CHAPTER 1

REVIEW OF ELEMENTARY PROBABILITY THEORY

This book assumes some familiarity with elementary probability theory. Good introductory texts are [63, 192]. We will begin by briefly reviewing some of these concepts in order to introduce notation and for future reference. We will also introduce basic notions of statistics that are particularly useful. See [106] for additional discussion of related material in a computer systems context.

1.1 SAMPLE SPACE, EVENTS, AND PROBABILITIES

Consider a random experiment resulting in an *outcome* (or "sample") represented by ω. For example, the experiment could be a pair of dice thrown onto a table and the outcome could be the exact orientation of the dice and their position on the table when they stop moving. The abstract space of all outcomes (called the *sample space*) is normally denoted by Ω, i.e., $\omega \in \Omega$.

An *event* is merely a subset of Ω. For example, in a dice throwing experiment, an event is "both dice land in a specific region of the table" or "the sum of the dots on the upward facing surfaces of the dice is 7." Clearly, many different individual outcomes ω belong to these events. We say that an event A has *occurred* if the outcome ω of the random experiment belongs to A, i.e., $w \in A$, where $A \subset \Omega$. Now consider two events A and B. We therefore say that A *and* B occur if the outcome $\omega \in A \cap B$. Also, we say that A *or* B occur if the outcome $\omega \in A \cup B$.

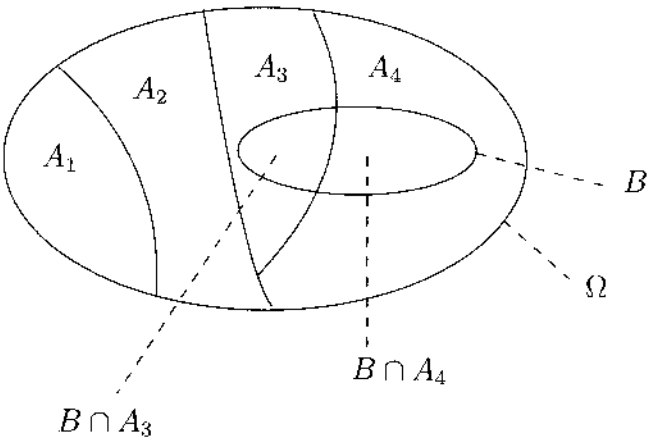

Figure 1.1 A partition of Ω.

A *probability measure* P maps each event $A \subset \Omega$ to a real number between zero and one inclusive, i.e., $\mathsf{P}(A) \in [0, 1]$. A probability measure has certain properties such as $\mathsf{P}(\Omega) = 1$ and

$$\mathsf{P}(A) \quad = \quad 1 - \mathsf{P}(\bar{A}),$$

where $\bar{A} = \{\omega \in \Omega \mid \omega \notin A\}$ is the complement of A. Also, if the events $\{A_i\}_{i=1}^n$ are disjoint (i.e., $A_i \cap A_j = \emptyset$ for all $i \neq j$), then

$$\mathsf{P}\left(\bigcup_{i=1}^{n} A_i\right) \quad = \quad \sum_{i=1}^{n} P(A_i),$$

i.e., P is finitely additive. Formally, a probability measure is defined to be countably additive. Also, $P(A)$ is defined only for events $A \subset \Omega$ that belong to a *σ-field* (sigma-field) or *σ-algebra* of events. These details are beyond the scope of this book.

The probability of an event A *conditioned on* (or "given that") another event B has occurred is

$$\mathsf{P}(A|B) \quad \equiv \quad \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)},$$

where $\mathsf{P}(B) > 0$ is assumed. Now suppose the events $A_1, A_2, ..., A_n$ form a *partition* of Ω, i.e.,

$$\bigcup_{i=1}^{n} A_i = \Omega \quad \text{and} \quad A_i \cap A_j = \emptyset \text{ for all } i \neq j.$$

Assuming that $\mathsf{P}(A_i) > 0$ for all i, the *law of total probability* states that, for any event B,

$$\mathsf{P}(B) \quad = \quad \sum_{i=1}^{n} \mathsf{P}(B|A_i)\mathsf{P}(A_i). \tag{1.1}$$

Note that the events $A_i \cap B$ form a partition of $B \subset \Omega$. See Figure 1.1, where $B = (B \cap A_4) \cup (B \cap A_3)$ and $B \cap A_i = \emptyset$ for $i = 1, 2$.

A group of events $A_1, A_2, ..., A_n$ are said to be *mutually independent* (or just "independent") if

$$\mathsf{P}\left(\bigcap_{i \in \mathcal{I}} A_i\right) = \prod_{i \in \mathcal{I}} \mathsf{P}(A_i)$$

for all subsets $\mathcal{I} \subset \{1, 2, ..., n\}$. Note that if events A and B are independent and $\mathsf{P}(B) > 0$, then $\mathsf{P}(A|B) = \mathsf{P}(A)$; therefore, knowledge that the event B has occurred has no bearing on the probability that the event A has occurred as well.

In the following, a comma between events will represent an intersection symbol, for example, the probability that A *and* B occur is

$$\mathsf{P}(A, B) \quad \equiv \quad \mathsf{P}(A \cap B).$$

1.2 RANDOM VARIABLES

A *random variable* X is a real-valued function with domain Ω. That is, for each outcome ω, $X(\omega)$ is a real number representing some feature of the outcome. For example, in a dice-throwing experiment, $X(\omega)$ could be defined as the sum of the dots on the upward-facing surfaces of outcome ω. Formally, random variables are defined to be *measurable* in the sense that the event $X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\}$ is an event (a member of the σ-field on Ω) for all events $B \subset \mathbb{R}$ (belonging to the *Borel* σ-field of subsets of $\mathbb{R}$). In this way, the quantity $\mathsf{P}(X \in B)$ is well defined for any set B that is of interest. Again, the details of this measurability condition are beyond the scope of this book. In the following, all functions are implicitly assumed to be measurable.

The strict range of X is defined to be the *smallest* subset R_X of $\mathbb{R}$ such that

$$\mathsf{P}(X \in R_X) \quad = \quad 1,$$

where $\mathsf{P}(X \in R_X)$ is short for $\mathsf{P}(\{\omega \in \Omega \mid X(\omega) \in R_X\})$.

Note that a (Borel-measurable) function g of a random variable X, $g(X)$, is also a random variable.

A group of random variables $X_1, X_2, ..., X_n$ are said to be mutually independent (or just "independent") if, for any collection $\{B_i\}_{i=1}^n$ of subsets of $\mathbb{R}$, the events $\{X_i \in B_i\}_{i=1}^n$ are independent; see Section 1.9.

1.3 CUMULATIVE DISTRIBUTION FUNCTIONS, EXPECTATION, AND MOMENT GENERATING FUNCTIONS

The probability *distribution* of a random variable X connotes the information $\mathsf{P}(X \in B)$ for all events $B \in \mathbb{R}$. We need only stipulate

$$\mathsf{P}(X \leq x) \quad \equiv \quad \mathsf{P}(X \in (-\infty, x])$$

for all $x \in \mathbb{R}$ to completely specify the distribution of X; see Equation (1.4). This leads us to define the *cumulative distribution function* (CDF) F_X of a random variable X as

$$F_X(x) = \mathsf{P}(X \leq x) \tag{1.2}$$

for $x \in \mathbb{R}$, where $\mathsf{P}(X \leq x)$ is, again, short for $\mathsf{P}(\{\omega \in \Omega \mid X(\omega) \leq x\})$. Clearly, a CDF F_X takes values in $[0, 1]$, is nondecreasing on $\mathbb{R}$, $F_X(x) \to 1$ as $x \to \infty$, and $F_X(x) \to 0$ as $x \to -\infty$.

The *expectation* of a random variable is simply its average (or "mean") value. We can define the expectation of a function g of a random variable X as

$$\mathsf{E}(g(X)) \quad = \quad \int_{-\infty}^{\infty} g(x)\, \mathrm{d}F_X(x), \tag{1.3}$$

where we have used a Stieltjes integral [133] that will be explained via explicit examples in the following. Note here that the expectation (when it exists) is simply a real number or $\pm\infty$. Also note that expectation is a linear operation over random variables. That is, for any two random variables X and Y and any two real constants a and b,

$$\mathsf{E}(aX + bY) \quad = \quad a\mathsf{E}X + b\mathsf{E}Y.$$

■ **EXAMPLE 1.1**

Suppose g is an *indicator function*, i.e., for some event $B \subset \mathbb{R}$

$$\begin{aligned} g(X(\omega)) \quad &\equiv \quad \mathbf{1}\{X(\omega) \in B\} \\ &\equiv \quad \begin{cases} 1 & \text{if } X(\omega) \in B, \\ 0 & \text{else.} \end{cases} \end{aligned}$$

In this case,

$$\mathsf{E}g(X) \quad = \quad \mathsf{P}(X \in B) \quad = \quad \int_B \mathrm{d}F_X(x), \tag{1.4}$$

where the notation refers to integration over the set B.

The nth *moment* of X is $\mathsf{E}(X^n)$ and the *variance* of X is

$$\sigma_X^2 \quad \equiv \quad \mathrm{var}(X) \equiv \mathsf{E}(X - \mathsf{E}X)^2,$$

i.e., the variance is the second *centered* moment. The *standard deviation* of X is the square root of the variance, $\sigma_X \geq 0$. The *moment generating function* (MGF) of X is

$$m_X(\theta) \quad = \quad \mathsf{E}e^{\theta X},$$

where θ is a real number. The moment generating function can also be used to completely describe the distribution of a random variable.

