Chapter 5
Random Variables

The generation of random numbers is too important to be left to chance.

– Robert R. Coveyou

WHAT IS COVERED IN THIS CHAPTER

- Definition of Random Variables and Their Basic Characteristics
- Discrete Random Variables: Bernoulli, Binomial, Poisson, Hypergeometric, Geometric, Negative Binomial, and Multinomial
- Continuous Random Variables: Uniform, Exponential, Gamma, Inverse Gamma, Beta, Double Exponential, Logistic, Weibull, Pareto, and Dirichlet
- Transformation of Random Variables
- Markov Chains

5.1 Introduction

Thus far we have been concerned with random experiments, events, and their probabilities. In this chapter we will discuss random variables and their probability distributions. The outcomes of an experiment can be associated with numerical values, and this association will help us arrive at the definition of a random variable.

A *random variable* is a variable whose numerical value is determined by the outcome of a random experiment.

Thus, a random variable is a mapping from the sample space of an experiment, $\mathcal{S}$, to a set of real numbers. In this respect, the term *random variable* is a misnomer. The more appropriate term would be *random function* or *random mapping*, given that X maps a sample space $\mathcal{S}$ to real numbers. We generally denote random variables by capital letters $X, Y, Z, \ldots$.

Example 5.1. **Three Coin Tosses.** Suppose a fair coin is tossed three times. We can define several random variables connected with this experiment. For example, we can set X to be the number of heads, Y the difference between the number of heads and the number of tails, and Z an indicator that heads appeared, etc.

Random variables X, Y, and Z are fully described by their probability distributions, associated with the sample space on which they are defined.

For random variable X the possible realizations are 0 (no heads in three flips), 1 (exactly one head), 2 (exactly two heads), and 3 (all heads). Fully describing random variable X amounts to finding the probabilities of all possible realizations. For instance, the realization $\{X = 2\}$ corresponds to either outcome in the event $\{HHT, HTH, THH\}$. Thus, the probability of X taking value 2 is equal to the probability of the event $\{HHT, HTH, THH\}$, which is equal to $3/8$. After finding the probabilities for other outcomes, we determine the distribution of random variable X:

X	0	1	2	3
Prob	$1/8$	$3/8$	$3/8$	$1/8$

.

The *probability distribution* of a random variable X is a table (assignment, rule, formula) that assigns probabilities to realizations of X, or sets of realizations.

Most random variables of interest to us will be the results of random sampling. There is a general classification of random variables that is based on the nature of realizations they can take. Random variables that take values from a finite or countable set are called *discrete random variables*. Random variable X from Example 5.1 is an example of a discrete random variable. Another type of random variable can take any value from an interval on a real line. These are called *continuous random variables*. The results of measurements are usually modeled by continuous random variables. Next, we will describe discrete and continuous random variables in a more structured manner.

5.2 Discrete Random Variables

Let random variable X take discrete values $x_1, x_2, \ldots, x_n, \ldots$ with probabilities $p_1, p_2, \ldots, p_n, \ldots$, $\sum_n p_n = 1$. The probability distribution function (PDF) is simply an assignment of probabilities to the realizations of X and is given by the following table.

$$\begin{array}{c|c} X & x_1 \; x_2 \; \cdots \; x_n \; \cdots \\ \hline \text{Prob} & p_1 \; p_2 \; \cdots \; p_n \; \cdots \end{array}.$$

The probabilities p_i sum up to 1: $\sum_i p_i = 1$. It is important to emphasize that discrete random variables can have an infinite number of realizations, as long as the infinite sum of the probabilities converges to 1. The PDF for discrete random variables is also called the probability mass function (PMF). The cumulative distribution function (CDF)

$$F(x) = P(X \leq x) = \sum_{n: x_n \leq x} p_n,$$

sums the probabilities of all realizations smaller than or equal to x. Figure 5.1a shows an example of a discrete random variable X with four values and a CDF as the sum of probabilities in the range $X \leq x$ shown in yellow.

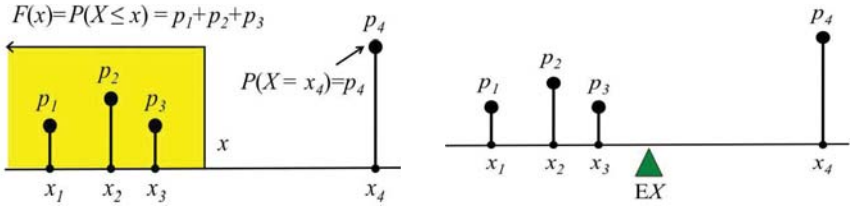


Fig. 5.1 (a) Example of a cumulative distribution function for discrete random variable X. The CDF is the sum of probabilities in the region $X \leq x$ (*yellow*). (b) Expectation as a point of balance for "masses" $p_1, \ldots, p_4$ located at the points $x_1, \ldots, x_4$.

The expectation of X is given by

$$\mathbb{E}X = x_1 p_1 + \cdots + x_n p_n + \cdots = \sum_n x_n p_n$$

and is a weighted average of all possible realizations with their probabilities as weights. Figure 5.1b illustrates the interpretation of the expectation as the point of balance for a system with weights $p_1, \ldots, p_4$ located at the locations $x_1, \ldots, x_4$.

The distribution and expectation of a function $g(X)$ are simple when X is discrete: one applies function g to realizations of X and retains the probabilities:

$$\begin{array}{c|cccccc} g(X) & g(x_1) & g(x_2) & \cdots & g(x_n) & \cdots \\ \hline \text{Prob} & p_1 & p_2 & \cdots & p_n & \cdots \end{array}$$

and

$$\mathbb{E}g(X) = g(x_1)p_1 + \cdots + g(x_n)p_n + \cdots = \sum_n g(x_n)p_n.$$

The kth moment of a discrete random variable X is defined as

$$m_k = \mathbb{E}X^k = \sum_n x_n^k p_n,$$

and the kth central moment is

$$\mu_k = \mathbb{E}(X - \mathbb{E}X)^k = \sum_n (x_n - \mathbb{E}X)^k p_n.$$

The first moment is the expectation, $m_1 = \mathbb{E}X$, and the second central moment is the variance, $\mu_2 = \mathbb{V}\text{ar}\,(X) = \mathbb{E}(X - \mathbb{E}X)^2$. Thus, the variance for a discrete random variable is

$$\mathbb{V}\text{ar}\,(X) = \sum_n (x_n - \mathbb{E}X)^2 p_n.$$

The skewness and kurtosis of X are defined via the central moments as

$$\gamma = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mathbb{E}(X - \mathbb{E}X)^3}{(\mathbb{V}\text{ar}\,(X))^{3/2}} \quad \text{and} \quad \kappa = \frac{\mu_4}{\mu_2^2} = \frac{\mathbb{E}(X - \mathbb{E}X)^4}{(\mathbb{V}\text{ar}\,(X))^2}. \tag{5.1}$$

The following properties are common for both discrete and continuous random variables:

For any set of random variables $X_1, X_2, \ldots, X_n$

$$\mathbb{E}(X_1 + X_2 + \cdots + X_n) = \mathbb{E}X_1 + \mathbb{E}X_2 + \cdots + \mathbb{E}X_n. \tag{5.2}$$

For any constant c, $\mathbb{E}(c) = c$ and $\mathbb{E}(cX) = c\,\mathbb{E}X$.

