Theta_1. \tag{3.70}$$

If Θ_0 and Θ_1 are of the same dimension (for example, $H_0 : \theta \leq 0$ and $H_1 : \theta > 0$), we may choose a prior density that assigns positive prior probability to Θ_0 and Θ_1.

Next, one calculates the posterior probabilities $P\{\Theta_0|\mathbf{X}\}$ and $P\{\Theta_1|\mathbf{X}\}$ as well as the **posterior odds ratio**, $P\{\Theta_0|X\}/P\{\Theta_1|X\}$. Here one might set a threshold like 1/9 for a 90% significance level or 1/19 for a 95% significance level to decide what constitutes evidence against H_0. Unlike classical tests of hypotheses, Bayesian testing treats H_0 and H_1 symmetrically.

Another way to formulate the problem is to let $\pi_0 = P(\Theta_0)$ and $1 - \pi_0 = P(\Theta_1)$ be the prior probabilities of Θ_0 and Θ_1. Further, let $g_i(\theta)$ be the prior p.d.f. of θ under Θ_i (or H_i), so that $\int_{\Theta_i} g_i(\theta)d\theta = 1$. This leads to $\pi(\theta) = \pi_0 g_0(\theta)I\{\theta \in \Theta_0\} + (1-\pi_0)g_1(\theta)I\{\theta \in \Theta_1\}$ where I is the indicator function. This formulation is quite general, allowing Θ_0 and Θ_1 to have different dimensions.

We can now calculate posterior probabilities or posterior odds. Letting $f(.|\theta)$ be a probability density, the marginal density of X under the prior π can be expressed as

$$m_\pi(x) = \int_\Theta f(x|\theta)\pi(\theta)\,d\theta = \pi_0 \int_{\Theta_0} f(x|\theta)g_0(\theta)\,d\theta + (1-\pi_0)\int_{\Theta_1} f(x|\theta)g_1(\theta)\,d\theta. \tag{3.71}$$

The posterior density of θ given the data $X = x$ is therefore $\pi_0 f(x|\theta)g_0(\theta)/m_\pi(x)$ or $(1-\pi_0)f(x|\theta)g_1(\theta)/m_\pi(x)$ depending on whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$. The posterior distributions of $\Theta_0|x$ and $\Theta_1|x$ are given respectively by

$$P^\pi(\Theta_0|x) = \frac{\pi_0}{m_\pi(x)}\int_{\Theta_0} f(x|\theta)g_0(\theta)\,d\theta \quad \text{and} \quad P^\pi(\Theta_1|x) = \frac{(1-\pi_0)}{m_\pi(x)}\int_{\Theta_1} f(x|\theta)g_1(\theta)\,d\theta. \tag{3.72}$$

This result can also be reported as the **Bayes factor** of H_0 relative to H_1,

$$BF_{01} = \frac{P(\Theta_0|x)}{P(\Theta_1|x)} \Big/ \frac{P(\Theta_0)}{P(\Theta_1)} = \frac{\int_{\Theta_0} f(x|\theta)g_0(\theta)\,d\theta}{\int_{\Theta_1} f(x|\theta)g_1(\theta)\,d\theta}. \tag{3.73}$$

Clearly, $BF_{10} = 1/BF_{01}$, and the posterior odds ratio of H_0 relative to H_1 is

$$\frac{P(\Theta_0|x)}{P(\Theta_1|x)} = \left(\frac{\pi_0}{1-\pi_0}\right) BF_{01}. \tag{3.74}$$

Thus, the posterior odds ratio of H_0 relative to H_1 is BF_{01} whenever $\pi_0 = \frac{1}{2}$. The smaller the value of BF_{01}, the stronger the evidence against H_0. This can easily be extended to any number of hypotheses.

3.8.5 Model selection and averaging

We discussed in Section 3.7 the challenging task of evaluating the relative merits of different models for the same dataset, both nested and nonnested. The Bayesian approach to these issues is often attractive. We illustrate Bayesian model selection with a model of a univariate dataset with an unknown number of normal (Gaussian) components. This is the **normal mixture model**.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with common density f given by a mixture of normals

$$f(x|\theta) = \sum_{j=1}^{k} p_j \phi(x|\mu_j, \sigma_j^2), \tag{3.75}$$

where $\phi(.|\mu_j, \sigma_j^2)$ denotes the normal density with mean μ_j and variance σ_j^2, k is the number of mixture components, and p_j is the weight given to the j-th normal component. Let M_k denote a k-component normal mixture model. Bayesian model selection procedures involve computing

$$m(x|M_k) = \int \pi(\theta_k) f(x|\theta_k)\, d\theta_k, \tag{3.76}$$

for each k of interest. We estimate the number of components $\hat{k}$ that gives the largest value.

