Chapter 2

Some Comments on Probability Theory

2.1 NOISE AND RANDOM VARIABLES

In the preceding chapter, we represented the results of an experiment as a vector **d** whose elements were individual measurements. Usually, however, a single number is insufficient results of an experiment represented the **d** whose elements as a vector to represent a single observation. Measurements contain noise, and if an observation were to be performed several times, each measurement would be different (Figure 2.1). Information about the range and shape of this scatter must also be provided to characterize the data completely.

The concept of a *random variable* is used to describe this property. Each random variable has definite and precise properties, governing the range and shape of the scatter of values one observes. These properties cannot be measured directly; however, one can only make individual measurements, or *realizations*, of the random variable and try to estimate its true properties from these data.

The true properties of the random variable d are specified by a *probability density function* $p(d)$ (abbreviated *p.d.f.*). This function gives the probability that a particular realization of the random variable will have a value in the neighborhood of d. The probability that the measurement is between d and $d+\mathrm{d}d$ is $p(d)\ \mathrm{d}d$ (Figure 2.2). (Our choice of the variable name "d" for "data" makes the differential "$\mathrm{d}d$" look a bit funny, but we will just have to live with it.)

Since each measurement must have some value, the probability that d lies somewhere between $-\infty$ and $+\infty$ is complete certainty (usually given the value of 100% or unity), which is written as

$$\int_{-\infty}^{+\infty} p(d)\mathrm{d}d = 1 \tag{2.1}$$

The probability P that d lies in some specific range, say between d_1 and d_2, is the integral of $p(d)$ over that range:

$$P(d_1, d_2) = \int_{d_1}^{d_2} p(d)\mathrm{d}d \tag{2.2}$$

Geophysical Data Analysis: Discrete Inverse Theory. DOI: 10.1016/B978-0-12-397160-9.00002-3

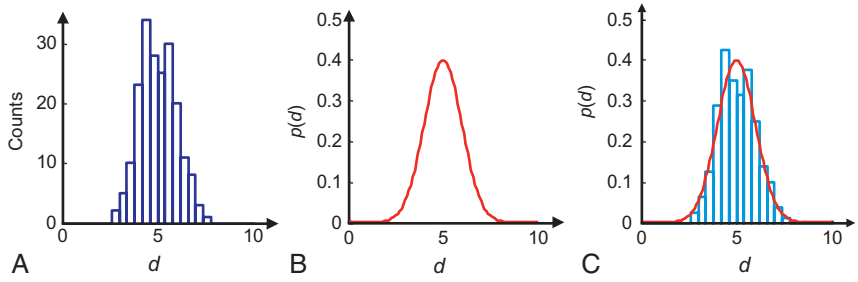

FIGURE 2.1 (A) Histogram showing data from 200 repetitions of an experiment in which datum d is measured. Noise causes observations to scatter about their mean value, $\langle d \rangle = 5$. (B) Probability density function (p.d.f.), $p(d)$, of the data. (C) Histogram (blue) and p.d.f. (red) superimposed. Note that the histogram has a shape similar to the p.d.f. *MatLab* script gda02_01.

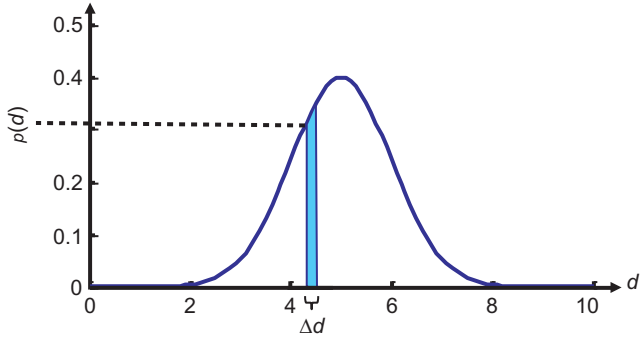

FIGURE 2.2 The shaded area $p(d)\Delta d$ of the probability density function $p(d)$ gives the probability, P, that the observation will fall between d and $d+\Delta d$. *MatLab* script gda02_02.

The special case of $d_1 = -\infty$, $d_2 = d$, which represents the probability that the value of the random variable is less than or equal to a given value d, is called the *cumulative distribution function* $P(d)$ (abbreviated *c.d.f.*). Note that the numerical value of P represents an actual probability, while the numerical value of p does not.

In *MatLab*, we use a vector `d` evenly spaced values with sampling `Dd` to represent the random variable and we use a vector `p` to represent the probability density function at corresponding values of d. The total probability `Ptotal` (which should be unity) and the cumulative probability distribution `P` are calculated as

```
Ptotal = Dd * sum(p);
P = Dd * cumsum(p);
```

(*MatLab* script gda02_03)

Here we are employing the Riemann approximation for an integral, $\int p(d)\,\mathrm{d}d \approx \Delta d \sum_i p(d_i)$. Note that the `sum()` function returns a scalar, the sum of the elements of `p`, whereas the `cumsum()` function returns a vector, the running sum of the elements of `p`.

The probability density function $p(d)$ completely describes the random variable, d. Unfortunately, it is a continuous function that may be quite complicated. A few numbers that summarize the major properties of the probability density function can be very helpful. One such kind of number indicates the typical numerical value of a measurement. The most likely measurement is the one with the highest probability, that is, the value of d at which $p(d)$ is peaked (Figure 2.3). However, if the distribution is skewed, this *maximum likelihood point* may not be a good indication of the typical measurement, since a wide range of other values also has high probability. In such instances, the *mean*, or *expected* measurement, $\langle d \rangle$, is a better characterization of a typical measurement. This number is the "balancing point" of the distribution and is given by

$$\langle d \rangle = E(d) = \int_{-\infty}^{+\infty} d\, p(d)\, \mathrm{d}d \tag{2.3}$$

Another property of a distribution is its overall width. Wide distributions imply very noisy data, and narrow ones imply relatively noise-free data. One way of measuring the width of a distribution is to multiply it by a function that is zero near the center (peak) of the distribution and that grows on either side of the peak (Figure 2.4). If the distribution is narrow, then the resulting function will be everywhere small; if the distribution is wide, then the result will be large.

A quantitative measure of the width of the peak is the area under the resulting function. If one chooses the parabola $(d - \langle d \rangle)^2$ as the function, where $\langle d \rangle = E(d)$ is the expected value of the random variable, then this measure is called the *variance* σ^2 of the distribution and is written as

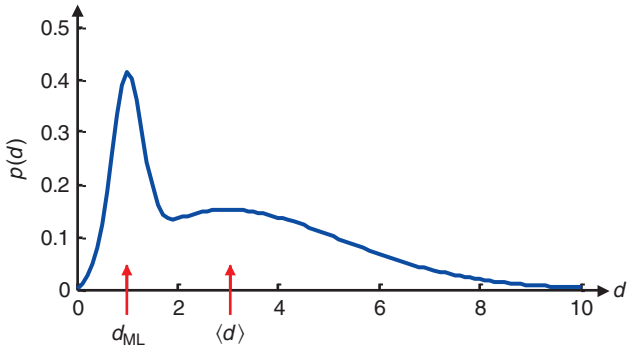

FIGURE 2.3 The maximum likelihood point d_{ML} of the probability density function $p(d)$ gives the most probable value of the datum d. In general, this value can be different than the mean datum $\langle d \rangle$ which is at the "balancing point" of the distribution. *MatLab* script gda02_04.

