APPENDIX B

PROBABILITY REVIEW

This review of the basics of probability theory focuses on random variables and their distributions. Few results are proven and we refer the reader to a standard undergraduate probability textbook for more complete results. Conditional probability and conditional distribution are discussed in Chapter 1.

Probability begins with a *random experiment*, which is loosely defined as an experiment for which the outcome is uncertain. Given such an experiment, the *sample space* Ω is the set of all possible outcomes. Individual outcomes, that is, the elements of the sample space, are denoted by ω. An *event* A is a subset of the sample space. Say that A occurs if the outcome of the experiment is contained in A.

A probability P is a function that assigns to each event a number between 0 and 1 in such a way that the following conditions are satisfied:

1. $0 \leq P(A) \leq 1$, for all $A \subseteq \Omega$.
2. $P(\Omega) = 1$.
3. Given a sequence of disjoint events $A_1, A_2, \ldots,$

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

We interpret $P(A)$ to mean *the probability that A occurs*.

A *random variable* is a real-valued function defined on a sample space. That is, the outcomes of a random variable are determined by a random experiment. For instance,

Introduction to Stochastic Processes with R, First Edition. Robert P. Dobrow.

assume that three coins are flipped. Let X be the number of heads. Then, X is a random variable that takes values 0, 1, 2, or 3, depending on the outcome of the coin flips.

Write $P(X = x)$ for the probability that X takes the value x, and $P(X \leq x)$ for the probability that X takes a value less than or equal to x. More generally, for $R \subseteq \mathbb{R}$, write $P(X \in R)$ for the probability that X takes a value that is contained in R. The notation $\{X \in R\}$ is shorthand for $\{\omega : X(\omega) \in R\}$, which is the set of all outcomes ω with the property that $X(\omega)$ is contained in R.

The *distribution* of a random variable X describes the set of values of X and their corresponding probabilities.

The function $F(x) = P(X \leq x)$ is the *cumulative distribution function (cdf) of X*. The cdf takes values between 0 and 1 and is defined for all real numbers. The cdf gives complete probabilistic information about a random variable in the sense that knowing the cdf is equivalent to knowing the distribution of the random variable.

Example B.1 For the random experiment of flipping a fair coin three times, let X denote the number of heads that occur. Letting H denote heads, T denote tails, and keeping track of the order of coin flips, the sample space is

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

If all outcomes are equally likely, then each of the eight outcomes occurs with probability $1/8$. This gives

$$P(X = x) = \begin{cases} P(\{TTT\}) & = 1/8, \quad \text{if } x = 0, \\ P(\{HTT, THT, TTH\}) & = 3/8, \quad \text{if } x = 1, \\ P(\{HHT, HTH, THH\}) & = 3/8, \quad \text{if } x = 2, \\ P(\{HHH\}) & = 1/8, \quad \text{if } x = 3, \end{cases}$$

and cdf

$$F(x) = P(X \leq x) = \begin{cases} 0, & \text{if } x < 0, \\ 1/8, & \text{if } 0 \leq x < 1, \\ 4/8, & \text{if } 1 \leq x < 2, \\ 7/8, & \text{if } 2 \leq x < 3, \\ 1, & \text{if } x \geq 3. \end{cases}$$

■

B.1 DISCRETE RANDOM VARIABLES

A random variable that takes values in a finite or countably infinite set is called a *discrete random variable*. As a function of x, the function $P(X = x)$ is the *probability mass function (pmf) of X*. The pmf describes the distribution of a discrete random variable. For $R \subseteq \mathbb{R}$,

$$P(X \in R) = \sum_{x \in R} P(X = x).$$

The *expectation*, or mean, of a discrete random variable X is defined as

$$E(X) = \sum_x xP(X = x).$$

The expectation is a *weighted average* of the values of X, with weights given by the pmf. Intuitively, the expectation of X is the long-run average value of X over repeated trials.

If g is a function and X is a random variable, then $Y = g(X)$ is a *function of a random variable*, which itself is a random variable that takes the value $g(x)$ whenever X takes the value x. A useful formula for computing the expectation of a function of a random variable is

$$E(Y) = E(g(X)) = \sum_x g(x)P(X = x). \tag{B.1}$$

The expectation is also computed as $E(Y) = \sum_y yP(Y = y)$, which requires knowledge of the distribution of Y.

Example B.2 The radius R of a circle is a random variable that takes values 1, 2, 4, and 8 with respective probabilities 0.1, 0.2, 0.3, and 0.4. Find the expected area of the circle.

