

Chapter 1

Probability

1.1 Definition of Probability

The notion of probability concerns the measure (“size”) of sets in a space. The space may be called a **Sample Space** or an **Event Space**. We define a probability as a set function according to:

Definition 1.1 *Let S be a (sample) space. Let $P(E)$ be a real additive set function defined on sets E in S ; P is called a **probability function** if the following conditions are satisfied:*

1. *If E is a subset (**event**) in S , then $P(E) \geq 0$.*
2. *$P(S) = 1$.*
3. *If $E, F \in S$ and $E \cap F = \emptyset$, then $P(E \cup F) = P(E) + P(F)$. If S is an infinite sample space, and E_1, E_2, E_3, \dots is an infinite sequence of disjoint events in S , then*

$$P(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i). \quad (1.1)$$

In the language of measure theory, this definition is equivalent to:

Definition 1.2 *Let S be a (sample) space. A **probability** P defined on S is a measure on S such that:*

$$P(S) = 1. \quad (1.2)$$

The motivation for this definition is the idea that we may have a set of possible outcomes, with each outcome having some likelihood of occurring, given by its probability measure. The sample space contains all possible outcomes, and the measure is normalized such that the probability over the whole space is unity.

Note that we shall use the terms *outcome*, *sampling*, and *event* interchangeably in the present context. We ignore here the issue of whether all possible subsets of S may be assigned a probability measure; the sets we shall be concerned with are measurable.

For example, consider $S_3 \equiv \{a, b, c\}$. The measure generated by $P(a) = P(b) = P(c) = 1/3$ defines a probability on S_3 . In particular, note that $P(S_3) = P(a) + P(b) + P(c) = 1$, according to the additive property of a measure.

Probability typically is used as the mathematical tool dealing with “randomness” or “uncertainty”. The concept of randomness is represented by the process of sampling an outcome from a set of possible outcomes, where the actual outcome is unknown until the sampling has occurred. In the above example with S_3 , the elements a, b, c of the sample space may be given a concrete realization in the form of three distinct balls to be drawn blindly from a container. We don’t know prior to the drawing which ball we will get. We only know that the probability of getting a specified ball is $1/3$.

The concept of uncertainty permits an extension of the application of probability beyond describing random processes. For example, it may be that in some trial process, only one outcome will actually occur. That is, the sample space in reality has only one element. However, we may be ignorant, prior to the trial, concerning the element. In this case, a larger sample space may be defined, with probabilities assigned to describe our uncertainty in what the true element is. Put another way, we may describe our “degree-of-belief” in the outcome as a probability measure on the enlarged sample space of possibilities.

When describing random processes, probability may be given a *frequency* interpretation. We may imagine repeating the sampling process many times. The outcome will vary from sampling to sampling, but we may keep a tally of the outcomes over a large number of samplings. The relative frequencies of the different outcomes will be given, in the limit of an infinite number of samplings, by the relative probabilities.

1.2 Random Variables and Notation

The sample space S is in general abstract. However, we may more-or-less naturally be able to define a correspondence between the elements of S and a set R_S of real numbers (including perhaps a vector of real numbers). In this case, the probability measure defines a measure on that set of real numbers. This definition of the measure may be extended to all real numbers by defining the measure to be 0 for any real number $X \notin R_S$. We will use the same symbol P to denote the probability measure on R_S as on the abstract sampling space S . The real number X that takes on values according to this equivalent measure is called a “random variable”. We will loosely use the terms **probability distribution** or **sampling distribution** to refer to the probability measure and its possible functional descriptions. Two probability distributions are considered equivalent (identical) if they differ on at most a set of measure (probability) zero.

There are a variety of notations in use for dealing with random variables and associated probability distributions. We'll adopt the following convention: Let X be a random variable. Let $F_X(x)$ be the probability that random variable X takes on a value that is less than or equal to x . We thus have:

$$F_X(x) \equiv P(X \leq x). \quad (1.3)$$

The function $F_X(x)$ is called the **cumulative distribution function**, or “CDF” for short. It is also referred to just as the **distribution function** in statistics literature. Specifying the CDF specifies the probability distribution. It is common practice to drop the X subscript and write just $F(x)$. We'll employ this convention except where it is useful to maintain the more formal notation. It is also common practice in physics usage to use the same symbol for the random variable as for a possible value the random variable could take on. This can lead to confusion, and we don't adopt this practice except in cases where it is convenient and should not cause confusion.

There are two common cases for our probability measure:

- Those where the sample space maps injectively into the integers;
- Those where the sample space maps injectively and continuously into the real numbers.

In the first case, we have a **discrete probability distribution**. For example, consider the “Poisson distribution”, arising in counting radioactive decays during a given time interval. The random variable N takes on values in the space of non-negative integers. The probability distribution (CDF) is

$$F_N(n) = \sum_{k=0}^n \frac{\theta^k}{k!} e^{-\theta}. \quad (1.4)$$

The real number $\theta \geq 0$ is a given parameter. Notice that as n becomes very large, $F_N(n)$ approaches one.

In the second case, we have a **continuous probability distribution**. For example, consider the “uniform distribution”, with CDF:

$$F_X(x) = \begin{cases} 0 & x \leq 0, \\ x/\theta & x \in (0, \theta), \\ 1 & x \geq \theta. \end{cases} \quad (1.5)$$

Again, θ is a given parameter. This distribution results when we sample in the interval $(0, \theta)$ with equal probability for each value. For simplicity, in general discussions we will typically use the notation for continuous distributions, with the understanding that the discrete measure is substituted for discrete distributions.

