1
Preliminary statistics

The solution of a geophysical inverse problem can be obtained by a combination of information from observed data, the theoretical relation between data and earth parameters (models), and prior information on data and models. Due to uncertainties in the data and model, probability theory can be used as a tool to describe the inverse problem. Excellent introductory books on the subject of probability theory are those of Feller (1968), Papoulis (1965), and Ross (1989). In this chapter we review probability theory and stochastic processes, concepts that are used later to describe the global optimization methods used in geophysical inverse problems. Readers familiar with the subject can proceed directly to Chapter 2.

1.1 Random variables

In simple language, a random variable is a variable used to represent the outcome of a random experiment. Familiar random experiments are the tossing of a die and the flipping of a coin. When a die is tossed, there are six possible outcomes, and it is not certain which one will occur. Similarly, when a coin is tossed, there are two possible outcomes, and it is not certain which one will occur. The outcome of a random experiment is usually represented by a point called a *sample point* $s$. The set that consists of all possible sample points is called the *sample space* $S$. Subsets of $S$ represent certain events such that an event $A$ consists of a certain collection of possible outcomes $s$. If two subsets contain no point $s$ in common, they are said to be *disjoint*, and the corresponding events are said to be *mutually exclusive*. Formally, any single-valued numerical function $X(s)$ defined on a sample space $S$ is called a *random variable*, and a unique real number is associated with each point $s$.

1.2 Random numbers

Most of the methods of geophysical inversion that we discuss in this book use probability or statistical theories that involve studying processes arising from...
Global optimization methods in geophysical inversion

random experiments. This means that we need to simulate random processes on a computer. In practice, this requires an algorithm to generate random numbers. Computer-generated random numbers have been used extensively in several applications. Most commonly, sampling methods using computer-generated random numbers have been used in situations where the mathematics became intractable. Metropolis and Ulam (1949) first proposed the application of random sampling to the solution of deterministic problems such as the evaluation of an integral of the type

$$I = \int_{x_{\text{min}}}^{x_{\text{max}}} f(x) \, dx.$$  \hspace{1cm} (1.1)

If this integral exists, it is given by the expected value of the function. For a random number $X$ that has a uniform distribution over the interval $(x_{\text{min}}, x_{\text{max}})$, the preceding integral can be replaced by the following sum:

$$I = \frac{1}{n} \sum_{i=1}^{n} f(X_i),$$  \hspace{1cm} (1.2)

where $X_i$ is a random sample from the uniform distribution. Techniques using this idea are referred to as Monte Carlo methods. In general, any method using a random walk is usually included in the category of Monte Carlo methods.

To apply any technique that involves random processes, computer-generated random numbers are required. Although a computer is a machine that produces output that is always predictable, computer-generated random numbers are used extensively and are called pseudorandom numbers. Most computers use a method called the congruential method to generate random samples from a uniform distribution (Kennedy and Gentle 1980; Rubinstein 1981). The machine produces a sequence of pseudorandom integers $I_1, I_2, I_3, \ldots$ between 0 and $N$ by the recurrence relation

$$I_{j+1} = \text{mod}(aI_j + c, N),$$  \hspace{1cm} (1.3)

where the mod operation is defined as

$$\text{mod}(a_1, a_2) = a_1 - \text{int}\left(\frac{a_1}{a_2}\right)a_2.$$  \hspace{1cm} (1.4)

In Eq. (1.3), $a$ and $c$ are positive integers called the multiplier and increment, respectively. The number thus generated will repeat itself with a period no greater than $N$. If $N$, $a$, and $c$ are properly chosen, then the period will be of maximum
Preliminary statistics

length; i.e., all possible integers between 0 and \( N \) will occur at some point. The highest possible value of \( N \) is controlled by the precision of the machine used. The real number that is returned is given by \( (I_j/N) \), so the number lies between 0 and 1. The random numbers thus generated are not free from sequential correlation in successive calls. Also, improper choice of the parameters \( N \), \( a \), and \( c \) causes further problems. Press et al. (1989) describe methods to improve the performance of random number generators. One of them is to do additional randomizing shuffles on the numbers generated by the machine. Given random samples from the standard uniform distribution, random numbers from some other distribution may be readily obtained by transformations (Rubinstein 1981).