Solution Let Y be the area of the circle. The expected area is

$$\begin{aligned} E(Y) &= E\left(\pi R^2\right) \\ &= \sum_{r=1}^{4} \pi r^2 P(R = r) \\ &= \pi\left(1(0.1) + 4(0.2) + 16(0.3) + 64(0.4)\right) \\ &= 31.3\pi. \end{aligned}$$

■

The expectation of a linear function of a random variable is a linear function of the expectation. From Equation (B.1), for constants a and b,

$$\begin{aligned} E(aX + b) &= \sum_x (ax + b)P(X = x) \\ &= a\sum_x xP(X = x) + b\sum_x P(X = x) \\ &= aE(X) + b, \end{aligned}$$

The *variance* of a random variable is a measure of variability or discrepancy from the mean. It is defined as

$$Var(X) = E\left((X - E(X))^2\right) = \sum_x (x - E(X))^2 P(X = x).$$

A computationally useful formula is $Var(X) = E\left(X^2\right) - (E(X))^2$. For constants a and b,

$$\begin{aligned} Var(aX + b) &= E\left(((aX + b) - (aE(X) + b))^2\right) \\ &= E\left(a^2(X - E(X))^2\right) \\ &= a^2 Var(X). \end{aligned}$$

The *standard deviation* of a random variable is defined as $SD(X) = \sqrt{Var(X)}$.

B.2 JOINT DISTRIBUTION

The *joint probability mass function* of X and Y is $P(X = x, Y = y)$, which is a function of x and y. The *joint cumulative distribution function* is

$$F(x, y) = P(X \leq x, Y \leq y) = \sum_{i \leq x} \sum_{j \leq y} P(X = i, Y = j).$$

From the joint pmf of X and Y one can obtain the individual, or marginal, distributions of each random variable. For instance,

$$P(X = x) = P(X = x, -\infty < Y < \infty) = \sum_y P(X = x, Y = y).$$

Similarly, $P(Y = y) = \sum_x P(X = x, Y = y)$.

The *covariance* is a measure of linear association between two random variables. It is defined as

$$\text{Cov}(X, Y) = E\left((X - E(X))(Y - E(Y))\right) = E(XY) - E(X)E(Y).$$

The *correlation* between X and Y is

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{SD(X)\, SD(Y)}.$$

The correlation satisfies $-1 \leq \text{Corr}(X, Y) \leq 1$ and is equal to ± 1 if one random variable is a linear function of the other.

If $g(x, y)$ is a function of two variables, and X and Y are random variables, then $g(X, Y)$ is a random variable whose expectation is

$$E(g(X, Y)) = \sum_x \sum_y g(x, y) P(X = x, Y = y).$$

In the case when $g(x, y) = x + y$,

$$\begin{aligned} E(X+Y) &= \sum_x \sum_y (x+y) P(X=x, Y=y) \\ &= \sum_x x \sum_y P(X=x, Y=y) + \sum_y y \sum_x P(X=x, Y=y) \\ &= \sum_x x P(X=x) + \sum_y y P(Y=y) \\ &= E(X) + E(Y), \end{aligned}$$

which gives the important *linearity* property of expectation. For random variables $X_1, \ldots, X_n$,

$$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n).$$

Example B.3 The solution of the classic *matching problem* is an elegant application of the linearity of expectation.

At the baseball stadium a group of n people wearing baseball hats all throw their hats into the air when their favorite player hits a home run. If the hats are mixed up at random when they fall to the ground, and each person picks up one hat, how many people, on average, will get their own hat back?

Solution Let $I_1, \ldots, I_n$ be a sequence of random variables where

$$I_k = \begin{cases} 1, & \text{if the } k\text{th person gets their hat back,} \\ 0, & \text{otherwise,} \end{cases}$$

for $k = 1, \ldots, n$. Then, $X = I_1 + \cdots + I_n$ is the number of people who get their hat back. For each k,

$$\begin{aligned} E(I_k) &= (1)P(k\text{th person gets their hat}) \\ &\quad + (0)P(k\text{th person does not get their hat}) \\ &= P(k\text{th person gets their hat}) \\ &= \frac{1}{n}, \end{aligned}$$

since there are n hats to choose from and exactly one belongs to the kth person. By linearity of expectation,

$$E(X) = E(I_1 + \cdots + I_n) = E(I_1) + \cdots + E(I_n) = \sum_{k=1}^{n} \frac{1}{n} = 1.$$

On average, one person gets their hat back. Remarkably, the solution does not depend on n. ■

For the variance of a sum of random variables,

$$Var(X+Y) = Var(X) + Var(Y) + 2\,\mathrm{Cov}(X,Y).$$

More generally,

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j).$$

Events A and B are *independent* if $P(A \cap B) = P(A)P(B)$. Intuitively, events are independent if knowledge of whether or not one occurs has no influence on the probability of whether or not the other occurs.

We say that discrete random variables X and Y are *independent* if

$$P(X = x, Y = y) = P(X = x)P(Y = y) \text{ for all } x, y.$$

Equivalently,

$$P(X \in R, Y \in S) = P(X \in R)P(Y \in S) \text{ for all } R, S \subseteq \mathbb{R}.$$

If X and Y are independent random variables, then

$$\begin{aligned} E(XY) &= \sum_x \sum_y xyP(X = x, Y = y) \\ &= \sum_x \sum_y xyP(X = x)P(Y = y) \\ &= \sum_x xP(X = x) \sum_y yP(Y = y) = E(X)E(Y), \end{aligned}$$

and thus $\mathrm{Cov}(X, Y) = 0$. Hence, for independent random variables, $Var(X + Y) = Var(X) + Var(Y)$.