1.4 DISCRETELY DISTRIBUTED RANDOM VARIABLES

For a *discretely distributed* (or just "discrete") random variable X, there is a set of countably many real numbers $\{a_i\}_{i=1}^{\infty}$ such that

$$\sum_{i=1}^{\infty} \mathsf{P}(X = a_i) = 1.$$

Assuming $\mathsf{P}(X = a_i) > 0$ for all i, the countable set $\{a_i\}_{i=1}^{\infty}$ is the strict range of X, i.e., $R_X = \{a_i\}_{i=1}^{\infty}$. So, a discrete random variable has a piecewise constant CDF with a countable number of jump discontinuities occurring at the a_i's. That is, if the a_i are defined so as to be an increasing sequence, F is constant on each open interval (a_i, a_{i+1}), $F(x) = 0$ for $x < a_1$, and

$$F(a_i) = \sum_{j=1}^{i} \mathsf{P}(X = a_j) = \sum_{j=1}^{i} p(a_j),$$

where p is the probability mass function (PMF) of the discrete random variable X, i.e.,

$$p(a_i) \equiv \mathsf{P}(X = a_i).$$

Note that we have dropped the subscript "X" on the PMF and CDF for notational convenience. Moreover, for any $B \subset \mathbb{R}$ and any real-valued function g over $\mathbb{R}$,

$$\mathsf{P}(X \in B) = \sum_{a_j \in B} p(a_j)$$

and

$$\mathsf{E}g(X) = \sum_{j=1}^{\infty} g(a_j)p(a_j) = \sum_{a \in R_X} g(a)p(a).$$

To see the connection between this expression and (1.3), note that

$$\begin{aligned} \mathrm{d}F(x) &= F'(x)\,\mathrm{d}x \\ &= \sum_{i=1}^{\infty} p(a_i)\delta(x - a_i)\,\mathrm{d}x, \end{aligned}$$

where δ is the Dirac delta function [164]. That is, δ is the unit impulse satisfying $\delta(t) = 0$ for all $t \neq 0$ and

$$\int_{-\infty}^{\infty} \delta(t)\,\mathrm{d}t = 1.$$

1.4.1 The Bernoulli distribution

A random variable X that is *Bernoulli* distributed has strict range consisting of two elements, typically $R_X = \{0, 1\}$. So, there is a real parameter $q \in (0, 1)$ such that $q = \mathsf{P}(X = 1) = 1 - \mathsf{P}(X = 0)$. Also,

$$\mathsf{E}g(X) = (1 - q) \cdot g(0) + q \cdot g(1)$$

with $\mathsf{E}X = q$ in particular.

1.4.2 The geometric distribution

A random variable X that is *geometrically* distributed has a single parameter $\lambda > 0$ and its strict range is the nonnegative integers, i.e.,

$$R_X = \mathbb{Z}^+ \equiv \{0, 1, 2, ...\}.$$

The parameter λ satisfies $0 < \lambda < 1$. The CDF of X is piecewise constant with

$$F(i) = 1 - \lambda^{i+1}$$

for all $i \in \mathbb{Z}^+$. The PMF of X is $p(i) = (1 - \lambda)\lambda^i$ for $i \in \mathbb{Z}^+$. To compute $\mathsf{E}X$, we rely on a little trick involving a derivative:

$$\begin{aligned}
\mathsf{E}X &= \sum_{i=0}^{\infty} ip(i) \\
&= (1-\lambda)\lambda \sum_{i=1}^{\infty} i\,\lambda^{i-1} \\
&= (1-\lambda)\lambda \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\sum_{i=1}^{\infty} \lambda^i\right) \\
&= (1-\lambda)\lambda \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{1}{1-\lambda} - 1\right) \\
&= (1-\lambda)\lambda \frac{1}{(1-\lambda)^2} \\
&= \frac{\lambda}{1-\lambda}.
\end{aligned}$$

Similarly, the moment generating function is

$$\begin{aligned}
m(\theta) &= (1-\lambda)\sum_{i=0}^{\infty} (\mathrm{e}^{\theta}\lambda)^i \\
&= \frac{1-\lambda}{1-\lambda \mathrm{e}^{\theta}}
\end{aligned}$$

for $\mathrm{e}^{\theta}\lambda < 1$, i.e., $\theta < -\log\lambda$.

1.4.3 The binomial distribution

A random variable Y is *binomially distributed* with parameters n and q if $R_Y = \{0, 1, ..., n\}$ and, for $k \in R_Y$,

$$\mathsf{P}(Y = k) = \binom{n}{k} q^k (1-q)^{n-k},$$

where $n \in \mathbb{Z}^+$, $0 < q < 1$, and

$$\binom{n}{k} \equiv \frac{n!}{k!(n-k)!}. \tag{1.5}$$

That is, $\binom{n}{k}$ is "n choose k," see Example 1.4. It is easy to see that, by the binomial theorem, $\sum_{k=0}^{n} P(Y = k) = 1$, i.e.,

$$\sum_{k=0}^{n} \binom{n}{k} q^k (1-q)^{n-k} = (q + (1-q))^n = 1.$$

Also,

$$\begin{aligned} m(\theta) &= \sum_{k=0}^{n} \binom{n}{k} q^k (1-q)^{n-k} \mathrm{e}^{\theta k} \\ &= (q\mathrm{e}^{\theta} + (1-q))^n. \end{aligned}$$

■ **EXAMPLE 1.2**

If we are given n independent Bernoulli distributed random variables, X_i, each having the same parameter q, then $Y = \sum_{i=1}^{n} X_i$ is binomially distributed with parameters n and q. That is, for $k \in \{0, 1, 2, ..., n\}$, the event $\{Y = k\}$ can be written as a union on disjoint component events, where k of the X_i equal 1 and $n - k$ of the X_i equal 0. Each such component event occurs with probability $p^k(1-p)^{n-k}$. The number of such events, i.e., the number of ways the random vector $(X_1, X_2, ..., X_n)$ has exactly k ones, is

$$\frac{n!}{k!(n-k)!} = \binom{n}{k},$$

where $n!$ is the number of *permutations* (ordered arrangements) of n different objects and the factors $k!$ and $(n-k)!$ in the denominator account for the k ones being indistinguishable and the $n - k$ zeros being indistinguishable.

1.4.4 The Poisson distribution

A random variable X is *Poisson* distributed with parameter $\lambda > 0$ if $R_X = \mathbb{Z}^+$ and the PMF is

$$p(i) = \frac{\lambda^i}{i!}\mathrm{e}^{-\lambda}$$

for $i \in \mathbb{Z}^+$. We can check that $\mathrm{E}X = \lambda$ as in the geometric case. The MGF is

$$\begin{aligned} m(\theta) &= \sum_{i=0}^{\infty} \mathrm{e}^{\theta n} \frac{\lambda^n}{n!} \mathrm{e}^{-\lambda} \\ &= \mathrm{e}^{-\lambda} \sum_{i=0}^{\infty} \frac{(\mathrm{e}^{\theta}\lambda)^n}{n!} \\ &= \exp((\mathrm{e}^{\theta} - 1)\lambda). \end{aligned}$$

1.4.5 The discrete uniform distribution

A discrete random variable X is *uniformly* distributed on a *finite* range

$$R_X \subset \mathbb{R}$$

if

$$\mathsf{P}(X = x) = \frac{1}{|R_X|}$$

for all $x \in R_X$, where $|R_X|$ is the size of (the number of elements in) R_X. Clearly, therefore, for any $A \subset R_X$,

$$\mathsf{P}(X \in A) = \frac{|A|}{|R_X|},$$

i.e., to compute this probability, one needs to *count* the number of elements in A and R_X.

■ **EXAMPLE 1.3**

Suppose that a random experiment consists of tossing two different six-sided dice on the floor. Consider the events consisting of all outcomes having the same numbers (d_1, d_2) on the upturned faces of the dice. Note that there are $6 \times 6 = 36$ such events. Assume that the probability of each such event is $\frac{1}{36}$, i.e., the dice are "fair." This implies that the random variables d_i are independent and uniformly distributed on their state space $\{1, 2, 3, 4, 5, 6\}$.

Suppose that we are interested in $\mathsf{P}(X \in \{7, 11\})$, where the random variable

$$X \equiv d_1 + d_2.$$

That is, we are interested in the event

$$(d_1, d_2) \in \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1), (5, 6), (6, 5)\}$$

with eight members. So, $\mathsf{P}(X \in \{7, 11\}) = \frac{8}{36}$.

■ **EXAMPLE 1.4**

Suppose that five cards (a poker hand) are drawn, without replacement, from a standard deck of 52 different playing cards. The random variable X enumerates each *combination* (not considering the order in which the individual cards were drawn) of poker hands beginning with 1 and ending with the total number of different poker

hands possible, $\binom{52}{5}$, where $\binom{n}{k}$ is given in (1.5). That is, $|R_X| = \binom{52}{5}$. Assume X is uniformly distributed (i.e., a fair hand is drawn) and suppose we wish to find the probability that a flush is drawn, i.e., $\mathsf{P}(X \in \text{flushes})$. As there are four suits each having 13 cards, the number of poker hands that are flushes is $4\binom{13}{5}$. So,

$$\mathsf{P}(X \in \text{flushes}) = \frac{4\binom{13}{5}}{\binom{52}{5}}.$$

Similarly, the probability of a drawing a poker hand with a pair of aces from a fair deck is

$$\frac{\binom{4}{2}\binom{48}{3}}{\binom{52}{4}},$$

where the term $\binom{4}{2} = 6$ is the number of ways to form a pair of aces and term $\binom{48}{3}$ is the number of ways to form the balance of the hand (i.e., 3 more cards) without using aces.

■ **EXAMPLE 1.5**

Consider the context of the previous example but now suppose that we care about the *order* in which the cards were drawn, again without replacement. To compute the number of different hands, we count the number of ways we can permute 5 cards chosen from a deck of 52 without replacement. This quantity is

$$52 \times 51 \times 50 \times 49 \times 48 = \frac{52!}{(52-5)!}.$$

Note that a single *combination* of 5 different cards will have $5! = 120$ different (ordered) permutations.

On the other hand, if the cards are drawn with replacement (i.e., each card is restored to the deck after it is drawn) and we continue to care about the order in which the cards are drawn, the number of possible hands is simply 52^5. Note that in this case, a hand may consist of several copies of the same card.

A group of example discrete distributions is given in Figure 1.2.