In both cases, we can define a **probability density function**, or PDF. The PDF corresponding to CDF $F_X(x)$ is usually denoted $f_X(x)$. For the discrete distribution it is given by:

$$f_X(x) = F_X(x) - F_X(x - 1), \quad (1.6)$$

or just $F_X(x)$ if x is the smallest value X can take on. That is, for a discrete distribution, we define the PDF to be $f_X(x) \equiv P(X = x)$. Thus, for the Poisson distribution, we have the PDF:

$$f_N(n) = \frac{\theta^n}{n!} e^{-\theta}, \quad n = 0, 1, \dots \quad (1.7)$$

For the continuous case:

$$f_X(x) = \frac{dF_X(x)}{dx}. \quad (1.8)$$

For the uniform distribution, this gives

$$f_X(x) = \begin{cases} 1/\theta, & x \in (0, \theta), \\ 0, & \text{otherwise.} \end{cases} \quad (1.9)$$

The constancy of $f_X(x)$ motivates the name “uniform”. Note that, whether continuous or discrete, the PDF is non-negative. In the continuous case, however, it may take on values greater than one. Figure 1.1 illustrates examples of CDFs and PDFs for discrete and continuous distributions.

We may make a technical remark here. It is convenient to consider our function space to be a space of equivalence classes. Two functions belong to the same equivalence class if they differ on at most a set of probability measure zero. Two functions in the same class are considered to be identical, and we adopt this notion of equality.

1.3 Some Basic Properties

We state, mostly leaving proof to the reader, some elementary properties of probabilities. First is the **rule of complementation**. We’ll denote the **complement** of event E with respect to space S by \tilde{E} . That is, $E \cup \tilde{E} = S$ and $E \cap \tilde{E} = \emptyset$. Then we have:

Theorem 1.1 (Rule of Complementation)

$$P(\tilde{E}) = 1 - P(E). \quad (1.10)$$

More generally, for a set of (not necessarily disjoint) events $\{E_1, E_2, \dots, E_k\}$:

$$P\left(\bigcap_{i=1}^k \tilde{E}_i\right) = 1 - P\left(\bigcup_{i=1}^k E_i\right). \quad (1.11)$$

This fact often comes in handy – sometimes it is much easier to compute the probability that \tilde{E} occurs. For example, suppose we want to know the probability that we will get at least one heads in n coin tosses. There are many sequences producing at least one heads. But there is only one sequence with no heads, with probability $1/2^n$. Then by the rule of complementation the probability to get at least one heads is $1 - 1/2^n$.

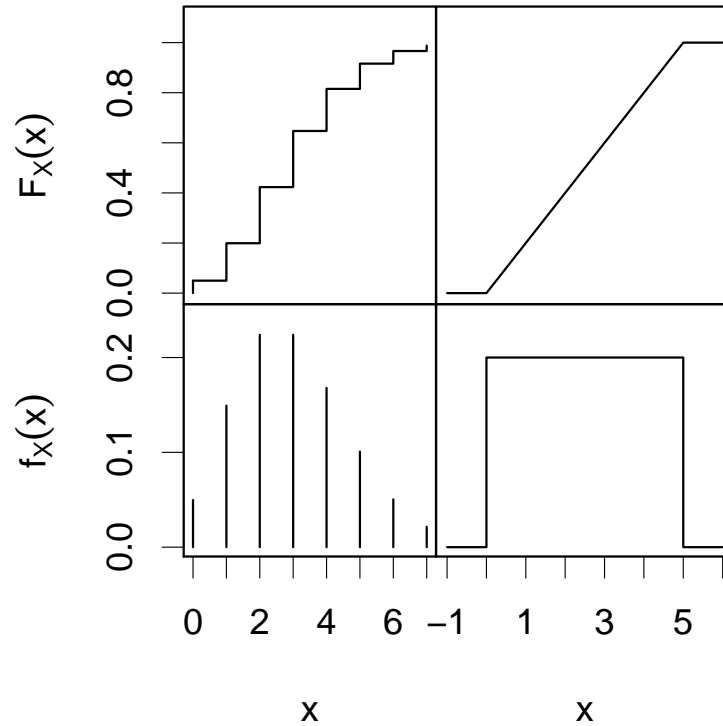


Figure 1.1: Examples of CDFs (top) and PDFs (bottom) for a discrete distribution (left) and a continuous distribution (right). The discrete distribution is a Poisson with mean three, and the continuous distribution is a uniform distribution from zero to five.

Let a and b be two distinct elements of sample space S , with probabilities $P(a)$ and $P(b)$ respectively. From the additive property of a measure, $P(\{a, b\}) = P(a) + P(b)$. Now suppose $A \subset S$ and $B \subset S$ are two measurable subsets of S , according to the probability measure. If A and B are disjoint, then we must have, again by the additive property:

$$P(A \cup B) = P(A) + P(B), \quad \text{if } A \cap B = \emptyset. \quad (1.12)$$

More generally, if A and B may not be disjoint,

$$\begin{aligned} P(A \cup B) &= P(A \cap \tilde{B}) + P(B) \\ &= P(A) + P(B) - P(A \cap B). \end{aligned} \quad (1.13)$$

The probability that both A and B occur, $P(A \cap B)$ is called the **joint probability** of A and B . We often write this as $P(A, B)$, particularly when discussing probabilities of possible values of random variables. The probability that A occurs independent of B is called the **marginal probability** for A .