Sequences of independent and identically distributed (i.i.d.) random variables are common models. For instance, an infinite sequence of fair coin flips can be modeled as an independent sequence $X_1, X_2, \ldots$, where for each k, $X_k = 1$, if the kth flip is heads, and 0, if the kth flip is tails. The random variable $S_n = X_1 + \cdots + X_n$ is the number of heads in the first n coin flips.

In statistics, one often models a simple random sample as an i.i.d. sequence of random variables from a common population.

B.3 CONTINUOUS RANDOM VARIABLES

A continuous random variable takes values in an uncountable set, most commonly $\mathbb{R}$, $(0, \infty)$ or (a, b), with $a < b$. For continuous random variables $P(X = x) = 0$ for all x, and probabilities are computed by integrating the *probability density function*. The density function plays a role analogous to the pmf for discrete variables for computing probabilities.

A function f is a probability density function of X if

1. $f(x) \geq 0$, for all x
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$,
3. For all $R \subseteq \mathbb{R}$, $P(X \in R) = \int_R f(x)\, dx$.

The cumulative distribution function of X is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt.$$

Differentiating with respect to x gives

$$\frac{d}{dx} F(x) = f(x).$$

That is, the density is the derivative of the cdf.

Example B.4 Let X be a continuous random variable with density function

$$f(x) = cx^2, \text{ for } 0 < x < 3.$$

(i) Find the constant c. (ii) Find $P(1 < X < 2)$. (iii) Find the density function of $Y = X^2$.

Solution

(i) To find c, solve

$$1 = \int_{-\infty}^{\infty} f(x)\, dx = \int_0^3 cx^2\, dx = 9c,$$

which gives $c = 1/9$.

(ii) The desired probability is

$$P(1 < X < 2) = \int_1^2 f(x)\, dx = \int_1^2 \frac{x^2}{9}\, dx = \frac{7}{27}.$$

(iii) To find the density of Y first find the cdf of Y and then differentiate. Since X takes values between 0 and 3, $Y = X^2$ takes values between 0 and 9. For $0 < y < 9$,

$$P(Y \leq y) = P\left(X^2 \leq y\right) = P\left(X \leq \sqrt{y}\right).$$

Taking the derivative with respect to y and applying the chain rule gives

$$f_Y(y) = \frac{d}{dy} P\left(X \leq \sqrt{y}\right) = \frac{1}{2\sqrt{y}} f_X\left(\sqrt{y}\right) = \frac{1}{2\sqrt{y}} \frac{y}{9} = \frac{\sqrt{y}}{18},$$

for $0 < y < 9$. ■

Formulas for expectation and variance are analogous to the discrete formulas with density function replacing pmf and integrals replacing sums. Thus, for a continuous random variable X with density function f,

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx \text{ and } Var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)\,dx.$$

If g is a function, then $E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)\,dx$.

For continuous random variables, the joint density of X and Y is the function $f(x, y)$ that satisfies

$$P((X, Y) \in R) = \iint_R f(x, y)\,dx\,dy, \text{ for all } R \subseteq \mathbb{R}^2.$$

The joint cdf of X and Y is the function

$$F(x, y) = P(X \le x, Y \le y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s, t)\,dt\,ds.$$

If X and Y are independent, then $P(X \le x, Y \le y) = P(X \le x)P(Y \le y)$ for all x and y. Equivalently, the joint density function factors into the product of the marginal densities. That is,

$$f(x, y) = f_X(x)\, f_Y(y), \text{ for all } x, y.$$

B.4 COMMON PROBABILITY DISTRIBUTIONS

Uniform (Discrete) Distribution

The simplest model for a random variable X taking values in a finite set is that all outcomes are equally likely. We say that X has a *uniform distribution*. Historically, probability began by considering equally likely outcomes, mostly in games of chance.

For a finite set R, let $|R|$ denote the number of elements of R. If X is uniformly distributed on $S = \{s_1, \ldots, s_k\}$, then

$$P(X = s_i) = \frac{1}{k}, \text{ for } i = 1, \ldots, k$$

and

$$P(X \in R) = \frac{|R|}{|S|} = \frac{|R|}{k}, \text{ for } R \subseteq S.$$

In the discrete uniform case, probability reduces to counting. The probability that X is contained in R is the number of elements of R divided by the number of elements of the sample space.

Mean and variance are

$$E(X) = \frac{s_1 + \cdots + s_k}{k} \text{ and } Var(X) = \left(\frac{s_1^2 + \cdots + s_k^2}{k}\right) - \left(\frac{s_1 + \cdots + s_k}{k}\right)^2.$$

For the case $S = \{1, \ldots, k\}$, this gives

$$E(X) = \frac{k+1}{2} \text{ and } Var(X) = \frac{k^2 - 1}{12}.$$

Bernoulli Distribution

A *Bernoulli* random variables takes values 1 and 0, with probabilities p and $1 - p$, respectively. It is common to refer to the dichotomous values of a Bernoulli variable as *success* and *failure*.