1.5 CONTINUOUSLY DISTRIBUTED RANDOM VARIABLES

The CDF of a continuously distributed (or just "continuous") random variable X has a piecewise-continuous and bounded derivative. The derivative $f = F'$ (i.e., $\mathrm{d}F(x) = f(x)\,\mathrm{d}x$) is known as the probability density function (PDF) of X. We clearly have

$$F(x) = \int_{-\infty}^{x} f(z)\,\mathrm{d}z.$$

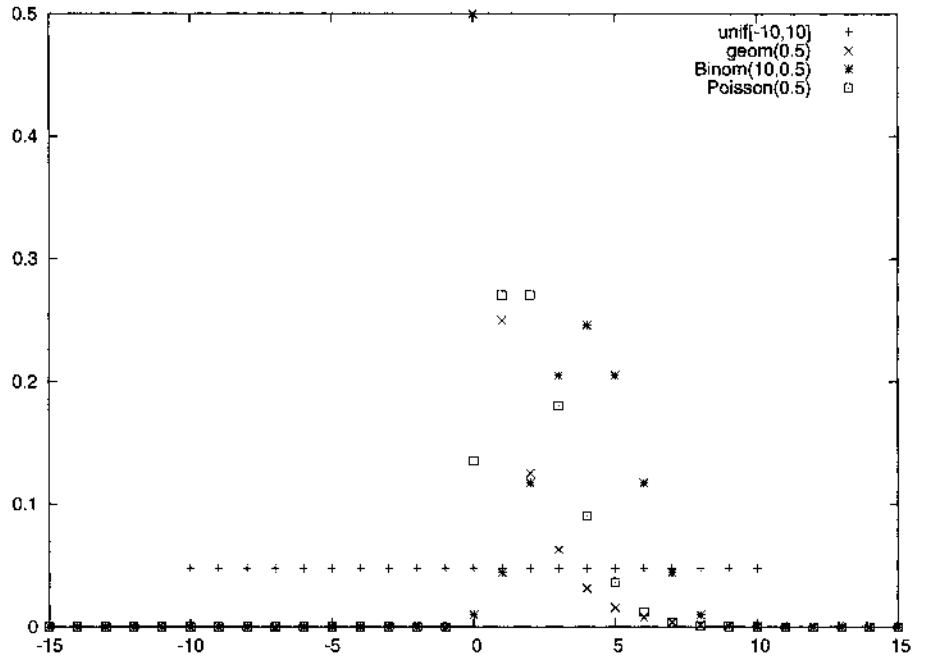

Figure 1.2 Example discrete distributions.

From this identity, we see that any PDF f is a nonnegative function satisfying

$$\int_{-\infty}^{\infty} f(z)\,\mathrm{d}z \;=\; 1.$$

Moreover,

$$\mathsf{E}g(X) \;=\; \int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{d}x$$

and, in particular, if $g(X) = \mathbf{1}\{X \in B\}$, then this reduces to

$$\mathsf{P}(X \in B) \;=\; \int_{B} f(z)\,\mathrm{d}z.$$

Finally, note that the range $R_X = \{x \in \mathbb{R} \mid f(x) > 0\}$.

1.5.1 The continuous uniform distribution

A random variable X is *uniformly* distributed over the interval $[a, b]$ if its PDF is

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \le x \le b, \\ 0 & \text{else,} \end{cases}$$

where $b > a$. Clearly, $\mathsf{E}X = (b + a)/2$. We can similarly define a uniformly distributed random variable X over any range $R_X \subset \mathbb{R}$ having finite total length, i.e., $|R_X| < \infty$.

1.5.2 The exponential distribution

A random variable X is *exponentially* distributed with real parameter $\lambda > 0$ if its PDF is

$$f(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{else.} \end{cases}$$

The CDF is $F(x) = 1 - \mathrm{e}^{-\lambda x}$ for $x \geq 0$, $\mathsf{E}X = 1/\lambda$, and the MGF is

$$m(\theta) = \frac{\lambda}{\lambda - \theta} \tag{1.6}$$

for $\theta < \lambda$; see Section 2.1.

1.5.3 The gamma distribution

A random variable X is *gamma* distributed with positive real parameters λ and r if its PDF is

$$f(x) = \begin{cases} \lambda^r x^{r-1} \mathrm{e}^{-\lambda x} / \Gamma_r & \text{if } x \geq 0, \\ 0 & \text{else,} \end{cases}$$

where the normalizing gamma function $\Gamma_r = \int_0^\infty z^r \mathrm{e}^{-z} \, \mathrm{d}z$. When r is a positive integer, we can integrate by parts to show that $\Gamma_r = (r-1)!$; in this case, the gamma distribution is sometimes called the *Erlang* distribution. Let μ be the mean and σ^2 be the variance associated with this distribution. We have the following identities:

$$\lambda = \frac{\mu}{\sigma^2} \quad \text{and} \quad r = \frac{\mu^2}{\sigma^2}. \tag{1.7}$$

Finally, the MGF is

$$m(\theta) = \left(\frac{\lambda}{\lambda - \theta} \right)^r \tag{1.8}$$

for $\theta < \lambda$. We see that a gamma distributed random variable with parameters λ and $r = 1$ is just an exponentially distributed random variable.

1.5.4 The Gaussian (or normal) distribution

A *Gaussian* (or *normally*) distributed random variable X with mean μ and variance σ^2 has PDF

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

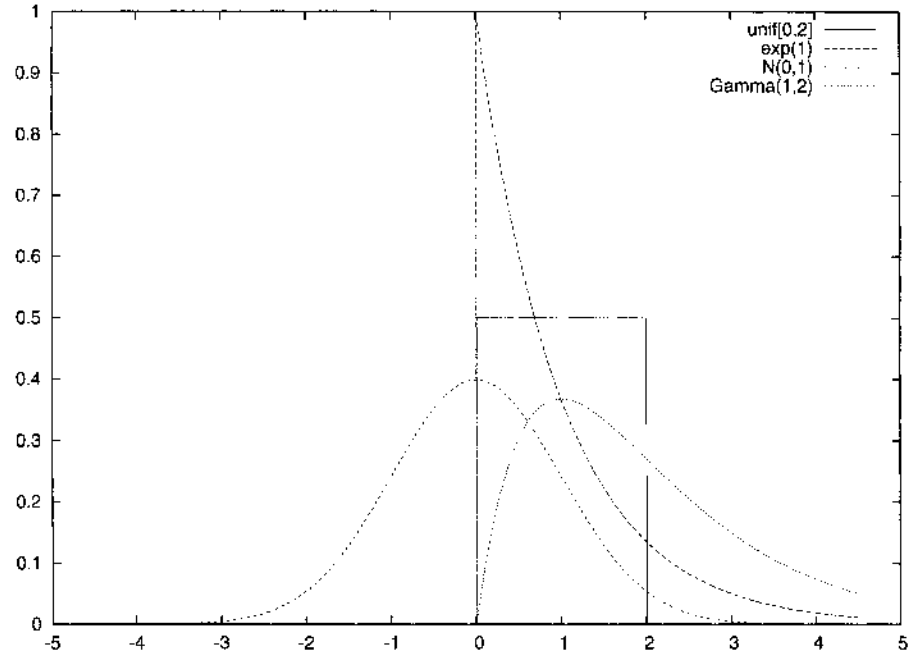

Figure 1.3 Example continuous distributions.

for $x \in \mathbb{R}$. The MGF is

$$m(\theta) = \exp(\mu\theta + \tfrac{1}{2}\sigma^2\theta^2). \tag{1.9}$$

See Figure 1.3 for example continuous distributions.

1.6 SOME USEFUL INEQUALITIES

Clearly, if the event $A_1 \subset A_2$, then $\mathsf{P}(A_1) \leq \mathsf{P}(A_2) = \mathsf{P}(A_1) + \mathsf{P}(A_2 \backslash A_1)$. For any collection of events $A_1, A_2, ..., A_n$, Boole's inequality holds:

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

Note that when the A_i are disjoint, equality holds simply by the additivity property of a probability measure P.

If two random variables X and Y are such that $\mathsf{P}(X \geq Y) = 1$ (i.e., $X \geq Y$ *almost surely*), then $\mathsf{E}X \geq \mathsf{E}Y$.

Consider a random variable X with $\mathsf{E}|X| < \infty$ and a real number $x > 0$. Since

$$|X| \geq |X|\mathbf{1}\{|X| \geq x\} \geq x\mathbf{1}\{|X| \geq x\} \text{ almost surely,}$$

we arrive at Markov's inequality:

$$\mathsf{E}|X| \;\geq\; \mathsf{E}x\mathbf{1}\{|X| \geq x\} \;=\; x\mathsf{E}\mathbf{1}\{|X| \geq x\} \;=\; x\mathsf{P}(|X| \geq x).$$

An alternative explanation for continuously distributed random variables X is

$$\begin{aligned}\mathsf{E}|X| \;&=\; \int_{-\infty}^{\infty} |z| f(z)\,\mathrm{d}z \;\geq\; \int_{-\infty}^{-x} (-z) f(z)\,\mathrm{d}z + \int_{x}^{\infty} z f(z)\,\mathrm{d}z \\ &\geq\; \int_{-\infty}^{-x} x f(z)\,\mathrm{d}z + \int_{x}^{\infty} x f(z)\,\mathrm{d}z \;=\; x\mathsf{P}(|X| \geq x).\end{aligned}$$

Now take $x = \varepsilon^2$, where $\varepsilon > 0$, and argue Markov's inequality with $(X - \mathsf{E}X)^2$ in place of $|X|$ to get Chebyshev's inequality

$$\mathrm{var}(X) \;\equiv\; E[(X - \mathsf{E}X)^2] \;\geq\; \varepsilon^2 \mathsf{P}(|X - \mathsf{E}X| \geq \varepsilon),$$

i.e.,

$$\mathsf{P}(|X - \mathsf{E}X| \geq \varepsilon) \;\leq\; \varepsilon^{-2}\mathrm{var}(X). \tag{1.10}$$

Noting that, for all $\theta > 0$, $\{X \geq x\} = \{\mathrm{e}^{\theta X} \geq \mathrm{e}^{\theta x}\}$ and arguing as for Markov's inequality gives the Chernoff (or Cramér) inequality:

$$\begin{aligned}\mathsf{E}\mathrm{e}^{\theta X} \;&\geq\; \mathrm{e}^{\theta x}\mathsf{P}(X \geq x) \\ \Rightarrow \mathsf{P}(X \geq x) \;&\leq\; \exp\left(-[x\theta - \log \mathsf{E}\mathrm{e}^{\theta X}]\right) \\ &\leq\; \exp\left(-\max_{\theta > 0}[x\theta - \log \mathsf{E}\mathrm{e}^{\theta X}]\right),\end{aligned} \tag{1.11}$$

where we have simply sharpened the inequality by taking the maximum over the free parameter θ. Note the log-MGF of X in the Chernoff bound.

