Appendix B

REVIEW OF PROBABILITY AND STATISTICS

Synopsis: A brief review is given of the topics in classical probability and statistics that are used in this book. Connections between probability theory and its application to the analysis of data with random measurement errors are highlighted. Note that some very different philosophical interpretations of probability theory are discussed in Chapter 11.

B.1 PROBABILITY AND RANDOM VARIABLES

The mathematical theory of probability begins with an **experiment**, which has a set S of possible outcomes. We will be interested in **events** which are subsets A of S.

■ **Definition B.1** The **probability function** P is a function defined on subsets of S with the following properties:

1. $P(S) = 1$
2. For every event $A \subseteq S$, $P(A) \geq 0$
3. If events $A_1, A_2, \ldots$ are pairwise mutually exclusive (that is, if $A_i \cap A_j$ is empty for all pairs i, j), then

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i). \tag{B.1}$$

■

The probability properties just given are fundamental to developing the mathematics of probability theory. However, applying this definition of probability to real-world situations frequently requires ingenuity.

■ **Example B.1** Consider the experiment of throwing a dart at a dart board. We will assume that our dart thrower is an expert who always hits the dart board. The sample space S consists of the points on the dart board. We can define an event A that consists of the points in the bullseye, so that $P(A)$ is the probability that the thrower hits the bullseye. ■

In practice, the outcome of an experiment is often a number rather than an event. Random variables are a useful generalization of the basic concept of probability.

■ **Definition B.2** A **random variable** X is a function $X(s)$ that assigns a value to each outcome s in the sample space S.

Each time we perform an experiment, we obtain a particular value of the random variable. These values are called **realizations** of the random variable. ■

■ **Example B.2** To continue our previous example, let X be the function that takes a point on the dart board and returns the associated score. Suppose that throwing the dart in the bullseye scores 50 points. Then for each point s in the bullseye, $X(s) = 50$. ■

In this book we deal frequently with experimental measurements that can include some random measurement error.

■ **Example B.3** Suppose we measure the mass of an object five times to obtain the realizations $m_1 = 10.1$ kg, $m_2 = 10.0$ kg, $m_3 = 10.0$ kg, $m_4 = 9.9$ kg, and $m_5 = 10.1$ kg. We will assume that there is one true mass m, and that the measurements we obtained varied because of random measurement errors e_i, so that

$$m_1 = m + e_1, \quad m_2 = m + e_2, \quad m_3 = m + e_3, \quad m_4 = m + e_4, \quad m_5 = m + e_5. \tag{B.2}$$

We can treat the measurement errors as realizations of a random variable E. Equivalently, since the true mass m is just a constant, we could treat the measurements $m_1, m_2, \ldots, m_5$ as realizations of a random variable M. In practice it makes little difference whether we treat the measurements or the measurement errors as random variables.

Note that, in a Bayesian approach, the mass m of the object would itself be a random variable. This is a viewpoint that we consider in Chapter 11. ■

The relative probability of realization values for a random variable can be characterized by a nonnegative **probability density function (PDF)**, $f_X(x)$, with

$$P(X \leq a) = \int_{-\infty}^{a} f_X(x)\, dx. \tag{B.3}$$

Because the random variable always has some value,

$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1. \tag{B.4}$$

The following definitions give some useful random variables that frequently arise in inverse problems.

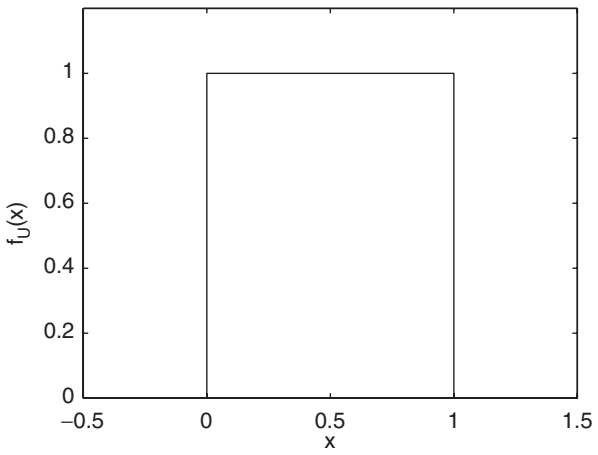

Figure B.1 The PDF for the uniform random variable on [0, 1].

■ **Definition B.3** The **uniform** random variable on the interval $[a, b]$ has the probability density function

$$f_U(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 & x < a \\ 0 & x > b \end{cases}. \tag{B.5}$$

See Figure B.1. ■

■ **Definition B.4** The **normal** or **Gaussian** random variable has the probability density function

$$f_N(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}. \tag{B.6}$$

See Figure B.2. The notation $N(\mu, \sigma^2)$ is used to denote a normal distribution with parameters μ and σ. The **standard normal** random variable, $N(0, 1)$, has $\mu = 0$ and $\sigma = 1$. ■

■ **Definition B.5** The **exponential** random variable has the probability density function

$$f_{exp}(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0 & x < 0 \end{cases}. \tag{B.7}$$

See Figure B.3. ■

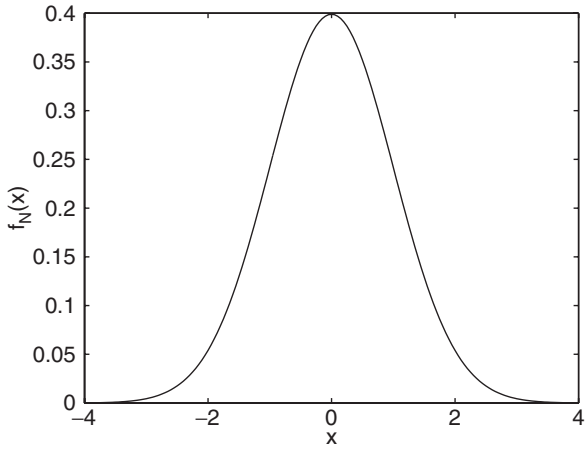

Figure B.2 The PDF of the standard normal random variable.