If X has a Bernoulli distribution with parameter $0 < p < 1$, then

$$E(X) = p \text{ and } Var(X) = p(1-p).$$

Binomial Distribution

Assume that $X_1, \ldots, X_n$ is an i.i.d. sequence of Bernoulli random variables with common parameter p, where each X_i represents success or failure on the ith trial. Let $X = X_1 + \cdots + X_n$. Then, X counts the number of successes in n trials, and has a *binomial distribution with parameters n and p*.

The pmf of the binomial distribution is derived by a counting argument using the fact that the event $\{X = k\}$ can be expressed as the set of all 0-1 sequences of length n with exactly k 1s. The number of such sequences is counted by the binomial coefficient $\binom{n}{k}$. It follows that

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \text{ for } k = 0, 1, \ldots, n,$$

with

$$E(X) = np \text{ and } Var(X) = np(1-p).$$

Example B.5 Whether or not tomato seeds germinate in Angel's garden is modeled as independent Bernoulli random variables with germination (success) probability $p = 0.8$. If 100 seeds are planted, find the probability that at least 75 germinate.

Solution Let X be the number of seeds that germinate. Then, X has a binomial distribution with parameters $n = 100$ and $p = 0.8$. The desired probability is

$$P(X \geq 75) = \sum_{k=75}^{100} \binom{100}{k} (0.8)^k (0.2)^{100-k} = 0.9125.$$

Since $P(X \geq 75) = 1 - P(X < 75) = 1 - P(X \leq 74)$, the probability is obtained in R by typing

```
> 1-pbinom(74,100,0.8)
[1] 0.9125246
```

■

Geometric Distribution

Given a sequence $X_1, X_2, \ldots$ of i.i.d. Bernoulli variables with parameter p, with X_i representing success or failure on the ith trial, let N be the index of the first trial in which success occurs. That is $\{N = k\}$ if and only if $X_i = 0$ for all $i < k$ and $X_k = 1$. Then, N has a *geometric distribution with parameter p* and

$$P(N = k) = (1 - p)^{k-1} p, \text{ for } k = 1, 2, \ldots,$$

with

$$E(N) = \frac{1}{p} \text{ and } Var(N) = \frac{1 - p}{p^2}.$$

Example B.6 In Texas hold 'em poker, players are initially dealt two cards. If the deck is reshuffled after each play, find the expected number of deals until a player gets at least one ace.

Solution The outcome of each deal is modeled as a Bernoulli variable with parameter

$$\begin{aligned} p &= P(\text{Player gets at least one ace}) \\ &= 1 - P(\text{Player gets no aces}) \\ &= 1 - \left(\frac{48}{52}\right)\left(\frac{47}{51}\right) = \frac{33}{221} = 0.1493. \end{aligned}$$

The number of deals required for a player to get at least one ace has a geometric distribution with parameter p. The expected number of required deals is $1/p = 221/33 = 6.697$. ■

Poisson Distribution

The Poisson distribution arises as a model for counts of independent events that occur in some fixed region of time or space. Examples include the number of traffic accidents along a stretch of highway, the number of births on a hospital ward, and the number of wrong numbers to your cell phone. The distribution is sometimes called the *law of rare events* and arises when the chance that some event occurs in a small interval of time or space is small.

The distribution depends on a parameter $\lambda > 0$, which can be interpreted as the *rate* of occurrence in a unit interval. A random variable X has a *Poisson distribution with parameter* λ, if

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \text{ for } k = 0, 1, \ldots$$

The distribution has the property that mean and variance are both equal to λ. For the mean,

$$E(X) = \sum_{k=0}^{\infty} kP(X = k) = \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-1)\ !}$$
$$= e^{-\lambda}\lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)\ !} = e^{-\lambda}\lambda \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda}\lambda e^{\lambda} = \lambda.$$

Example B.7 During the peak month of May, tornados hit Oklahoma at the rate of about 21.7 per month, according to the National Weather Service. Find the probability that there will be fewer than 15 tornados next May.

Solution If it is assumed that successive tornado hits are independent events, then the Poisson distribution is a reasonable model. Let X be the number of tornados which hit Oklahoma next May. Then,

$$P(X < 15) = P(X \leq 14) = \sum_{k=0}^{14} \frac{e^{-21.7}(21.7)^k}{k!} = 0.054.$$

In R, type

```
> ppois(14,21.7)
[1] 0.05400056
```

■

The Poisson distribution is closely related to the binomial distribution and arises as a limiting distribution when the number of trials is large and the success probability is small. If X has a binomial distribution with large n and small p, then the distribution of X will be approximately equal to a Poisson distribution with parameter $\lambda = np$.