From the Bayesian point of view, a natural approach to model uncertainty is to include all models M_k under consideration for later decisions. This is suitable for prediction, eliminating the underestimation of uncertainty that results from choosing a single best model $M_{\hat{k}}$. This leads to Bayesian **model averaging**. Consider the parameter space encompassing all models under consideration, $\Theta = \cup_k \Theta_k$. Let the likelihood h and prior π be given by

$$h(.|\theta) = f_k(.|\theta_k) \text{ and } \pi(\theta) = p_k g_k(\theta_k) \text{ if } \theta \in \Theta_k, \tag{3.77}$$

where $p_k = P_\pi(M_k)$ is the prior probability of M_k and g_k integrates to 1 over Θ_k. Given X, the posterior is the weighted average of the model posteriors,

$$\begin{aligned} \pi(\theta|X) &= \frac{h(X|\theta)\pi(\theta)}{m(X)} = \sum_k \frac{p_k}{m(X)} f_k(X|\theta_k) g_k(\theta_k) I_{\Theta_k}(\theta_k) \\ &= \sum_k P(M_k|X) g_k(\theta_k|X) I_{\Theta_k}(\theta_k), \end{aligned} \tag{3.78}$$

where the normalizing constant based on the data is given by

$$\begin{aligned} m(X) &= \int_{\Theta} h(X|\theta)\pi(\theta)\,d\theta, \quad P(M_k|X) = p_k m_k(X)/m(X), \\ g_k(\theta_k|X) &= f_k(X|\theta_k)g_k(\theta_k)/m_k(X), \quad \text{and } m_k(X) = \int_{\theta_k} f_k(X|\theta_k)g_k(\theta_k)\,d\theta_k. \end{aligned} \tag{3.79}$$

The needed predictive density $\ell(y|X)$ of **future** y values given X is obtained by integrating out the **nuisance** parameter θ. That is,

$$\begin{aligned} \ell(y|X) &= \int_{\Theta} h(y|\theta)\pi(\theta|X)\,d\theta \\ &= \sum_k P(M_k|X) \int_{\Theta_k} f_k(y|\theta_k)g_k(\theta_k|X)\,d\theta_k \\ &= \sum_k P(M_k|X)\ell_k(y|X), \end{aligned} \tag{3.80}$$

averages of predictive densities where $\ell_k(y|X) = \int_{\Theta_k} f_k(y|\theta_k)g_k(\theta_k|X)\,d\theta_k$. This is obtained by averaging over all models.

3.8.6 Bayesian computation

Least-squares estimation can usually be calculated as a solution to a system of linear equations and maximum likelihood estimation can be accomplished with an optimization algorithm for nonlinear equations such as the EM algorithm. But characterizing the multivariate posterior distribution, finding the highest posterior density (HPD), or the maximum a priori (MAP) model in Bayesian estimation requires integrating over the full space of possible models. While this is not demanding for simple models such as estimation of the binomial proportion or the mean and variance of a univariate normal function, it can be quite difficult in realistic models of scientific interest. For example, a ΛCDM cosmological model fit to unpolarized and polarized fluctuations of the Cosmic Microwave Background radiation and to the galaxy clustering correlation function can have ∼15 parameters with nonlinear effects on different statistics derived from the datasets (Trotta 2008; see also the interactive model simulations of M. Tegmark at http://space.mit.edu/home/tegmark/movies.html). Complete coverage of the ∼15-dimensional model space is not computationally feasible even if the likelihood is not difficult to calculate.

The practical solution to this problem is to obtain a limited number of independent draws from model parameter space, $\theta \in \Omega$, so that the desired results are estimated with reasonable reliability and accuracy. A commonly desired product is the distribution of the posterior density, $P(\theta|x)$, for a range of model parameter values. For inferences around the peak of low-dimensional unimodal models, a small number of simulations (say, ∼100−1000) may suffice. But for models with complex likelihood functions, hierarchical model structures, high dimensionality, and/or interest in the tails of the model distributions, efficient strategies are needed to sample the model space to address the scientific question.