1.3 Probability

The concept of probability is used to analyze the experiment represented by set \( S \). Let us assume that \( P(A) \) denotes the probability of an event (collection of outcomes) \( A \); then we make the following axioms:

\[
P\{A\} \geq 0 \quad \text{for any event } A; \tag{1.5}
\]

i.e., the probabilities of all the events are non-negative and

\[
P\{S\} = 1; \tag{1.6}
\]

i.e., the probability of the whole sample space is unity.

Thus, by convention or axiom, the probability that one of a sequence of mutually exclusive events \( \{A_i\} \) will occur is given by

\[
P\{A_1 \text{ or } A_2 \text{ or } \ldots \} = \sum_i P(A_i). \tag{1.7}
\]

There exist different interpretations of the concept of probability. They can be classified into the following categories: (1) the classical interpretation, (2) the relative-frequency interpretation, and (3) the Bayesian interpretation. The classical definition of probability (Papoulis 1965) is as follows: Let \( N \) be the total number of possible outcomes of an experiment. If in \( N_a \) of all these outcomes the event \( A \) occurs, then \( P(A) \) is given by

\[
P(A) = \frac{N_a}{N}, \tag{1.8a}
\]

provided that the occurrence of all the events is equally likely. The main criticism of this definition is that it is circular because \textit{equally likely} also means \textit{equally probable}. 

© in this web service Cambridge University Press
www.cambridge.org
Global optimization methods in geophysical inversion

According to the relative-frequency interpretation (von Mises 1957), the probability of an event $A$ is the following limit of the relative frequency:

$$P(A) = \lim_{N \to \infty} \frac{N_A}{N}, \quad \text{(1.8b)}$$

where $N_A$ is the number of occurrences of $A$ in $N$ number of trials. In the relative-frequency definition, the concept of equally likely events is completely avoided. Here a trial is used to mean repetition of an experiment under identical circumstances. The problems with the relative-frequency interpretations are as follows: The limit (Eq. 1.8b) can only be assumed to exist, the number of trials is always finite, and the definition gives no meaning to the probability of a hypothesis (Jeffreys 1939).

The Bayesian interpretation of probability is due to Bayes (1763), who defined the probability to be a degree of belief. Thus the probability theory can be viewed as an extension of deductive logic and is called inductive logic. In deductive logic, a proposition can either be true or false, but in inductive logic, the probability of a proposition constitutes a degree of belief, with proof or disproof as extremes. The Bayesian interpretation can again be classified into two categories: the logical interpretation and the subjective interpretation. In the logical interpretation, the probability is objective, an aspect of the state of affairs. In the subjective interpretation, the degree of belief is a personal degree of belief such that the axioms of probability theory are not violated.

The Bayesian interpretation appears to be conceptually very clear.

### 1.4 Probability distribution, distribution function, and density function

A discrete random variable $X(s)$ assumes a finite number of values. For each value $x_i$ there is a unique probability that the random variable assumes the value $x_i$:

$$P\{X(s) = x_i\} = p_i, \quad i = 0, 1, 2, \ldots \quad \text{(1.9)}$$

The sequence $\{p_i\}$ is called the probability distribution of $X(s)$, and the cumulative probability

$$P\{X(s) \leq x\} = \sum_{x_i \leq x} p_i = F(x), \quad -\infty < x < \infty, \quad \text{(1.10)}$$
Preliminary statistics

is called the distribution function of \( X(s) \). For example, consider that for a random variable assuming values \( x_i, i = 0, 1, 2, 3, 4, 5, \ldots \), the probability distribution is given as \( p_0 = 0, p_1 = 0.1, p_2 = 0.3, p_3 = 0.2, p_4 = 0.25, \) and \( p_5 = 0.15 \) and 0 elsewhere. The distribution function \( F(x) \) can be computed using Eq. (1.10). The probability distribution and corresponding distribution function are shown in Figure 1.1.

The following properties are satisfied by a distribution function: \( F(-\infty) = 0, F(\infty) = 1, \) and it is a non-decreasing function of \( x \): \( F(x_1) = F(x_2) \) for \( x_1 < x_2 \). If \( F(x) \) is differentiable, the probability density function (pdf) \( p(x) \) is given by

\[
p(x) = \frac{dF(x)}{dx}.
\]

If \( F(m) \) denotes the distribution function of a vector of variables \( m \), with \( m = [m_1, m_2, \ldots, m_n]^T \), then if \( F(m) \) is differentiable, the probability density function (pdf) \( p(m) \) of \( m \) is given by

\[
p(m) = \frac{\partial^n F(m)}{\partial m_1 \partial m_2 \cdots \partial m_n}.
\]

The probability of \( m \) being in a certain region or volume \( A \) is given by

\[
P(m \in A) = \int_A p(m) \partial m_1 \partial m_2 \cdots \partial m_n.
\]

Figure 1.1 An example of a probability distribution (open squares) and its corresponding distribution function (open circles). Color version available online at www.cambridge.org/sen_stoffa.
The strict definition of the pdf is that it is normalized; i.e.,
\[ \int p(m) \, dm = 1. \]  
(1.13)
An unnormalized probability density function is simply called the *density function*.