The Cauchy-Schwarz inequality states that

$$\mathsf{E}|XY| \;\leq\; \sqrt{\mathsf{E}(X^2)}\sqrt{\mathsf{E}(Y^2)}$$

for all random variables X and Y with the inequality strict whenever $X \neq cY$ or $Y = 0$ almost surely for some constant c. This inequality is an immediate consequence of the fact that

$$\mathsf{E}\left(\frac{X}{\sqrt{\mathsf{E}(X^2)}} - \frac{Y}{\sqrt{\mathsf{E}(Y^2)}}\right)^2 \;\geq\; 0$$

whenever $X \neq 0$ and $Y \neq 0$ almost surely. Also, note that if we take $Y = 1$ almost surely, the Cauchy-Schwarz simply states that the variance of a random variable X is not negative, i.e., that

$$\mathsf{E}(X^2) - (\mathsf{E}X)^2 \;\geq\; 0.$$

This is also an immediate consequence of Jensen's inequality. A real-valued function g on $\mathbb{R}$ is said to be *convex* if

$$g(px + (1-p)y) \quad \leq \quad pg(x) + (1-p)g(y) \tag{1.12}$$

for any $x, y \in \mathbb{R}$ and any real fraction $p \in [0,1]$. If the inequality is reversed in this definition, the function g would be *concave*. For any convex function g and random variable X, we have Jensen's inequality:

$$g(\mathsf{E}X) \quad \leq \quad \mathsf{E}(g(X)). \tag{1.13}$$

1.7 JOINT DISTRIBUTION FUNCTIONS

For the case of two random variables X and Y the *joint* CDF is

$$F_{X,Y}(x,y) = \mathsf{P}(X \leq x, Y \leq y) \text{ for } x, y \in \mathbb{R}.$$

We can similarly define a joint CDF for more than two random variables.

1.7.1 Joint PDF

If X and Y are both continuously distributed, we can define their joint PDF as

$$f_{X,Y} \quad = \quad \frac{\partial^2 F_{X,Y}}{\partial x \, \partial y}.$$

For any (Borel measurable) real-valued function g over $\mathbb{R}^2$,

$$\mathsf{E}g(X,Y) \quad = \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y) \, \mathrm{d}x \, \mathrm{d}y$$

and, in particular, if

$$g(X,Y) = \mathbf{1}\left\{ \begin{pmatrix} X \\ Y \end{pmatrix} \in A \right\},$$

for some (Borel) $A \subset \mathbb{R}^2$, then this reduces to

$$\mathsf{P}\left(\begin{pmatrix} X \\ Y \end{pmatrix} \in A \right) \quad = \quad \iint_A f_{X,Y}(x,y) \, \mathrm{d}x \, \mathrm{d}y.$$

1.7.2 Marginalizing a joint distribution

Beginning with the joint CDF $F_{X,Y}$, we can obtain either *marginal* CDF F_Y or F_X by simply taking limits:

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x,y) \quad \text{and} \quad F_Y(y) = \lim_{x\to\infty} F_{X,Y}(x,y).$$

To see why, note that

$$\begin{aligned}\lim_{x\to\infty} F_{X,Y}(x,y) &= \mathsf{P}(X \le x, Y < \infty)\\ &= \mathsf{P}(\{\omega \in \Omega \mid X(\omega) \le x\} \cap \Omega)\\ &= \mathsf{P}(X \le x),\end{aligned}$$

where we used the fact that, by definition of a (real-valued) random variable, the event $Y < \infty$ is the whole sample space Ω. Similarly, one can recover either *marginal* PDF from the joint PDF:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x.$$

Marginal PMFs are similarly obtained by summation of a joint PMF.

1.8 CONDITIONAL EXPECTATION

Consider an event A such that $\mathsf{P}(A) > 0$ and a random variable X. The *conditional expected value of* X *given* A, denoted $\mu(X|A)$, is computed by simply using the conditional distribution of X given A:

$$F_{X|A}(z) \equiv \mathsf{P}(X \le z|A),$$

i.e.,

$$\mu(X|A) = \int_{-\infty}^{\infty} x\,\mathrm{d}F_{X|A}(x).$$

In the case of a discretely distributed random variable X, simply

$$\mu(X|A) = \sum_{j=1}^{\infty} a_j \mathsf{P}(X = a_j|A),$$

where $\{a_j\}_{j=1}^{\infty} = R_X$. The conditional PMF of X given A is denoted $p_{X|A}$, i.e., $p_{X|A}(a_j) = \mathsf{P}(X = a_j|A)$ for all j.

In a similar way we can define the conditional PDF of a continuously distributed random variable X given the event A,

$$f_{X|A}(x) \equiv \frac{\mathrm{d}}{\mathrm{d}x} F_{X|A}(x),$$

and thereby compute

$$\mu(X|A) = \int_{-\infty}^{\infty} x f_{X|A}(x)\,\mathrm{d}x. \tag{1.14}$$

Consider now two discretely distributed random variables X and Y. We will define the *conditional expectation of X given the random variable Y* denoted $\mathsf{E}(X|Y)$. The quantity $\mathsf{E}(X|Y)$ is a random variable itself. Indeed, suppose $\{b_j\}_{j=1}^{\infty} = R_Y$ and, for *all* samples

$$\omega_j \in \{\omega \in \Omega \mid Y(\omega) = b_j\} \equiv B_j$$

define

$$\mathsf{E}(X|Y)(\omega_j) \equiv \mu(X|B_j),$$

where the conditional expected value on the right-hand side can be denoted $\mu(X|Y = b_j)$. That is, the conditional expectation $\mathsf{E}(X|Y)$ maps all samples in the event B_j to the conditional expected value $\mu(X|B_j)$. Therefore, the random variable $\mathsf{E}(X|Y)$ is almost surely a *function of* Y.[1]

Now consider two random variables X and Y which are continuously distributed with joint PDF $f_{X,Y}$. For $f_Y(y) > 0$, we can define the conditional *density*:

$$f_{X|Y}(x|y) \equiv \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

for all $x \in \mathbb{R}$. Note that $f_{X|Y}(\cdot|y)$ is itself a PDF and, with this conditional density, the following conditional expected value can be computed

$$\mu(X|Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\,\mathrm{d}x,$$

where, unlike (1.14), the event $\{Y = y\}$ has zero probability. Again, note that $\mu(X|Y = y)$ is a function of y, say $h(y)$, and that the conditional expectation $\mathsf{E}(X|Y) = h(Y)$.

When X and Y are independent, the conditional distribution of X given Y is just the distribution of X and, therefore, $\mathsf{E}(X|Y) = \mathsf{E}X$. In other words, if X and Y are independent, then knowledge of Y (i.e., *given* Y) does not affect the *remaining uncertainty* of the random variable X.

■ **EXAMPLE 1.6**

For the purposes of a simple graphical example, suppose that the sample space $\Omega = [0, 1] \subset \mathbb{R}$ (but recall that, in general, Ω can be a completely abstract space without ordering). In Figure 1.4, the previous case of jointly discrete random variables X, Y and $\mathsf{E}(X|Y)$ are plotted as functions from Ω to $\mathbb{R}$. To further simplify the graph, we assume that these random variables are piecewise constant functions over Ω and that the probability P is the (Lebesgue) measure corresponding to Euclidean length.

[1] Equivalently, $\mathsf{E}(X|Y)$ is $\sigma(Y)$-*measurable*, where $\sigma(Y)$ is the smallest σ-algebra over Ω containing the events B_j. In this concrete way, $\sigma(Y)$ quantifies the information content of Y.

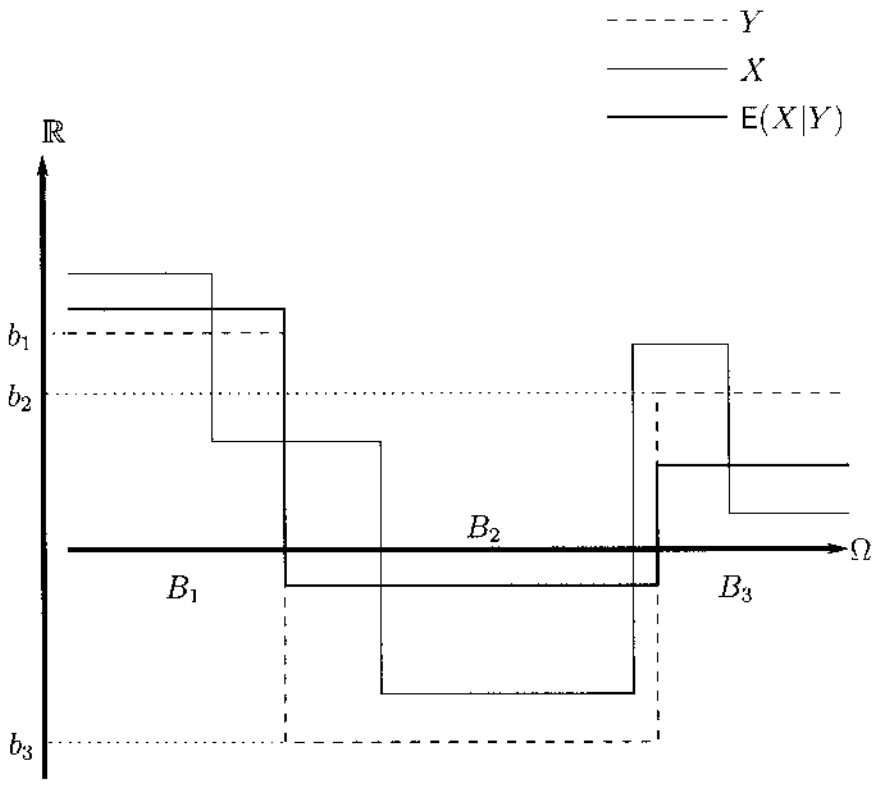

Figure 1.4 Smoothing effect of conditional expectation.

Figure 1.4 shows that $\mathsf{E}(X|Y)$ is a *smoothed* (less "uncertain") version of X. Also $\mathsf{E}(X|Y)$ depends on Y only through the events

$$B_j \equiv \{Y = b_j\},$$

i.e., the extent to which the random variable Y discriminates the samples $\omega \in \Omega$. That is, consider a discrete random variable Z with range $R_Z = \{c_j\}_{j=1}^{\infty}$ and define the events

$$C_j \equiv \{Z = c_j\}$$

for all j. If the collections of events $\{B_j\}_{j=1}^{\infty}$ and $\{C_j\}_{j=1}^{\infty}$ are the same (allowing for differences of probability zero), then $\mathsf{E}(X|Y) = \mathsf{E}(X|Z)$ almost surely. Note that R_Y can be different from R_Z, in which case

$$\mathsf{E}(X|Y) = h(Y) = \mathsf{E}(X|Z) = g(Z) \text{ almost surely,}$$

with $g \not\equiv h$.