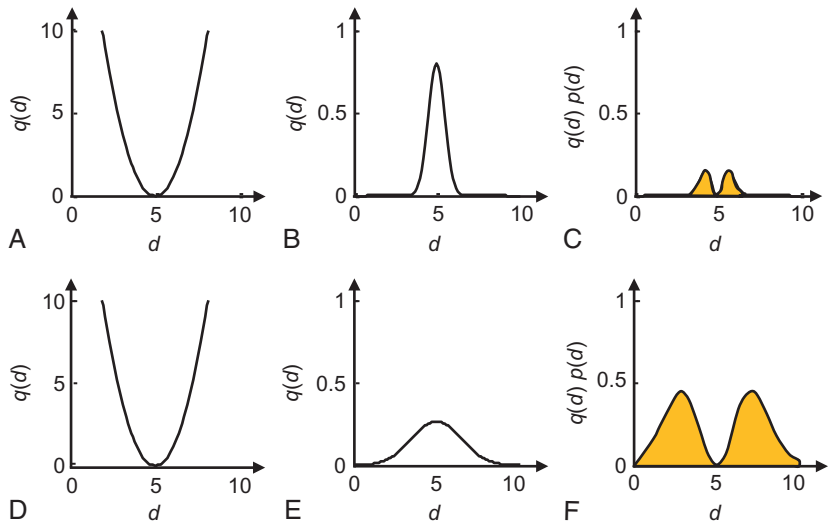

FIGURE 2.4 (A and D) Parabola of the form $q(d) = (d - \langle d \rangle)^2$ is used to measure the width of two probability density functions $p(d)$ (B and E), which have the same mean $\langle d \rangle$ but different widths. The product qp is everywhere small for the narrow function (C) but had two large peaks for the wider distribution (F). The area (shaded orange) under qp is a measure of the width of the function $p(d)$ and is called the variance. The variances of (A) and (F) are $(0.5)^2$ and $(1.5)^2$, respectively. *MatLab* script gda02_05.

$$\sigma^2 = \int_{-\infty}^{+\infty} (d - \langle d \rangle)^2 p(d) \mathrm{d}d \tag{2.4}$$

The square root of the variance, σ, is a measure of the width of the distribution. In *MatLab*, `the` expected value and variance are computed as

```
Ed = Dd * sum(d.*p);
sigma2 = Dd * sum(((d-Ed).^2).*p);
```

(*MatLab* script gda02_05)

Here `d` is a vector of equally spaced values of the random variable d, with spacing `Dd`, and `p` is the corresponding value of the probability density function.

As we will discuss further in Chapter 5, the mean and variance can be estimated from a set of N realizations of data d_i as

$$\langle d \rangle^{\text{est}} = \frac{1}{N} \sum_{i=1}^{N} d_i \quad \text{and} \quad (\sigma^2)^{\text{est}} = \frac{1}{N-1} \sum_{i=1}^{N} (d_i - \langle d \rangle^{\text{est}})^2 \tag{2.5}$$

The quantity $\langle d \rangle^{\text{est}}$ is called the *sample mean* and the quantity σ^{est} is called the *sample standard deviation*. In *MatLab*, these estimates can be computed using the `mean(dr)` and `std(dr)` functions, where `dr` is a vector of N realizations of the random variable d.

2.2 CORRELATED DATA

Experiments usually involve the collection of more than one datum. We therefore need to quantify the probability that a set of random variables will take on a given value. The joint probability density function $p(\mathbf{d})$ is the probability that the first datum will be in the neighborhood of d_1, that the second will be in the neighborhood of d_2, etc. If the data are independent—that is, if there are no patterns in the occurrence of the values between pairs of random variables—then this joint distribution is just the product of the individual distributions (Figure 2.5)

$$p(\mathbf{d}) = p(d_1)\, p(d_2)\, p(d_3) \cdots p(d_N) \tag{2.6}$$

The probability density function for a single random variable, say d_i, irrespective of all the others, is computed by integrating $p(\mathbf{d})$ over all the other variables:

$$p(d_i) = \underbrace{\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty}}_{(N-1 \text{ integrals})} p(\mathbf{d}) \mathrm{d}d_j \mathrm{d}d_k \cdots \mathrm{d}d_l \tag{2.7}$$

In some experiments, measurements *are* correlated. High values of one datum tend to occur consistently with either high or low values of another datum (Figure 2.6). The joint distribution for such data must be constructed to take this correlation into account. Given a joint distribution $p(d_1, d_2)$ for two random variables d_1 and d_2, one can test for correlation by selecting a function that divides the (d_1, d_2) plane into four quadrants of alternating sign, centered on the mean of the distribution (Figure 2.7). If one multiplies the distribution by this function, and then sums up the area, the result will be zero for uncorrelated distributions, since they tend to lie equally in all four quadrants. Correlated distributions will have either positive or negative area, since they tend to be concentrated in two

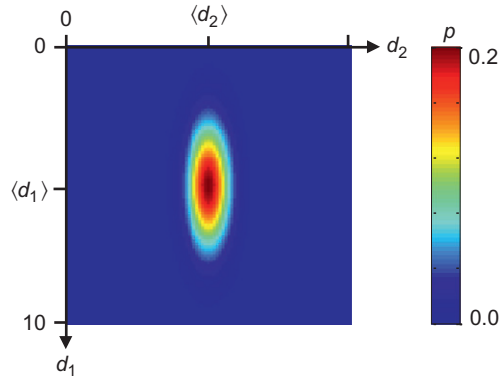

FIGURE 2.5 The probability density function $p(d_1,d_2)$ is displayed as an image, with values given by the accompanying color bar. These data are uncorrelated, since especially large values of d_2 are no more or less likely if d_1 is large or small. In this example, the variance of d_1 and d_2 are $\sigma_1^2 = (1.5)^2$ and $\sigma_2{}^2 = (0.5)^2$, respectively. *MatLab* script gda02_06.

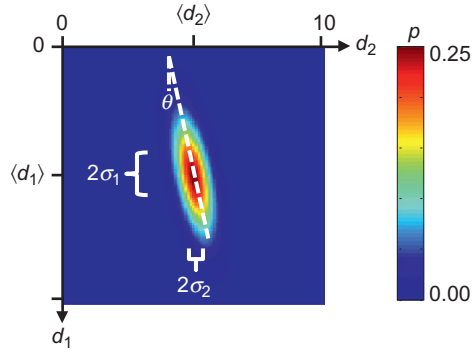

FIGURE 2.6 The probability density function $p(d_1,d_2)$ is displayed as an image, with values given by the accompanying color bar. These data are positively correlated, since large values of d_2 are especially probable if d_1 is large. The function has means $\langle d_1\rangle = 5$ and $\langle d_2\rangle = 5$ and widths in the coordinate directions $\sigma_1 = 1.5$ and $\sigma_2 = 0.5$. The angle θ is a measure of the degree of correlation and is related to the covariance $\text{cov}(d_1,d_2) = 0.4$. *MatLab* script gda02_07.

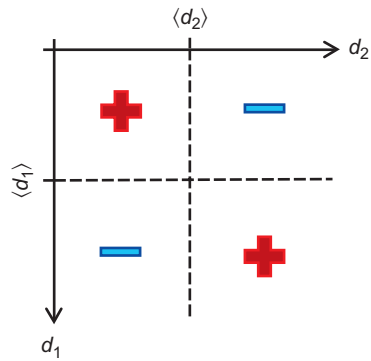

FIGURE 2.7 The function $q(d_1,d_2) = (d_1 - \langle d_2\rangle)(d_2 - \langle d_2\rangle)$ divides the (d_1,d_2) plane into four quadrants of alternating sign.

opposite quadrants (Figure 2.8). If $[d_1 - \langle d_1\rangle][d_2 - \langle d_2\rangle]$ is used as the function, the resulting measure of correlation is called the covariance:

$$\text{cov}(d_1, d_2) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} [d_1 - \langle d_1\rangle][d_2 - \langle d_2\rangle] p(d_1, d_2)\,\mathrm{d}d_1\,\mathrm{d}d_2 \tag{2.8}$$

Note that the covariance of a datum with itself is just the variance. The covariance, therefore, characterizes the basic shape of the joint distribution.

When there are many data given by the vector $\mathbf{d}$, it is convenient to define a vector of expected values and a matrix of covariances as

$$\begin{aligned} \langle d\rangle_i &= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} d_i p(\mathbf{d})\,\mathrm{d}d_1\cdots\mathrm{d}d_N \\ [\text{cov}\,\mathbf{d}]_{ij} &= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} [d_i - \langle d_i\rangle][d_j - \langle d_j\rangle] p(\mathbf{d})\,\mathrm{d}d_1\cdots\mathrm{d}d_N \end{aligned} \tag{2.9}$$

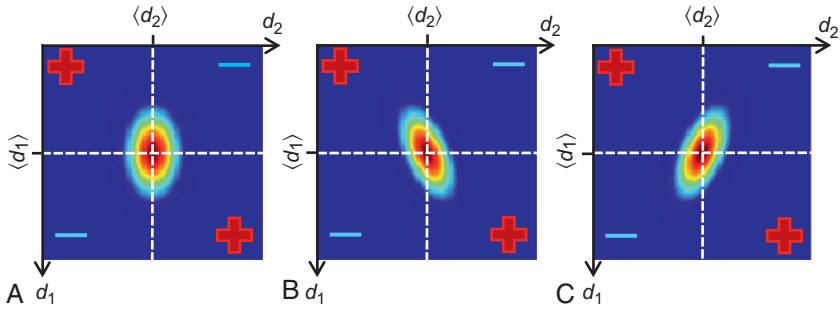

FIGURE 2.8 The probability density function $p(d_1,d_2)$ displayed as an image, when the data are (A) uncorrelated, (B) positively correlated, and (C) negatively correlated. The dashed lines indicated the four quadrants of alternating sign used to determine the correlation (see Figure 2.7). *MatLab* script gda02_08.