The independence of two random variables is defined via the independence of events. Two random variables X and Y are independent if for arbitrary intervals A and B, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent, that is, when

$$\mathbb{P}(X \in A,\ Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B),$$

holds.

If the random variables $X_1, X_2, \ldots, X_n$ are independent, then

$$\begin{aligned} \mathbb{E}(X_1 \cdot X_2 \cdot \ \ldots \ \cdot X_n) &= \mathbb{E}X_1 \cdot \mathbb{E}X_2 \cdot \ \ldots \ \cdot \mathbb{E}X_n, \text{ and} \\ \mathbb{V}\text{ar}\,(X_1 + X_2 + \cdots + X_n) &= \mathbb{V}\text{ar}\,X_1 + \mathbb{V}\text{ar}\,X_2 + \cdots + \mathbb{V}\text{ar}\,X_n. \end{aligned} \tag{5.3}$$

For a constant c, $\mathbb{V}\text{ar}\,(c) = 0$, and $\mathbb{V}\text{ar}\,(cX) = c^2\mathbb{V}\text{ar}\,X$.

If $X_1, X_2, \ldots, X_n, \ldots$ are independent and identically distributed random variables, we will refer to them as i.i.d. random variables.

The arguments behind these properties involve the linearity of the sums (for discrete variables) and integrals (for continuous variables). The independence of the X_is is critical for (5.3).

Moment-Generating Function. A particularly useful function for finding moments and for more advanced operations with random variables is the *moment-generating function.* For a random variable X, the moment-generating function is defined as

$$m_X(t) = \mathbb{E}e^{tX} = \sum_n p_n e^{tx_n}, \tag{5.4}$$

which for discrete random variables has the form $m_X(t) = \sum_n p_n e^{tx_n}$. When the moment-generating function exists, it uniquely determines the distribution. If X has distribution F_X and Y has distribution F_Y, and if $m_X(t) = m_Y(t)$ for all t, then it follows that $F_X = F_Y$.

The name "moment-generating" is motivated by the fact that the kth derivative of $m_X(t)$ evaluated at $t = 0$ results in the kth moment of X, that is, $m_X^{(k)}(t) = \sum_n p_n x_n^k e^{tx_n}$, and $m_X^{(k)}(0) = \sum_n p_n x_n^k = \mathbb{E}X^k$. For example, if

X	0	1	3
Prob	0.2	0.3	0.5

,

then $m_X(t) = 0.2 + 0.3\ e^t + 0.5\ e^{3t}$. Since $m'_X(t) = 0.3\ e^t + 1.5\ e^{3t}$, the first moment is $\mathbb{E}X = m'(0) = 0.3 + 1.5 = 1.8$. The second derivative is $m''_X(t) = 0.3\ e^t + 4.5\ e^{3t}$, the second moment is $\mathbb{E}X^2 = m''(0) = 0.3 + 4.5 = 4.8$, and so on.

In addition to generating the moments, moment-generating functions satisfy

$$\begin{aligned} m_{X+Y}(t) &= m_X(t)\ m_Y(t), \\ m_{cX}(t) &= m_X(ct), \end{aligned} \tag{5.5}$$

which helps in identifying distributions of linear combinations of random variables whenever their moment-generating functions exist.

The properties in (5.5) follow from the properties of expectations. When X and Y are independent, e^{tX} and e^{tY} are independent as well, and by (5.3) $\mathbb{E}e^{t(X+Y)} = \mathbb{E}e^{tX}e^{tY} = \mathbb{E}e^{tX} \cdot \mathbb{E}e^{tY}$.

Example 5.2. **Apgar Score.** In the early 1950s, Dr. Virginia Apgar proposed a method to assess the health of a newborn child by assigning a grade referred to as the Apgar score (Apgar, 1953). It is given twice for each newborn, once at 1 min after birth and again at 5 min after birth.

Possible values for the Apgar score are 0, 1, 2, $\cdots$, 9, and 10. A child's score is determined by five factors: muscle tone, skin color, respiratory effort, strength of heartbeat, and reflex, with a high score indicating a healthy infant. Let the random variable X denote the Apgar score of a randomly selected newborn infant at a particular hospital. Suppose that X has a given probability distribution:

X	0	1	2	3	4	5	6	7	8	9	10
Prob	0.002	0.001	0.002	0.005	0.02	0.04	0.17	0.38	0.25	0.12	0.01

.

The following MATLAB program calculates (a) $\mathbb{E}X$, (b) $\mathbb{V}\mathrm{ar}(X)$, (c) $\mathbb{E}X^4$, (d) $F(x)$, (e) $\mathbb{P}(X < 4)$, and (f) $\mathbb{P}(2 < X \leq 3)$:

```
X = 0:10;
p = [0.002  0.001  0.002  0.005  0.02  ...
      0.04  0.17 0.38  0.25  0.12  0.01];
EX = X * p'               %(a) EX = 7.1600
VarX = (X-EX).^2 * p'     %(b) VarX = 1.5684
EX4 = X.^4 * p'           %(c) EX4 = 3.0746e+003
ps = [0 cumsum(p)];
Fx = @(x)   ps( min(max( floor(x)+2, 1),12) ); %handle
Fx(3.45)                  %(d) ans = 0.0100
sum(p(X < 4))             %(e) ans = 0.0100
sum(p(X > 2 & X <= 3))    %(f) ans = 0.0050
```

Note that the CDF F is expressed as function handle `Fx` to a custom-made function.

Example 5.3. **Cells.** Randomly observed circular cells on a plate have a diameter D that is a random variable with the following PMF:

D	8	12	16
Prob	0.4	0.3	0.3

.

(a) Find the CDF for D.

(b) Find the PMF for the random variable $A = D^2\pi/4$ (the area of a cell). Show that $\mathbb{E}A \neq (\mathbb{E}D)^2\pi/4$. Explain.

(c) Find the variance $\mathbb{V}\mathrm{ar}\,(A)$.

(d) Find the moment-generating functions $m_D(t)$ and $m_A(t)$. Find $\mathbb{V}\mathrm{ar}\,(A)$ using its moment-generating function.

(e) It is known that a cell with $D > 8$ is observed. Find the probability of $D = 12$ taking into account this information.

Solution:

(a)

$$F_D(d) = \begin{cases} 0, & d < 8 \\ 0.4, & 8 \le d < 12 \\ 0.7, & 12 \le d < 16 \\ 1, & d \ge 16 \end{cases}$$

(b)

A	$8^2\,\pi/4$	$12^2\,\pi/4$	$16^2\,\pi/4$
Prob	0.4	0.3	0.3

A	$16\,\pi$	$36\,\pi$	$64\,\pi$
Prob	0.4	0.3	0.3

$\mathbb{E}A = 16\,\pi(\frac{4}{10}) + 36\,\pi(\frac{3}{10}) + 64\,\pi(\frac{3}{10}) = \frac{364\,\pi}{10} = 114.3540.$

$\mathbb{E}D = 8(\frac{4}{10}) + 12(\frac{3}{10}) + 16(\frac{3}{10}) = 116/10 = 11.6$

$\frac{(\mathbb{E}D)^2\,\pi}{4} = \frac{3364\,\pi}{100} \neq \frac{364\,\pi}{10}$.

The expectation is a linear operator, and such a "plug-in" operation would work only if the random variable A were a linear function of D, that is, if $A = \alpha D + \beta$, $\mathbb{E}A = \alpha\mathbb{E}D + \beta$. In our case, A is quadratic in D, and "passing" the expectation through the equation is not valid.