A variety of computational tools is available for computing posterior densities for a model space. Under some conditions where the bounds of the probability density can be estimated, **rejection sampling** and **importance sampling** can be applied (Press *et al.* 2007). More generally, **Markov chain Monte Carlo (MCMC)** simulations are used. The calculation begins with initial draws from the probability distribution from random or selected locations in the parameter space. Sequential samples are drawn; simple Markov chains are random walks that depend only on the previous location in the space. The goal is to design the Markov chain to converge efficiently to the desired posterior distribution.

The **Gibbs sampler** and the **Metropolis–Hastings algorithm** are often found to be effective in MCMC simulations in multidimensional parameter spaces. First, one divides the full parameter vector θ into lower dimensional subvectors. At a given iteration of the Markov chain, draws from these subvectors are made with other subvectors kept fixed. When each subvector is updated separately, a new iteration is begun. This technique can be combined with the Metropolis–Hastings algorithm that uses a rule to accept or reject the next location in the Markov chain depending on the ratio of the probability density function at the two locations. Steps forward are accepted only if they increase the posterior density. A jumping distribution is constructed to map the parameter space in an efficient fashion. Jumps can be made either within or between subspaces defined by the Gibbs sampler subvectors.

Many strategies have been developed to improve the convergence and accuracy of these methods. Initial iterations of the MCMC that depend strongly on the initial starting points in parameter space can be ignored; this is the **burn-in** fraction. Convergence can be assessed by comparing chains based on different starting points. Jumping distributions can be weighted by local gradients in the posterior distribution, and designed to have covariance structure similar to the posterior yet be more easily computed. Complex posterior distributions can be approximated by simpler functions, such as normal mixture models.

The technologies of Bayesian computation are rapidly developing, and specialized study is needed to be well-informed. The monographs by Gelman *et al.* (2003) and Gemerman & Lopes (2006) give thorough reviews of these issues. Recent developments are described by Brooks *et al.* (2011). Astrophysical applications of these methods for modeling planetary orbits from sparse, irregularly spaced radial velocity time series are described by Clyde *et al.* (2007) and Ford & Gregory (2007). Cosmological applications are described by Feroz *et al.* (2009).

3.9 Remarks

Mathematical statisticians have worked industriously for decades to establish the properties (biasedness, consistency, asymptotical normality and so forth) of many statistics derived using various procedures under various conditions. Those statistics that perform well are widely promulgated for practical use.

However, the logical inverse of this situation is also true: when mathematical statisticians have not established the properties of a statistic, then the properties of that statistic cannot be assumed to be known. Indeed, even subtle changes in the assumptions (e.g. a sample is i.i.d. but not normally distributed) can invalidate a critical theorem. There is no guarantee

that a new statistic, or a standard statistic examined under nonstandard conditions, will have desirable properties such as unbiasedness and asymptotic normality. Astronomers are acting in a valid manner scientifically when they construct a new functionality (i.e. statistic) of their data that measures a characteristic of scientific interest. But they may not then assume that the mathematical behavior of that statistic is known, even if it superficially resembles a standard and well-studied statistic. As a result, many statements of estimated values (which assume unbiasedness) or confidence intervals in the astronomical literature have uncertain reliability. When novel statistics are considered, and even with standard statistics applied to small or non-Gaussian samples, the bootstrap approach to estimation and confidence intervals is recommended, especially when there is no closed-form expression for the estimator.

The twentieth century has witnessed long arguments on the relative value of Fisher's MLE and Bayesian approaches to statistical modeling. Today, these debates are muted, and most statisticians agree that neither frequentist nor Bayesian approaches are self-evidently perfect for all situations. If prior information is available and can be effectively stated as a mathematical prior distribution, then Bayesian statistics may be more appropriate. In many situations the frequentist and Bayesian solutions are similar or identical. Essays on this issue particularly oriented towards astrophysicists, some of which frankly advocate a Bayesian orientation, include Loredo (1992), Cousins (1995) and Efron (2004). Instructive and balanced discussions of different approaches to statistical inference appear in the monographs by Barnett (1999), Cox (2006) and Geisser (2006).