We may introduce the notion of a **conditional probability**: Think of $B \subset S$ as defining a new sample space, from which we may draw events with certain probabilities. Let us use P_B to denote the probability measure on B . If these probabilities are governed by the same probability measure as on S , then the relative probabilities among samplings from B are the same as the relative probabilities on S . In order to define a probability measure on B , we must satisfy the condition

$$P_B(B) = 1. \quad (1.14)$$

Therefore,

$$P_B(A) = P(A)/P(B), \quad (1.15)$$

for any $A \subset B$. The measure $P_B(A)$ is referred to as the “probability of A given B ”, and is called a “conditional probability”, denoted by

$$P(A|B) \equiv P_B(A). \quad (1.16)$$

We may extend the definition of a conditional probability to any set $A \subset S$. We partition A according to:

$$A = (A \cap B) \cup (A \cap \tilde{B}). \quad (1.17)$$

Then we obtain

$$P_B(A) = P(A|B) = P(A \cap B)/P(B). \quad (1.18)$$

We may now quote the important theorem relating conditional probabilities:

Theorem 1.2 (Bayes) *Let $A \subset S$ and $B \subset S$. Then*

$$P(B|A)P(A) = P(A|B)P(B). \quad (1.19)$$

The proof is obvious from our definition of a conditional probability. Assuming $P(A) \neq 0$ this theorem is generally used in the form

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}. \quad (1.20)$$

For an example of the application of Bayes theorem, consider the following problem: You are having a disagreement with a colleague over the value of a fundamental parameter $\theta \in [1, 2]$. You are correctly sure that θ is between one and two, but any value in between is just as good as any other as far as you know. You design an experiment to measure θ . The sampling distribution describing your experiment is:

$$P(x; \theta)dx = \begin{cases} \frac{1}{\theta} & x \in [1, \theta] \\ 0 & x > \theta. \end{cases} \quad (1.21)$$

That is, x is sampled uniformly on the interval $[1, \theta]$. Given the result, x , of the experiment you make a bet with your colleague that $\theta > a$. What should a be for a break even bet?

This example raises the interesting subject of Bayesian statistics, but we will leave that discussion for later. For now, we imagine a model $[P(\theta)]$ encapsulating your opinion about θ , before the experiment is performed, in which θ has been sampled uniformly on the interval $[1, 2]$. The experiment is then performed with this sampled value of θ . We are given x and wish to know $P(\theta|x)d\theta$. Note that $P(x|\theta) = P(x; \theta)$. Then according to Bayes theorem we have

$$P(\theta|x)d\theta = \frac{P(x|\theta)dxP(\theta)d\theta}{\int_{\theta=1}^{\theta=2} P(x|\theta)dxP(\theta)d\theta} \quad (1.22)$$

$$= \begin{cases} \frac{d\theta}{\theta \ln(2/x)}, & \theta \geq x, \\ 0 & \theta < x. \end{cases} \quad (1.23)$$

Letting $0.5 = \int_a^2 P(\theta|x)d\theta$, we obtain the result $a = \sqrt{2x}$. For example, if $x = 1.5$, we'll make our break-even bet at $\theta > \sqrt{3}$. Presumably, the bet must eventually be settled by performing further experiments.

1.4 Statistical Independence

If an event in our sample space may be described by n random variables $X = \{X_1, X_2, \dots, X_n\}$, we refer to the distribution function $F_X(x) \equiv P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n)$ as the **joint distribution function** for X . We may define a notion of independence of random variables according to:

Definition 1.3 *Two random variables, X and Y , are statistically independent iff:*

$$f_{XY}(x, y) = f_X(x)f_Y(y),$$

where f_{XY} is the joint distribution (PDF) for X and Y , f_X is the marginal distribution for X , and f_Y is the marginal distribution for Y .

An important case of statistical independence is when we sample multiple times from the same distribution. That is, we imagine sampling n times from f_X . Our sample space consists of vectors in an n -dimensional vector space, with random variables X_1, X_2, \dots, X_n . Each sampling from f_X is independent of all other samplings. That is, the probability distribution is

$$f_{\{X\}}(\{x\}) = \prod_{k=1}^n f_{X_k}(x_k). \quad (1.24)$$

Thus, X_i and X_j are statistically independent if $i \neq j$. We refer to these X_i as **Independent, Identically Distributed** and use the acronym IID.

1.5 Expectation Values

The **Expectation Value** of a function, u , of a random variable X , is defined by:

$$\langle u(X) \rangle = \int_{\text{all } x} u(x)f(x)dx, \quad (1.25)$$

with the obvious generalization to joint PDFs. Other common notations are:

$$\langle u(X) \rangle = Eu(X) = \overline{u(X)}. \quad (1.26)$$

Theorem 1.3 *If X and Y are two statistically independent RVs, then*

$$\langle u(X)v(Y) \rangle = \langle u(X) \rangle \langle v(Y) \rangle. \quad (1.27)$$

Proof:

$$\begin{aligned} \langle u(X)v(Y) \rangle &= \int_{\text{all } x} \int_{\text{all } y} u(x)v(y)f_{XY}(x,y)dx dy \\ &= \int_{\text{all } x} \int_{\text{all } y} u(x)v(y)f_X(x)f_Y(y)dx dy \quad (\text{independence}) \\ &= \int_{\text{all } x} u(x)f_X(x)dx \int_{\text{all } y} v(y)f_Y(y)dy \\ &= \langle u(X) \rangle \langle v(Y) \rangle. \end{aligned} \quad (1.28)$$

1.6 Mean and Variance

Definition 1.4 *The mean of a random variable is its expectation value.*

Definition 1.5 *The variance of a random variable x is the square of the standard deviation, and is the expectation value:*

$$\text{Var}(X) = \sigma_X^2 = \langle (X - \langle X \rangle)^2 \rangle \quad (1.29)$$

$$= \langle X^2 \rangle - \langle X \rangle^2. \quad (1.30)$$

The variance plays a special role in statistics because it provides a measure of the fluctuations that can be expected in samples from the distribution.

The variance generalizes in the multivariate case to the **Moment Matrix**, with elements:

$$M_{ij} = \langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle \quad (1.31)$$

$$= \langle X_i X_j \rangle - \langle X_i \rangle \langle X_j \rangle. \quad (1.32)$$

Notice that the diagonal elements are simply the individual variances. The off-diagonal elements are called **covariances**.