Henceforth, we will abbreviate these multidimensional integrals as $\int \mathrm{d}^N d$. The diagonal elements of the covariance matrix are variances. They are measures of the scatter in the data. The off-diagonal elements are covariances. They indicate the degree to which pairs of data are correlated. Notice that the integral for the mean can be written in terms of the univariate probability density function $p(d_i)$ and the integral for the variance can be written in terms of the bivariate probability density function $p(d_i,d_j)$, since the other dimension of $p(\mathbf{d})$ are just "integrated away" to unity:

$$\begin{aligned} \langle d \rangle_i &= \int_{-\infty}^{+\infty} d_i p(d_i) \mathrm{d}d_i \\ [\mathrm{cov}\ \mathbf{d}]_{ij} &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [d_i - \langle d_i \rangle][d_j - \langle d_j \rangle] p(d_i, d_j) \mathrm{d}d_i \mathrm{d}d_j \end{aligned} \tag{2.10}$$

The covariance matrix can be estimated from a set of N realizations of data. Suppose that there are N different types of data and that K realizations of them have been observed. The data can be organized into a matrix $\mathbf{D}$, with the N columns referring to the different data types and the K rows to the different realizations. The *sample covariance* is then

$$[\mathrm{cov}\ \mathbf{d}]_{ij}^{\mathrm{est}} = \frac{1}{K} \sum_{k=1}^{K} (D_{ki} - \langle D_i \rangle^{\mathrm{est}})(D_{kj} - \langle D_j \rangle^{\mathrm{est}}) \tag{2.11}$$

Here $\langle D_i \rangle^{\mathrm{est}}$ is the sample mean of the ith data type. The *MatLab* function `cov(D)` implements this formula.

2.3 FUNCTIONS OF RANDOM VARIABLES

The basic premise of inverse theory is that the data and model parameters are related. Any method that solves the inverse problem—that estimates a model parameter on the basis of data—will map errors from the data to the estimated

model parameters. Thus the *estimates* of the model parameters are themselves random variables, which are described by a distribution $p(\mathbf{m}^{\text{est}})$. Whether or not the *true* model parameters are random variables depends on the problem. It is appropriate to consider them deterministic quantities in some problems and random variables in others. *Estimates* of the model parameters, however, are always random variables.

We need the tools to transform probability density functions from $p(\mathbf{d})$ to $p(\mathbf{m})$ when the relationship $\mathbf{m}(\mathbf{d})$ is known. We start simply and consider just one datum and one model parameter, related by the simple function $m(d) = 2d$. Now suppose that $p(d)$ is *uniform* on the interval (0,1); that is, d has equal probability of being anywhere in this range. The probability density function is constant and must have amplitude $p(d) = 1$, since the total probability must be unity (width $\times$ height $= 1 \times 1 = 1$). The probability density function $p(m)$ is also uniform, but on the interval (0, 2), since m is twice d. Thus, $p(m) = ½$, since its total probability must also be unity (width $\times$ height $= 2 \times ½ = 1$) (Figure 2.9). This result shows that $p(m)$ is not merely $p[d(m)]$, but rather must include a factor that accounts for the stretching (or shrinking) of the m-axis with respect to the d-axis.

This stretching factor can be derived by transforming the integral for total probability:

$$1 = \int_{d_{\min}}^{d_{\max}} p(d)\mathrm{d}d = \int_{d(m_{\min})}^{d(m_{\max})} p[d(m)] \frac{\mathrm{d}d}{\mathrm{d}m} \mathrm{d}m = \int_{m_{\min}}^{m_{\max}} p(m)\mathrm{d}m \tag{2.12}$$

By inspection, $p(m) = p[d(m)]\mathrm{d}d/\mathrm{d}m$, so the stretching factor is $\mathrm{d}d/\mathrm{d}m$. The limits $(d_{\min}, d_{\max})$ transform to $(m_{\min}, m_{\max})$. However, depending upon the function $m(d)$, we may find that $m_{\min} > m_{\max}$; that is, the direction of integration might be reversed ($m(d) = 1/d$ would be one such case). We handle this problem by adding an absolute value sign

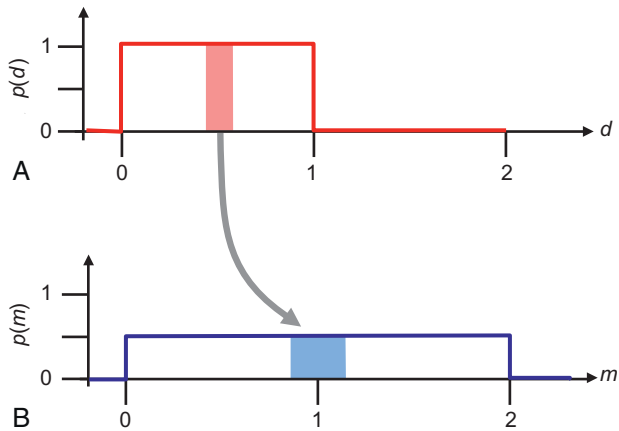

FIGURE 2.9 (A) The uniform probability density function $p(d) = 1$ on the interval $0 < d < 1$. (B) The transformed probability density function $p(m)$, given the relationship $m = 2d$. Note that a patch (shaded rectangle) of probability in m is wider and lower than the equivalent patch in d.

$$p(m) = p[d(m)]\left|\frac{\mathrm{d}d}{\mathrm{d}m}\right| \tag{2.13}$$

together with the understanding that the integration is always performed in the direction of positive m. Note that in the case above, with $p(d) = 1$ and $m(d) = 2d$, we find $\mathrm{d}d/\mathrm{d}m = ½$ and (as expected) $p(m) = 1 \times ½ = ½$.

In general, probability density functions change shape when transformed from d to m. Consider, for example, the uniform probability density function $p(d) = 1$ on the interval (0, 1) together with the function $m(d) = d^2$ (Figure 2.10). We find $d = m^{½}$, $\mathrm{d}d/\mathrm{d}m = ½m^{-½}$, and $p(m) = ½m^{-½}$, with m defined on the interval (0, 1). Thus, while $p(d)$ is uniform, $p(m)$ has a peak (actually an integrable singularity) at $m = 0$ (Figure 2.10B).

The general case of transforming $p(\mathbf{d})$ to $p(\mathbf{m})$, given the functional relationship $\mathbf{d}(\mathbf{m})$, is more complicated but is derived using the rule for transforming multidimensional integrals that is analogous to Equation (2.13). This rule states that the volume element transforms as $\mathrm{d}^N d = J(\mathbf{m})\mathrm{d}^N m$ where $J(\mathbf{m}) = |\det(\partial\mathbf{d}/\partial\mathbf{m})|$ is the *Jacobian determinant*, that is, the absolute value of the determinant of the matrix whose elements are $[\partial\mathbf{d}/\partial\mathbf{m}]_{ij} = \partial d_i/\partial m_j$:

$$\begin{aligned} 1 &= \int p(\mathbf{d})\mathrm{d}^N d = \int p[\mathbf{d}(\mathbf{m})]\left|\det\left[\frac{\partial\mathbf{d}}{\partial\mathbf{m}}\right]\right|\mathrm{d}^N m = \int p[\mathbf{d}(\mathbf{m})]J(\mathbf{m})\mathrm{d}^N m \\ &= \int p(\mathbf{m})\mathrm{d}^N m \end{aligned} \tag{2.14}$$

Hence, by inspection, we find that the probability density function transforms as

$$p(\mathbf{m}) = p[\mathbf{d}(\mathbf{m})]\left|\det\left[\frac{\partial\mathbf{d}}{\partial\mathbf{m}}\right]\right| = p[\mathbf{d}(\mathbf{m})]J(\mathbf{m}) \tag{2.15}$$

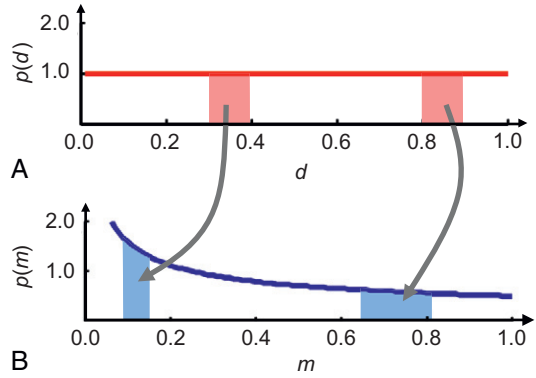

FIGURE 2.10 (A) The uniform probability density function $p(d) = 1$ on the interval $0 < d < 1$. (B) The transformed probability density function $p(m)$, given the relationship $m = d^2$. Note areas of equal probability, which are of equal height and width in the variable d are transformed into areas of unequal height and width in the variable, m. *MatLab* script gda02_09.