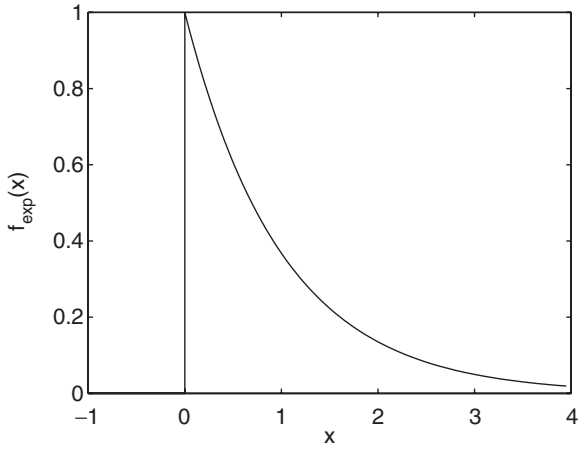

Figure B.3 The exponential probability density function ($\lambda = 1$).

■ **Definition B.6** The **double–sided exponential** random variable has the probability density function

$$f_{\text{dexp}}(x) = \frac{1}{2^{3/2}\sigma} e^{-\sqrt{2}|x-\mu|/\sigma}. \tag{B.8}$$

See Figure B.4. ■

■ **Definition B.7** The χ^2 random variable has the probability density function

$$f_{\chi^2}(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\frac{1}{2}\nu-1} e^{-x/2} \tag{B.9}$$

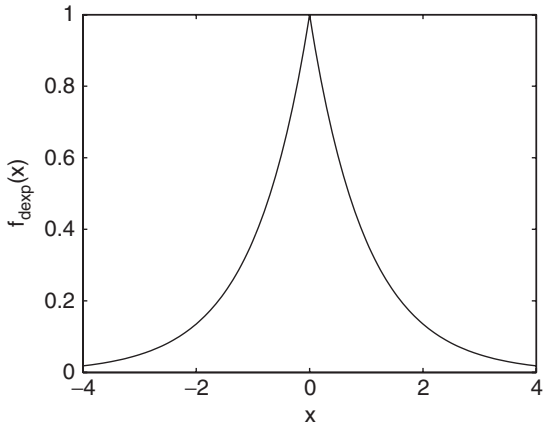

Figure B.4 The double-sided exponential probability density function ($\mu = 0$, $\lambda = 1$).

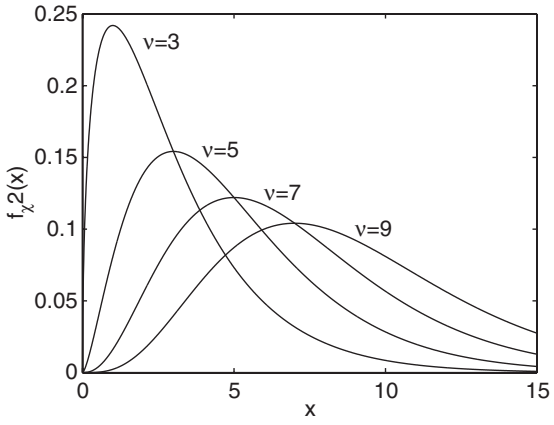

Figure B.5 The χ^2 probability density function for several values of ν.

where the **gamma function** is

$$\Gamma(x) = \int_0^\infty \xi^{x-1} e^{-\xi} d\xi \tag{B.10}$$

and the parameter ν is called the **number of degrees of freedom**. See Figure B.5.

It can be shown that for n independent random variables, X_i with standard normal distributions, the random variable

$$Z = \sum_{i=1}^{n} X_i^2 \tag{B.11}$$

is a χ^2 random variable with $\nu = n$ degrees of freedom [40]. ■

■ **Definition B.8** The **Student's *t* distribution** with ν degrees of freedom has the probability density function

$$f_t(x) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{\nu\pi}} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}. \tag{B.12}$$

■

See Figure B.6. The Student's t distribution is so named because W. S. Gosset used the pseudonym "Student" in publishing the first paper in which the distribution appeared. In the limit as ν goes to infinity, Student's t distribution approaches a standard normal distribution.

The **cumulative distribution function (CDF)** $F_X(a)$ of a one–dimensional random variable X is given by the definite integral of the associated PDF:

$$F_X(a) = P(X \leq a) = \int_{-\infty}^{a} f_X(x)\, dx. \tag{B.13}$$

Note that $F_X(a)$ must lie in the interval $[0, 1]$ for all a, and is a nondecreasing function of a because of the unit area and nonnegativity of the PDF.