(c)

$$\mathbb{V}\mathrm{ar}\,A = \mathbb{E}A^2 - (\mathbb{E}A)^2 = 1720\,\pi^2 - 1324.96\,\pi^2 = 395.04\,\pi^2,$$

since

A^2	$16^2\,\pi^2$	$36^2\,\pi^2$	$64^2\,\pi^2$
Prob	0.4	0.3	0.3

and $EA^2 = 1720\,\pi^2$.

(d) $m_D(t) = \mathbb{E}e^{tD} = 0.4e^{8t} + 0.3e^{12t} + 0.3e^{16t}$, and $m_A(t) = \mathbb{E}e^{tA} = 0.4e^{16\pi t} + 0.3e^{36\pi t} + 0.3e^{64\pi t}$.

From $m'_A(t) = 6.4e^{16\pi t} + 10.8e^{36\pi t} + 19.2e^{64\pi t}$, and $m''_A(t) = 6.4e^{16\pi t} + 10.8e^{36\pi t} + 19.2e^{64\pi t}$, we find $m'_A(0) = 36.4\pi$ and $m''_A(0) = 1720\pi$, leading to the result in (c).

(e) When $D > 8$ is true, only two values for D are possible, 12 and 16. These values are equally likely. Thus, the distribution for $D|\{D > 8\}$ is

$D\|\{D > 8\}$	12	16
Prob	0.3/0.6	0.3/0.6

,

and $\mathbb{P}(D = 12|D > 8) = 1/2$. We divided 0.3 by 0.6 since $\mathbb{P}(D > 8) = 0.6$. From the definition of the conditional probability it follows that,

$\mathbb{P}(D = 12|D > 8) = \mathbb{P}(D = 12, D > 8)/\mathbb{P}(D > 8) = \mathbb{P}(D = 12)/\mathbb{P}(D > 8) = 0.3/0.6 = 1/2$.
✎

There are important properties of discrete distributions in which the realizations $x_1, x_2, \ldots, x_n$ are irrelevant and the focus is on the probabilities only, such as the measure of *entropy*. For a discrete random variable where the probabilities are $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$ the (Shannon) entropy is defined as

$$\mathcal{H}(\boldsymbol{p}) = -\sum_i p_i \log(p_i).$$

Entropy is a measure of the uncertainty of a random variable and for finite discrete distributions achieves its maximum when the probabilities of realizations are equal, $\boldsymbol{p} = (1/n, 1/n, \ldots, 1/n)$.

For the distribution in Example 5.2, the entropy is 1.5812.

```
ps = [.002  .001  .002  .005  .02  .04   .17   .38  .25 .12 .01]
entropy = @(p)   -sum( p(p>0) .* log(p(p>0)))
entropy(ps) %1.5812
```

The maximum entropy for distributions with 11 possible realizations is 2.3979.

Jointly Distributed Discrete Random Variables. So far we have discussed probability distributions of a single random variable. As we delve deeper into this subject, a two-dimensional extension will be needed.

When two or more random variables constitute the coordinates of a random vector, their joint distribution is often of interest. For a random vector (X, Y) the joint distribution function is defined via the probability of the event $\{X \le x, Y \le y\}$,

$$F(x, y) = \mathbb{P}(X \le x, Y \le y).$$

The univariate case $\mathbb{P}(a \le X \le b) = F(b) - F(a)$ takes the bivariate form

$$\mathbb{P}(a_1 \le X \le a_2, b_1 \le Y \le b_2) = F(a_2, b_2) - F(a_1, b_2) - F(a_2, b_1) + F(a_1, b_1).$$

Marginal CDFs F_X and F_Y are defined as follows: for X, $F_X(x) = F(x, \infty)$ and for Y as $F_Y(y) = F(\infty, y)$.

For a discrete bivariate random variable, the PMF is

$$p(x, y) = \mathbb{P}(X = x, Y = y), \ \sum_{x,y} p(x, y) = 1,$$

while for marginal random variables X and Y the PMFs are

$$p_X(x) = \sum_y p(x, y), \quad p_Y(y) = \sum_x p(x, y).$$

The conditional distribution of X given $Y = y$ is defined as

$$p_{X|Y}(x|y) = p(x,y)/p_Y(y),$$

and, similarly, the conditional distribution for Y given $X = x$ is

$$p_{Y|X}(y|x) = p(x,y)/p_X(x).$$

When X and Y are independent, for any "cell" (x,y), $p(x,y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y) = p_X(x)\,p_Y(y)$, that is, the joint probability of (x,y) is equal to the product of the marginal probabilities. If $p(x,y) = p_X(x)p_Y(y)$ holds for every (x,y), then X and Y are independent. The independence of two discrete random variables is fundamental for the inference in contingency tables (Chapter 12) and will be revisited later.

Example 5.4. PMF of a two-dimensional discrete random variable is given by the following table:

		Y		
		5	10	15
X	1	0.1	0.2	0.3
	2	0.25	0.1	0.05

The marginal distributions for X and Y are

X	1	2
Prob	0.6	0.4

and

Y	5	10	15
Prob	0.35	0.3	0.35

while the conditional distribution for X when $Y = 10$ and the conditional distribution for Y when $X = 2$ are

$X\|Y = 10$	1	2
Prob	$\frac{0.2}{0.3}$	$\frac{0.1}{0.3}$

and

$Y\|X = 2$	5	10	15
Prob	$\frac{0.25}{0.4}$	$\frac{0.1}{0.4}$	$\frac{0.05}{0.4}$

,

respectively. Here X and Y are not independent since

$$0.1 = \mathbb{P}(X = 1, Y = 5) \neq \mathbb{P}(X = 1)\mathbb{P}(Y = 5) = 0.6 \cdot 0.35 = 0.21.$$

✎

For two independent random variables X and Y, $\mathbb{E}XY = \mathbb{E}X \cdot \mathbb{E}Y$; that is, the expectation of a product of random variables is equal to the product of their expectations.

The *covariance* of two random variables X and Y is defined as

$$\mathbb{C}\text{ov}(X,Y) = \mathbb{E}((X - \mathbb{E}X)\cdot(Y - \mathbb{E}Y)) = \mathbb{E}XY - \mathbb{E}X\cdot\mathbb{E}Y.$$

For a discrete random vector (X,Y), $\mathbb{E}XY = \sum_x \sum_y xyp(x,y)$, and the covariance is expressed as

$$\mathbb{C}\text{ov}(X,Y) = \sum_x \sum_y xyp(x,y) - \sum_x xp_X(x) \sum_y yp_Y(y).$$

It is easy to see that the covariance satisfies the following properties:

$$\begin{aligned}
&\mathbb{C}\text{ov}(X,X) = \mathbb{V}\text{ar}\,(X),\\
&\mathbb{C}\text{ov}(X,Y) = \mathbb{C}\text{ov}(Y,X), \text{ and}\\
&\mathbb{C}\text{ov}(aX+bY,Z) = a\mathbb{C}\text{ov}(X,Z) + b\mathbb{C}\text{ov}(Y,Z).
\end{aligned}$$

For (X,Y) from Example 5.4 the covariance between X and Y is -1. The calculation is provided in the following MATLAB code. Note that the distribution of the product XY is found in order to calculate $\mathbb{E}XY$.

```
X = [1 2]; pX = [0.6 0.4]; EX = X * pX'                    %EX = 1.4000
Y = [5 10 15]; pY = [0.35  0.3  0.35]; EY = Y*pY'          %EY =10
XY = [5 10 15 20 30];
pXY = [0.1   0.2+0.25   0.3  0.1  0.05];  EXY = XY * pXY'  %EXY = 13
CovXY = EXY - EX * EY                                      %CovXY = -1
```

The *correlation* between random variables X and Y is the covariance normalized by the standard deviations:

$$\mathbb{C}\text{orr}(X,Y) = \frac{\mathbb{C}\text{ov}(X,Y)}{\sqrt{\mathbb{V}\text{ar}\,X\cdot\mathbb{V}\text{ar}\,Y}}.$$

In Example 5.4, the variances of X and Y are $\mathbb{V}\text{ar}\,X = 0.24$ and $\mathbb{V}\text{ar}\,Y = 17.5$. Using these values, we find that the correlation $\mathbb{C}\text{orr}(X,Y)$ is $-1/\sqrt{0.24\cdot 17.5} = -0.488$. Thus, the random components in (X,Y) are negatively correlated.