Note that for the linear transformation $\mathbf{m} = \mathbf{Md}$, the Jacobian is constant, with the value $J = |\det(\mathbf{M}^{-1})| = |\det(\mathbf{M})|^{-1}$. As an example, consider a two-dimensional probability density function that is uniform on the intervals (0, 1) for d_1 and (0, 1) for d_2, together with the transformation $m_1 = d_1 + d_2$, $m_2 = d_1 - d_2$. As is shown in Figure 2.11, $p(\mathbf{d})$ corresponds to a square of unit area in the (d_1, d_2) plane and $p(\mathbf{m})$ corresponds to a square of area 2 in the (m_1, m_2) plane. In order that the total area be unity in both cases, we must have $p(\mathbf{d}) = 1$ and $p(\mathbf{m}) = ½$. The transformation matrix is

$$M = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad \text{so} \quad |\det(M)| = 2 \quad \text{and} \quad J = \frac{1}{2} \tag{2.16}$$

Thus, by Equation (2.15), we find that $\mathrm{p}(\mathbf{m}) = p(\mathbf{d})J = 1 \times ½ = ½$, which agrees with our expectations.

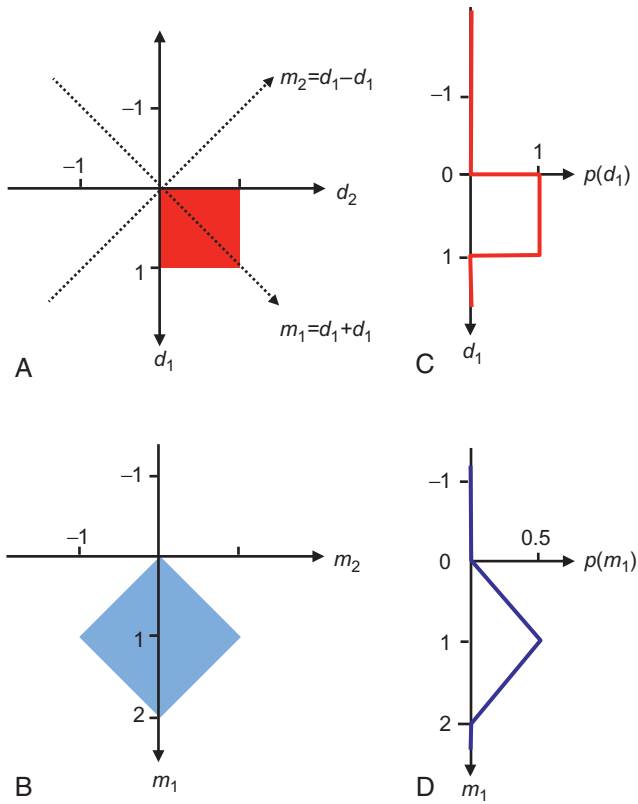

FIGURE 2.11 (A) The uniform probability density function $p(d_1,d_2) = 1$ on the interval $0 < d_1 < 1$, $0 < d_2 < 1$. Also shown are (m_1,m_2) axes, where $m_1 = d_1 + d_2$ and $m_2 = d_1 - d_2$. (B) The transformed probability density function $p(m_1,m_2) = 0.25$. (C) The univariate distribution $p(d_1)$ is formed by integrating $p(d_1,d_2)$ over d_2. It is a uniform distribution of amplitude 1. (D) The univariate distribution $p(m_1)$ is formed by integrating $p(m_1,m_2)$ over m_2. It is a triangular distribution of peak amplitude 0.5.

Note that we can convert $p(\mathbf{d})$ to a univariate distribution $p(d_1)$ by integrating over d_2. Since the sides of the square are parallel to the coordinate axes, the integration yields the uniform probability density function, $p(d_1) = 1$ (Figure 2.11C). Similarly, we can convert $p(\mathbf{m})$ to a univariate distribution $p(m_1)$ by integrating over m_2. However, because the sides of the square are oblique to the coordinate axes, $p(m_1)$ is a triangular—not a uniform—probability density function (Figure 2.11D).

Transforming a probability density function $p(\mathbf{d})$ is straightforward, but tedious. Fortunately, in the case of the linear function $\mathbf{m} = \mathbf{M}\mathbf{d} + \mathbf{v}$, where $\mathbf{M}$ and $\mathbf{v}$ are an arbitrary matrix and vector, respectively, it is possible to make some statements about the properties of the results without explicitly calculating the transformed probability density function $p(\mathbf{m})$. In particular, the mean and covariance can be shown, respectively, to be

$$\langle \mathbf{m} \rangle = \mathbf{M}\langle \mathbf{d} \rangle + \mathbf{v} \tag{2.17a}$$

and

$$[\text{cov } \mathbf{m}] = \mathbf{M}[\text{cov } \mathbf{d}]\mathbf{M}^{\mathrm{T}} \tag{2.17b}$$

These rules are derived by transforming the definition of the mean and variance:

$$\begin{aligned} \langle m \rangle_i &= \int m_i p(\mathbf{m}) \mathrm{d}^N m = \int \sum_j M_{ij} d_j p[\mathbf{d}(\mathbf{m})] \left| \det \left[\frac{\partial \mathbf{d}}{\partial \mathbf{m}} \right] \right| \left| \det \left[\frac{\partial \mathbf{m}}{\partial \mathbf{d}} \right] \right| \mathrm{d}^N d \\ &= \sum_j M_{ij} \int d_j p(\boldsymbol{d}) \mathrm{d}^N d = \sum_j M_{ij} \langle d_j \rangle \end{aligned} \tag{2.18}$$

$$\begin{aligned} [\text{cov } \mathbf{m}]_{ij} &= \\ &= \int (m_i - \langle m \rangle_i)(m_j - \langle m_j \rangle) p(\mathbf{m}) \mathrm{d}^N m \\ &= \int \sum_p (M_{ip} d_p - M_{ip} \langle d_p \rangle) \sum_q (M_{jq} d_q - M_{jq} \langle d_q \rangle) \\ &\quad p[\boldsymbol{d}(\mathbf{m})] \left| \det \left[\frac{\partial \mathbf{d}}{\partial \mathbf{m}} \right] \right| \left| \det \left[\frac{\partial \mathbf{m}}{\partial \mathbf{d}} \right] \right| \mathrm{d}^N d \\ &= \sum_p \sum_q M_{ip} M_{jq} \int (d_p - \langle d_p \rangle)(d_q - \langle d_q \rangle) p(\boldsymbol{d}) \mathrm{d}^N d \\ &= \sum_p \sum_q M_{ip} [\text{cov } \mathbf{d}]_{pq} M_{jq} \end{aligned} \tag{2.19}$$

Equation (2.17b) is very important, because the covariance of the data is a measure of the amount of measurement error. The equation functions as a rule for *error propagation*; that is, given [cov **d**] representing measurement error, it

provides a way to compute [cov **m**] representing the corresponding error in the model parameters. While the rule requires that the data and the model parameters be linearly related, it is independent of the functional form of the probability density function $p(\mathbf{d})$. Furthermore, it can be shown to be correct even when the matrix **M** is not square.

As an example, consider a model parameter m_1, which is linearly related to the data by

$$m_1 = \frac{1}{N}\sum_{i=1}^{N} d_i = \frac{1}{N}[1, 1, 1, \ldots, 1]\mathbf{d} \tag{2.20}$$

Note that this formula is the sample mean, as defined in Equation (2.5). This formula implies that matrix $\mathbf{M} = [1, 1, 1, \ldots, 1]/N$ and vector $\mathbf{v} = 0$. Suppose that the data are uncorrelated and all have the same mean $\langle d \rangle$ and variance σ_d^2. Then we see that $\langle m_1 \rangle = \mathbf{M}\langle \mathbf{d} \rangle + \mathbf{v} = \langle d \rangle$ and $\text{var}(m_1) = \mathbf{M}\ [\text{cov}\ \mathbf{d}]\ \mathbf{M}^T = \sigma_d^2/N$. The model parameter m_1 has a probability density function $p(m_1)$ with the same mean as **d**; that is, $\langle m_1 \rangle = \langle d \rangle$. Hence it is an estimate of the mean of the data. Its variance $\sigma_m^2 = \sigma_d^2/N$ is less than the variance of **d**. The square root of the variance, which is a measure of the width of the $p(m_1)$, is proportional to $N^{-1/2}$. Thus, accuracy of determining the mean of a group of data increases as the number of observations increases, albeit slowly (because of the square root).