Poisson Approximation of Binomial Distribution

Consider the binomial distribution with parameters n and p_n. Assume that

$$\lim_{n\to\infty} p_n = 0 \text{ and } \lim_{n\to\infty} np_n = \lambda > 0.$$

Then, the binomial pmf converges to the pmf of a Poisson distribution with parameter λ. That is, for $k = 0, 1, \ldots,$

$$\lim_{n\to\infty} \frac{n!}{k!(n-k)\,!} p_n^k (1-p_n)^{n-k} = \frac{e^{-\lambda}\lambda^k}{k!}.$$

Proof. Consider the binomial probability with $p = \lambda/n$, which gives

$$\begin{aligned}
\frac{n!}{k!(n-k)\,!} p^k (1-p)^{n-k} &= \frac{n!}{k!(n-k)\,!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{n^k \left(1-\frac{1}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{-k} \left(1-\frac{\lambda}{n}\right)^{n} \\
&= \frac{\lambda^k}{k!} \left[\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\right] \left(1-\frac{\lambda}{n}\right)^{-k} \left(1-\frac{\lambda}{n}\right)^{n}. \qquad \text{(B.2)}
\end{aligned}$$

Take the limit as $n \to \infty$, and consider the four factors on the right-hand side of Equation (B.2).

(i) Since λ and k are constants, $\lambda^k/k!$ stays unchanged in the limit.

(ii) For fixed k,

$$\lim_{n\to\infty} \left(1-\frac{1}{n}\right)\cdots\left(1-\frac{k-1}{n}\right) = 1^{k-1} = 1.$$

(iii)

$$\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^{-k} = 1^{-k} = 1.$$

(iv) Recall that the constant $e = 2.71827\ldots$ is defined as the limit

$$\lim_{x\to\infty} \left(1+\frac{1}{x}\right)^x = e.$$

Make the substitution $1/x = -\lambda/n$, so that $n = -\lambda x$. This gives

$$\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^n = \lim_{x\to\infty} \left(1+\frac{1}{x}\right)^{-\lambda x} = \left[\lim_{x\to\infty} \left(1+\frac{1}{x}\right)^x\right]^{-\lambda} = e^{-\lambda}.$$

Plugging in the four limits in Equation (B.2) gives the result. ∎

Example B.8 Mutations in DNA sequences occur from environmental factors and mistakes when a cell copies its DNA in preparation for cell division. The mutation rate per nucleotide of human DNA has been estimated at about 2.5×10^{-8}. There are about 3.3×10^9 nucleotide bases in the human DNA genome. Find the probability that exactly 80 DNA sites mutate in a random person's DNA.

Solution Let X be the number of mutations. If successive mutations are independent, then X has a binomial distribution with $n = 3.3 \times 10^9$ and $p = 2.5 \times 10^{-8}$. The distribution is approximated by a Poisson distribution with parameter

$$\lambda = np = (33 \times 10^9)(2.5 \times 10^{-8}) = 82.5.$$

The desired probability is

$$P(X = 80) \approx \frac{e^{-82.5}(82.5)^{80}}{80!} = 0.043.$$

Note that the exact probability using the binomial distribution is

$$P(X = 80) = \binom{3.3 \times 10^9}{80} (2.5 \times 10^{-8})^{80}(1 - 2.5 \times 10^{-8})^{3.3\times10^9-80}.$$

The approximate probability using the Poisson distribution can be compared with the exact probability in R. We display 12 significant digits, and see that the approximation is good to nine digits.

```
> options(digits=12)
> dpois(80,82.5)
[1] 0.0428838140788
> dbinom(80,3.3*10^9,2.5*10^(-8))
[1] 0.042883814558
```

■

Multinomial Distribution

The multinomial distribution generalizes the binomial distribution. Consider a sequence of n i.i.d. random variables, where each variable takes one of k possible values. Assume that the ith value occurs with probability p_i, with $p_1 + \cdots + p_k = 1$. For $i = 1, \ldots, k$, let X_i denote the number of times outcome i occurs. Then, $(X_1, \ldots, X_k)$ has a *multinomial distribution with parameters* $n, p_1, \ldots, p_k$. The joint pmf is

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}, \text{ for } x_1 + \cdots + x_k = n.$$

Marginally, each X_i has a binomial distribution with parameters n and p_i.

Example B.9 According to the Red Cross, the distribution of blood types in the United States is as follows: O: 44%, A: 42%, B: 10%, and AB: 4%. In a sample of six people, find the probability that three are type O, two are A, one is B, and none are AB.

Solution Let X_O, X_A, X_B, and X_{AB} denote the number of people in the sample with the respective blood types. Then, (X_O, X_A, X_B, X_{AB}) has a multinomial distribution with parameters $6, 0.44, 0.42, 0.10, 0.04$. The desired probability is

$$\begin{aligned} &P(X_O = 3, X_A = 2, X_B = 1, X_{AB} = 0) \\ &\quad = \frac{6!}{3!2!1!0!}(0.44)^3(0.42)^2(0.10)^1(0.04)^0 \\ &\quad = 0.09016. \end{aligned}$$

■

Uniform (Continuous) Distribution

For $a < b$, the uniform distribution on (a, b) is a continuous model for equally likely outcomes on a bounded interval. The density function is constant. A random variable X is *uniformly distributed on* (a, b) if the density of X is

$$f(x) = \frac{1}{b-a}, \text{ for } a < x < b.$$

Mean and variance are

$$E(X) = \frac{a+b}{2} \text{ and } Var(X) = \frac{(b-a)^2}{12}.$$

For $a < c < d < b$,

$$P(c < X < d) = \frac{d-c}{b-a} = \frac{\text{Length of } (c,d)}{\text{Length of } (a,b)}.$$

Exponential Distribution

The exponential distribution is a positive continuous distribution which often arises as a model for arrival, or waiting, times. Applications include the time when customers arrive at a queue, when electronic components fail, and when calls come in to a call center. The distribution depends on a parameter $\lambda > 0$, which has the interpretation of the arrival rate.