In general, $\mathsf{E}(X|Y)$ is the function of Y which minimizes the *mean-square error* (MSE),

$$\mathsf{E}[(X - h(Y))^2],$$

among all (measurable) functions h. So, $\mathsf{E}(X|Y)$ is the best approximation of X given Y. Again, $\mathsf{E}(X|Y)$ and X have the same expectation in particular, i.e.,

$$\mathsf{E}(\mathsf{E}(X|Y)) = \mathsf{E}X.$$

It is left as a simple exercise to check this property for the cases of jointly discrete or jointly continuous random variables X and Y considered above.

1.9 INDEPENDENT RANDOM VARIABLES

A collection of continuously distributed random variables $X_1, X_2, ..., X_n$ are said to be *mutually independent* (or just "independent") if and only if their joint PDF is equal to the product of the marginal PDFs, i.e.,

$$f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

for all real $x_1, x_2, ..., x_n$. This definition is consistent with that at the end of Section 1.2.

When X and Y are both discretely distributed, we can similarly define their joint PMF $p_{X,Y}$ and there is a similar condition for independence (the joint PMF is the product of the marginal PMFs). In general, a condition for independence is that the joint CDF is the product of the marginal CDFs.

If the random variables $\{X_i\}_{i=1}^n$ are independent, then

$$\mathsf{E}\prod_{i=1}^{n} X_i = \prod_{i=1}^{n} \mathsf{E}X_i. \tag{1.15}$$

In particular, for $n = 2$, this means that if X_1 and X_2 are independent, then they are *uncorrelated*, i.e., their *covariance* equals zero:

$$0 = \mathrm{cov}(X_1, X_2) \equiv \mathsf{E}\left((X_1 - \mathsf{E}X_1)(X_2 - \mathsf{E}X_2)\right) = \mathsf{E}(X_1X_2) - \mathsf{E}X_1\mathsf{E}X_2.$$

The converse is, however, not true in general; see Problem 1.4 at the end of the chapter.

1.9.1 Sums of independent random variables

In this section we will consider sums of mutually independent continuous random variables. Our objective is to find the PDF of the sum given the PDF of the component random variables.

To this end, consider two independent random variables X_1 and X_2 with PDFs f_1 and f_2 respectively; so, $f_{X_1,X_2} = f_1 f_2$. Thus, the CDF of the sum is

$$F(z) = \mathsf{P}(X_1 + X_2 \le z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x_1} f_1(x_1) f_2(x_2)\,\mathrm{d}x_2\,\mathrm{d}x_1.$$

Exchanging the first integral on the right-hand side with a derivative with respect to z, the PDF of $X_1 + X_2$ is

$$f(z) = \frac{\mathrm{d}}{\mathrm{d}z} F(z) = \int_{-\infty}^{\infty} f_1(x_1) f_2(z - x_1)\,\mathrm{d}x_1 \quad \text{for all } z \in \mathbb{R}.$$

Thus, f is the *convolution* of f_1 and f_2 which is denoted $f = f_1 * f_2$.

In this context, moment generating functions can be used to simplify calculations. Let the MGF of X_i be

$$m_i(\theta) = \mathsf{E}\mathrm{e}^{\theta X_i} = \int_{-\infty}^{\infty} f_i(x)\mathrm{e}^{\theta x}\,\mathrm{d}x.$$

Note that m_i is basically the (bilateral) Laplace transform [164] of f_i. The MGF of $X_1 + X_2$ is

$$m(\theta) = \mathsf{E}\mathrm{e}^{\theta(X_1+X_2)} = \mathsf{E}\mathrm{e}^{\theta X_1}\mathrm{e}^{\theta X_2} = m_1(\theta) m_2(\theta), \tag{1.16}$$

where the last equation holds because of the independence of X_1 and X_2. So, convolution of PDFs corresponds to simple multiplication of MGFs (which, in turn, corresponds to addition of independent random variables).

■ EXAMPLE 1.7

As an example, suppose X_1 and X_2 are independent and both exponentially distributed with parameter λ. The PDF of $X_1 + X_2$ is f, where $f(z) = 0$ for $z < 0$ and, for $z \ge 0$,

$$f(z) = \int_0^z f_1(x_1) f_2(z - x_1)\,\mathrm{d}x_1 = \lambda^2 z \mathrm{e}^{-\lambda z}.$$

The MGF of $X_1 + X_2$ is, by (1.6) and (1.16),

$$m(\theta) = \left(\frac{\lambda}{\lambda - \theta}\right)^2,$$

which is consistent with the PDF just computed. There is a one-to-one relationship between PDFs and MGFs of nonnegative random variables.[2] We can therefore use the MGF approach to find the PDF of the sum of n independent random variables $\{X_i\}_{i=1}^n$ each having an exponential distribution with parameter λ. Indeed, the MGF of $\sum_{i=1}^n X_i$ is easily computed as

$$m(\theta) = \left(\frac{\lambda}{\lambda - \theta}\right)^n. \tag{1.17}$$

[2] In this case, the MGF is a *unilateral* Laplace transform [164]: $m(\theta) = \int_0^\infty f(z)\mathrm{e}^{\theta z}\,\mathrm{d}z$.

This is the MGF of a gamma (Erlang) distributed random variable with parameters λ and $n \in \mathbb{Z}^+$ and PDF

$$f_n(z) = \frac{\lambda^n z^{n-1} e^{-\lambda z}}{(n-1)!} \tag{1.18}$$

for $z \geq 0$.

■ **EXAMPLE 1.8**

Suppose that, for $i = 1, 2$, the random variable X_i is Gaussian distributed with mean μ_i and variance σ_i^2. Also suppose that X_1 and X_2 are independent. By (1.9) and (1.16), the MGF of $X_1 + X_2$ is

$$\begin{aligned} m(\theta) &= \exp\left(\mu_1\theta + \tfrac{1}{2}\sigma_1^2\theta^2\right) \times \exp\left(\mu_2\theta + \tfrac{1}{2}\sigma_2^2\theta^2\right) \\ &= \exp\left((\mu_1 + \mu_2)\theta + \tfrac{1}{2}(\sigma_1^2 + \sigma_2^2)\theta^2\right), \end{aligned}$$

which we recognize as a Gaussian MGF. Thus, $X_1 + X_2$ is Gaussian distributed with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$; see Problems 1.5 and 1.6.

1.10 CONDITIONAL INDEPENDENCE

If

$$\mathsf{P}(A \mid B, C) = \mathsf{P}(A \mid B), \tag{1.19}$$

the events A and C are said to be independent *given* B. This is a natural extension of the unqualified notion of independent events, i.e., events A and C are (unconditionally) independent if

$$\mathsf{P}(A \mid C) = \mathsf{P}(A).$$

Note that (1.19) implies $\mathsf{P}(C \mid B, A) = \mathsf{P}(C \mid B)$.

Similarly, random variables X and Y are conditionally independent given Z if

$$\mathsf{P}(X \in A \mid Z \in B,\ Y \in C) = \mathsf{P}(X \in A \mid Z \in B)$$

for all $A, B, C \subset \mathbb{R}$. Conditional independence does not imply (unqualified) independence, as we will see in the following chapter.

1.11 A LAW OF LARGE NUMBERS

In this section, we describe the basic connection between statistics and probability through the laws of large numbers (LLNs) [44, 62, 63]. Suppose we have an IID sequence of random

variables $X_1, X_2, X_3, \ldots$. Also suppose that the common distribution has finite variance, i.e.,

$$\begin{aligned} \sigma^2 &\equiv \mathrm{var}(X) \\ &\equiv \mathsf{E}(X - \mathsf{E}X)^2 \\ &< \infty, \end{aligned}$$

where $X \sim X_i$. Finally, suppose that the mean exists and is finite, i.e.,

$$\mu \equiv \mathsf{E}X < \infty.$$

Define the sum

$$S_n = X_1 + X_2 + \cdots + X_n$$

for $n \geq 1$ and note that $\mathsf{E}S_n = n\mu$ and $\mathrm{var}(S_n) = n\sigma^2$. The quantity S_n/n is called the *empirical* mean of X after n samples and is an *unbiased* estimate of μ, i.e.,

$$\mathsf{E}\left(\frac{S_n}{n}\right) = \mu.$$

Also, because of the following *weak* LLN, S_n/n is said to be a *weakly consistent* estimator of μ.

Theorem 1.11.1. *For all* $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathsf{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) = 0.$$

Proof: By Chebyshev's inequality (1.10),

$$\mathsf{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\mathrm{var}(S_n/n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

Therefore,

$$\lim_{n\to\infty} \mathsf{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) = 0$$

as desired. □

The *strong* LLN asserts that, if $\mathsf{E}|X| < \infty$, then

$$\mathsf{P}\left(\lim_{n\to\infty} \frac{S_n}{n} = \mu\right) = 1.$$

In other words, $S_n/n \to \mu$ almost surely. So, S_n/n is said to be a *strongly consistent* estimator of μ. A proof of the strong LLN, which implies the weak LLN, is given in [44, 62].

1.12 FIRST-ORDER AUTOREGRESSIVE ESTIMATORS

Given a series of samples X_n at "times" $n \in \{0, 1, 2, ...\}$, one may wish to iteratively estimate their mean. To this end, we can define $\overline{X}_n$ as the "current" estimate of the mean. A typical problem in this context is that the underlying distribution of the samples is slowly changing with time. For example, the Transmission Control Protocol (TCP) estimates average packet round-trip times (RTTs) [139] (and "absolute" variation about the estimated average) to calibrate a time-out mechanism for acknowledgements of transmitted packets. Also, for packet queues in Internet routers, proposed active queue management (AQM) techniques such as [105, 165, 231] estimate packet backlogs and the number of long-term active TCP sessions.

Since the distribution of X_n is changing, one could want to more significantly weight the recent samples X_k (i.e., $k \leq n$ and $k \approx n$) in the computation of $\overline{X}_n$. For example, one might use a *moving average* (MA) of order 2, i.e., for $n \geq 2$,

$$\overline{X}_n = \beta_0 X_n + \beta_1 X_{n-1} + \beta_2 X_{n-2},$$

where $0 < \beta_i$ for all i and $\sum_i \beta_i = 1$ (e.g., all $\beta_i = \frac{1}{3}$); clearly, no older samples X_k with $k < n - 2$ affect the MA $\overline{X}_n$.