In the case of uncorrelated data with uniform variance (that is, $[\text{cov}\ \mathbf{d}] = \sigma_d^2\mathbf{I}$), the covariance of the model parameters is $[\text{cov}\ \mathbf{m}] = \mathbf{M}[\text{cov}\ \mathbf{d}]\mathbf{M}^T = \sigma_d^2\ \mathbf{M}\mathbf{M}^T$. In general, $\mathbf{M}\mathbf{M}^T$, while symmetric, is not diagonal. Not only do the model parameters have unequal variance, but they are also correlated. Strongly correlated model parameters are usually undesirable, but (as we will discuss later) good experimental design can sometimes eliminate them.

2.4 GAUSSIAN PROBABILITY DENSITY FUNCTIONS

The probability density function for a particular random variable can be arbitrarily complicated, but in many instances, data possess the rather simple *Gaussian* (or *Normal*) probability density function

$$p(d) = \frac{1}{(2\pi)^{1/2}\sigma}\exp\left[-\frac{(d - \langle d \rangle)^2}{2\sigma^2}\right] \tag{2.21}$$

This probability density function has mean $\langle d \rangle$ and variance σ^2 (Figure 2.12). The Gaussian probability density function is so common because it is the limiting probability density function for the sum of random variables. The *central limit theorem* shows (with certain limitations) that regardless of the probability density function of a set of independent random variables, the probability density function of their sum tends to a Gaussian distribution as the number of summed variables increases. As long as the noise in the data comes from several sources of comparable size, it will tend to follow a Gaussian probability density function. This behavior is exemplified by the sum of the two uniform

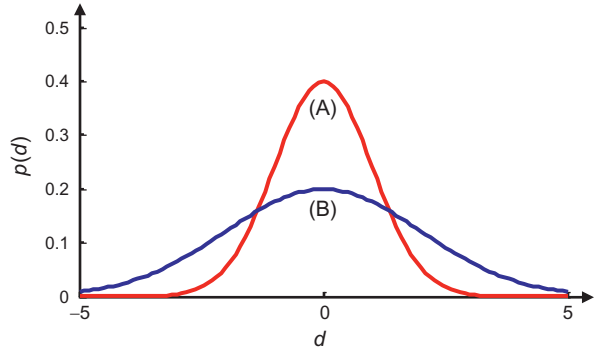

FIGURE 2.12 Gaussian (Normal) distribution with zero mean and $\sigma = 1$ for curve (A) and $\sigma = 2$ for curve (B). *MatLab* script gda02_10.

probability density functions in Section 2.3. The probability density function of their sum is more nearly Gaussian than the individual probability density functions (it being triangular instead of rectangular).

The joint probability density function for two independent Gaussian variables is just the product of two univariate probability density functions. When the data are correlated (say, with mean $\langle \mathbf{d} \rangle$ and covariance [cov **d**]), the joint probability density function is more complicated, since it must express the degree of correlation. The appropriate generalization can be shown to be

$$p(\mathbf{d}) = \frac{1}{(2\pi)^{N/2}(\det[\text{cov } \mathbf{d}])^{1/2}} \exp\left(-\frac{1}{2}[\mathbf{d} - \langle \mathbf{d} \rangle]^{\mathrm{T}}[\text{cov } \mathbf{d}]^{-1}[\mathbf{d} - \langle \mathbf{d} \rangle]\right) \tag{2.22}$$

Note that this probability density function reduces to Equation (2.21) in the special case of $N = 1$ (where [cov **d**] becomes σ_d^2). It is perhaps not apparent that the general case has an area of unity, a mean of $\langle \mathbf{d} \rangle$ and a covariance matrix of [cov **d**]. However, these properties can be derived by inserting Equation (2.22) into the relevant integral and by transforming to the new variable $\mathbf{y} = [\text{cov } \mathbf{d}]^{-½}$ $[\mathbf{d} - \langle \mathbf{d} \rangle]$ (whence the integral becomes substantially simplified).

When $p(\mathbf{d})$ (Equation 2.22) is transformed using the linear rule $\mathbf{m} = \mathbf{M}\mathbf{d}$, the resulting $p(\mathbf{m})$ is also Gaussian in form with mean $\langle \mathbf{m} \rangle = \mathbf{M}\langle \mathbf{d} \rangle$ and covariance matrix $[\text{cov } \mathbf{m}] = \mathbf{M}[\text{cov } \mathbf{d}]\mathbf{M}^{\mathrm{T}}$. Thus, all linear functions of Gaussian random variables are themselves Gaussian.

In Chapter 5, we will show that the information contained in each of two probability density functions can be combined by multiplying the two distributions. Interestingly, the product of two Gaussian probability density functions is itself Gaussian (Figure 2.13). Given Gaussian $p_{\mathrm{A}}(\mathbf{d})$ with mean $\langle \mathbf{d}_{\mathrm{A}} \rangle$ and covariance $[\text{cov } \mathbf{d}]_{\mathrm{A}}$ and Gaussian $p_{\mathrm{B}}(\mathbf{d})$ with mean $\langle \mathbf{d}_{\mathrm{B}} \rangle$ and covariance $[\text{cov } \mathbf{d}]_{\mathrm{B}}$, the product $p_{\mathrm{C}}(\mathbf{d}) = p_{\mathrm{A}}(\mathbf{d})\, p_{\mathrm{B}}(\mathbf{d})$ is Gaussian with mean and variance (e.g., Menke and Menke, 2011, their Section 5.4)

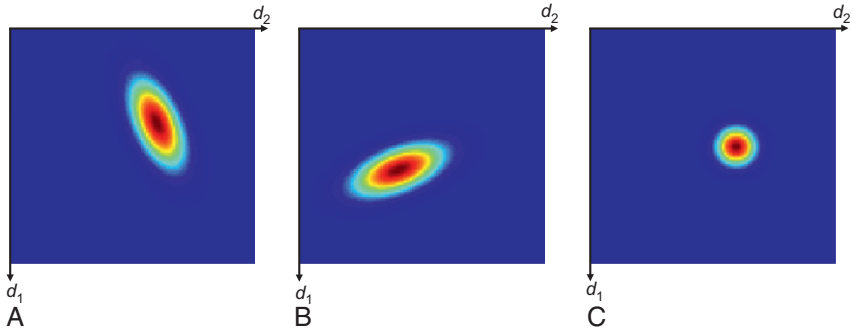

FIGURE 2.13 (A) A Normal probability density function $p_A(d_1,d_2)$. (B) Another Normal probability density function $p_B(d_1,d_2)$. (C) The product of these two functions $p_C(d_1,d_2) = p_A(d_1,d_2)p_B(d_1,d_2)$ is Normal. *MatLab* script gda02_11.

$$\langle \mathbf{d}_C \rangle = \left([\text{cov } \mathbf{d}]_A^{-1} + [\text{cov } \mathbf{d}]_B^{-1} \right)^{-1} \left([\text{cov } \mathbf{d}]_A^{-1} \langle \mathbf{d}_A \rangle + [\text{cov } \mathbf{d}]_B^{-1} \langle \mathbf{d}_B \rangle \right)$$
$$[\text{cov } \mathbf{d}]_C^{-1} = [\text{cov } \mathbf{d}]_A^{-1} + [\text{cov } \mathbf{d}]_B^{-1} \tag{2.23}$$

The idea that the model and data are related by an explicit relationship $\mathbf{g}(\mathbf{m}) = \mathbf{d}$ can be reinterpreted in light of this probabilistic description of the data. We can no longer assert that this relationship can hold for the data themselves, since they are random variables. Instead, we assert that this relationship holds for the mean data: $\mathbf{g}(\mathbf{m}) = \langle \mathbf{d} \rangle$. The distribution for the data can then be written as

$$p(\mathbf{d}) = \frac{1}{(2\pi)^{N/2} (\det[\text{cov } \mathbf{d}])^{1/2}} \exp\left(-\frac{1}{2} [\mathbf{d} - \mathbf{g}(\mathbf{m})]^T [\text{cov } \mathbf{d}]^{-1} [\mathbf{d} - \mathbf{g}(\mathbf{m})] \right) \tag{2.24}$$

The model parameters now have the interpretation of a set of unknown quantities that define the shape of the distribution for the data. One approach to inverse theory (which will be pursued in Chapter 5) is to use the data to determine the distribution and thus the values of the model parameters.

For the Gaussian distribution (Equation 2.24) to be sensible, $\mathbf{g}(\mathbf{m})$ must not be a function of any random variables. This is why we differentiated between data and auxiliary variables in Chapter 1; the latter must be known exactly. If the auxiliary variables are themselves uncertain, then they must be treated as data and the inverse problem becomes an implicit one with a much more complicated distribution than the above problem exhibits.