5.3 Some Standard Discrete Distributions

5.3.1 Discrete Uniform Distribution

A random variable X that takes values from 1 to n with equal probabilities of $1/n$ is called a discrete uniform random variable. In MATLAB `unidpdf`

and `unidcdf` are the PDF and CDF of X, while `unidinv` is its quantile. For example,

```
 unidpdf(1:5, 5)
%ans =   0.2000    0.2000    0.2000    0.2000    0.2000

 unidcdf(1:5, 5)
%ans =   0.2000    0.4000    0.6000    0.8000    1.0000
```

are the PDF and CDF of the discrete uniform distribution on $\{1,2,3,4,5\}$. From $\sum_{i=1}^{n} i = n(n+1)/2$, and $\sum_{i=1}^{n} i^2 = n(n+1)(2n+1)/6$, one can derive $\mathbb{E}X = (n+1)/2$ and $\mathbb{V}\mathrm{ar}\, X = (n^2-1)/12$. One of the important uses of discrete uniform distribution is in nonparametric statistics (page 894).

Example 5.5. **Discrete Uniform: A Basis for Random Sampling.** Suppose that a population is finite and that we need a sample such that every subject in the population has an equal chance of being selected.

If the population size is N and a sample of size n is needed, then if replacement is allowed (each sampled object is recorded and then returned back to the population), there would be N^n possible equally likely samples. If replacement is not allowed or possible (all subjects in the selected sample are to be different, i.e., sampling is without replacement), then there would be $\binom{N}{n}$ different equally likely samples (see Section 3.5 for a definition of $\binom{N}{n}$).

The theoretical model for random sampling is the discrete uniform distribution. If replacement is allowed, each of $\{1,2,\ldots,N\}$ has a probability of $1/N$ of being selected. In the case of no replacement, possible subsets of n subjects can be indexed as $\{1,2,\ldots,\binom{N}{n}\}$ and each subset has a probability of $1/\binom{N}{n}$ of being selected.

In MATLAB, random sampling is achieved by the function `randsample`. If the population has n indexed subjects (from 1 to n), the indices in a random sample of size k are found as `indices=randsample(n,k)`.

If it is possible to code the entire population as a vector `population`, then taking a sample of size k is done by `y=randsample(population,k)`.

The default is set to sampling without replacement. For sampling with replacement, the flag for replacement should be `'true'`. If the sampling is done with replacement, it can be weighted with a nonnegative weight assigned to each subject in the population: `y=randsample(population,k,true,w)`. The size of weight vector `w` should be the same as that of `population`.

For instance,

```
 randsample(['A' 'C' 'G' 'T'],50,true,[1 1.5 1.4  0.9])
%ans =  GCCTAGGGCATCCAAGTCGCGGCCGAGAATCAACGTTGCAGTGCTCAAAT
```

5.3.2 Bernoulli and Binomial Distributions

A simple Bernoulli random variable Y is dichotomous with $\mathbb{P}(Y=1)=p$ and $\mathbb{P}(Y=0)=1-p$ for some $0\le p\le 1$ and is denoted as $Y\sim\mathcal{B}er(p)$. It is named after Jakob Bernoulli (1654–1705), a prominent Swiss mathematician and astronomer. Suppose that an experiment consists of n independent trials $(Y_1,\ldots,Y_n)$ in which two outcomes are possible (e.g., success or failure), with $\mathbb{P}(\text{success})=\mathbb{P}(Y=1)=p$ for each trial. If $X=x$ is defined as the number of successes (out of n), then $X=Y_1+Y_2+\cdots+Y_n$, and there are $\binom{n}{x}$ arrangements of x successes and $n-x$ failures, each having the same probability $p^x(1-p)^{n-x}$. X is a *binomial* random variable with the PMF

$$p_X(x)=\binom{n}{x}p^x(1-p)^{n-x},\ x=0,1,\ldots,n.$$

This is denoted by $X\sim\mathcal{B}in(n,p)$. From the moment-generating function $m_X(t)=(pe^t+(1-p))^n$, we obtain $\mu=\mathbb{E}X=np$ and $\sigma^2=\mathbb{V}\text{ar}\,X=np(1-p)$.

The cumulative distribution for a binomial random variable is not simplified beyond the sum, that is, $F(x)=\sum_{i\le x}p_X(i)$. However, interval probabilities can be computed in MATLAB using `binocdf(x,n,p)`, which computes the CDF at value x. The PMF can also be computed in MATLAB using `binopdf(x,n,p)`. In WinBUGS, the binomial distribution is denoted as `dbin(p,n)`. Note the reversed order of parameters n and p.

Example 5.6. **Left-Handed Families.** About 10% of the world's population is left-handed. Left-handedness is more prevalent in men (1/9) than in women (1/13). Studies have shown that left-handedness is linked to the gene LRRTM1, which affects the symmetry of the brain. In addition to its genetic origins, left-handedness also has developmental origins. When both parents are left-handed, a child has a probability of 0.26 of being left-handed.

Ten families in which both parents are left-handed and have a single child are selected, and the ten children are inspected for left-handedness. Let X be the number of left-handed children among the inspected. What is the probability that X

(a) Is equal to 3?

(b) Falls anywhere between 3 and 6, inclusive?

(c) Is at most 4?

(d) Is not less than 4?

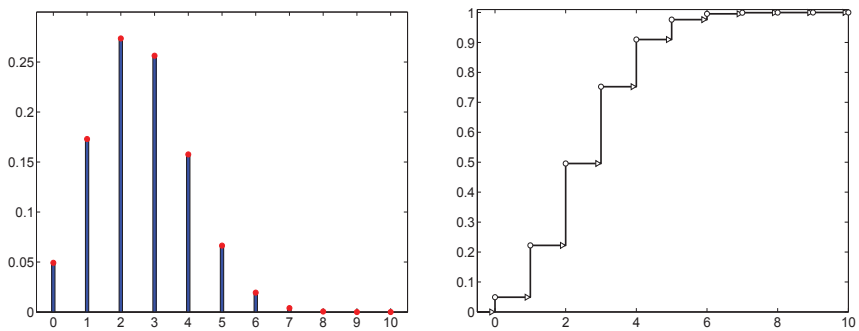
(e) Would you be surprised if the number of left-handed children among the ten inspected was eight or more? Why or why not?