The **covariance coefficients**, measuring the degree of linear correlation, are given by:

$$\rho_{ij} = \frac{M_{ij}}{\sqrt{M_{ii}M_{jj}}}.$$

1.7 Some Important Probability Distributions

Name	Event space	$f(x)$	$\langle x \rangle$	$\sigma(x)$
Binomial	$\{0, 1, \dots, n\}$	$\binom{n}{x} \theta^x (1 - \theta)^{n-x}$	nq	$\sqrt{n\theta(1 - \theta)}$
Cauchy ⁽¹⁾	real numbers	$\frac{2}{\pi\Gamma} \frac{1}{1+4(x-\theta)^2/\Gamma^2}$	undefined	undefined
Chisquare	positive real	$\frac{e^{-x/2} x^{n/2-1}}{\Gamma(n/2) 2^{n/2}}$	n	$\sqrt{2n}$
Exponential	positive real	$\theta e^{-\theta x}$	$1/\theta$	$1/\theta$
Gamma ⁽²⁾	positive real	$\frac{\theta^n x^{n-1} e^{-\theta x}}{\Gamma(n)}$	n/θ	\sqrt{n}/θ
Normal ⁽³⁾	real numbers	$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$	θ	σ
Poisson	$\{0, 1, 2, \dots\}$	$\theta^x e^{-\theta}/x!$	θ	$\sqrt{\theta}$
Uniform	reals in $\{0, \theta\}$	$1/\theta$	$\theta/2$	$\theta/\sqrt{12}$

(1) Also known as Lorentz or Breit-Wigner.

(2) x = time to observe n events in a Poisson process.

(3) Also known as Gaussian.

1.8 Convergence

Consider a sequence of random variables, $\{X_1, X_2, \dots\}$. Such a sequence may approach a limiting value as the index approaches infinity. We say that the sequence “converges” on the limiting value. However, there are multiple notions of convergence. The most important for us are:

Definition 1.6 Let X_1, X_2, \dots be a sequence of random variables, distributed according to cumulative distribution functions F_1, F_2, \dots , respectively. Further, let X be a random variable distributed according to cumulative distribution function F . If

$$\lim_{n \rightarrow \infty} F_n(x) = F(x), \quad (1.33)$$

except possibly at discontinuities of F , then the sequence X_1, X_2, \dots is said to **converge in distribution** (or **weakly**) to X .

An important application of this notion of convergence appears in the central limit theorem (section 1.9).

Let us illustrate this notion of convergence with the relationship between two distributions. Suppose we are estimating the efficiency to detect events in an experiment. We start with a sample of n events (perhaps simulated according to the methods in chapter 2). We pass these events through our selection criteria, and count how many, k , pass. The sampling distribution for k , assuming the events are independently identically distributed, is the **binomial distribution**

(as the reader should verify):

$$P(k; n, p) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}, \quad (1.34)$$

where p is the efficiency we are trying to measure.

Let us investigate the behavior of this distribution as n becomes large, but $\theta \equiv p_n n$ given and finite, where we have written p_n for p to explicitly refer to its dependence on n in this limit. Thus, we consider the sequence of distributions:

$$F_n(k) = \sum_{j \leq k} \frac{n!}{k!(n-k)!} \left(\frac{\theta}{n}\right)^k \left(1 - \frac{\theta}{n}\right)^{n-k}. \quad (1.35)$$

Note that we are using the same symbol k to stand for a value of the random variable of each of these distributions. We want to see what happens as we let $n \rightarrow \infty$, for any given finite value of k :

$$\begin{aligned} \lim_{n \rightarrow \infty} F_n(k) &= \sum_{j \leq k} \lim_{n \rightarrow \infty} \frac{n!}{k!(n-k)!} \left(\frac{\theta}{n}\right)^k \left(1 - \frac{\theta}{n}\right)^{n-k} \\ &= \sum_{j \leq k} \frac{\theta^k}{k!} \lim_{n \rightarrow \infty} \frac{n(n-1) \cdots (n-k+1)}{n^k} \frac{\left(1 - \frac{\theta}{n}\right)^n}{\left(1 - \frac{\theta}{n}\right)^k} \\ &= \sum_{j \leq k} \frac{\theta^k}{k!} e^{-\theta}. \end{aligned} \quad (1.36)$$

This result is the Poisson distribution, which we have already seen in Section 1.2.

A stronger notion of convergence is the following:

Definition 1.7 Let $\epsilon > 0$ and let $P(|X_n - X| \geq \epsilon)$ be the probability that the difference between random variable X_n and a number X is greater than or equal to ϵ . If, for any given $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|X_n - X| \geq \epsilon) = 0, \quad (1.37)$$

then the sequence X_1, X_2, \dots is said to **converge in probability** to X .

It is important to understand what this limit means. We have a sequence of random variables, X_1, X_2, \dots , each distributed according to some distribution, say F_n for X_n . The definition of this limit says that the probability of sampling an X_n much different from X becomes vanishingly small as n becomes very large.

We may see that convergence in probability implies weak convergence. According to convergence in probability,

$$\lim_{n \rightarrow \infty} F_n(x) = F(x) = \begin{cases} 0 & x < X, \\ 1 & x > X. \end{cases} \quad (1.38)$$

Otherwise we would have a non-zero $P(|X_n - X| \geq \epsilon)$ for any given $\epsilon > 0$.

A still stronger notion of convergence is the following:

Definition 1.8 Let X_1, X_2, \dots be a sequence of random variables. If

$$P\left(\lim_{n \rightarrow \infty} X_n = X\right) = 1, \quad (1.39)$$

then the sequence X_n is said to converge **almost surely** (or **strongly**) to X .

Note that the reason for the “almost” in almost sure convergence is that we don’t eliminate convergence to something other than X , if it occurs with zero probability. We can drop the word “almost” if there are no such exceptions.