As an example of constructing the distribution for a set of data, consider an experiment in which the temperature d_i in some small volume of space is measured N times. If the temperature is assumed not to be a function of time and space, the experiment can be viewed as the measurement of N realizations of the same random variable or as the measurement of one realization of N distinct

random variables that all have the same distribution. We adopt the second viewpoint.

If the data are independent Gaussian random variables with mean $\langle \mathbf{d} \rangle$ and variance σ_d^2 so that $[\text{cov } \mathbf{d}] = \sigma_d^2 \mathbf{I}$, then we can represent the assumption that all the data have the same mean by an equation of the form $\mathbf{Gm} = \mathbf{d}$:

$$\begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} [m_1] = \begin{bmatrix} d_1 \\ d_2 \\ \cdots \\ d_N \end{bmatrix} \tag{2.25}$$

where m_1 is a single model parameter. We can then compute explicit formulas for the expressions in $p(\mathbf{d})$ as

$$\begin{aligned} \left(\det[\text{cov } \mathbf{d}]^{-1}\right)^{1/2} &= \left(\sigma_\text{d}^{-2N}\right)^{1/2} = \sigma_\text{d}^{-N} \\ [\mathbf{d} - \mathbf{Gm}]^\text{T} [\text{cov } \mathbf{d}]^{-1} [\mathbf{d} - \mathbf{Gm}] &= \sigma_\text{d}^{-2} \sum_{i=1}^{N} (d_i - m_1)^2 \end{aligned} \tag{2.26}$$

The joint distribution is therefore

$$p(\mathbf{d}) = \frac{\sigma_\text{d}^{-N}}{(2\pi)^{N/2}} \exp\left[-\frac{1}{2} \sigma_\text{d}^{-2} \sum_{i=1}^{N} (d_i - m_1)^2 \right] \tag{2.27}$$

2.5 TESTING THE ASSUMPTION OF GAUSSIAN STATISTICS

In the following chapters, we shall derive methods of solving inverse problems that are applicable whenever the data exhibit Gaussian statistics. In many instances, the assumption that the data follow this distribution is a reasonable one; nevertheless, having some means to test it is important.

First, consider a set of N Gaussian random variables x_i each with zero mean and unit variance. Suppose we construct a new random variable

$$\chi_K^2 = \sum_{i=1}^{K} x_i^2 \tag{2.28}$$

by summing squares of x_i. The function relating the x_i to $\chi^2{}_K$ is nonlinear, so $\chi^2{}_K$ does not have a Gaussian probability density function, but rather a different one (which we will not derive here) with the functional form

$$p(\chi_K^2) = \frac{1}{2^{K/2}\left(\frac{K}{2} - 1\right)!} \left[\chi_K^2\right]^{(K/2)-1} \exp\left(-\frac{1}{2}\chi_K^2\right) \tag{2.29}$$

It is called the *chi-squared probability density function*. It can be shown to be unimodal with mean K and variance $2K$. We shall make use of it in the discussion to follow.

We begin by supposing that we have some method of solving the inverse problem for the estimated model parameters. Assuming further that the model is explicit, we can compute the variation of the data about its estimated mean—a quantity we refer to as the error $\mathbf{e} = \mathbf{d} - \mathbf{g}(\mathbf{m}^{\text{est}})$. Does this error follow an uncorrelated Gaussian distribution with uniform variance?

To test the hypothesis that it does, we first make a histogram of the N errors e_i, in which the histogram intervals have been chosen so that there are about the same number of errors e_i in each interval. This histogram is then normalized to unit area, and the area A_i^{est} of each of the, say, p intervals is noted. We then compare these areas (which are all in the range from zero to unity) with the areas A_i predicted by a Gaussian distribution with the same mean and variance as the e_i. The overall difference between these areas can be quantified by using

$$\left(\chi_K^2\right)^{\text{est}} = N \sum_{i=1}^{p} \frac{\left(A_i^{\text{est}} - A_i\right)^2}{A_i} \tag{2.30}$$

If the data followed a Gaussian distribution exactly, then $(\chi^2{}_K)^{\text{est}}$ should be close to zero (it will not *be* zero since there are always random fluctuations). We therefore need to inquire whether the $(\chi^2{}_K)^{\text{est}}$ measured for any particular data set is sufficiently far from zero that it is improbable that the data follow the Gaussian distribution. This is done by computing the theoretical distribution of $(\chi^2{}_K)^{\text{est}}$ and testing whether $(\chi^2{}_K)^{\text{est}}$ is probable. The usual rule for deciding that the data do not follow the assumed distribution is that values greater than or equal to $(\chi^2{}_K)^{\text{est}}$ occur less than 5% of the time (if many realizations of the entire experiment were performed).

The quantity $(\chi^2{}_K)^{\text{est}}$ can be shown to follow approximately a $\chi^2{}_K$ distribution with $K = p - 3$ degrees of freedom, regardless of the type of distribution involved. The reason that the degrees of freedom are $p - 3$ rather than p is that three constraints have been introduced into the problem: that the area of the histogram is unity and that the mean and variance of the Gaussian distribution match those of the data. This test is known as *Pearson's chi-squared test* (Figure 2.14). In *MatLab*, the probability P that $\chi^2{}_K$ is greater than or equal to $(\chi^2{}_K)^{\text{est}}$ is computed as

```
P = 1-chi2cdf( x2est, K );
```

(*MatLab* script gda02_12)

Here `chi2cdf()` is the cumulative chi-squared distribution, that is, the probability that $\chi^2{}_K$ is less than or equal to $(\chi^2{}_K)^{\text{est}}$.

2.6 CONDITIONAL PROBABILITY DENSITY FUNCTIONS

Consider a scenario in which we are measuring the diameter d_1 and weight d_2 of sand grains drawn randomly from a pile of sand. We can consider d_1 and d_2 random variables described by a joint probability density function $p(d_1, d_2)$. The variables d_1 and d_2 will be correlated, since large grains will also tend to be heavy. Now, suppose that $p(d_1, d_2)$ is known. Once we draw a sand grain from the pile and weigh it, we already know something about its diameter, since

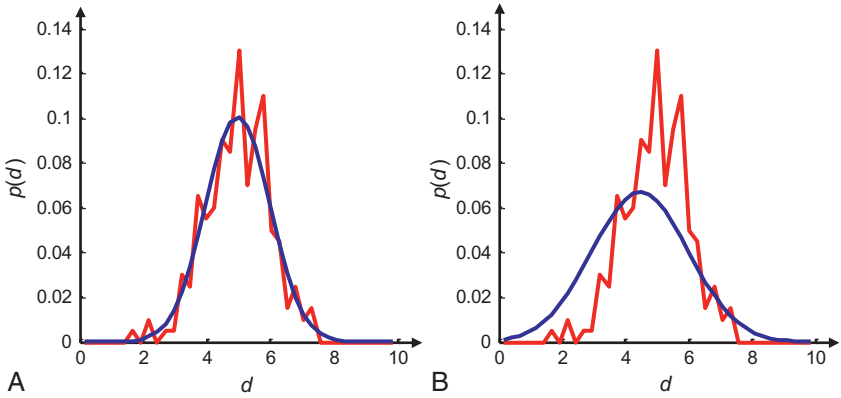

FIGURE 2.14 Example of Pierson's chi-squared test. The red curve is a probability density function (p.d.f.) estimated by binning 200 realizations of a random variable d drawn from a Gaussian population with a mean of 5 and a variance of 1^2. (A) Gaussian p.d.f. with the same mean and variance as the empirical one. (B) Gaussian p.d.f. with a mean of 4.5 and a variance of 1.5^2. According to the test, χ^2 values exceeding the observed value occur extremely frequently (75% of the time) for (A) but extremely infrequently (0.003%) for (B). *MatLab* script gda02.12.

diameter is correlated with weight. The quantity that embodies this information is called the *conditional* probability density function of d_1, given d_2, and is written $p(d_1|d_2)$.