The solution is given by the following annotated MATLAB script:

```
% Solution
```

```
disp('(a) Bin(10, 0.26): P(X = 3)');
binopdf(3, 10, 0.26)
 % ans = 0.2563
disp('(b) Bin(10, 0.26): P(3 <= X <= 6)');
 %  using binopdf(x, n, p)
disp('(b)-using PDF');   binopdf(3, 10, 0.26) + ...
binopdf(4, 10, 0.26) + binopdf(5, 10, 0.26)+ binopdf(6, 10, 0.26)
 % using binocdf(x, n, p)
disp('(b)-using CDF'); binocdf(6, 10, 0.26) - binocdf(2, 10, 0.26)
 % ans = 0.4998
 %(c) at most four i.e., X <= 4
disp('(c) Bin(10, 0.26): P(X <= 4)'); binocdf(4, 10, 0.26)
 % ans = 0.9096
 %(d) not less than 4 is 4,5,...,10, or complement of <=3
disp('(d)  Bin(12, 0.7): P(X >= 4)'); 1-binocdf(3, 10, 0.26)
 % ans = 0.2479
disp('(e) Bin(10, 0.26): P(X >= 8)');
1-binocdf(7, 10, 0.26)
 % ans = 5.5618e-04
 % Yes, this would be a surprising outcome since
 % the probability of such an event is rather small
```

Panels (a) and (b) in Figure 5.2 show, respectively, the PMF and CDF for the binomial $\mathcal{B}in(10, 0.26)$ distribution.



(a) (b)

Fig. 5.2 Binomial $\mathcal{B}in(10, 0.26)$: (a) PMF and (b) CDF.

How does one recognize that random variable X has a binomial distribution?

(a) It allows an interpretation as the sum of "successes" in n Bernoulli trials, for n fixed.

(b) The Bernoulli trials are independent.
(c) The Bernoulli probability p is constant for all n trials.

Next we discuss how to deal with a binomial-like framework in which condition (c) is violated.

Generalized Binomial Sampling*. Suppose that n independent experiments are performed and that an event A has a probability of p_i of appearing in the ith experiment.

We are interested in the probability that A appeared exactly k times in the n experiments. The binomial setup is not directly applicable since the probabilities of A differ from experiment to experiment. However, the binomial setup is useful as a hint on how to solve the general case. In the binomial setup the probability of k events A in n experiments is equal to the coefficient of z^k in the expansion of $G(z) = (pz+q)^n$. Indeed, $(pz+q)^n = p^n q^0 z^n + \cdots + \binom{n}{k} p^k q^{n-k} z^k + \cdots + npq^{n-1}z + p^0 q^n$.

The polynomial $G(z)$ is called the *probability-generating function*. If X is a discrete integer-valued random variable such that $p_n = \mathbb{P}(X = n)$, then its probability-generating function is defined as

$$G_X(z) = \mathbb{E}z^X = \sum_n p_n z^n.$$

Note that in the polynomial $G_X(z)$, the probability $p_n = \mathbb{P}(X = n)$ is the coefficient of the power z^n. Also, $G_X(e^z)$ is the moment-generating function $m_X(z)$.

In the general binomial setup, the polynomial $(pz+q)^n$ becomes

$$G_X(z) = (p_1 z + q_1) \times (p_2 z + q_2) \times \cdots \times (p_n z + q_n) = \sum_{i=0}^{n} a_i z^i, \qquad (5.6)$$

and the probability that there are k events A in n experiments is equal to the coefficient a_k of z^k in the polynomial $G_X(z)$. This follows from the two properties of $G(z)$: (i) When X and Y are independent, $G_{X+Y}(z) = G_X(z)\, G_Y(z)$, and (ii) if X is a Bernoulli $\mathcal{B}er(p)$, then $G_X(z) = pz + q$.

Example 5.7. **System with Unreliable Components.** Let S be a system consisting of ten unreliable components that work and fail independently of each other. The components are operational in some fixed time interval $[0, T]$ with the probabilities

```
ps =[0.5 0.3 0.2 0.5 0.6 0.4 0.2 0.4 0.7 0.8];
```

Let a random variable X represent the number of components that remain operational after time T.

Find (a) the distribution for X and (b) $\mathbb{E}X$ and $\mathbb{V}\mathrm{ar}\, X$.

```
ps =[0.5 0.3 0.2 0.5 0.6 0.4 0.2 0.4 0.7 0.8];
qs = 1- ps;
all = [ps' qs'];
[m n]= size(all);
Gz = [1];  %initial
for i = 1:m
    Gz = conv(Gz, all(i,:) );
    % conv as polynomial multiplication
end
%at the end, Gz is the product of p_i x + q_i
%
sum(Gz) %the sum is 1
probs = Gz(end:-1:1);
k = 0:10
% probs=[0.0010  0.0117  0.0578  0.1547  0.2507 ...
% 0.2582  0.1716  0.0727  0.0188  0.0027  0.0002]
EX = k * probs' %expectation 4.6
EX2 = k.^2 * probs';
VX = EX2 - (EX)^2  %variance 2.12
```

Note that in the above script we used the convolution operation `conv` to multiply polynomials, as in

```
conv([2 -1],[1 3 2])
% ans =  2   5   1  -2,
```

which is interpreted as $(2z-1)\cdot(z^2+3z+2) = 2z^3+5z^2+z-2$.

From the MATLAB calculations we find that the probability-generating function $G(z)$ from (5.6) is

$$\begin{aligned} G(z) = {} & 0.00016128z^{10} + 0.00268992z^9 + 0.01883264z^8 + 0.07273456z^7 \\ & + 0.17155808z^6 + 0.25816544z^5 + 0.25070848z^4 + 0.15470576z^3 \\ & + 0.05777184z^2 + 0.01170432z + 0.00096768, \end{aligned}$$

and the random variable X, the number of operational items, has the following distribution (after rounding to four decimal places):

X	0	1	2	3	4	5	6	7	8	9	10
Prob	0.0010	0.0117	0.0578	0.1547	0.2507	0.2582	0.1716	0.0727	0.0188	0.0027	0.0002

The answers to (b) are $\mathbb{E}X = 4.6$ and $\mathbb{V}\mathrm{ar}\,X = 2.12$.

Note that a "solution" in which one finds the average of the component probabilities, `ps`, as $\bar{p} = \frac{1}{10}(0.5+0.3+\cdots+0.8) = 0.46$, and then applies the standard binomial calculation, will lead to the correct expectation, 4.6, because of linearity. However, the variance and probabilities for X would be different. For example, the probability $\mathbb{P}(X=4)$ would be `binopdf(4,10,0.46)=0.2331,` while the correct value is 0.2507.

Example 5.8. **Surviving Pairs.**  Daniel Bernoulli (1700–1782), a nephew of Jacob Bernoulli, posed and solved the following problem. If among N

married pairs there are m random deaths, what is the expected number of intact marriages?

Suppose that there are N pairs of balls denoted by 1,1, 2,2, ..., N,N. If m balls are selected at random and removed, what is the expected number of intact pairs? Consider the pair i. Define a Bernoulli random variable Y_i equal to 1 if pair i remains intact after the removal of m balls, and 0 otherwise. Then the number of unaffected pairs N_m would be the sum of all Y_i, for $i = 1,\ldots,N$.