### 1.4.1 Examples of distribution and density functions

#### 1.4.1.1 Normal or Gaussian distribution

A continuous random variable \( x \) is said to be *normally distributed* if its density function \( p(x) \) is given as
\[ p(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{(x-\langle x \rangle)^2}{2\sigma^2} \right], \]  
(1.14)
where \( \langle x \rangle \) and \( \sigma^2 \) are called the *mean* and the *variance*, respectively. Note that
\[ \int dx \, p(x) = 1. \]  
(1.15)
The corresponding distribution function is given by
\[ F(x) = \int_{-\infty}^{x} dp(x') = \frac{1}{2} + \text{erf} \left( \frac{x-\langle x \rangle}{\sigma} \right), \]  
(1.16)
where \( \text{erf}(x) \) is called the *error function* (Abramowitz and Stegun 1972). The plots of the Gaussian probability density and distribution functions are shown in Figure 1.2. Note that the mean and the variance completely describe a normal distribution.

#### 1.4.1.2 Cauchy distribution

The *Cauchy density function* for a continuous random variable \( x \) is defined as
\[ p(x) = \frac{\alpha}{\pi} \frac{1}{\alpha^2 + x^2}, \]  
(1.17)
where \( \alpha \) is a constant that controls the width of the distribution (Figure 1.3). The peak value of the distribution is equal to \( (1/\pi\alpha) \); i.e., it is inversely proportional to the value of \( \alpha \). Thus, with decreasing values of \( \alpha \), the distribution becomes narrower, and the height of the peak increases.
Figure 1.2 A Gaussian probability density function for a mean = 0.0 and a standard deviation = 10.0. Its corresponding distribution function is also shown. Color version available online at www.cambridge.org/sen_stoffa.

Figure 1.3 A Cauchy probability density function for different values of \( \alpha \). Color version available online at www.cambridge.org/sen_stoffa.
Global optimization methods in geophysical inversion

1.4.1.3 Gibbs’ distribution

For a continuous random variable $x$, the Gibbs probability density function (also called the Boltzmann pdf) is defined as

$$p(x) = \frac{\exp \left[ \frac{E(x)}{T} \right]}{\int_{-\infty}^{\infty} \exp \left[ \frac{E(x')}{T} \right] dx'},$$

(1.18)

where $E(x)$ is a function of the random variable $x$, and $T$ is a control parameter. Note that the shape of the distribution is controlled by changing the parameter $T$. This parameter is called temperature in thermodynamics. This density function is not analytically integrable except in some very simple cases. This density function will be discussed in much greater detail in subsequent chapters.

1.5 Joint and marginal probability distributions

Let $X$ and $Y$ be two discrete random variables defined on the same sample space. Their joint (bivariate) probability distribution $\{\Gamma_{ij}\}$ is defined by

$$\Gamma_{ij} = P\{X = x_i \text{ and } Y = y_j\}, \quad i, j = 0, 1, \ldots$$

(1.19)

with

$$\sum_i \sum_j \Gamma_{ij} = 1.$$  

(1.20)

The sequence $\{p_i\}$ and $\{q_j\}$, called the marginal distributions of $X$ and $Y$, respectively, are calculated from the joint distribution by the following formulas:

$$p_i = \sum_j \Gamma_{ij} \quad \text{and} \quad q_j = \sum_i \Gamma_{ij},$$

(1.21)

When two random variables $X$ and $Y$ are independently distributed, or stochastically independent, we have

$$\Gamma_{ij} = p_i q_j,$$

(1.22)

for all $i$ and $j$.