The *conditional* probability density function of $p(d_1|d_2)$ is not the same as $p(d_1, d_2)$, although it is related to it. The key difference is that $p(d_1|d_2)$ is really only a probability density function in the variable d_1, with the variable d_2 just providing auxiliary information. Thus, the integral of $p(d_1|d_2)$ with respect to d_1 needs to be unity, regardless of the value of d_2. Thus, we must normalize $p(d_1, d_2)$ by dividing it by the total probability that d_1 occurs, given a specific value for d_2

$$p(d_1|d_2) = \frac{p(d_1, d_2)}{\int p(d_1, d_2)\mathrm{d}d_1} = \frac{p(d_1, d_2)}{p(d_2)} \tag{2.31}$$

Here we have used the fact that $p(d_2) = \int p(d_1, d_2)\mathrm{d}d_1$. The same logic allows us to calculate the conditional probability density function for d_2, given d_1

$$p(d_2|d_1) = \frac{p(d_1, d_2)}{\int p(d_1, d_2)\mathrm{d}d_2} = \frac{p(d_1, d_2)}{p(d_1)} \tag{2.32}$$

See Figure 2.15 for an example. Combining these two equations yields

$$p(d_1, d_2) = p(d_1|d_2)p(d_2) = p(d_2|d_1)p(d_1) \tag{2.33}$$

This result shows that the two conditional probability density functions are related but that they are *not* equal: $p(d_1|d_2) \neq p(d_2|d_1)$. The two equations can be further rearranged into a result called *Bayes Theorem*

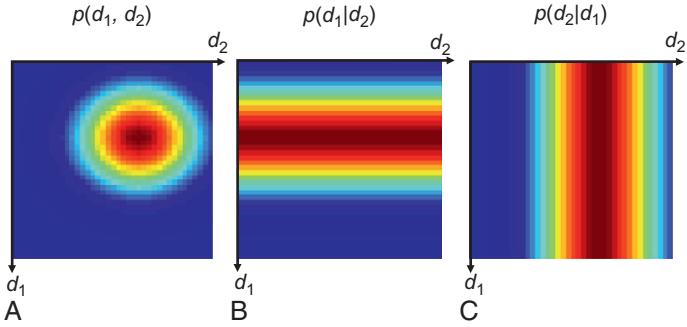

FIGURE 2.15 Example of conditional probability density functions. (A) A Gaussian joint probability density function $p(d_1, d_2)$. (B) The corresponding conditional probability density function $p(d_1|d_2)$. (C) The corresponding conditional probability density function $p(d_2|d_1)$. *MatLab* script gda02_13.

$$p(d_1|d_2) = \frac{p(d_2|d_1)p(d_1)}{p(d_2)} = \frac{p(d_2|d_1)p(d_1)}{\int p(d_1,d_2)\mathrm{d}d_1} = \frac{p(d_2|d_1)p(d_1)}{\int p(d_2|d_1)p(d_1)\mathrm{d}d_1}$$
$$p(d_2|d_1) = \frac{p(d_1|d_2)p(d_2)}{p(d_1)} = \frac{p(d_1|d_2)p(d_2)}{\int p(d_1,d_2)\mathrm{d}d_2} = \frac{p(d_1|d_2)p(d_2)}{\int p(d_1|d_2)p(d_2)\mathrm{d}d_2} \tag{2.34}$$

Note that only the denominators of the three fractions in each equation are different. They correspond to three different but equivalent ways of writing $p(d_1)$ and $p(d_2)$.

As an example, consider the case where diameter can take on only two discrete values, small (S) and big (B), and when weight can take on only two values, light (L) and heavy (H). A hypothetical joint probability function is

$$P(d_1,d_2) = \begin{bmatrix} d_1|d_2 & L & H \\ S & 0.8000 & 0.0010 \\ B & 0.1000 & 0.0990 \end{bmatrix} \tag{2.35}$$

In this scenario, about 90% of the small sand grains are light, about 99% of the large grains are heavy, and small/light grains are much more common than big/heavy ones. Univariate distributions are computed by summing over rows or columns, and in Equation (2.7):

$$P(d_1) = \begin{bmatrix} d_1 & \\ S & 0.8010 \\ B & 0.1990 \end{bmatrix} \quad \text{and} \quad P(d_2) = \begin{bmatrix} d_2 & L & H \\ & 0.9000 & 0.1000 \end{bmatrix} \tag{2.36}$$

According to Equation (2.34), the conditional distributions are

$$P(d_1|d_2) = \begin{bmatrix} d_1|d_2 & L & H \\ S & 0.8888 & 0.0100 \\ B & 0.1111 & 0.9900 \end{bmatrix} \text{ and } P(d_2|d_1) = \begin{bmatrix} d_1|d_2 & L & H \\ S & 0.9986 & 0.0012 \\ B & 0.5025 & 0.4974 \end{bmatrix} \tag{2.37}$$

Now suppose that we pick one sand grain from the pile, measure its diameter, and determine that it is big. What is the probability that it is heavy? We may be tempted to think that the probability is very high, since weight is highly correlated to size. But this reasoning is incorrect because heavy grains are about equally divided between the big and small size categories. The correct probability is given by $P(H|B)$, which is 49.74%.

Bayes theorem offers some insight into what is happening. Equation (2.34), adapted for discrete values by interpreting the integral as a sum, becomes

$$P(H|B) = \frac{P(B|H)P(H)}{P(B|L)P(L) + P(B|H)P(H)} = \frac{0.9900 \times 0.1000}{0.1111 \times 0.9000 + 0.9900 \times 0.1000}$$
$$= \frac{0.0990}{0.1000 + 0.0990} = \frac{0.0990}{0.1990} = 0.4974 \tag{2.38}$$

The numerator of Equation (2.38) represents the big, heavy grains and the denominator represents *all* the ways that one can get big grains, that is, the sum of big, heavy grains and big, light grains. In the scenario, light grains are extremely common, and although only a small fraction of them are heavy, their number affects the probability very significantly.

The above analysis, called *Bayesian Inference*, allows us to assess the importance of any given measurement. Before having measured the size of the sand grain, our best estimate of whether it is heavy is 10%, because heavy grains make up 10% of the total population (that is, $P(H) = 0.10$). After the measurement, the probability rises to 49.74%, which is about a factor of five more certain. As we will see in Chapter 5, Baysian Inference plays an important role in the solution of inverse problems.

2.7 CONFIDENCE INTERVALS

The confidence of a particular observation is the probability that one realization of the random variable falls within a specified distance of the true mean. Confidence is therefore related to the distribution of area in $p(d)$. If most of the area is concentrated near the mean, then the interval for, say, 95% confidence will be very small; otherwise, the confidence interval will be large. The width of the confidence interval is related to the variance. Distributions with large variances will also tend to have large confidence intervals. Nevertheless, the relationship is not direct, since variance is a measure of width, not area. The relationship is easy to quantify for the simplest univariate distributions. For instance, Gaussian probability density functions have 68% confidence intervals 1σ wide and 95% confidence intervals 2σ wide. Other types of simple distributions have similar relationships. If one knows that a particular Gaussian random variable has $\sigma = 1$, then if a realization of that variable has the value 50, one can state that there is a 95% chance that the mean of

the random variable lies between 48 and 52. One might symbolize this by $\langle d \rangle = 50 \pm 2$ (95%).

The concept of confidence intervals is more difficult to work with when one is dealing with joint probability density functions of several correlated random variables. One must define some volume in the space of data and compute the probability that the true means of the data are within the volume. One must also specify the shape of that volume. The more complicated the distribution, the more difficult it is to choose an appropriate shape and calculate the probability within it.

Even in the case of the Gaussian multivariate probability density functions, statements about confidence levels need to be made carefully, as is illustrated by the following scenario. Suppose that the Gaussian probability density function $p(d_1, d_2)$ represents two measurements, say the length and diameter of a cylinder, and suppose that these measurements are uncorrelated with equal variance, σ_d^2. As we might expect, the univariate probability density function $p(d_1) = \int p(d_1, d_2) \mathrm{d}d_2$ has variance, σ_d^2, and so the probability, P_1, that d_1 falls between $d_1 - \sigma_\mathrm{d}$ and $d_1 + \sigma_\mathrm{d}$, is 0.68 or 68%. Similarly, the probability, P_2, that d_2 falls between $d_2 - \sigma_\mathrm{d}$ and $d_2 + \sigma_\mathrm{d}$, is also 68%. But P_1 represents the probability of d_1, irrespective of the value of d_2, and P_2 represents the probability of d_2, irrespective of the value of d_1. The probability, P, that *both* d_1 and d_2 simultaneously fall within their respective one-sigma confidence intervals is $P = P_1 P_2 = (0.68)^2 = 0.46$ or 46%, which is significantly smaller than 68%.

One occasionally encounters a journal article containing a table of many (say 100) estimated parameters, each one with a stated 2σ error bound. The probability that *all one hundred* measurements fall within their respective bounds is $(0.95)^{100}$ or 0.6%—which is pretty close to zero!

2.8 COMPUTING REALIZATIONS OF RANDOM VARIABLES

The ability to create a vector of realizations of a random variable is very important. For instance, it can be used to simulate noise when testing a data analysis method on synthetic data (that is, artificially prepared data with well-controlled properties). And it can be used to generate a suite of possible models, to test against data.