For the uniform PDF on the unit interval, for example, the CDF is a ramp function

$$F_U(a) = \int_{-\infty}^{a} f_u(z)\, dz \tag{B.14}$$

$$F_U(a) = \begin{cases} 0 & a \leq 0 \\ a & 0 \leq a \leq 1 \\ 1 & a > 1 \end{cases}. \tag{B.15}$$

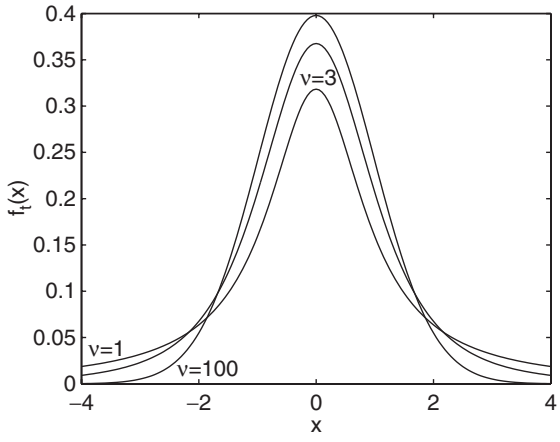

Figure B.6 The Student's t probability density function for $\nu = 1, 3, 100$.

The PDF, $f_X(x)$, or CDF, $F_X(a)$, completely determine the probabilistic properties of a random variable. The probability that a particular realization of X will lie within a general interval $[a, b]$ is

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) \tag{B.16}$$

$$= \int_{-\infty}^{b} f(x)\,dx - \int_{-\infty}^{a} f(x)\,dx = \int_{a}^{b} f(x)\,dx. \tag{B.17}$$

B.2 EXPECTED VALUE AND VARIANCE

■ **Definition B.9** The **expected value** of a random variable X, denoted by $E[X]$ or μ_X, is

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx. \tag{B.18}$$

In general, if $g(X)$ is some function of a random variable X, then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx. \tag{B.19}$$

■

Some authors use the term "mean" for the expected value of a random variable. We will reserve this term for the average of a set of data. Note that the expected value of a random variable is not necessarily identical to the **mode** [the value with the largest value of $f(x)$] nor is it necessarily identical to the **median**, the value of x for which the value of the CDF is $F(x) = 1/2$.

■ **Example B.4** The expected value of an $N(\mu, \sigma)$ random variable X is

$$E[X] = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx \tag{B.20}$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} (x+\mu) e^{-\frac{x^2}{2\sigma^2}}\,dx \tag{B.21}$$

$$= \mu \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}\,dx + \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} x e^{-\frac{x^2}{2\sigma^2}}\,dx. \tag{B.22}$$

The first integral term is μ because the integral of the entire PDF is 1, and the second term is zero because it is an odd function integrated over a symmetric interval. Thus

$$E[X] = \mu. \tag{B.23}$$

■

■ **Definition B.10** The **variance** of a random variable X, denoted by $\text{Var}(X)$ or σ_X^2, is given by

$$\begin{aligned}\text{Var}(X) &= \sigma_X^2 \\ &= E[(X - \mu_X)^2] \\ &= E[X^2] - \mu_X^2 \\ &= \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)\, dx. \end{aligned} \tag{B.24}$$

The **standard deviation** of X, often denoted σ_X, is

$$\sigma_X = \sqrt{\text{Var}(X)}. \tag{B.25}$$

■

The variance and standard deviation serve as measures of the spread of the random variable about its expected value. Since the units of σ are the same as the units of μ, the standard deviation is generally more practical as a measure of the spread of the random variable. However, the variance has many properties that make it more useful for certain calculations.

B.3 JOINT DISTRIBUTIONS

■ **Definition B.11** If we have two random variables X and Y, they *may* have a **joint probability density function (JDF)**, $f(x, y)$ with

$$P(X \leq a \quad \text{and} \quad Y \leq b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x, y)\, dy\, dx. \tag{B.26}$$

■

If X and Y have a joint probability density function, then we can use it to evaluate the expected value of a function of X and Y. The expected value of $g(X, Y)$ is

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y)\, dy\, dx. \tag{B.27}$$

■ **Definition B.12** Two random variables X and Y are **independent** if a JDF exists and is defined by

$$f(x, y) = f_X(x) f_Y(y). \tag{B.28}$$

■

■ **Definition B.13** If X and Y have a JDF, then the **covariance** of X and Y is

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]. \tag{B.29}$$

■

If X and Y are independent, then $E[XY] = E[X]E[Y]$, and $\text{Cov}(X, Y) = 0$. However if X and Y are dependent, it is still possible, given some particular distributions, for X and Y to have $\text{Cov}(X, Y) = 0$. If $\text{Cov}(X, Y) = 0$, X and Y are called **uncorrelated**.

■ **Definition B.14** The **correlation** of X and Y is

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}. \tag{B.30}$$

Correlation is thus a scaled covariance. ■

■ **Theorem B.1** The following properties of Var, Cov, and correlation hold for any random variables X and Y and scalars s and a.

1. $\text{Var}(X) \geq 0$
2. $\text{Var}(X + a) = \text{Var}(X)$
3. $\text{Var}(sX) = s^2\text{Var}(X)$
4. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$
5. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
6. $\rho(X, Y) = \rho(Y, X)$
7. $-1 \leq \rho_{XY} \leq 1$ ■

The following example demonstrates the use of some of these properties.