The probability that pair i is not affected by the removal of m balls is

$$\frac{\binom{2N-2}{m}}{\binom{2N}{m}} = \frac{\frac{(2N-2)(2N-3)\ldots(2N-2-m+2)(2N-2-m+1)}{m!}}{\frac{2N(2N-1)\ldots(2N-m+2)(2N-m+1)}{m!}} = \frac{(2N-m)(2N-m-1)}{2N(2N-1)},$$

and it is equal to $\mathbb{E}Y_i$. If N_m is the number of unaffected pairs, then

$$\begin{aligned} N_m &= Y_1 + Y_2 + \cdots + Y_N \\ \mathbb{E}N_m &= \mathbb{E}Y_1 + \mathbb{E}Y_2 + \cdots + \mathbb{E}Y_N = N\mathbb{E}Y_i = \frac{(2N-m)(2N-m-1)}{2(2N-1)}. \end{aligned}$$

For example, if among $N = 1000$ couples there are 100 random deaths, then the expected number of unaffected couples is 902.4762. If among $N = 1000$ couples there are 1936 deaths, then a single couple is expected to remain intact.

Even though N_m is the sum of N Bernoulli random variables Y_i, each with the same probability $p = \frac{(2N-m)(2N-m-1)}{2N(2N-1)}$, it does not have a binomial distribution due to the dependence among Y_is.

5.3.3 Hypergeometric Distribution

Suppose a box contains m balls, k of which are white and $m-k$ of which are black. Suppose we randomly select and remove n balls from the box *without replacement*, so that when sampling is finished, there are only $m-n$ balls left in the box. If X is the number of white balls in n selected, then the probability that $X = x$ is

$$p_X(x) = \frac{\binom{k}{x}\binom{m-k}{n-x}}{\binom{m}{n}}, \quad x \in \{0,1,\ldots,\min\{n,k\}\}.$$

Random variable X is called hypergeometric and denoted by $X \sim \mathcal{HG}(m,k,n)$, where m,k, and n are integer parameters.

This PMF can be deduced by counting rules. There are $\binom{m}{n}$ different ways of selecting the n balls from a box with a total of m balls. From these (each equally likely), there are $\binom{k}{x}$ ways of selecting x white balls from the k white balls in the box and, similarly, $\binom{m-k}{n-x}$ ways of choosing the black balls. The probability $\mathbb{P}(X = x)$ is the ratio of these two numbers. The PDF and CDF of $\mathcal{HG}(40,15,10)$ are shown in Figure 5.3.

It can be shown that the mean and variance for the hypergeometric distribution are, respectively,

$$\mathbb{E}X = n\frac{k}{m} \quad \text{and} \quad \mathbb{V}\text{ar}\, X = n\frac{k}{m}\left(1 - \frac{k}{m}\right)\frac{m-n}{m-1}.$$

The MATLAB commands for hypergeometric CDF, PDF, quantile, and a random number are `hygecdf`, `hygepdf`, `hygeinv`, and `hygernd`. WinBUGS does not have a built-in command for a hypergeometric distribution.

Example 5.9. **CASES.** In a group of 40 people, 15 are "CASES" and 25 are "CONTROLS." A sample of 10 subjects is selected [(A) with replacement and (B) without replacement]. Find the probability $\mathbb{P}$(at least 2 subjects are CASES).

```
%Solution
%(A) - with replacement (binomial case);
%Let X be the number of CASES. The event
%X is at least 2 is the complement of X <= 1.
disp('(A) Bin(10, 15/40): P(X >= 2)');  1 - binocdf(1, 10, 15/40)
% ans = 0.9363
% or
 1 - binopdf(0, 10, 15/40) - binopdf(1, 10, 15/40)
% ans = 0.9363

%B - without replacement (hypergeometric case) hygecdf(x, m, k, n)
% where m size of population,
% k - number of cases among m,  and n sample size.
disp('(B) HyGe(40,15,10): P(X >=2)'); 1 - hygecdf(1, 40, 15, 10)
% ans = 0.9600,  or
 1 - hygepdf(0, 40, 15, 10)- hygepdf(1, 40, 15, 10)
% ans = 0.9600
```


Example 5.10. **Capture–Recapture Models.** Suppose that an unknown number m of animals inhabit a particular region. To assess the population size, ecologists often apply the following capture–recapture scheme. They catch k animals, tag them, and release them back into the region. After some time, when the tagged animals are expected to be mixed well with the untagged, a second catch of size n is made. Suppose that x animals in the second sample are found to be tagged.

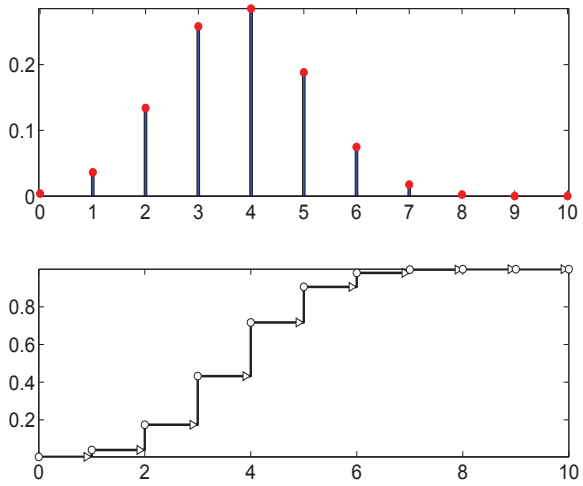


Fig. 5.3 The PDF and CDF for hypergeometric distribution with $m = 40, k = 15$, and $n = 10$.

If catching any animal is assumed equally likely, the number x of tagged animals in the second sample is hypergeometric $\mathcal{HG}(m,k,n)$. Ecologists use the observed ratio x/n as an approximation to k/m, from which m is estimated as

$$\hat{m} = \frac{k \times n}{x}.$$

A statistically better estimator of m (known as the Schnabel formula) is given as

$$\hat{m} = \frac{(k+1) \times (n+1)}{(x+1)} - 1.$$

In epidemiology and public health, capture–recapture methods use multiple, routinely collected, computerized data sources to estimate various population indexes.

For example, Gjini et al. (2004) investigated the number of matching records of pneumococcal meningitis among adults in England by comparing data from Hospital Episode Statistics (HES) and the Public Health Laboratory Services reconciled laboratory records (RLR). The time period covered was April 1996 to December 1999. The authors found 646 records in RLR and 737 in HES, and matching based on demographic information was possible in 296 cases.

By the capture–recapture method the estimated incidence is $\hat{m} = 646 \cdot 737/296 = 1608.5 \approx 1609$. If Schnabel's formula is used, then $\hat{m} \approx 1607$.

Thus, the total incidence of of pneumococcal meningitis in England between April 1996 to December 1999 is estimated to be 1607.


For large m, the hypergeometric distribution is close to binomial. More precisely, when $m \to \infty$ and $k/m \to p$, the hypergeometric distribution with parameters (m,k,n) approaches a binomial with parameters (n,p) for any value of x between 0 and n. It is also instructive to compare expressions for $\mathbb{E}X$ and $\mathbb{V}\mathrm{ar}\,X$ for the two distributions.

```
format long
disp('(A)=(B) for large population');
1 - binocdf(1, 10, 150000/400000)   %ans = 0.936335370875895
1 - hygecdf(1, 400000, 150000, 10)  %ans = 0.936337703719839
```

We will use the hypergeometric distribution later in the book (i) in the Fisher exact test (page 602) and in Logrank test (page 818).