Alternatively, an order 1 autoregressive (AR) estimate could be used, i.e., for $n > 1$,

$$\begin{aligned} \overline{X}_n &= \alpha \overline{X}_{n-1} + (1 - \alpha) X_n \\ \Rightarrow \overline{X}_n - \overline{X}_{n-1} &= (1 - \alpha)(X_n - \overline{X}_{n-1}), \end{aligned} \tag{1.20}$$

where $0 < \alpha < 1$ is the "forgetting factor" and $\overline{X}_0 = X_0$. Note that all past values of X contribute to the current value of this autoregressive processes according to weights that exponentially diminish:

$$\overline{X}_n = \alpha^n \overline{X}_0 + (1 - \alpha)[\alpha^{n-1} X_1 + \alpha^{n-2} X_2 + ... + \alpha X_{n-1} + X_n].$$

Also note that if $1 - \alpha$ is a power of 2, then the autoregressive update (1.20) is simply implemented with two additive operations and one bit-shift (the latter to multiply by $1 - \alpha$). There is a simple trade-off in the choice of α. A small α implies that $\overline{X}_n$ is more responsive to the current samples X_k, but this can lead to undesirable oscillations in the AR process $\overline{X}$. A large value of α means that the AR process will have diminished oscillations ("low-pass" filter) but will be less responsive to changes in the distribution of the samples X_k [136, 164].

■ **EXAMPLE 1.9**

Consider a sequence of independent random variables X_n. Initially, the distribution is uniform on the interval [0,1] (i.e., $EX = 0.5$), but for $n \geq 20$ the distribution is uniform on the interval [3,4] (i.e., EX changes to 3.5). We see from Figure 1.5 that for forgetting factor $\alpha = 0.2$, a sample path of the first-order AR process $\overline{X}$ responds much more quickly to the change in mean (at $n = 20$) but is more oscillatory than the corresponding sample path of the AR process using forgetting factor $\alpha = 0.8$.

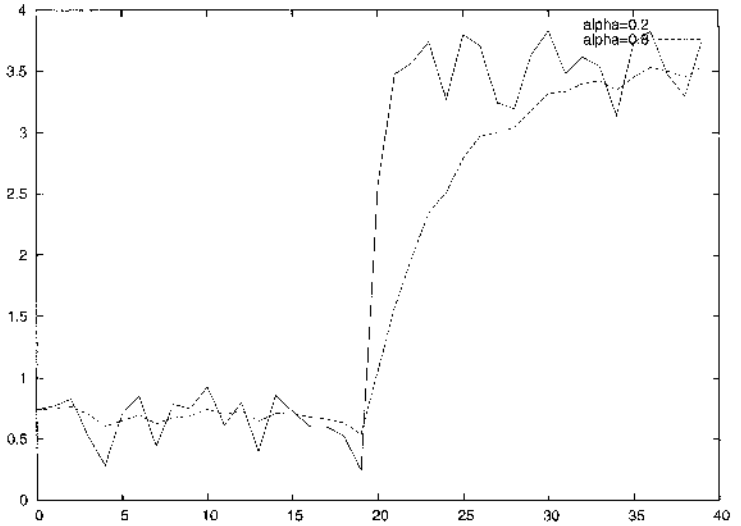

Figure 1.5 Performance of first order AR estimators.

1.13 MEASURES OF SEPARATION BETWEEN DISTRIBUTIONS

Consider two cumulative distribution functions F_1 and F_2. Suppose we wish to measure the degree to which they are different. More formally, we wish to stipulate a kind of *metric* $m(F_1, F_2) \in \mathbb{R}^+$ on the set of such distributions satisfying certain properties:

- $m(F_1, F_2) = 0$ when $F_1 \equiv F_2$ and
- m increases as the amount of "separation" between F_1 and F_2 increases.

In networking, such metrics have recently been proposed (in, e.g., [67]) to detect anomalous activity. That is, suppose a CDF F_1 representing the nominal distribution of an attribute of a flow of packets (e.g., of their destination port number attributes [41]) is known. The CDF F_2 represents the corresponding distribution as measured online. When $m(F_1, F_2)$ exceeds a certain threshold, the network operator (or some automatic intrusion detection system (IDS)) may deem that the packet flow is exhibiting abnormal behavior and decide to take action (e.g., filter-out/firewall any identified offending packets).

Some specific ways to measure the separation between two distributions include the Kolmogorov-Smirnov distance between two CDFs:

$$m(F_1, F_2) \quad = \quad \max_{x \in \mathbb{R}} |F_1(x) - F_2(x)|,$$

For a parameter $\alpha \geq 1$, we can also define

$$m(F_1, F_2) \quad = \quad \left(\int_{-\infty}^{\infty} |F_1(x) - F_2(x)|^{\alpha} \, \mathrm{d}x \right)^{1/\alpha},$$

where we note that, as $\alpha \to \infty$, this metric converges to that of Kolmogorov-Smirnov.

Given two PMFs p_1 and p_2 on the same state space we can define their *chi-squared separation* as

$$\sum_x \frac{(p(x) - q(x))^2}{p(x)}.$$

A similar definition exists for PDFs, i.e., for continuous distributions.

Given the means μ and variances σ of two distributions, the Fisher separation between them is

$$\frac{|\mu_1 - \mu_2|}{\sigma_1^2 + \sigma_2^2};$$

alternatively, the standard deviation could be used instead of the variance in the denominator leading to a dimensionless Fisher separation metric. Note how the Fisher separation is increasing in the difference of the means but decreasing in either variance. There are other potentially significant features of a distribution besides the mean and variance, e.g., median and mode that can be used as a basis of a measure of separation ([106], p. 185).

In [67], the *entropy* of a distribution was argued to be a significant feature for detection of distributed denial-of-service (DDoS) attacks. The entropy of a PMF p with strict range R is defined to be

$$\sum_{x \in R} p(x) \log p(x).$$

Thus one can consider the difference between the entropies of two distributions with the same range R, p_1 and p_2, as a measure of their separation. Another useful measure of separation is the Kullback-Leibler distance [49]:

$$\sum_{x \in R} p_1(x) \log \frac{p_1(x)}{p_2(x)}.$$

Again, similar definitions exist for PDFs.

1.14 STATISTICAL CONFIDENCE

Central limit theorems (CLTs) date back to Laplace and DeMoivre. They demonstrated that scaled sums of IID random variables of rather *arbitrary* distribution converge *in distribution* to a Gaussian. Convergence in distribution does not necessarily involve the existence of a limiting random variable and is a notion of convergence that is weaker than that of the weak LLN. A common use of the CLT is to evaluate the degree of "confidence" in the result of a group of experimental trials such as those obtained from a simulation study or a poll.

1.14.1 A central limit theorem

Suppose we have an IID sequence of random variables $X_1, X_2, X_3, \ldots$. Also, as in the case of the weak LLN, suppose that the common distribution has finite variance, i.e.,

$$\sigma^2 \equiv \text{var}(X) \equiv \mathsf{E}(X - \mathsf{E}X)^2 < \infty,$$

where $X \sim X_i$ and $\sigma > 0$. Finally, suppose that the mean exists and is finite, i.e.,

$$\mu \equiv \mathsf{E}X < \infty.$$

Define the cumulative sum

$$S_n = X_1 + X_2 + \cdots + X_n$$

for $n \geq 1$ and note again that $\mathsf{E}S_n = n\mu$ and $\mathrm{var}(S_n) = n\sigma^2$. Thus, for all n,

$$\mathsf{E}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}}\right) = 0$$

and

$$\mathrm{var}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}}\right) = 1.$$

Theorem 1.14.1. *For all* $x \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathsf{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2}\, \mathrm{d}z.$$

That is,

$$\frac{S_n - n\mu}{\sigma\sqrt{n}}$$

converges in distribution to a standard (mean 0 and variance 1) Gaussian. A proof for this central limit theorem is given in Section 5.2 of [63]; see Problem 1.16 at the end of this chapter.

1.14.2 Confidence intervals

Suppose that n identically distributed samples $X_1, X_2, X_3, \ldots, X_n$ have been generated by repeated trials of some experiment. Let μ and $\sigma > 0$ be the mean and standard deviation, respectively, of X_k. Assuming a central limit theorem for the sequence $\{X_k\}$, we have that

$$\frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu) \tag{1.21}$$

is approximately distributed as a standard Gaussian random variable, where the *sample mean* (or *empirical mean*) is

$$\overline{X}_n \equiv \frac{1}{n}\sum_{k=1}^{n} X_k.$$

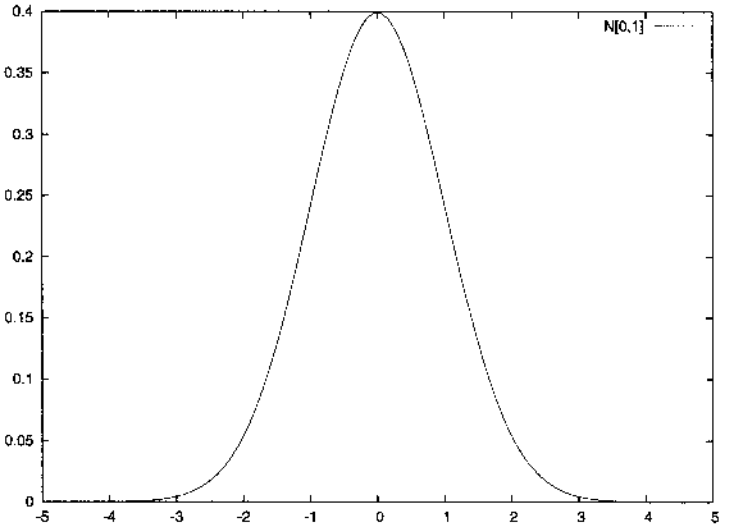

Figure 1.6 PDF of a standard Gaussian random variable $\exp(-z^2/2)/\sqrt{2\pi}$.