■ **Example B.5** Suppose that Z is a standard normal random variable. Let

$$X = \mu + \sigma Z. \tag{B.31}$$

Then

$$E[X] = E[\mu] + \sigma E[Z] \tag{B.32}$$

so

$$E[X] = \mu. \tag{B.33}$$

Also,

$$\begin{aligned} \text{Var}(X) &= \text{Var}(\mu) + \sigma^2\text{Var}(Z) \\ &= \sigma^2. \end{aligned} \tag{B.34}$$

Thus if we have a program to generate random numbers with the standard normal distribution, we can use it to generate normal random numbers with any desired expected value and standard deviation. The MATLAB command **randn** generates independent realizations of an $N(0, 1)$ random variable. ■

■ **Example B.6** What is the CDF (or PDF) of the sum of two independent random variables $X + Y$? To see this, we write the desired CDF in terms of an appropriate integral over the JDF, $f(x, y)$, which gives

$$F_{X+Y}(z) = P(X + Y \leq z) \tag{B.35}$$

$$= \iint_{x+y\leq z} f(x, y)\, dx\, dy \tag{B.36}$$

$$= \iint_{x+y\leq z} f_X(x) f_Y(y)\, dx\, dy \tag{B.37}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f_X(x) f_Y(y)\, dx\, dy \tag{B.38}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f_X(x)\, dx\ f_Y(y)\, dy \tag{B.39}$$

$$= \int_{-\infty}^{\infty} F_X(z - y) f_Y(y)\, dy. \tag{B.40}$$

See Figure B.7. The associated PDF is

$$f_{X+Y}(z) = \frac{d}{dz} \int_{-\infty}^{\infty} F_X(z - y) f_Y(y)\, dy \tag{B.41}$$

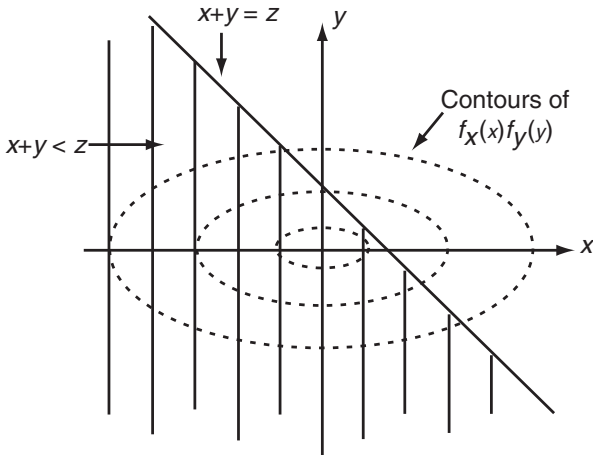

Figure B.7 Integration of a joint probability density function for two independent random variables, X, and Y, to evaluate the CDF of $Z = X + Y$.

$$= \int_{-\infty}^{\infty} \frac{d}{dz} F_X(z-y) f_Y(y)\, dy \tag{B.42}$$

$$= \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\, dy \tag{B.43}$$

$$= f_X(z) * f_Y(z). \tag{B.44}$$

Adding two independent random variables thus produces a new random variable that has a PDF given by the convolution of the PDF's of the two individual variables. ■

The JDF can be used to evaluate the CDF or PDF arising from a general function of jointly distributed random variables. The process is identical to the previous example except that the specific form of the integral limits is determined by the specific function.

■ **Example B.7** Consider the product of two independent, identically distributed, standard normal random variables,

$$Z = XY \tag{B.45}$$

with a JDF given by

$$f(x, y) = f(x) f(y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}. \tag{B.46}$$

The CDF of Z is

$$F(z) = P(Z \le z) = P(XY \le z). \tag{B.47}$$

For $z \le 0$, this is the integral of the JDF over the exterior of the hyperbolas defined by $xy \le z \le 0$, whereas for $z \ge 0$, we integrate over the interior of the complementary hyperbolas $xy \le z \ge 0$. At $z = 0$, the integral covers exactly half of the (x, y) plane (the second and fourth quadrants) and, because of the symmetry of the JDF, has accumulated half of the probability, or 1/2.

The integral is thus

$$F(z) = 2 \int_{-\infty}^{0} \int_{z/x}^{\infty} \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}\, dy\, dx \quad (z \le 0) \tag{B.48}$$

and

$$F(z) = 1/2 + 2 \int_{-\infty}^{0} \int_{0}^{z/x} \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}\, dy\, dx \quad (z \ge 0). \tag{B.49}$$

As in the previous example for the sum of two random variables, the PDF may be obtained from the CDF by differentiating with respect to z. ■

B.4 CONDITIONAL PROBABILITY

In some situations we will be interested in the probability of an event happening given that some other event has also happened.

■ **Definition B.15** The **conditional probability** of A given that B has occurred is given by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{B.50}$$

■

Arguments based on conditional probabilities are often very helpful in computing probabilities. The key to such arguments is the **law of total probability**.

■ **Theorem B.2** Suppose that $B_1, B_2, \ldots, B_n$ are mutually disjoint and exhaustive events. That is, $B_i \cap B_j = \emptyset$ (the empty set) for $i \neq j$, and

$$\bigcup_{i=1}^{n} B_i = S. \tag{B.51}$$

Then

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i). \tag{B.52}$$

■

It is often necessary to reverse the order of conditioning in a conditional probability. Bayes' theorem provides a way to do this.

■ **Theorem B.3 Bayes' Theorem**

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}. \tag{B.53}$$

■

■ **Example B.8** A screening test has been developed for a very serious but rare disease. If a person has the disease, then the test will detect the disease with probability 99%. If a person does not have the disease, then the test will give a false positive detection with probability 1%. The probability that any individual in the population has the disease is 0.01%. Suppose that a randomly selected individual tests positive for the disease. What is the probability that this individual actually has the disease?