1.9 Characteristic Functions and the Central Limit Theorem

An important tool in probability theory is the **characteristic function** (or the related moment generating function), which is simply the Fourier transform of the PDF:

$$\phi_X(k) = \int_{-\infty}^{\infty} e^{ikx} f_X(x) dx \quad (1.40)$$

$$= \langle e^{ikX} \rangle. \quad (1.41)$$

The $f(x)dx$ is to be interpreted in general as the probability measure. That is, we define the characteristic function generally as:

$$\phi_X(k) \equiv \int_{-\infty}^{\infty} e^{ikx} dF_X(x). \quad (1.42)$$

For example, consider the Poisson distribution:

$$\begin{aligned} \phi_N(k) &= \int_{-\infty}^{\infty} e^{ikn} dF_N(n) \\ &= \sum_{n=0}^{\infty} e^{ikn} \frac{\theta^n}{n!} e^{-\theta} \\ &= e^{-\theta} \sum_{n=0}^{\infty} \frac{(e^{ik}\theta)^n}{n!} \\ &= e^{\theta(e^{ik}-1)}. \end{aligned} \quad (1.43)$$

For another example, consider the important (as we shall soon understand) normal distribution:

$$N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}}, \quad (1.44)$$

typically given the abbreviated label $N(\mu, \sigma)$. It is straightforward to demonstrate that the mean is μ and the variance is σ^2 . The characteristic function for

$N(0, 1)$ may be found by completing the square in the exponent:

$$\begin{aligned}\phi_X(k) &= \int_{-\infty}^{\infty} e^{ikx} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{k^2}{2}} \int_{-\infty}^{\infty} e^{-\frac{(x-ik)^2}{2}} dx \\ &= e^{-\frac{k^2}{2}} \quad \text{for } N(0, 1).\end{aligned}\tag{1.45}$$

The characteristic function for $N(\mu, \sigma)$ may readily be found from this to be:

$$\phi_X(k) = e^{ik\mu - \frac{\sigma^2 k^2}{2}} \quad \text{for } N(\mu, \sigma).\tag{1.46}$$

Note that for $\mu \neq 0$, the characteristic function is complex-valued, a result of the lack of symmetry of the PDF about $X = 0$.

A related notion is the **moment generating function**:

$$m_X(k) \equiv \int_{-\infty}^{\infty} e^{kx} f(x) dx = \langle e^{kX} \rangle.\tag{1.47}$$

However, while the characteristic function always exists, the moment generating function may not. The moments of the distribution are readily computed if the moment generating function is known, according to:

$$\langle X^\ell \rangle = \lim_{k \rightarrow 0} \left(\frac{d^\ell}{dk^\ell} e^{kX} \right)\tag{1.48}$$

$$= \lim_{k \rightarrow 0} \frac{d^\ell}{dk^\ell} m_X(k).\tag{1.49}$$

For example, the moment generating function for the Poisson distribution is

$$m_N(k) = e^{\theta(e^k - 1)},\tag{1.50}$$

as may be readily seen from our result for the characteristic function. The first moment, or mean, of the Poisson distribution is therefore

$$\begin{aligned}\langle N \rangle &= \lim_{k \rightarrow 0} \frac{d}{dk} e^{\theta(e^k - 1)} \\ &= \lim_{k \rightarrow 0} (\theta e^k) e^{\theta(e^k - 1)} \\ &= \theta.\end{aligned}\tag{1.51}$$

For the higher moments, it is often more interesting to know the “central moments”, $\langle (N - \langle N \rangle)^\ell \rangle$, defined relative to the mean.¹ The moment generating

¹The central moments are related to, but distinct from, another set of quantities called the “cumulants” of the distribution. The ℓ th cumulant is defined as the coefficient of $k^\ell/\ell!$ in the expansion of $\log m_X(k)$.

1.9. CHARACTERISTIC FUNCTIONS AND THE CENTRAL LIMIT THEOREM 13

function for the central moments of the Poisson distribution is readily found from the moment generating function:

$$m_{N-\theta}(k) = e^{\theta(e^k - 1 - k)}. \quad (1.52)$$

The ℓ th central moment is

$$\begin{aligned} \langle (N - \theta)^\ell \rangle &= \lim_{k \rightarrow 0} \frac{d^\ell}{dk^\ell} e^{\theta(e^k - 1 - k)} \\ &= \lim_{k \rightarrow 0} e^{\theta(e^k - 1 - k)} \theta (\ell e^k - 1) \\ &= (\ell - 1)\theta. \end{aligned} \quad (1.53)$$

Theorem 1.4 (Lévy-Cramér) *Let X, X_1, X_2, \dots be random variables with corresponding characteristic functions $\phi_X, \phi_{X_1}, \phi_{X_2}, \dots$. The sequence X_n converges in distribution to X iff $\lim_{n \rightarrow \infty} \phi_{X_n}(k) = \phi_X(k)$ for all k .*

We omit the proof of this theorem here; the reader is referred to texts on probability theory. It is at least plausible from our understanding of Fourier transforms that the mapping between probability distributions and characteristic functions should in some sense be continuous and invertible.

We are ready now for the celebrated:

Theorem 1.5 (Central Limit Theorem) *Let (X_1, X_2, \dots, X_n) be a sequence of IID random variables from a distribution, F , having finite mean μ and finite variance σ^2 . Then, if $S_n/n = \frac{1}{n} \sum_{i=1}^n X_i$ is the **sample mean**, the distribution of S_n/n approaches the normal distribution as $n \rightarrow \infty$, with mean $\langle S_n/n \rangle = \mu$ and variance $\langle (S_n/n - \langle S_n/n \rangle)^2 \rangle = \sigma^2/n$.*

Proof: Remark: Proofs found on popular web sites often implicitly assume that the higher moments of F exist. We wish to avoid this assumption, hence will work a bit harder.