The mean μ is taken to be an *unknown quantity* that is to be estimated from the samples X_k. Note that the sample mean is an *unbiased* estimate of μ, i.e.,

$$\mathsf{E}\overline{X}_n = \mu.$$

This, in turn, allows us to assume that the quantity in (1.21) is Gaussian distributed for relatively small n and use this to compute "error bounds" on the law of large numbers. If Y is a standard Gaussian random variable,

$$\begin{aligned} \mathsf{P}(|Y| \leq 2) &= \Phi(2) - \Phi(-2) \\ &= \int_{-2}^{2} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2}\, \mathrm{d}z \\ &\approx 0.95; \end{aligned}$$

see Figure 1.6. Assuming that the standard deviation σ is known, by the central limit approximation, for all sufficiently large n,

$$0.95 \approx \mathsf{P}\left(\left|\frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu)\right| \leq 2\right) = \mathsf{P}\left(\mu \in \left[\overline{X}_n - \frac{2\sigma}{\sqrt{n}},\ \overline{X}_n + \frac{2\sigma}{\sqrt{n}}\right]\right). \quad (1.22)$$

So, with probability 0.95 (i.e., "19 times out of 20"), the true mean μ resides in the interval

$$\left[\overline{X}_n - \frac{2\sigma}{\sqrt{n}},\ \overline{X}_n + \frac{2\sigma}{\sqrt{n}}\right].$$

Consequently, this interval is called the *95% confidence interval* for μ.

Typically, in practice, the standard deviation σ is also not known and must also be estimated from the samples X_k. The *sample variance* is

$$\begin{aligned}\overline{\sigma_n^2} &\equiv \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2 \\ &= \left(\frac{1}{n}\sum_{k=1}^{n} X_k^2\right) - (\overline{X}_n)^2.\end{aligned}$$

The implicit assumption is that a more general form of central limit theorem holds when the sample standard deviation,

$$\overline{\sigma}_n \equiv \sqrt{\overline{\sigma_n^2}} \geq 0,$$

is used instead of the true, but unknown, standard deviation σ above [194]. That is, the 95% confidence interval for μ is taken to be

$$\left[\overline{X}_n - \frac{2\overline{\sigma}_n}{\sqrt{n}},\ \overline{X}_n + \frac{2\overline{\sigma}_n}{\sqrt{n}}\right]. \tag{1.23}$$

It turns out that the sample variance defined above is a consistent but *biased* estimator for σ^2. In fact,

$$\mathsf{E}\overline{\sigma_n^2} = \frac{n-1}{n}\sigma^2. \tag{1.24}$$

Thus, for a small number n of samples, it may prove more accurate to use

$$\sqrt{\frac{n}{n-1}\overline{\sigma_n^2}}$$

instead of $\bar{\sigma}_n$ in the 95% confidence interval formulation above. That is,

$$\frac{n}{n-1}\overline{\sigma_n^2}$$

is both a consistent and unbiased estimator of σ^2. Note that this quantity and $\overline{\sigma_n^2}$ are both *consistent* estimators of σ^2, i.e., by the LLN, they both converge to σ^2 as $n \to \infty$.

When the number of available samples n is small (less than 30), the quantity in (1.21) approximately follows a Student's t distribution [106]. Thus, instead of (1.22), the following is used to define the confidence interval for small sample sizes:

$$\mathsf{P}\left(\left|\frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu)\right| \leq \zeta_n(0.95)\right) = 0.95, \tag{1.25}$$

where the function ζ_n is defined by the Student's t distribution with n degrees of freedom (samples).

Confidence intervals are discussed in [155, 194], Section 5.3 of [63], and Section 2.3 of [74].

1.14.3 Recursive formulas and a stopping criterion

In terms of simulation code, a simple *recursive* formula in the number of observed samples for the sample mean and sample standard deviation may be convenient [194]. For the sample mean, we clearly have

$$\overline{X}_n = \frac{1}{n}((n-1)\overline{X}_{n-1} + X_n).$$

Also, the sample variance satisfies

$$\overline{\sigma_n^2} = \overline{\sigma^2}_{n-1} + \frac{1}{n}\left(-\overline{\sigma^2}_{n-1} + (X_n - \overline{X}_n)^2\right). \tag{1.26}$$

Let

$$\gamma_n \equiv \frac{n}{n-1}\overline{\sigma_n^2}$$

be the *unbiased* estimate of the variance. From (1.26), one can derive the following identity:

$$\gamma_n = \gamma_{n-1} - \frac{1}{n-1}\gamma_{n-1} + \frac{1}{n}(X_n - \overline{X}_{n-1})^2. \tag{1.27}$$

Recall that a basic assumption in the computation of the confidence interval above was that the sample points X_i are independent. In practice, a simulation may produce sequences of highly dependent X_i. In this case, several independent *batches* of consecutive samples may be obtainable by simulation [106]. The confidence interval is taken using the sample mean and standard deviation of each batch; note that the "straight" sample mean $\overline{X}_n$ would equal that obtained from the batches.

Finally, in order to arrive at a criterion to terminate a simulation, define the *relative error*

$$\xi_n \equiv \frac{\sqrt{\mathrm{var}(\overline{X}_n)}}{|\overline{X}_n|} \tag{1.28}$$

$$= \frac{\overline{\sigma}_n}{\sqrt{n}|\overline{X}_n|}. \tag{1.29}$$

Here we have assumed $\mu \neq 0$ and $\sigma^2 < \infty$. A simulation may be terminated when the relative error reaches, say, 0.1, i.e., the variance of the sample standard deviation is 10% of the sample mean. Note that we can express the statement of the confidence interval (1.23) as

$$\mathsf{P}\left(\frac{\mu}{\overline{X}_n} \in [1 - 2\xi_n,\ 1 + 2\xi_n]\right) \approx 0.95, \tag{1.30}$$

where we could use $\zeta_n(0.95)$, defined using the Student's t distribution in (1.25), instead of "2" in (1.30) if the number of samples n is small. So, using the stopping criterion $\xi_n \leq 0.1$, the claim is then made that the sample mean $\overline{X}_n$ is accurate to within 20% of the true mean with 95% probability. Also, for a relative error of 0.1, the *number of required samples* is, by (1.28),

$$n = 100\left(\frac{\overline{\sigma}_n}{\overline{X}_n}\right)^2 \approx 100\left(\frac{\sigma}{\mu}\right)^2. \tag{1.31}$$

In summary, a simulation produces IID samples X_i from which a sequence of sample means $\overline{X}_n$ and sample variances $\overline{\sigma^2}_n$ are computed. The simulation may be terminated when the computed relative error ξ_n reaches a predefined threshold. The accuracy of the resulting estimate $\overline{X}_n$ of the quantity of interest μ can be interpreted in terms of a confidence interval (implicitly invoking a central limit theorem or Student's t distribution).

■ **EXAMPLE 1.10**

A great deal of current research activity in networking involves *comparisons* of the performance of competing mechanisms (devices, algorithms, protocols, etc.) by simulation or through prototypical deployment on the Internet. Suppose that n trials are conducted for each of two mechanisms in order to compare their performance, leading to a dataset

$$\{D_{k,i}\}_{i=1}^{n}$$

for mechanism $k \in \{1, 2\}$. Also suppose that, for all i, the ith trial (yielding data $D_{1,i}$ and $D_{2,i}$) was conducted under arguably equal environmental conditions for both mechanisms. Let

$$X_i \equiv D_{1,i} - D_{2,i}$$

be the difference in the performance of the two mechanisms for the common environmental conditions of trial i. That is, for each trial, an "apple-to-apple" comparison is made using *coupled* or *paired* observations.

To assess whether $\mathsf{E}X > 0$ (respectively, $\mathsf{E}X = 0$), we simply compute the confidence interval according to (1.22) or (1.25) and determine whether the origin is to the left of it (respectively, contained by it).

Given uncoupled observations that are different in number, a procedure for deciding whether $\mathsf{E}X = 0$ (i.e., a "t test") is summarized in Section 13.4.2 of [106].

1.15 DECIDING BETWEEN TWO ALTERNATIVE CLAIMS

Suppose that a series of measurements X_i are drawn from the Internet. A problem is that the Internet may be in different "states" of operation leading to samples X_i that are independent but *not* identically distributed. As a great simplification, suppose that the network can be in only one of two states indexed by $j \in \{1, 2\}$. Also suppose that a batch of n samples X_i, $1 \leq i \leq n$, is taken while the network is in a single state but that state is not known. Finally, suppose it is known that the probability that the network is in state 1 is p_1 and that

$$\mu_j = \mathsf{E}(X_i \mid \text{network state } j)$$

for *known* values μ_j, $j \in \{1, 2\}$. Without loss of generality, take $\mu_1 < \mu_2$.

We wish to infer from the sample data X_i the state of the network. More specifically, we wish to minimize the probability of error P_e in our decision. To do this, note that by the

central limit theorem, *given* that the network is in state j, the sample mean

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is approximately normally distributed with mean μ_j and variance $\sigma^2_{j,n}$. An unbiased, consistent estimate of this variance is

$$\overline{\sigma^2_{j,n}} \equiv \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

To determine the probability of decision error, we will condition on the state of the network:

$$\mathsf{P}_e = \sum_{k=1}^{2} \mathsf{P}(\text{error} \mid \text{network state } k)p_k, \tag{1.32}$$

where $p_2 = 1 - p_1$. To this point, we have not described the method by which a decision on the network state is made. Consider a decision based on the comparison of the sample mean $\overline{X}_n$ with a threshold θ, where

$$\mu_1 \leq \theta \leq \mu_2,$$

so that the network is deemed to be in state 2 (having the higher mean) if $\overline{X}_n > \theta$; otherwise, the network is deemed to be in state 1. So, a more concrete expression for the error probability (1.32) ensues because we can approximate

$$\begin{aligned}\mathsf{P}(\text{error} \mid \text{network in state 2}) &= \mathsf{P}(\overline{X}_n \leq \theta \mid \text{network in state 2}) \\ &\approx \Phi\left(\frac{\mu_2 - \theta}{\sqrt{\overline{\sigma^2_{2,n}}}}\right),\end{aligned}$$

where Φ is 1 minus the CDF of a standard Gaussian distribution. Using a similar argument when conditioning on the network being in state 1, we arrive at the following expression:

$$P_e = \Phi\left(\frac{\theta - \mu_1}{\sqrt{\overline{\sigma^2_{1,n}}}}\right)p_1 + \Phi\left(\frac{\mu_2 - \theta}{\sqrt{\overline{\sigma^2_{2,n}}}}\right)p_2. \tag{1.33}$$

Thus, to determine the optimal value of θ, we need to minimize P_e over θ. This approach can be easily generalized to make decisions among more than two alternative (mutually exclusive) hypotheses. In [113], Wald's classical framework to decide between two alternative claims based on *sequential* independent observations is used on failed scan (connection set-up attempt) data to detect the presence of Internet worms.