5.3.4 Poisson Distribution

This discrete distribution is named after Simeon Denis Poisson (1781–1840), French mathematician, geometer, and physicist.

The PMF for the Poisson distribution is

$$p_X(x) = \frac{\lambda^x}{x!}e^{-\lambda}, \quad x = 0,1,2,\ldots,$$

which is denoted by $X \sim \mathcal{P}oi(\lambda)$. From the moment-generating function $m_X(t) = \exp\{\lambda(e^t - 1)\}$ we have $\mathbb{E}X = \lambda$ and $\mathbb{V}\mathrm{ar}\,X = \lambda$; the mean and the variance coincide.

The sum of a finite independent set of Poisson variables is also Poisson. Specifically, if $X_i \sim \mathcal{P}oi(\lambda_i)$, then $Y = X_1 + \cdots + X_k$ is distributed as $\mathcal{P}oi(\lambda_1 + \cdots + \lambda_k)$. If $X_1 \sim \mathcal{P}oi(\lambda_1)$ and $X_2 \sim \mathcal{P}oi(\lambda_2)$ are independent, then the distribution of X_1 given that $X_1 + X_2 = n$ is binomial $\mathcal{B}in\left(n, \frac{\lambda_1}{\lambda_1+\lambda_2}\right)$ (Exercise 5.7).

Furthermore, the Poisson distribution is a limiting form for a binomial model, i.e.,

$$\lim_{n,np\to\infty,\lambda} \binom{n}{x} p^x(1-p)^{n-x} = \frac{1}{x!}\lambda^x e^{-\lambda}. \tag{5.7}$$

The MATLAB commands for Poisson CDF, PDF, quantile, and random number are `poisscdf`, `poisspdf`, `poissinv`, and `poissrnd`. In WinBUGS the Poisson distribution is denoted as `dpois(lambda)`.

Example 5.11. **Poisson Model for EBs.** After 7 days of aggregation, the microscopy images of 2000 embryonic bodies (EBs) are used to assess their surface area size. The probability that the area of a randomly selected EB exceeds the critical size S_c is 0.001.

(a) Find the probability that the areas of exactly three EBs, among the 2000, exceed the critical size.

(b) Find the probability that the number of EBs exceeding the critical size is between three and eight, inclusively.

We use a Poisson approximation to the binomial probabilities since n is large, p is small, and product np is moderate.

```
%Solution
disp('Poisson(2): P(X=3)'); poisspdf(3, 2)
  %ans= 0.1804
disp('Poisson(2): P(3 <= X <= 8)'); poisscdf(8, 2)-poisscdf(2, 2)
  %ans= 0.3231
```

Figure 5.4 shows the PMF and CDF of the $\mathcal{P}oi(2)$ distribution.

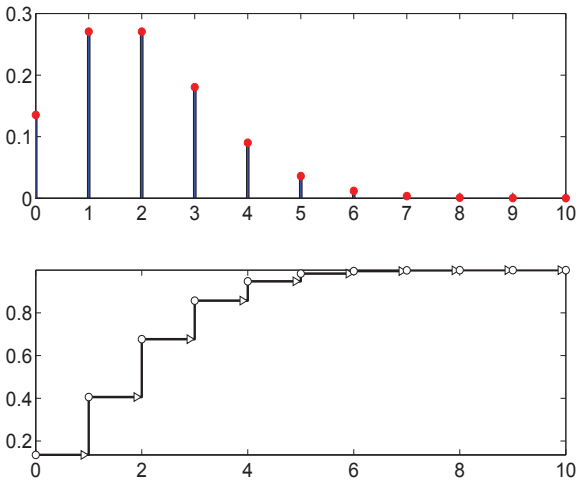


Fig. 5.4 (Top) Poisson probability mass function. (Bottom) Cumulative distribution function for $\lambda = 2$.

In the binomial sampling scheme, when $n \to \infty$ and $p \to 0$, so that $np \to \lambda$, binomial probabilities converge to Poisson probabilities.

The following MATLAB simulation demonstrates the convergence. In the binomial distribution, n is increasing as 2,000, 200,000, 20,000,000 and p is decreasing as 0.001, 0.00001, 0.0000001, so that np remains constant

and equal to 2. Then, the binomial probabilities of $X = 3$ are compared to probability of $X = 3$ when X is distributed as Poisson with parameter $\lambda = 2$.

```
disp('P(X=3) for Bin(2000, 0.001), Bin(200000, 0.00001), ');
disp(' Bin(20000000, 0.0000001), and Poi(2) ');
format long
binopdf(3, 2000, 0.001)           % 0.180537328031786
binopdf(3, 200000, 0.00001)       % 0.180447946554779
binopdf(3, 20000000, 0.0000001)   % 0.180447058859339
poisspdf(3, 2)                    % 0.180447044315484
format short
```

Example 5.12. **Cold.** Suppose that the number of times during a year that an individual catches a cold can be modeled by a Poisson random variable with an expectation of 4. Further, suppose that a new drug based on vitamin C reduces this expectation to 3 (but the distribution still remains Poisson) for 90% of the population but has no effect on the remaining 10% of the population. We will calculate

(a) the probability that an individual taking the drug has two colds in a year if that individual is in part of the population that benefits from the drug;

(b) the probability that a randomly chosen individual has two colds in a year if that individual takes the drug; and

(c) the conditional probability that a randomly chosen individual is in the part of the population that benefits from the drug, given that the individual had two colds in the year during which he/she took the drug.

```
poisspdf(2,3)   %(Cold (a))
     %ans =  0.2240
poisspdf(2,3)*0.90 + poisspdf(2,4)*0.10   %(Cold (b))
     %ans = 0.2163
poisspdf(2,3)*0.90/(poisspdf(2,3)*0.90 + ...
                    poisspdf(2,4)*0.10) %(Cold (c))
     %ans = 0.9323
```

Example 5.13. **Imperfectly Observed Poisson.** Suppose that the number of particular experimental events in time interval $[0, T]$ has a Poisson distribution $\mathcal{P}oi(\lambda T)$. A student who is observing the experiment may fail to count some of the events. An event is counted with probability equal to p and missing one event is independent of missing or counting the others. What is the distribution of the number of events in $[0, T]$ that are counted?

By total probability formula,

$$\begin{aligned}\mathbb{P}(n \text{ events counted}) &= \sum_{k=n}^{\infty} (\mathbb{P}(n \text{ events counted} | k \text{ events happened}) \\ &\quad \times \mathbb{P}(k \text{ events happened}) \\ &= \sum_{k=n}^{\infty} \binom{k}{n} p^n (1-p)^{k-n} (\lambda T)^k \exp\{-\lambda T\}/k! \\ &= \exp\{-\lambda T\}(p\lambda T)^n/n! \sum_{k=n}^{\infty} \frac{[(1-p)\lambda T]^{k-n}}{(k-n)!} \\ &= (p\lambda T)^n \exp\{-p\lambda T\}/n!\end{aligned}$$

after representing $\binom{k}{n}$ by factorials and observing that $\sum_{k=n}^{\infty} \frac{[(1-p)\lambda T]^{k-n}}{(k-n)!} = \sum_{v=0}^{\infty} \frac{[(1-p)\lambda T]^v}{v!} = \exp\{(1-p)\lambda T\}$. Thus, the number of counted events is again Poisson but with the rate $p\lambda T$.
✎