For simplicity, we'll prove the theorem for the special case $\mu = 0$ and $\sigma = 1$. The general case follows readily from this. We'll also consider explicitly the case of a continuous distribution. Our approach is to show that the characteristic function for S_n/\sqrt{n} approaches the $N(0, 1)$ characteristic function as $n \rightarrow \infty$ (noting that the multiplication by \sqrt{n} should give a unit variance according to the theorem). Then the theorem follows from the Lévy-Cramér theorem. Let $\phi_X(k)$ be the characteristic function for F . The characteristic function for S_n/\sqrt{n} is:

$$\begin{aligned} \phi_{S_n/\sqrt{n}} &\equiv \langle e^{ikS_n/\sqrt{n}} \rangle = \langle e^{ik(X_1 + \dots + X_n)/\sqrt{n}} \rangle \\ &= [\phi_X(k/\sqrt{n})]^n, \end{aligned} \quad (1.54)$$

since the X_i are IID. We wish to compare this, in the limit $n \rightarrow \infty$, with the characteristic function for $N(0, 1)$, that is, with $e^{-k^2/2}$.

There are some facts that we will use, which we gather here:

1. The magnitude of the characteristic function is bounded by 1:

$$\begin{aligned} |\phi_X(k)| &= \left| \int_{-\infty}^{\infty} e^{ikx} f(x) dx \right| \\ &\leq \int_{-\infty}^{\infty} |e^{ikx}| f(x) dx \\ &\leq 1. \end{aligned} \tag{1.55}$$

2. For real numbers $-1 \leq a \leq 1$ and $-1 \leq b \leq 1$, and integer n :

$$\begin{aligned} |a^n - b^n| &= \left| (a - b) \sum_{m=0}^{n-1} a^m b^{n-1-m} \right| \\ &\leq n |a - b|. \end{aligned} \tag{1.56}$$

3. If $x \geq 0$, then $e^{-x} \leq 1 - x + x^2/2$, as may be demonstrated by first showing that $e^{-x} \geq 1 - x$, and using the result to show that $e^{-x} \leq 1 - x + x^2/2$. The idea is to note the inequality for small values of x , then compare derivatives of the two sides for all values to demonstrate that the functions never cross.

4. It may similarly be demonstrated that, for any real x :

$$(a) \quad |e^{ix} - 1 - ix| \leq \frac{x^2}{2}. \tag{1.57}$$

$$(b) \quad \left| e^{ix} - 1 - ix - \frac{(ix)^2}{2} \right| \leq \left| \frac{x^3}{6} \right|. \tag{1.58}$$

Let us now make our comparison:

$$\begin{aligned} \left| \phi_X(k/\sqrt{n})^n - e^{-k^2/2} \right| &= \left| \phi_X(k/\sqrt{n})^n - (e^{-k^2/2n})^n \right| \\ &\leq n \left| \phi_X(k/\sqrt{n}) - e^{-k^2/2n} \right| \quad \text{facts (1) and (2)} \\ &\leq n \left[\left| \phi_X(k/\sqrt{n}) - (1 - k^2/2n) \right| + \left| (1 - k^2/2n) - e^{-k^2/2n} \right| \right] \quad \text{triangle inequality} \\ &\leq n \left| \phi_X(k/\sqrt{n}) - (1 - k^2/2n) \right| + \frac{k^4}{8n} \quad \text{fact (3)}. \end{aligned} \tag{1.59}$$

The last term approaches zero as $n \rightarrow \infty$, for any given k .

It remains to demonstrate that the first term also approaches zero. The presence of the overall factor of n complicates things a bit. We'll use both parts of fact (4) to break things into appropriate pieces. We have:

$$\begin{aligned} n \left| \phi_X(k/\sqrt{n}) - (1 - k^2/2n) \right| &= n \left| \langle e^{ikX/\sqrt{n}} - (1 - ikX/\sqrt{n} + (ikX)^2/2n) \rangle \right|, \\ &\leq n \left| \langle e^{ikX/\sqrt{n}} - (1 - ikX/\sqrt{n} + (ikX)^2/2n) \rangle \right|. \end{aligned}$$

1.9. CHARACTERISTIC FUNCTIONS AND THE CENTRAL LIMIT THEOREM 15

The insertion of the ikX term on the first line is harmless since $\langle X \rangle = 0$, and the insertion of X^2 is also harmless, because $\langle X^2 \rangle = 1$. Using the triangle inequality and fact 4a we find

$$\begin{aligned} \left| e^{ikX/\sqrt{n}} - (1 - ikX/\sqrt{n} + (ikX)^2/2n) \right| &\leq \left| e^{ikX/\sqrt{n}} - (1 - ikX/\sqrt{n}) \right| + (kX)^2/2n \\ &\leq (kX)^2/n. \end{aligned} \quad (1.60)$$

Using instead fact 4b, we find

$$\left| e^{ikX/\sqrt{n}} - (1 - ikX/\sqrt{n} + (ikX)^2/2n) \right| \leq |kX|^3/6n^{3/2}. \quad (1.61)$$

We want n times the expectation value of this.