A random variable X has an *exponential distribution with parameter* λ, if the density function of X is

$$f(x) = \lambda e^{-\lambda x}, \text{ for } x > 0.$$

The cdf is

$$F(x) = P(X \leq x) = \int_0^x \lambda e^{-\lambda t}\, dt = 1 - e^{-\lambda x}, \text{ for } x > 0.$$

Mean and variance are

$$E(X) = \frac{1}{\lambda} \text{ and } Var(X) = \frac{1}{\lambda^2}.$$

The exponential distribution plays a prominent role in probability models because of its *memoryless* property. A random variable X is memoryless if

$$P(X > s + t | X > s) = P(X > t), \text{ for all } s, t > 0.$$

The exponential distribution is the only continuous distribution which is memoryless.

Example B.10 The lifetime of an electronic component is modeled with an exponential distribution. If components fail on average after 1,200 hours, find the probability that the component lasts more than 1,300 hours.

Solution Let X be the time until the component fails. Since the parameter of an exponential distribution is the reciprocal of the mean, model X with an exponential distribution with parameter $\lambda = 1/1200$. The desired probability is

$$\begin{aligned} P(X > 1300) &= \int_{1300}^{\infty} f(x)\, dx \\ &= \int_{1300}^{\infty} \frac{1}{1200} e^{-x/1200}\, dx \\ &= e^{-1300/1200} = 0.3385. \end{aligned}$$

In R, type

```
> 1-pexp(1300,1/1200)
[1] 0.3384654
```

■

Normal Distribution

The normal distribution plays a central role in statistics, is a common model for numerous natural and biological phenomenon, and arises as the limit for many random processes and distributions. It is also called the Gaussian distribution after Carl Friedrich Gauss who discovered its utility as a model for astronomical measurement errors. The distribution is parameterized by two numbers μ and σ^2, which are the mean and variance, respectively, of the distribution.

A random variable X has a *normal distribution with parameters* μ *and* σ^2, if the density function of X is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ for } -\infty < x < \infty.$$

The shape of the density is the famous *bell curve*. The density is symmetric about the line $x = \mu$. The inflection points, where the curvature of the density changes sign, occur one standard deviation from the mean, at the points $\mu \pm \sigma$.

The normal distribution has the property that a linear function of a normal random variable is normal. If X is normally distributed with mean μ and variance σ^2, then $Y = aX + b$ is normally distributed with mean

$$E(Y) = E(aX + b) = aE(X) + b = a\mu + b$$

and variance

$$Var\,(Y) = Var\,(aX + b) = a^2 Var\,(X) = a^2\sigma^2.$$

A common heuristic for working with the normal distribution is the *68–95–99.7 rule*, which says that for a normal distribution, the probability of being one, two, and three standard deviations from the mean is, respectively about 0.68, 0.95, and 0.997.

The *standard normal distribution* is a normal distribution with mean 0 and variance 1.

Example B.11 Assume that babies' birth weights are normally distributed with mean 120 ounces and standard deviation 20 ounces. (i) Find the approximate probability that a random baby's birth weight is between 100 and 140 ounces. (ii) Find the exact probability that a baby's birth weight is less than 136 ounces. (iii) *Low birth weight* is defined as the 5th percentile of the birth weight distribution. At what weight is a baby's birth weight considered low?

Solution

(i) Let X be a random baby's birth weight. The desired probability is $P(100 < X < 140)$. Observe that 100 is one standard deviation below the mean and 140 is one standard deviation about the mean. The event $\{100 < X < 140\}$ is the event that the birth weight is within one standard deviation of the mean. By the 68–95–99.7 rule, the probability is about 0.68.

(ii) The desired probability is

$$P(X < 136) = \int_{-\infty}^{136} f(x)\,dx = \int_{-\infty}^{136} \frac{1}{\sqrt{2\pi(20^2)}} e^{-\frac{(x-120)^2}{2(20)^2}}\,dx.$$

There is no elementary closed form for the cdf of the normal distribution, and numerical methods are needed to solve integrals such as this. In R, type

```
> pnorm(136,120,20)
[1] 0.78814460
```

(iii) The problem asks for the 5th percentile of the normal distribution. Equivalently, we seek the number q such that $P(X \leq q) = 0.05$. In R, use the quantile function.

```
> qnorm(0.05,120,20)
[1] 87.102927461
```

Babies with birth weights less than 87 ounces are considered to have low birth weight. ■

Bivariate Normal Distribution

The bivariate normal distribution for X and Y, and the multivariate normal distribution for $X_1, \ldots, X_m$, are generalizations of the normal distribution to higher dimensions. The bivariate normal distribution is specified by five parameters: $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho$. These are the means and variances of X and Y, and their correlation. If $\mu_X = \mu_Y = 0$ and $\sigma_X^2 = \sigma_Y^2 = 1$, this gives the bivariate standard normal distribution with joint density

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}},$$

for $-\infty < x, y < \infty$ and $-1 < \rho < 1$.