MatLab provides a function `random()` that can generate realizations drawn from many different probability density functions. For instance,

```
m = random('Normal',mbar,sigma,N,1);
```

(*MatLab* script gda02_14)

creates a vector `m` of `N` Gaussian-distributed (Normally distributed) data with mean `mbar` and variance `sigma^2`.

In cases where no predefined function is available, it is possible to transform an available distribution, say $p(d)$, into the desired distribution, say $q(m)$, using the transformation rule

$$p[d(m)]\frac{\mathrm{d}d}{\mathrm{d}m} = q(m) \tag{2.39}$$

Most software environments provide a predefined function for realizations of a uniform distribution on the interval (0,1). Then, since $p(d) = 1$, Equation (2.39) is a differential equation for $d(m)$

$$\frac{\mathrm{d}d}{\mathrm{d}m} = q(m) \quad \text{or} \quad d = \int q(m)\mathrm{d}m = Q(m) \tag{2.40}$$

Here, $Q(m)$ is the cumulative probability distribution corresponding to $q(m)$. The transformation is then $m = Q^{-1}(d)$; that is, one must invert the cumulative probability distribution to give the value of m for which the probability d occurs. Thus, the transformation requires that the inverse cumulative probability distribution be known.

MatLab provides a `norminv()` function that calculates the inverse cumulative probability distribution in the Gaussian case, as well as a `random('unif',...)` function that returns realizations of the uniform probability density function. Thus,

```
d = random('unif',0,1,N,1);
m = norminv(d,mbar,sigma);
```

(*MatLab* script gda02_14)

creates `N` realizations of a Gaussian probability density function (Figure 2.16). Such an approach offers no advantage in the Gaussian case, since the `random('Normal',...)` function is available. It is of practical use in cases not supported by *MatLab*, as long as an appropriate `Qinv()` function can be provided.

Another method of producing a vector of realizations of a random variable is the *Metropolis-Hastings algorithm*. It is a useful alternative to the transformation method described above, especially since it requires evaluating only the probability density function $p(d)$ and not its cumulative inverse. It is an iterative algorithm that builds the vector $\mathbf{d}$ element by element. The first element, d_1, is set to an arbitrary number, such as zero. Subsequent elements are generated in sequence, with an element d_i, generating a successor d_{i+1} according to this algorithm: First, randomly draw a *proposed* successor d' from a conditional probability density function $q(d'|d_i)$. The exact form of $q(d'|d_i)$ is arbitrary; however, it must be chosen to that d' is typically in the neighborhood of d_i. One possible choice is the Gaussian function

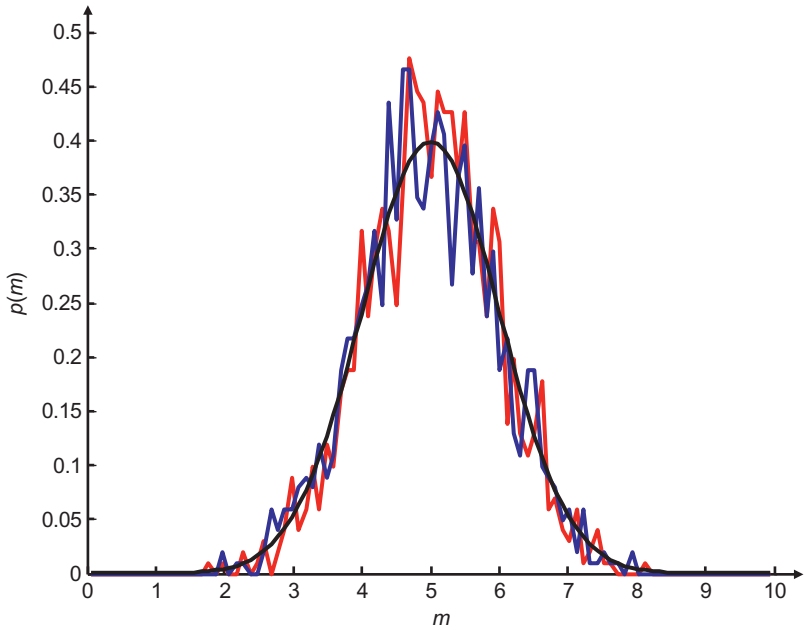

FIGURE 2.16 Gaussian probability density function $p(m)$ with mean 5 and variance 1^2. (Red curve) Computed by binning 1000 realizations of a random variable generated using *MatLab*'s random ("Normal",...) function. (Blue) Computed by binning 1000 realizations of a random variable generated by transforming a uniform distribution. (Black) Exact formula. *MatLab* script gda02_14.

$$q(d'|d_i) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left\{-\frac{(d'-d_i)^2}{2\sigma^2}\right\} \tag{2.41}$$

Here, σ represents the size of the neighborhood, that is, the typical deviation of d' away from d_i. Second, generate a random number α drawn from a uniform distribution on the interval (0,1). Third, accept the proposed successor and set $d_{i+1} = d'$ if

$$\alpha < \frac{p(d')q(d_i|d')}{p(d_i)q(d'|d_i)} \tag{2.42}$$

Otherwise, set $d_{i+1} = d_i$. When repeated many times, this algorithm leads to a vector $\mathbf{d}$ that has approximately the probability density function $p(d)$ (Figure 2.17). Note that the conditional probability density functions cancels from Equation (2.42) when $q(d'|d_i) = q(d_i|d')$, as is the case for the Gaussian in Equation (2.41).

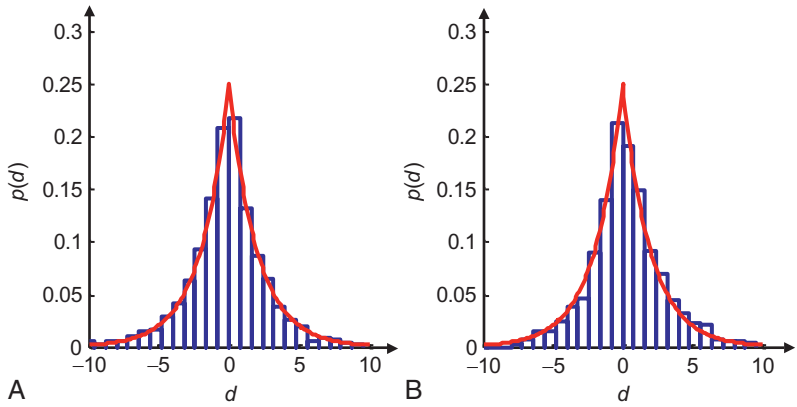

FIGURE 2.17 Histograms (blue curves) of 5000 realizations of a random variable d for the probability density function (red curves) $p(d) = ½c\exp(-/d//c)$ with $c = 2$. (A) Realizations computed by transforming data drawn from a uniform distribution and (B) realizations computed using the Metropolis-Hastings algorithm. *MatLab* script gda02_14.

2.9 PROBLEMS

2.1. What is the mean and variance of the uniform distribution $p(d) = 1$ on the interval (0,1)?

2.2. Suppose d is a Gaussian random variable with zero mean and unit variance. What is the probability density function of $E = e^2$? Hint: Since the sign of d gets lost when it is squared, you can assume that $p(d)$ is one-sided, that is, defined for only $d \geq 0$ and with twice the amplitude of the usual Gaussian.

2.3. Write a *MatLab* script that uses the `random()` function to create a vector $\mathbf{d}$ of $N = 1000$ realizations of a Gaussian-distributed random variable with mean $\langle d \rangle = 4$ and variance $\sigma_d^2 = 2^2$. Count up the number of instances where $d_i > (\langle \mathrm{d} \rangle + 2\sigma_\mathrm{d})$. Is this about the number you expected?

2.4. Suppose that the data are uncorrelated with uniform variance, $[\mathrm{cov}\ \mathbf{d}] = \sigma_d^2\mathbf{I}$, and that the model parameters are linear functions of the data, $\mathbf{m} = \mathbf{M}\mathbf{d}$. (A) What property must $\mathbf{M}$ have for the model parameters to be uncorrelated with uniform variance σ_m^2? (B) Express this property in terms of the rows of the $\mathbf{M}$.

2.5. Use the transformation method to compute realizations of the probability density function $p(m) = 3m^2$ on the interval (0,1), starting from realizations of the uniform distribution $p(d) = 1$. Check your results by plotting a histogram.

REFERENCES

Menke, W., Menke, J., 2011. Environmental Data Analysis with MatLab. Academic Press, Elsevier Inc, Oxford, UK 263pp.