Definitions for marginal distributions of continuous variables can be written down simply by replacing the sums in Eq. (1.21) with integrals. Also, equations similar to Eq. (1.21) can be easily written down for joint and marginal distribution and density functions.
1.6 Mathematical expectation, moments, variances, and covariances

The expectation, moments, variances, and covariances are the quantities usually required to describe a distribution. The mathematical expectation (or mean value, or expected value) of a discrete random variable $X$ is defined as

$$ e(X) = < x > = \sum_i x_i p_i $$  \hspace{1cm} (1.23)

and of a continuous random variable as

$$ e(X) = < x > = \int dx \, x \, p(x). $$  \hspace{1cm} (1.24)

The mean $e(X)$ exists if the sum or the integral converges absolutely. The mean or the mathematical expectation is also called the first moment. A moment of order $r$ is given by

$$ e\{X^r\}. $$

The central moment of order $r$ is given by

$$ e\{X - e(X)\}^r, \quad r = 1, 2, \ldots $$

That is, the second central moment (also called the variance of the distribution) is given by

$$ \sigma_x^2 = e\{(X - e(X))^2\} = \int_{-\infty}^{\infty} dx (x - <x>)^2 p(x). $$  \hspace{1cm} (1.25)

The square root of the variance is called the standard deviation.

Let us now consider two continuous random variables $X$ and $Y$ having a joint distribution such that the moments of the marginal distributions describe the properties of $X$ and $Y$, respectively. $X$ and $Y$ are now related using the covariance of $X$ and $Y$ as given by

$$ \sigma_{XY} = e(XY) - e(X)e(Y). $$

Now consider two random variables $X$ and $Y$ with standard deviations $\sigma_x$ and $\sigma_y$, respectively, and define the following:

$$ U = \frac{X - e(X)}{\sigma_x} \quad \text{and} \quad V = \frac{Y - e(Y)}{\sigma_y}; $$  \hspace{1cm} (1.27)

i.e., $e(U) = 0$ and $e(V) = 0$ and $\sigma_u^2 = e(U^2) = 1$ and $\sigma_v^2 = e(V^2) = 1$. $U$ and $V$ are called standardized random variables. The correlation coefficient between $X$ and $Y$ is defined as the covariance between $U$ and $V$; i.e.,

$$ \rho_{XY} = \sigma_{U,V} = e(UV) = E\left\{ \frac{X - e(X)}{\sigma_x} \left[ \frac{Y - e(Y)}{\sigma_y} \right]\right\}. $$
Using Eq. (1.26), we obtain
\[ \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \] (1.28)

The value of the correlation coefficient is not affected by changes of scale or origin because it is calculated from standardized random variables.

The covariance function is a measure of correlation between the values of the random variables; i.e., it determines if the values taken by one random variable depend on the values of another. Consider a vector
\[ \mathbf{x} = [x_1, x_2, x_3, \ldots, x_n]^T, \] (1.29)
of \( n \) random variables \( x_1, x_2, x_3, \ldots, x_n \) with means \( <x_1>, <x_2>, <x_3>, \ldots, <x_n> \), respectively. Assuming that these random variables are continuous, the mean values are given by the following integral:
\[ <\mathbf{x}> = \int \mathbf{x} p(\mathbf{x}), \] (1.30)
or, for each variable, as
\[ <x_i> = \int \mathcal{d}x_i \int \mathcal{d}x_2 \cdots \int \mathcal{d}x_n x_i p(\mathbf{x}). \] (1.31)

Similarly, the covariance matrix is given by
\[ \text{cov}(\mathbf{x}) = \int \mathbf{x}(\mathbf{x} - <\mathbf{x}>)(\mathbf{x} - <\mathbf{x}>)^T p(\mathbf{x}), \] (1.32)
or each component will be given by the following equation:
\[ \text{cov}_{ij} = \int \mathcal{d}x_1 \int \mathcal{d}x_2 \cdots \int \mathcal{d}x_n \left( x_i - <x_i> \right) \left( x_j - <x_j> \right) p(\mathbf{x}). \] (1.33)

The joint distribution of two independent random variables is simply the product of two univariate distributions. When the variables are correlated (characterized by the covariance matrix), the distribution is more complicated, and a general form of the normal distribution for a vector of random variable \( \mathbf{x} \) is defined as (Menke 1984)
\[ p(\mathbf{x}) = \frac{\text{det} \text{cov}(\mathbf{x})^{-1/2}}{(2\pi)^{n/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - <\mathbf{x}>)^T \left[ \text{cov}(\mathbf{x}) \right]^{-1} (\mathbf{x} - <\mathbf{x}>) \right\}. \] (1.34)