The past decade has witnessed a surge of Bayesian applications in astronomy. When prior information is available from astrophysical constraints and can be readily incorporated into a prior distribution for the model, this is an excellent approach. Bayesian inference is also powerfully applied when the astronomer seeks more than simple estimators and confidence intervals. Astrophysical insights can emerge from examination of complicated multimodal posterior distributions or from marginalization of ancillary variables. However, Bayesian methods can be overused. Specification of priors is often tricky, and computation in multidimensional parameter space can require millions of iterations without guarantee of convergence.

Inference problems encountered in astronomy are incredibly varied, so no single approach can be recommended for all cases. When a parametric model, either heuristic or astrophysical, can be reasonably applied, then point estimation can be pursued. For many problems, least-squares and maximum likelihood estimation provide good solutions. Bayesian inference can be pursued when the situation is known to have external constraints.

3.10 Recommended reading

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2003) *Bayesian Data Analysis*, 2nd ed., Chapman & Hall, London

A comprehensive and practical volume on Bayesian methodology. Topics include fundamentals of Bayesian inference, estimation of single- and multiple-parameter modes,

hierarchical models, model comparison, sample inadequacies, Markov chain simulation techniques, regression models, hierarchical and generalized linear models, robust inference, mixture models, multivariate models and nonlinear models.

Hogg, R. V., McKean, J. W. & Craig, A. T. (2005) *Introduction to Mathematical Statistics*, 6th ed., Prentice Hall, Englewood Cliffs

A widely used textbook for graduate students in statistics with broad coverage on the theory and practice of inference. Topics include probability distributions, unbiasedness and consistency, hypothesis testing and optimality, χ^2 tests, bootstrap procedures, maximum likelihood methods, sufficiency, nonparametric statistics, Bayesian statistics and linear modeling.

James, F. (2006) *Statistical Methods in Experimental Physics*, 2nd ed., World Scientific, Singapore

This excellent volume for physical scientists covers many of the topics on statistical inference discussed here: likelihood functions, information and sufficiency, bias and consistency, asymptotic normality, point estimation methods, confidence intervals, hypothesis tests, goodness-of-fit tests and various maximum likelihood methods. Additional topics include concepts of probability, probability distributions, information theory, decision theory and Bayesian inference.

Rice, J. A. (2007) *Mathematical Statistics and Data Analysis*, 2nd ed., Duxbury Press

A high-quality text with broad coverage at an upper-level undergraduate level. Topics on inference include the method of moments, maximum likelihood, sufficiency, hypothesis testing and goodness-of-fit.

Wasserman, L. (2004) *All of Statistics: A Concise Course in Statistical Inference*, Springer, Berlin

A slim, sophisticated volume with broad scope intended for computer science graduate students needing a background in statistics, it is also valuable for physical scientists. Topics include theory of probability and random variables, bootstrap resampling, methods of parametric inference, hypothesis testing, Bayesian inference, decision theory, linear and loglinear models, multivariate models, graph theory, nonparametric density estimation, multivariate classification, stochastic processes and Bayesian computational methods.

3.11 R applications

We do not provide **R** scripts for inference here because most of the examples presented in later chapters implement some type of statistical inference. A few particular applications can be highlighted. In Section 4.7.1, we compare the performance of several point estimation methods – method of moments, least squares, weighted least squares (minimum χ^2), MLE and MVUE – on simulated datasets following the Pareto (power-law) distribution.

Estimation of parameters for a real dataset using a normal model follows, with hypothesis tests of normality. Chapter 5 implements a variety of nonparametric hypothesis tests including Wilcoxon tests, two-sample tests, and e.d.f.-based tests such as the Kolgomorov–Smirnov test. Chapter 6 includes some nonparametric and semi-parametric model-fitting methods including spline fitting, kernel density estimators and local regression. Chapter 7 on regression covers important inferential procedures including least-squares and weighted least-squares estimation, maximum likelihood estimation, robust estimation and quantile regression. Multivariate models are treated in Chapter 8, and some multivariate classifiers that can be considered to be examples of inferential modeling are treated in Chapter 9. A variety of specialized inferential modeling tasks is shown for censored, time series and spatial data in Chapters 10–12.

Elementary Bayesian computations using **R** are presented in the volumes by Albert (2009) and Kruschke (2011). Scripts often call **CRAN** packages such as *rbugs* that provide links to independent codes specializing in MCMC and other computational methods used in Bayesian inference.