Let A be the event "the person tests positive." Let B be the event "the person has the disease." We then want to compute $P(B|A)$. By Bayes' theorem,

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}. \tag{B.54}$$

We have that $P(A|B)$ is 0.99, and that $P(B)$ is 0.0001. To compute $P(A)$, we apply the law of total probability, considering separately the probability of a diseased individual testing positive and the probability of someone without the disease testing positive.

$$P(A) = 0.99 \times 0.0001 + 0.01 \times 0.9999 = 0.010098. \tag{B.55}$$

Thus

$$P(B|A) = \frac{0.99 \times 0.0001}{0.010098} = 0.0098. \tag{B.56}$$

In other words, even after a positive screening test, it is still unlikely that the individual will have the disease. The vast majority of those individuals who test positive will in fact not have the disease. ■

The concept of conditioning can be extended from simple events to distributions and expected values of random variables. If the distribution of X depends on the value of Y, then we can work with the **conditional PDF** $f_{X|Y}(x)$, the **conditional CDF** $F_{X|Y}(a)$, and the **conditional expected value** $E[X|Y]$.

In this notation, we can also specify a particular value of Y by using the notation $f_{X|Y=y}$, $F_{X|Y=y}$, or $E[X|Y = y]$. In working with conditional distributions and expected values, the following versions of the law of total probability can be very useful.

■ **Theorem B.4** Given two random variables X and Y, with the distribution of X depending on Y, we can compute

$$P(X \leq a) = \int_{-\infty}^{\infty} P(X \leq a|Y = y) f_Y(y)\, dy \tag{B.57}$$

and

$$E[X] = \int_{-\infty}^{\infty} E[X|Y = y] f_Y(y)\, dy. \tag{B.58}$$

■

■ **Example B.9** Let U be a random variable uniformly distributed on $(1, 2)$. Let X be an exponential random variable with parameter $\lambda = U$. We will find the expected value of X:

$$E[X] = \int_1^2 E[X|U = u] f_U(u)\, du. \tag{B.59}$$

Since the expected value of an exponential random variable with parameter λ is $1/\lambda$, and the PDF of a uniform random variable on $(1, 2)$ is $f_U(u) = 1$,

$$\begin{aligned} E[X] &= \int_1^2 \frac{1}{u}\, du \\ &= \ln 2. \end{aligned} \tag{B.60}$$

■

B.5 THE MULTIVARIATE NORMAL DISTRIBUTION

■ **Definition B.16** If the random variables $X_1, \ldots, X_n$ have a **multivariate normal** (MVN) **distribution**, then the joint probability density function is

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sqrt{\det(\mathbf{C})}} e^{-(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2} \tag{B.61}$$

where $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_n]^T$ is a vector containing the expected values along each of the coordinate directions of $X_1, \ldots, X_n$, and $\mathbf{C}$ contains the covariances between the random variables:

$$C_{i,j} = \mathrm{Cov}(X_i, X_j). \tag{B.62}$$

Notice that if $\mathbf{C}$ is singular, then the joint probability density function involves a division by zero and is simply not defined. ■

The vector $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{C}$ completely characterize the MVN distribution. There are other multivariate distributions that are not completely characterized by the expected values and covariance matrix.

■ **Theorem B.5** Let $\mathbf{X}$ be a multivariate normal random vector with expected values defined by the vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$, and let $\mathbf{Y} = \mathbf{AX}$. Then $\mathbf{Y}$ is also multivariate normal, with

$$E[\mathbf{Y}] = \mathbf{A}\boldsymbol{\mu} \tag{B.63}$$

and

$$\mathrm{Cov}(\mathbf{Y}) = \mathbf{ACA}^T. \tag{B.64}$$

■

■ **Theorem B.6** If we have an n–dimensional MVN distribution with covariance matrix $\mathbf{C}$ and expected value $\boldsymbol{\mu}$, and the covariance matrix is of full rank, then the quantity

$$Z = (\mathbf{X} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \tag{B.65}$$

has a χ^2 distribution with n degrees of freedom. ■

■ **Example B.10** We can generate vectors of random numbers according to an MVN distribution with known mean and covariance matrix by using the following process, which is very similar to the process for generating random normal scalars.

1. Find the Cholesky factorization $\mathbf{C} = \mathbf{L}\mathbf{L}^T$.
2. Let $\mathbf{Z}$ be a vector of n independent $N(0, 1)$ random numbers.
3. Let $\mathbf{X} = \boldsymbol{\mu} + \mathbf{L}\mathbf{Z}$.

Because $E[\mathbf{Z}] = \mathbf{0}$, $E[\mathbf{X}] = \boldsymbol{\mu} + \mathbf{L}\mathbf{0} = \boldsymbol{\mu}$. Also, since $\text{Cov}(\mathbf{Z}) = \mathbf{I}$ and $\text{Cov}(\boldsymbol{\mu}) = \mathbf{0}$, $\text{Cov}(\mathbf{X}) = \text{Cov}(\boldsymbol{\mu} + \mathbf{L}\mathbf{Z}) = \mathbf{L}\mathbf{I}\mathbf{L}^T = \mathbf{C}$. ■

B.6 THE CENTRAL LIMIT THEOREM

■ **Theorem B.7** Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (IID) random variables with a finite expected value μ and variance σ^2. Let

$$Z_n = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n}\sigma}. \tag{B.66}$$

In the limit as n approaches infinity, the distribution of Z_n approaches the standard normal distribution. ■

The central limit theorem shows why quasinormally distributed random variables appear so frequently in nature; the sum of numerous independent random variables produces an approximately normal random variable, regardless of the distribution of the underlying IID variables. In particular, this is one reason that measurement errors are often normally distributed. As we saw in Chapter 2, having normally distributed measurement errors leads us to consider least squares solutions to parameter estimation and inverse problems.

B.7 TESTING FOR NORMALITY

Many of the statistical procedures that we will use assume that data are normally distributed. Fortunately, the statistical techniques that we describe are generally robust in the face of small deviations from normality. Large deviations from the normal distribution can cause problems. Thus it is important to be able to examine a data set to see whether or not the distribution is approximately normal.