Problems

1.1 For any two events A and B show that

$$\mathsf{P}(A) \leq \mathsf{P}(A \cap B) + \mathsf{P}(\overline{B}).$$

1.2 Prove the law of total probability (1.1).

1.3 Consider two independent random variables X_1 and X_2.

(a) Show that they are uncorrelated, i.e., $\mathrm{cov}(X_1, X_2) = 0$.

(b) Show that $\mathrm{var}(X_1 + X_2) = \mathrm{var}(X_1) + \mathrm{var}(X_2)$.

1.4 Suppose that X is a continuous random variable uniformly distributed on the interval $[-1, 1]$ and that $Y = X^2$. Show that X and Y are uncorrelated but that they are (clearly) dependent.

1.5 Two random variables X_1 and X_2 are said to be *jointly Gaussian* distributed if their joint PDF is

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2\pi \det(\mathbf{C})} \exp\left(-(\underline{x} - \underline{\mu})^{\mathrm{T}} \mathbf{C}^{-1} (\underline{x} - \underline{\mu})\right), \qquad (1.34)$$

where the (symmetric) *covariance matrix* is

$$\mathbf{C} = \begin{bmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) \\ \mathrm{cov}(X_1, X_2) & \mathrm{var}(X_2) \end{bmatrix}, \qquad \mu = \begin{bmatrix} \mathsf{E}X_1 \\ \mathsf{E}X_2 \end{bmatrix}$$

and $\det(\mathbf{C})$ is the *determinant* of $\mathbf{C}$. Show that if two jointly Gaussian random variables are uncorrelated, then they are also independent.[3]

1.6 For *jointly* Gaussian X_1 and X_2 show:

(a) For scalars α_1 and $\alpha 2$, $\alpha_1 X_1 + \alpha_2 X_2$ (a "linear combination" of Gaussian random variables) is Gaussian distributed.

(b) $\mathsf{E}(X|Y)$ is Gaussian distributed.

1.7 Show that a Bernoulli random variable X with $R_X = \{0, 1\}$ can always be represented as an indicator function $\mathbf{1}_A$ for some event A (express A in terms of X). Prove Jensen's inequality (1.13) for a Bernoulli random variable.

1.8 Prove Jensen's inequality for any discretely distributed random variable.

1.9 Find the MGF of a geometrically distributed random variable and verify the expression given for the MGF corresponding to a Poisson distribution.

1.10 Compute the variance of the distributions described in Sections 1.4 and 1.5.

1.11 If $\mathsf{P}(X \geq 0) = 1$ (i.e., X is nonnegative almost surely), show

$$\mathsf{E}X = \int_0^\infty \mathsf{P}(X > x)\, \mathrm{d}x.$$

1.12 Suppose X is Cauchy distributed with PDF

$$f(x) = \frac{1}{\pi(1 + x^2)}$$

[3]The joint PDF of $n > 2$ jointly Gaussian random variables has the same form as (1.34) except that the 2π term is replaced by the more general term $(2\pi)^{n/2}$.

for $x \in \mathbb{R}$.

(a) Verify that $\mathsf{E}(X\mathbf{1}\{X > 0\}) = \infty$ and that $\mathsf{E}(-X1\{X < 0\}) = \infty$

(b) Conclude that $\mathsf{E}X$ does not exist. Note that the question of *existence* is different from the question of *finiteness*.

1.13 The *linear least square error* (LLSE) estimator of a random variable X given a random variable Y is a linear function $h(Y) = aY + b$ which minimizes the MSE

$$\mathsf{E}[(X - h(Y))^2];$$

i.e., for the LLSE estimator, the constants a and b are chosen so as to minimize the MSE. Show that the LLSE estimator is the conditional expectation $\mathsf{E}(X|Y)$ when the random variables X and Y are jointly Gaussian with $\mathrm{var}(Y) > 0$ (so, in this case, the best estimator is a linear one).

1.14 For the gamma distribution, verify (1.7) and (1.8) and show that

$$\Gamma_r = (r-1)!$$

when $r \in \mathbb{Z}^+$.

1.15 Use MGFs to prove that the binomial distribution with parameters n and q converges to the Poisson distribution with parameter λ as $q \to 0$ and $n \to \infty$ in such a way that $nq \to \lambda$. This is Poisson's theorem (sometimes called the law of small numbers). Hint: $(1+x)^{1/x} \to \mathrm{e}$ as $x \downarrow 0$.

1.16 Consider a random variable Y_λ that is Poisson distributed with parameter λ, i.e., $\mathsf{E}Y_\lambda = \lambda$.

(a) Show that $\mathrm{var}(Y_\lambda) = \lambda$ too.

(b) Using MGFs, show that, as $\lambda \to \infty$, the distribution of

$$\frac{Y_\lambda - \lambda}{\sqrt{\lambda}}$$

converges to a standard (mean 0 and variance 1) Gaussian distribution.

This is a kind of CLT for Poisson processes; see Chapter 2. The CLT has been generalized to many other contexts, including functional CLTs on the convergence of stochastic processes to Brownian motion, see [21].

1.17 Prove (1.24).

1.18 Prove (1.26) and (1.27).

1.19 If U is a continuous random variable that is uniformly distributed over $[0, 1]$, show that $F^{-1}(U)$ has CDF F, where

$$F^{-1}(u) \equiv \inf\{x \mid F(x) = u\}.$$

1.20 Suppose a group of (distinct) persons each has an independent birthday among the 365 possible ones. Determine the minimum number n of such persons required so that the

probability that at least two share a birthday is greater than or equal to 0.5.
Hints: Compute the probability of the complementary event and the answer is between 20 and 25.
The surprisingly small result is known as the birthday paradox.

1.21 Suppose a hacker wishes to have a domain name server (DNS) associate the domain name www.kesidis.com with one of his own 32-bit Internet Protocol (IP) addresses so that he can intercept some of the critically important correspondences that are directed to this site. The hacker simultaneously transmits q identical queries for www.kesidis.com to the targeted DNS. Further suppose that the targeted DNS naively forwards each query to an authoritative DNS using IID transaction identifiers (used to authenticate the authoritative DNS's response) which are 16 bits long and not known to the hacker. Shortly thereafter, and before the authoritative DNS can reply, the hacker also transmits s responses to the targeted DNS spoofing those of the authoritative DNS, where each such response associates www.kesidis.com with the hacker's chosen IP address and contains a guess at one of the transaction identifiers generated by the targeted DNS. Assuming

$$s = q \equiv n,$$

find the value of n so that the probability that a forwarded query and spoofed response have the same transaction identifier is 0.5, i.e., the probability that the hacker guesses correctly and thereby poisons the targeted DNS's cache.

1.22 In an idealized network using ALOHA medium access control, each of n nodes attempts to transmit a packet in a time slot with probability p, after the initial packet transmission failed due to interference from another host. Such retransmission decisions are independent. Suppose all hosts are synchronized to common time slot boundaries and that they are always in "retransmission" mode (with a packet to transmit).

(a) Find the probability that a packet is successfully transmitted in a given time slot by a given node.

(b) Find the probability that a packet is successfully transmitted in a given time slot by any node.

(c) Find the expected number of successfully transmitted packets per time slot by the group of nodes, i.e., the network's throughput.

(d) Show that the throughput is maximized by the choice of $p = 1/n$ and that, as $n \to \infty$, the throughput converges to $1/e \approx 0.37$ packets per time slot.

(d) Show that the maximum throughput of *un*slotted ALOHA is $1/(2e)$.

1.23 Show that the minimizing value of the decision threshold θ in (1.33) is $(\mu_1 + \mu_2)/2$ when the sample variances are equal, i.e., $\overline{\sigma^2_{1,n}} = \overline{\sigma^2_{2,n}}$.

1.24 Consider a link carrying packets to a Web server. Suppose that, under normal operating conditions, the link will subject the Web server to a data rate of 4 Mbps. However, when the Web server is under a DDoS attack, the link will carry an average of 6 Mbps to the server. An IDS samples the link's data rate and determines whether the server is under attack. Assume known standard deviations in the data rate of 1 Mbps under normal conditions and

of 1.5 Mbps under attack conditions. Finally, assume attack conditions exist 25% of the time. Find the value of the optimal decision threshold θ (compared against the sample mean) that minimizes the probability of decision error.

1.25 In the previous problem, instead of minimizing the probability of decision error, suppose the probability that it was decided that the network is under attack when, in fact, it was not can be no more than 0.10, i.e.,

$$\mathsf{P}_n(\theta) \equiv \mathsf{P}(\text{decision error} \mid \text{no attack}) \quad \leq \quad 0.10.$$

Again, this decision is based on the sample mean. Note that this event is called a *false positive* or *type II* error. Such false positives can be caused by legitimate "flash crowds" of temporary but excessive demand. Subject to this bound we wish to minimize the probability of missed detection ("type I" error) of an actual attack:

$$\mathsf{P}_a(\theta) \equiv \mathsf{P}(\text{decision error} \mid \text{attack}).$$

That is, find

$$\arg \min_{\theta \mid \mathsf{P}_n(\theta) \leq 0.10} \mathsf{P}_a(\theta).$$

This is called a *Neyman-Pearson test* [174] (of the sample mean with the computed value of θ).

1.26 Consider a simple *binary symmetric channel* model wherein a single bit $X \in \{0, 1\}$ is transmitted and a single bit Y is received such that the transmission error probabilities are equal, i.e., $\mathsf{P}(Y = 1|X = 0) = \mathsf{P}(Y = 0|X = 1) \equiv \epsilon_e$. For $i \in \{0, 1\}$, find $\mathsf{P}(X = i|Y = i)$ (i.e., the probability that bit i was transmitted given bit i was received) in terms of ϵ_e and $s_0 \equiv \mathsf{P}(X = 0)$.

1.27 Suppose that the reported outcome of a poll regarding a presidential race is, "Candidate X will collect 48% of the vote with an error of $\pm 2\%$ 19 times out of 20." Explain this outcome in the terms previously used to describe statistical confidence based on the CLT. Also, what are the basic underlying assumptions about which specific individuals, among the entire voting population, were polled for this statement to hold?

1.28 In the context of Example 1.10, suppose that we have compiled only the separate empirical distributions p_1 and p_2, respectively, based on the data sets $\{D_{1,i}\}_{i=1}^n$ and $\{D_{2,i}\}_{i=1}^n$, but *not* the empirical distribution of $X = D_1 - D_2$ on a trial-by-trial basis. Given p_1 and p_2:

(a) Can we obtain the empirical mean $\overline{X}_n$ of X?

(b) Can we obtain the empirical variance of X?