The marginal and conditional distributions of a bivariate normal distribution are normal. In particular, the conditional distribution of X given $Y = y$ is normal with mean ρy and variance $1 - \rho^2$. Similarly, the conditional distribution of Y given $X = x$ is normal with mean ρx and variance $1 - \rho^2$.

If X and Y have a bivariate normal distribution, then $aX + bY$ is normally distributed for all nonzero constants a and b.

If $\rho = 0$, we say that X and Y are *uncorrelated*. If normal random variables are uncorrelated, then they are independent. Note that this is not true for random variables in general. If two random variables X and Y are uncorrelated it does not necessarily mean that they are independent.

Gamma Distribution

The gamma distribution is a nonnegative continuous distribution which depends on two parameters. The distribution encompasses a large family of unimodal, skewed and symmetric distribution shapes and is a popular distribution for modeling positive continuous processes. It also arises naturally as the distribution of a sum of i.i.d. exponential random variables.

A random variable X has a *gamma distribution with parameters* $r > 0$ *and* $\lambda > 0$, if the density function of X is

$$f(x) = \frac{1}{\Gamma(r)} \lambda^r x^{r-1} e^{-\lambda x}, \text{ for } x > 0,$$

where $\Gamma(r)$ is the *gamma function*

$$\Gamma(r) = \int_0^\infty t^{r-1} e^{-t}\, dt.$$

If r is a positive integer, then $\Gamma(r) = (r-1)!$ For a gamma random variable,

$$E(X) = \frac{r}{\lambda} \text{ and } Var(X) = \frac{r}{\lambda^2}.$$

A sum of n independent exponential random variables with common parameter λ has a gamma distribution with parameters n and λ.

Example B.12 Bella is fishing and the time it takes to catch a fish is exponentially distributed with mean 20 minutes. Every time she catches a fish, she throws it back in the water, and continues fishing. Find the probability that she will catch five fish in the first hour.

Solution Let X_1 denote how long it takes for Bella to catch her first fish. For $k = 2, \ldots, 5$, let X_k be the time from when she throws her $(k-1)$th fish in the water to when she catches her kth fish. Then, $X_1, \ldots, X_5$ are i.i.d. exponential random variables with parameter $\lambda = 1/20$, and $S = X_1 + \cdots + X_5$ is the total time it takes for Bella to catch five fish. The random variable S has a gamma distribution with parameters $r = 5$ and $\lambda = 1/20$. The desired probability is

$$P(S \le 60) = \int_0^{60} \frac{1}{\Gamma(5)} \left(\frac{1}{20}\right)^5 x^4 e^{-x/20}\, dx = 0.185.$$

■

Beta Distribution

The beta distribution is a continuous distribution on $(0, 1)$ that depends on two parameters. It is a generalization of the uniform distribution. A random variable X *has a beta distribution with parameters* $a > 0$ *and* $b > 0$, if the density function of X is

$$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, \text{ for } 0 < x < 1,$$

where

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

For a beta random variable,

$$E(X) = \frac{a}{a+b} \text{ and } Var(X) = \frac{ab}{(a+b+1)(a+b)^2}.$$

The uniform distribution on $(0,1)$ is obtained for $a = b = 1$.

B.5 LIMIT THEOREMS

The classic limit theorems of probability are concerned with sequences of i.i.d. random variables. If $X_1, X_2, \ldots$ is such a sequence with common mean $\mu = E(X_1) < \infty$, let $S_n = X_1 + \cdots + X_n$. The *law of large numbers* says that the sequence of averages S_n/n converges to μ, as $n \to \infty$. There are two versions of the law—weak and strong.

Theorem B.1. (Weak law of large numbers). *For any* $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| < \epsilon\right) = 1.$$

While the weak law asserts that for large n, the average S_n/n is with high probability close to μ, it does not say that having come close to μ, the sequence of averages will always stay close to μ.

If a sequence of numbers $x_1, x_2, \ldots$ converges to a limit x then eventually, for n sufficiently large, the terms $x_n, x_{n+1}, x_{n+2}, \ldots$ will all be arbitrarily close to x. That is, for any $\epsilon > 0$, there is some index N such that $|x_n - x| \le \epsilon$, for all $n \ge N$.

The strong law of large numbers asserts that with probability 1 the sequence of averages $S_1/1, S_2/2, S_3/3, \ldots$ behaves precisely in this way.

Theorem B.2. (Strong law of large numbers).

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

By the law of large numbers, $S_n/n - \mu$ converges to 0, as $n \to \infty$. The *central limit theorem* asserts that $(S_n/n - \mu)/(\sigma/\sqrt{n})$ converges to a normally distributed random variable. Remarkably, this is true for *any* i.i.d. sequence $X_1, X_2, \ldots$ with finite mean and variance.