Plotting a histogram of the data provides a quick view of the distribution. The histogram should show a roughly "bell shaped" distribution, symmetrical around a single peak. If the histogram shows that the distribution is obviously skewed, then it would be unwise to assume that the data are normally distributed.

The **Q–Q plot** provides a more precise graphical test of whether a set of data could have come from a particular distribution. The data points

$$\mathbf{d} = [d_1, d_2, \ldots, d_n]^T \tag{B.67}$$

are first sorted in numerical order from smallest to largest into a vector $\mathbf{y}$, which is plotted versus

$$x_i = F^{-1}((i - 0.5)/n) \quad (i = 1, 2, \ldots, n) \tag{B.68}$$

where $F(x)$ is the CDF of the distribution against which we wish to compare our observations.

If we are testing to see if the elements of $\mathbf{d}$ could have come from the normal distribution, then $F(x)$ is the CDF for the standard normal distribution:

$$F_N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}z^2}\, dz. \tag{B.69}$$

If the elements of $\mathbf{d}$ are normally distributed, the points (y_i, x_i) will follow a straight line with a slope and intercept determined by the standard deviation and expected value, respectively, of the normal distribution that produced the data.

■ **Example B.11** Figure B.8 shows the histogram from a set of 100 data points. The characteristic bell-shaped curve in the histogram makes it seem that these data might be normally distributed. The sample mean is 0.20 and the sample standard deviation is 1.81.

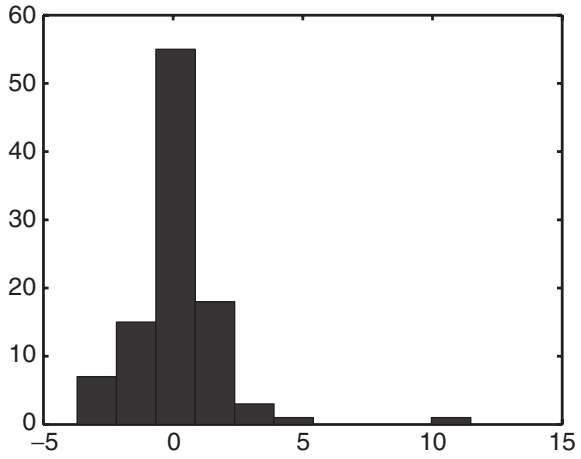

Figure B.8 Histogram of a sample data set.

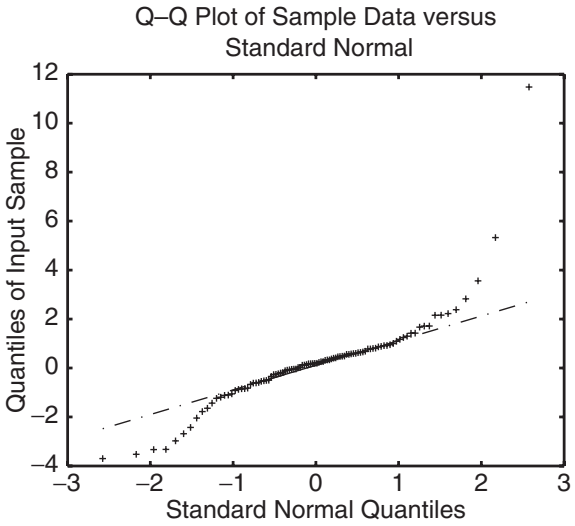

Figure B.9 $Q-Q$ plot for the sample data set.

Figure B.9 shows the $Q-Q$ plot for our sample data set. It is apparent that the data set contains more extreme values than the normal distribution would predict. In fact, these data were generated according to a t distribution with five degrees of freedom, which has broader tails than the normal distribution. See Figure B.6. Because of these deviations from normality, it would be wise not to treat these data as if they were normally distributed. ■

There are a number of statistical tests for normality. These tests, including the Kolmogorov–Smirnov test, Anderson–Darling test, and Lilliefors test, each produce probabilistic measures called p-values. A small p-value indicates that the observed data would be unlikely if the distribution were in fact normal, while a larger p-value is consistent with normality.

B.8 ESTIMATING MEANS AND CONFIDENCE INTERVALS

Given a collection of noisy measurements $m_1, m_2, \ldots, m_n$ of some quantity of interest, how can we estimate the true value m, and how uncertain is our estimate? This is a classic problem in statistics.

We will assume first that the measurement errors are independent and normally distributed with expected value 0 and some unknown standard deviation σ. Equivalently, the measurements themselves are normally distributed with expected value m and standard deviation σ.

We begin by computing the measurement average

$$\bar{m} = \frac{m_1 + m_2 + \cdots + m_n}{n}. \tag{B.70}$$

This **sample mean** $\bar{m}$ will serve as our estimate of m. We will also compute an estimate s of the standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(m_i - \bar{m})^2}{n-1}}. \tag{B.71}$$

The key to our approach to estimating m is the following theorem.

■ **Theorem B.8** (The Sampling Theorem) Under the assumption that measurements are independent and normally distributed with expected value m and standard deviation σ, the random quantity

$$t = \frac{m - \bar{m}}{s/\sqrt{n}} \tag{B.72}$$

has a **Student's *t* distribution** with $n-1$ degrees of freedom. ■

If we had the true standard deviation σ instead of the estimate s, then t would in fact be normally distributed with expected value 0 and standard deviation 1. This does not quite work out because we have used an estimate s of the standard deviation. For smaller values of n, the estimate s is less accurate, and the t distribution therefore has fatter tails than the standard normal distribution. As n goes to infinity, s becomes a better estimate of σ and it can be shown that the t distribution converges to a standard normal distribution [40].