Let $\delta > 0$ and let R_δ be the set $\{|X| < \delta\sqrt{n}\}$. Let I_S be the “**indicator function**” for set S . That is,

$$I_{R_\delta}(X) \equiv \begin{cases} 1 & \text{if } X \in R_\delta \\ 0 & \text{otherwise.} \end{cases} \quad (1.62)$$

Likewise, the indicator function for the complement is

$$I_{\tilde{R}_\delta}(X) \equiv \begin{cases} 1 & \text{if } X \notin R_\delta \\ 0 & \text{otherwise.} \end{cases} \quad (1.63)$$

Then we may write

$$n \left\langle \left| e^{ikX/\sqrt{n}} - (1 - ikX/\sqrt{n} + (ikX)^2/2n) \right| \right\rangle \leq \langle (kX)^2 I_{\tilde{R}_\delta}(X) \rangle + \left\langle \frac{|kX|^3}{6\sqrt{n}} I_{R_\delta}(X) \right\rangle. \quad (1.64)$$

Consider first the last term:

$$\begin{aligned} \left\langle \frac{|kX|^3}{6\sqrt{n}} I_{R_\delta}(X) \right\rangle &= \frac{|k|^3}{6\sqrt{n}} \int_{R_\delta} |x|^3 f(x) dx \\ &\leq \frac{|k|^3}{6\sqrt{n}} \int_{R_\delta} \delta\sqrt{n} x^2 f(x) dx \\ &\leq \frac{|k|^3}{6} \delta, \end{aligned} \quad (1.65)$$

since $\langle X^2 \rangle = 1$. Thus, this term can be made arbitrarily small. For any given ϵ , we choose $\delta > 0$ such that

$$\frac{|k|^3 \delta}{6} \leq \epsilon/2. \quad (1.66)$$

Note that our choice for δ is independent of n .

The other term is

$$\langle (kX)^2 I_{\tilde{R}_\delta}(X) \rangle = k^2 \int_{\tilde{R}_\delta} x^2 f(x) dx \quad (1.67)$$

For our given ϵ and above choice of δ , we chose integer N such that whenever $n \geq N$:

$$k^2 \int_{\tilde{R}_\delta} x^2 f(x) dx = k^2 \left(1 - \int_{x \leq \delta\sqrt{n}} x^2 f(x) dx \right) \leq \epsilon/2. \quad (1.68)$$

This is possible, since the integrand is nowhere negative, and hence the integral on the right approaches $\langle X^2 \rangle = 1$. This concludes our proof.

It may be remarked that we have given the simplest form of the central limit theorem. For example, with additional conditions, it may be generalized to the case where (X_1, X_2, \dots, X_n) are not IID.

1.10 Chebyshev's Inequality

Chebyshev's inequality gives us an amusing bound on the probability of sampling in the tails of a distribution. If X is a random variable with mean $\langle X \rangle = \mu$ and standard deviation σ , then

$$P(|X - \mu| \geq n\sigma) \leq \frac{1}{n^2}. \quad (1.69)$$

To see why this is true, suppose that F_X corresponds to a continuous distribution with PDF $f_X(x)$. We have

$$P(|X - \mu| \geq n\sigma) = \int_{-\infty}^{\mu - n\sigma} f_X(x) dx + \int_{\mu + n\sigma}^{\infty} f_X(x) dx \quad (1.70)$$

$$\leq \frac{1}{(n\sigma)^2} \left[\int_{-\infty}^{\mu - n\sigma} (x - \mu)^2 f_X(x) dx + \int_{\mu + n\sigma}^{\infty} (x - \mu)^2 f_X(x) dx \right] \quad (1.71)$$

$$\leq \frac{1}{(n\sigma)^2} \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx \quad (1.72)$$

$$\leq \frac{1}{n^2}. \quad (1.73)$$

The case of a discrete distribution is left as an exercise.

1.11 Laws of Large Numbers

The weight of experience provides the intuition that the sample mean of n samples from a distribution will more and more likely be near to the mean as the sample size is increased. This notion is embodied in two "Laws of Large Numbers". The "weak" version tells us that the probability that the sample mean will differ from the mean by an amount greater than any given number approaches zero for large n :

Theorem 1.6 (Weak Law of Large Numbers) *Let (X_1, X_2, \dots, X_n) be a sequence of IID random variables from distribution F , having finite mean μ and*

finite variance σ^2 . Let $M = \frac{1}{n} \sum_{i=1}^n X_i$ be the sample mean. Then, given any $\epsilon > 0$

$$\lim_{n \rightarrow \infty} P(|M - \mu| > \epsilon) = 0. \quad (1.74)$$

Proof: Remark: The distributions actually need not be identical, as long as they have the same mean and variance.

The expectation value of the sample mean is:

$$\langle M \rangle = \left[\prod_{i=1}^n \int_{-\infty}^{\infty} dF(x_i) \right] \frac{1}{n} \sum_{i=1}^n x_i \quad (1.75)$$

$$= \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{\infty} dF(x_i) x_i \quad (1.76)$$

$$= \mu. \quad (1.77)$$

Likewise it can be shown that the variance of the sample mean is σ^2/n . If we make the substitution $\epsilon = n\sigma$ in Chebyshev's inequality we have:

$$P(|X - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2}. \quad (1.78)$$

Applying this to M gives:

$$P(|M - \mu| > \epsilon) \leq \frac{\sigma^2}{n^2 \epsilon^2}. \quad (1.79)$$

For any given $\epsilon > 0$ this probability may be made arbitrarily small by taking n large enough.

The **Strong Law of Large Numbers** looks similar, stating (under the same conditions) that:

$$P\left(\lim_{n \rightarrow \infty} |M - \mu| = 0\right) = 1. \quad (1.80)$$

However, it is not the same. The weak law is a statement of convergence in probability, i.e., that the probability of a large fluctuation away from the mean approaches zero for large sample size. The strong law is a statement of almost sure convergence, that in the limit $n \rightarrow \infty$ the sample mean will precisely equal μ , except on a set of measure zero. The set of measure zero exists because, for example, it is possible that every sample will yield a value bigger than μ , but with vanishing probability as $n \rightarrow \infty$. The strong law implies the weak law, because almost sure convergence implies convergence in probability. For a proof of the strong law, the reader is referred to texts on probability theory.