Theorem B.3. (Central limit theorem). *For all t,*

$$\lim_{n\to\infty} P\left(\frac{S_n/n - \mu}{\sigma/\sqrt{n}} \le t\right) = P(Z \le t),$$

where Z has a standard normal distribution.

The central limit theorem gives that for large n, $X_1 + \cdots + X_n$ has an approximate normal distribution with mean $n\mu$ and variance $n\sigma^2$.

Example B.13 The number of accidents per week in a stretch of highway has a Poisson distribution with parameter $\lambda = 2$. If the number of accidents is independent from week to week, what is the probability that over a year's time there will be more than 100 accidents?

Solution Let $X_1, \ldots, X_{52}$ be the number of accidents, respectively, during each week of the year. Then, $S_{52} = X_1 + \cdots + X_{52}$ is the total number of accidents in a year. The X_i are i.i.d. with common mean and variance $\lambda = 2$. By the central limit theorem,

$$P(S_{52} > 100) = P\left(\frac{S_{52}/52 - \mu}{\sigma/\sqrt{52}} > \frac{100/52 - 2}{\sqrt{2}/\sqrt{52}}\right)$$
$$\approx P(Z > -0.392) = 0.652.$$

We can compare the central limit approximation with the exact result. The sum of independent Poisson random variables has a Poisson distribution, and $X_1 + \cdots + X_{52}$ has a Poisson distribution with parameter $52\lambda = 104$. The exact probability is

$$P(S_{52} > 100) = 1 - P(S_{52} \le 100) = 1 - \sum_{k=0}^{100} \frac{e^{-104}104^k}{k!} = 0.629. \quad \blacksquare$$

B.6 MOMENT-GENERATING FUNCTIONS

Moment-generating functions are remarkably versatile tools for proving results involving sums and limits of random variables. Let X be a random variable. The *moment-generating function (mgf) of X* is the function $m(t) = E\left(e^{tX}\right)$, defined for all t for which the expectation exists.

The name *moment-generating* comes from the fact that the moments of X can be derived by taking successive derivatives of the mgf. In particular,

$$m'(t) = \frac{d}{dt}E\left(e^{tX}\right) = E\left(\frac{d}{dt}e^{tX}\right) = E\left(Xe^{tX}\right),$$

and $m'(0) = E(X)$. In general, the kth derivative of the mgf gives

$$m^{(k)}(0) = E\left(X^k\right), \text{ for } k = 1, 2, \ldots$$

Example B.14 Let X be a Bernoulli random variable with success parameter p. Find the mgf of X.

Solution The mgf of X is

$$m(t) = E\left(e^{tX}\right) = e^{t(1)}p + e^{t(0)}(1-p) = 1 - p + pe^t. \quad \blacksquare$$

Example B.15 Let X have a standard normal distribution. Find the mgf.

Solution The mgf is

$$\begin{aligned} m(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x^2-2tx)/2}\, dx \\ &= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2}\, dx = e^{t^2/2}, \end{aligned}$$

as the last integral gives the density of a normal distribution with mean t and variance 1 which integrates to 1. Check that $m'(0) = 0 = E(X)$ and $m''(0) = 1 = E\left(X^2\right)$. $\blacksquare$

Here are four key properties of the mgf.

Properties of Moment-Generating Functions

1. If X and Y are independent, then the mgf of $X + Y$ is the product of their respective mgfs. That is,

$$\begin{aligned} m_{X+Y}(t) &= E\left(e^{t(X+Y)}\right) = E\left(e^{tX}e^{tY}\right) = E\left(e^{tX}\right)E\left(e^{tY}\right) \\ &= m_X(t)m_Y(t). \end{aligned}$$

2. For constant c,

$$m_{cX}(t) = m_X(ct).$$

3. Moment-generating functions uniquely determine the underlying probability distribution. That is, if $m_X(t) = m_Y(t)$ for all t, then the distributions of X and Y are the same.

4. *Continuity Theorem:* Let $X_1, X_2, \ldots$ be a sequence of random variables with corresponding mgfs $m_{X_1}, m_{X_2}, \ldots$ Assume that X is a random variable such that for all t, $m_{X_n}(t) \to m_X(t)$, as $n \to \infty$. Then,

$$\lim_{n\to\infty} P(X_n \le x) = P(X \le x),$$

at each x for which $P(X \le x)$ is continuous.

Example B.16 Let X and Y be independent Poisson random variables with respective parameters λ_1 and λ_2. Use moment-generating functions to show that $X + Y$ has a Poisson distribution with parameter $\lambda_1 + \lambda_2$.

Solution The mgf of a Poisson random variable with parameter λ is

$$m(t) = \sum_{k=0}^{\infty} e^{tk} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t-1)}.$$

Hence, the mgf of $X + Y$ is

$$m_{X+Y}(t) = m_X(t) m_Y(t) = e^{\lambda_1(e^t-1)} e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)},$$

which is the mgf of a Poisson random variable with parameter $\lambda_1 + \lambda_2$. ■