Let $t_{n-1,0.975}$ be the 97.5%-tile of the t distribution and let $t_{n-1,0.025}$ be the 2.5%-tile of the t distribution. Then

$$P\left(t_{n-1,0.025} \leq \frac{m - \bar{m}}{s/\sqrt{n}} \leq t_{n-1,0.975}\right) = 0.95. \tag{B.73}$$

This can be rewritten as

$$P\left(\left(t_{n-1,0.025}s/\sqrt{n}\right) \leq (m - \bar{m}) \leq \left(t_{n-1,0.975}s/\sqrt{n}\right)\right) = 0.95. \tag{B.74}$$

We can construct the 95% **confidence interval** for m as the interval from $\bar{m} + t_{n-1,0.025}s/\sqrt{n}$ to $\bar{m} + t_{n-1,0.975}s/\sqrt{n}$. Because the t distribution is symmetric, this can also be written as $\bar{m} - t_{n-1,0.975}s/\sqrt{n}$ to $\bar{m} + t_{n-1,0.975}s/\sqrt{n}$.

As we have seen, there is a 95% probability that when we construct the confidence interval, that interval will contain the true mean, m. Note that we have not said that, given a particular set of data and the resulting confidence interval, there is a 95% probability that m is in the confidence interval. The semantic difficulty here is that m is not a random variable, but is rather some true fixed quantity that we are estimating; the measurements $m_1, m_2, \ldots, m_n$, and the calculated $\bar{m}$, s and confidence interval are the random quantities.

■ **Example B.12** Suppose that we want to estimate the mass of an object and obtain the following ten measurements of the mass (in grams):

$$\begin{array}{ccccc} 9.98 & 10.07 & 9.94 & 10.22 & 9.98 \\ 10.01 & 10.11 & 10.01 & 9.99 & 9.92 \end{array}. \tag{B.75}$$

The sample mean is $\bar{m} = 10.02$ g. The sample standard deviation is $s = 0.0883$. The 97.5%-tile of the t distribution with $n - 1 = 9$ degrees of freedom is (from a t-table or function) 2.262. Thus our 95% confidence interval for the mean is

$$\left[\bar{m} - 2.262s/\sqrt{n}, \bar{m} + 2.262s/\sqrt{n}\right] \text{ g}. \tag{B.76}$$

Substituting the values for $\bar{m}$, s, and n, we get an interval of

$$\left[10.02 - 2.262 \times 0.0883/\sqrt{10},\ 10.02 + 2.262 \times 0.0883/\sqrt{10}\right] \text{ g} \tag{B.77}$$

or

$$[9.96, 10.08] \text{ g}. \tag{B.78}$$

■

The foregoing procedure for constructing a confidence interval for the mean using the t distribution was based on the assumption that the measurements were normally distributed. In situations where the data are not normally distributed this procedure can fail in a very dramatic fashion. See Exercise B.8. However, it may be safe to generate an approximate confidence interval using this procedure if (1) the number n of data is large (50 or more) or (2) the distribution of the data is not strongly skewed and n is at least 15.

B.9 HYPOTHESIS TESTS

In some situations we want to test whether or not a set of normally distributed data could reasonably have come from a normal distribution with expected value μ_0. Applying the sampling theorem, we see that if our data did come from a normal distribution with expected value μ_0, then there would be a 95% probability that

$$t_{\text{obs}} = \frac{\mu_0 - \bar{m}}{s/\sqrt{n}} \tag{B.79}$$

would lie in the interval

$$[F_t^{-1}(0.025),\ F_t^{-1}(0.975)] = [t_{n-1,0.025}, t_{n-1,0.975}] \tag{B.80}$$

and only a 5% probability that t would lie outside this interval. Equivalently, there is only a 5% probability that $|t_{\text{obs}}| \geq t_{n-1,0.975}$.

This leads to the ***t*–test**: If $|t_{\text{obs}}| \geq t_{n-1,0.975}$, then we reject the hypothesis that $\mu = \mu_0$. On the other hand, if $|t| < t_{n-1,0.975}$, then we cannot reject the hypothesis that $\mu = \mu_0$. Although the 95% confidence level is traditional, we can also perform the t–test at a 99% or some other confidence level. In general, if we want a confidence level of $1 - \alpha$, then we compare $|t|$ to $t_{n-1,1-\alpha/2}$.

In addition to reporting whether or not a set of data passes a t–test it is good practice to report the associated ***t*–test *p*–value** . The p-value associated with a t–test is the largest value of α for which the data passes the t–test. Equivalently, it is the probability that we could have gotten a greater t value than we have observed, given that all of our assumptions are correct.

■ **Example B.13** Consider the following data:

$$\begin{array}{rrrrr} 1.2944 & -0.3362 & 1.7143 & 2.6236 & 0.3082 \\ 1.8580 & 2.2540 & -0.5937 & -0.4410 & 1.5711 \end{array}. \tag{B.81}$$

These appear to be roughly normally distributed, with a mean that seems to be larger than 0. We will test the hypothesis $\mu = 0$. The t statistic is

$$t_{\text{obs}} = \frac{\mu_0 - \bar{m}}{s\sqrt{n}}, \tag{B.82}$$

which for this data set is

$$t_{\text{obs}} = \frac{0 - 1.0253}{1.1895/\sqrt{10}} \approx -2.725. \tag{B.83}$$

Because $|t_{\text{obs}}|$ is larger than $t_{9,0.975} = 2.262$, we reject the hypothesis that these data came from a normal distribution with expected value 0 at the 95% confidence level. ■

The t–test (or any other statistical test) can fail in two ways. First, it could be that the hypothesis that $\mu = \mu_0$ is true, but our particular data set contained some unlikely values and failed the t–test. Rejecting the hypothesis when it is in fact true is called a **type I error** . We can control the probability of a type I error by decreasing α.