1.12 The Exponential Family

An important class of probability distributions is the "Exponential Family":

$$f_X(x; \theta) = \exp[A(\theta)D(x) + B(\theta) + C(x)], \quad (1.81)$$

where X labels a random variable (or a vector of random variables, $X = (X_1, X_2, \dots, X_n)$), and θ labels a parameter of the distribution. In the vector case, the mappings C, D are to real numbers. More generally, we could have a vector of parameters, $\theta = (\theta_1, \theta_2, \dots, \theta_r)$, in which case the exponential family takes the form:

$$f_X(x; \theta) = \exp \left[\sum_{i=1}^k A_i(\theta) D_i(x) + B(\theta) + C(x) \right]. \quad (1.82)$$

Here, functions A_i, B, C , and D_i all map vectors to real numbers.

1.13 Transformations

We are often interested in the probability distribution, g , for random variables $Y = (Y_1, Y_2, \dots, Y_n) = h(X)$, given the probability distribution, f , for the (perhaps measured) random variables $X = (X_1, X_2, \dots, X_n)$. If the Y_i 's are linearly independent, the new pdf for Y is simply found by:

$$g(y) d^n(y) = g[h(x)] \left| \frac{\partial h}{\partial x} \right| d^n(x) \quad (1.83)$$

$$= f(x) d^n(x), \quad (1.84)$$

where $\left| \frac{\partial h}{\partial x} \right|$ is the absolute value of the Jacobian determinant. Hence,

$$g(y) = \frac{f[h^{-1}(y)]}{\left| \frac{\partial h}{\partial x} \right| [h^{-1}(y)]} \quad (1.85)$$

Rather than determining the entire transformation, we are often content to learn the new moment matrix.

1.14 Propagation of Errors

If $y = (y_1, y_2, \dots, y_k)$ is linearly dependent on $x = (x_1, x_2, \dots, x_n)$, *i.e.*,

$$y = Tx + a, \quad (1.86)$$

where T is a $k \times n$ transformation matrix, then it is easily shown that the moment matrix for y is given by:

$$M_y = TM_x T^\dagger. \quad (1.87)$$

If y is non-linearly dependent on x , we often make the linear approximation anyway, letting

$$T_{ij} = \left. \frac{\partial y_i}{\partial x_j} \right|_{x \sim \langle x \rangle}.$$

It should be kept in mind though, that this corresponds to taking the first term in a Taylor series expansion, and may not be a good approximation for some transformations, or far away from $\langle x \rangle$.

1.15 Propagation of Errors - Special Case

Example: Suppose $k = 1$. Then, in the linear approximation:

$$M_y = \sigma_y^2 = TM_x T^\dagger \quad (1.88)$$

$$= \sum_{i=1}^n \sum_{j=1}^n \frac{\partial y}{\partial x_i} \Big|_{\mathbf{x} \sim \langle x \rangle} \frac{\partial y}{\partial x_j} \Big|_{\mathbf{x} \sim \langle x \rangle} (M_x)_{ij}. \quad (1.89)$$

If the x_i 's are statistically independent, then

$$(M_x)_{ij} = \sigma_{x_i}^2 \delta_{ij}, \quad (1.90)$$

and hence,

$$\sigma_y^2 = \sum_{i=1}^n \left(\frac{\partial y}{\partial x_i} \Big|_{x \sim \langle x \rangle} \sigma_{x_i} \right)^2. \quad (1.91)$$

This is our most commonly-used form for propagating errors. Just remember the assumptions of linearity and independence, as well as the typically approximate knowledge of $\langle x \rangle$!

1.16 Exercises

1. Prove that $P(\emptyset) = 0$.
2. Prove the Rule of Complementation:

$$P(\cap_{i=1}^n \widetilde{E}_i) = 1 - P(\cup_{i=1}^n E_i).$$

3. Arbitrary union Prove:

$$P(\cup_{i=1}^n E_i) = \sum_{i=1}^n P(E_i) - \sum_{j>i}^n P(E_i \cap E_j) \quad (1.92)$$

$$+ \sum_{k>j>i}^n P(E_i \cap E_j \cap E_k) \quad (1.93)$$

$$\dots \quad (1.94)$$

$$+ (-)^{n-1} P(E_1 \cap E_2 \dots \cap E_n). \quad (1.95)$$

4. Derive the Poisson distribution for radioactive decays. That is, given an average rate for decays, Γ , and an interval of time T , what is the probability to observe n decays? You may neglect any depletion in the radioactive source.
5. There are two radioactive sources. One emits gamma rays in 75% of the decays, and beta rays in the other 25%. The other emits gammas 1/3 of the time, and betas 2/3. A source is chosen at random, and the first decay observed is a gamma. What is the probability that a gamma will be observed on the second decay? How about the 83rd decay?

6. Consider a collision process which produces charged particles. The charged particles are tracked in a tracking device, such as a drift chamber. However, the track detection is not completely efficient. What is the probability of losing at least one track in a six track event, if the single-track inefficiency is 5%? State any assumptions.
7. Prove the Chebyshev inequality for a discrete distribution.
8. Generation of Normal Distribution: Suppose you have a source of uniformly distributed random numbers on the interval $(0, 1)$. Using the Central Limit Theorem, give an algorithm to produce random numbers which are approximately normally distributed, with mean 0 and variance 1.
9. If $p(x) = \frac{1}{\theta}e^{-x/\theta}$, what is the PDF for $y = 1/x$?
10. **Puzzle:** There are three boxes. One of them contains a million dollars and the other two are empty. Select one of them at random, but don't open it. Of the two non-selected boxes, you are then told a box which does not contain the money. To optimize your chances of becoming rich, should you change your original selection?
11. Relevant to our discussion of the central limit theorem, invent a probability distribution for which the mean and variance are finite, but not the higher moments. [The "relevance" to the central limit theorem comes in the fact that the theorem applies to such a distribution. This fact is often not recognized in popular proofs of the theorem, which assume the existence of all moments.]
12. Derive Eqn. 1.87.