The second way in which the t–test can fail is more difficult to control. It could be that the hypothesis $\mu = \mu_0$ was false, but the sample mean was close enough to μ_0 to pass the t–test. In this case, we have a **type II error**. The probability of a type II error depends very much on how close the true mean is to μ_0. If the true mean $\mu = \mu_1$ is very close to μ_0, then a type II error is quite likely. If the true mean $\mu = \mu_1$ is very far from μ_0, then a type II error will be less likely. Given a particular alternative hypothesis, $\mu = \mu_1$, we call the probability of a type II error $\beta(\mu_1)$ and call the probability of not making a type II error $(1 - \beta(\mu_1))$ the **power** of the test. We can estimate $\beta(\mu_1)$ by repeatedly generating sets of n random numbers with $\mu = \mu_1$ and performing the hypothesis test on the sets of random numbers. See Exercise B.9.

The results of a hypothesis test should always be reported with care. It is important to discuss and justify any assumptions (such as the normality assumption made in the t–test) underlying the test. The p-value should always be reported along with whether or not the

hypothesis was rejected. If the hypothesis was not rejected and some particular alternative hypothesis is available, it is good practice to estimate the power of the hypothesis test against this alternative hypothesis. Confidence intervals for the mean should be reported along with the results of a hypothesis test.

It is important to distinguish between the statistical significance of a hypothesis test and the actual magnitude of any difference between the observed mean and the hypothesized mean. For example, with very large n it is nearly always possible to achieve statistical significance at the 95% confidence level, even though the observed mean may differ from the hypothesis by only 1% or less.

B.10 EXERCISES

B.1 Compute the expected value and variance of a uniform random variable in terms of the parameters a and b.

B.2 Compute the CDF of an exponential random variable with parameter λ.

B.3 Show that

$$\text{Cov}(aX, Y) = a\text{Cov}(X, Y) \tag{B.84}$$

and that

$$\text{Cov}(X + Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z). \tag{B.85}$$

B.4 Show that the PDF for the sum of two independent uniform random variables on $[a, b] = [0, 1]$ is

$$f(x) = \begin{cases} 0 & (x \leq 0) \\ x & (0 \leq x \leq 1) \\ 2 - x & (1 \leq x \leq 2) \\ 0 & (x \geq 0) \end{cases}. \tag{B.86}$$

B.5 Suppose that X and Y are independent random variables. Use conditioning to find a formula for the CDF of $X + Y$ in terms of the PDF's and CDF's of X and Y.

B.6 Suppose that $\mathbf{x} = (X_1, X_2)^T$ is a vector composed of two random variables with a multivariate normal distribution with expected value $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$, and that $\mathbf{A}$ is a 2 by 2 matrix. Use properties of expected value and covariance to show that $\mathbf{y} = \mathbf{Ax}$ has expected value $\mathbf{A}\boldsymbol{\mu}$ and covariance $\mathbf{ACA}^T$.

B.7 Consider the following data, which we will assume are drawn from a normal distribution:

$$\begin{matrix} -0.4326 & -1.6656 & 0.1253 & 0.2877 & -1.1465 \\ 1.1909 & 1.1892 & -0.0376 & 0.3273 & 0.1746 \end{matrix}.$$

Find the sample mean and standard deviation. Use these to construct a 95% confidence interval for the mean. Test the hypothesis $H_0 : \mu = 0$ at the 95% confidence level. What do you conclude? What was the corresponding p-value?

B.8 Using MATLAB, repeat the following experiment 1000 times. Use the Statistics Toolbox function **exprnd()** to generate five exponentially distributed random numbers with $\mu = 10$. Use these five random numbers to generate a 95% confidence interval for the mean. How many times out of the 1000 experiments did the 95% confidence interval cover the expected value of 10? What happens if you instead generate 50 exponentially distributed random numbers at a time? Discuss your results.

B.9 Using MATLAB, repeat the following experiment 1000 times. Use the **randn** function to generate a set of 10 normally distributed random numbers with expected value 10.5 and standard deviation 1. Perform a t–test of the hypothesis $\mu = 10$ at the 95% confidence level. How many type II errors were committed? What is the approximate power of the t–test with $n = 10$ against the alternative hypothesis $\mu = 10.5$? Discuss your results.

B.10 Using MATLAB, repeat the following experiment 1000 times. Using the **exprnd()** function of the Statistics Toolbox, generate five exponentially distributed random numbers with expected value 10. Take the average of the five random numbers. Plot a histogram and a probability plot of the 1000 averages that you computed. Are the averages approximately normally distributed? Explain why or why not. What would you expect to happen if you took averages of 50 exponentially distributed random numbers at a time? Try it and discuss the results.

B.11 NOTES AND FURTHER READING

Most of the material in this appendix can be found in virtually any introductory textbook in probability and statistics. Some recent textbooks include [5, 22]. The multivariate normal distribution is a somewhat more advanced topic that is often ignored in introductory courses. [142] has a good discussion of the multivariate normal distribution and its properties. Numerical methods for probability and statistics are a specialized topic. Two standard references include [84, 161].