5.3.5 *Geometric Distribution*

Suppose that independent trials are repeated and that in each trial the probability of a success is equal to $0 < p < 1$. We are interested in the number of failures X before the first success. The number of failures is a random variable with a geometric $\mathcal{G}e(p)$ distribution. Its PMF is given by

$$p_X(x) = p(1-p)^x, \quad x = 0,1,2,\ldots.$$

The expected value is $\mathbb{E}X = (1-p)/p$ and the variance is $\mathbb{V}\mathrm{ar}\, X = (1-p)/p^2$. The moments can be found either directly or by the moment-generating function, which is

$$m_X(t) = \frac{p}{1-(1-p)e^t}.$$

The geometric random variable possesses a "memoryless" property. That is, if we condition on the event $X \geq m$, for some nonnegative integer m, then for $n \geq m$, $\mathbb{P}(X \geq n | X \geq m) = \mathbb{P}(X \geq n-m)$ (Exercise 5.25). The MATLAB commands for geometric CDF, PDF, quantile, and random number are `geocdf`, `geopdf`, `geoinv`, and `geornd`. There are no special names for the geometric distribution in WinBUGS; the negative binomial can be used as `dnegbin(p,1)`.

If instead of the number of failures before the first success (X) one is interested in the total number of experiments until the first success (Y), then the relationship is simple: $Y = X + 1$. In this formulation of the geometric

distribution, $Y \sim \mathcal{G}eom(p)$, $\mathbb{E}Y = \mathbb{E}X + 1 = q/p + 1 = 1/p$, and $\mathbb{V}\text{ar}\, Y = \mathbb{V}\text{ar}\, X = (1-p)/p^2$.

Example 5.14. **CASES I.** Let a subject constitute either a CASE or a CONTROL depending on whether the subject's level of LDL cholesterol is >160 mg/dL or ≤160 mg/dL, respectively. According to a recent National Health and Nutrition Examination Survey (NHANES III), the prevalence of CASES among white male Americans aged 20 and older (target population) is $p = 20\%$. Subjects are sampled (when the population is large, it is unimportant if the sampling is done with or without replacement) until the first CASE is found. The number of CONTROLS sampled before finding the first CASE is a geometric random variable with parameter $p = 0.2$ (Fig. 5.5).

(a) Find the probability that seven CONTROLS will be sampled before we come across the first CASE.

(b) Find the probability that the number of CONTROLS before the first CASE will fall between four and eight, inclusively.

```
disp('X ~ Geometric(0.2):P(X=7)');
geopdf(7, 0.2)
     %ans=0.0419
disp('X ~ Geometric(0.2):P(4 <= X <= 8)');
geocdf(8, 0.2) - geocdf(3,0.2)
     %ans=0.2754
```

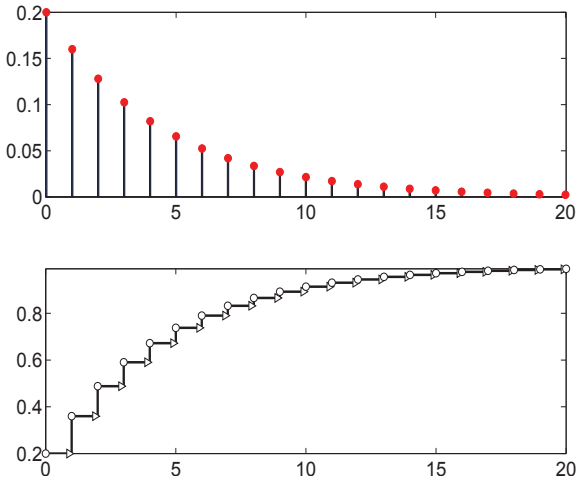


Fig. 5.5 Geometric (Top) PMF and (Bottom) CDF for $p = 0.2$.

Example 5.15. **Mingling Trees.** The degree to which the individual trees of two species are mingled together is an intrinsic property of a two-species population. Two species are said to be segregated if the individuals of each tend to have a member of their own species as nearest neighbor, rather than a member of the other species. To assess the segregation, Pielou (1961) developed a field experiment in which alternating uninterrupted runs of *Pseudotsuga menziesii* and *Pinus ponderosa* are measured along a narrow long belt.

The data in table give lengths of runs Y and their frequency.

Run length, Y	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	21	20	21	4	6	6	2	3	3	1	0	1

(a) Assuming geometric $\mathcal{G}eom(1/3)$ for Y, or equivalently $\mathcal{G}e(1/3)$ distribution for $X = Y - 1$, find the mean $\mathbb{E}Y$, variance $\mathbb{V}\text{ar}\,Y$, and $\mathbb{P}(Y > 5|Y > 2)$. Is this probability the same as $\mathbb{P}(Y > 5 - 2)$ (memoryless property)?

(b) What are sample counterparts of quantities from (a)?

```
%mingling.m
%(a)
[ex varx] = geostat(1/3)
ey = ex+1    %E(Y)=1/(1/3)=3
vary = varx  %Var(Y)=((1-1/3)/((1/3)^2)=6
%P(Y>5|Y>2)=P(Y>5)/P(Y>2)=P(X>=5)/P(X>=2)
(1-geocdf(4, 1/3))/(1-geocdf(1, 1/3))   %0.2963
% memoryles
%P(Y>5-2)=P(Y>3)=P(X>=3)
1-geocdf(2, 1/3)    %0.2963

%(b) empirical counterparts to (a)
Y=1:12;
freq   =[21  20  21   4   6   6  2   3   3   1   0   1];
n=sum(freq)      %88
ybar = sum(Y .* freq)/n  %3.3295
s2y = sum((Y -ybar).^2  .* freq)/(n-1)  %5.9936
sum(freq(Y >5))/sum(freq(Y>2))    %0.3404
sum(freq(Y>3))/n    %0.2955
```

Note that geometric $\mathcal{G}eom(1/3)$ distribution provides a good model for the data, as evidenced by the closeness of empirical moments and probabilities to their theoretical counterparts. Later in the text (Chapter 7 and Chapter 17) we will learn how to, given the data, set the model, estimate parameters, and assess the goodness of model fit.

5.3.6 Negative Binomial Distribution

The negative binomial distribution was formulated by Montmort (1714). Here we are dealing with independent trials again. This time we count the number of failures observed until a fixed number of successes ($r \geq 1$) occur. Let p be the probability of success in a single trial.

If we observe r consecutive successes at the start of the experiment, then the count of failures is $X = 0$ and $\mathbb{P}(X = 0) = p^r$. If $X = x$, then we have observed x failures and r successes in $x + r$ trials. There are $\binom{x+r}{x}$ different ways of arranging x failures in those $x + r$ trials, but we can only be concerned with those arrangements in which the last trial ended in a success. So there are really only $\binom{x+r-1}{x}$ equally likely arrangements. For any particular arrangement, the probability is $p^r(1-p)^x$. Therefore, the PMF is

$$p_X(x) = \binom{r+x-1}{x} p^r(1-p)^x, \quad x = 0,1,2,\ldots.$$

Sometimes this PMF is stated with $\binom{r+x-1}{r-1}$ in place of the equivalent $\binom{r+x-1}{x}$. This distribution is denoted as $X \sim \mathcal{NB}(r,p)$.

From its moment-generating function

$$m_X(t) = \left(\frac{p}{1-(1-p)e^t}\right)^r,$$

the expectation of a negative binomial random variable is $\mathbb{E}X = r(1-p)/p$ and its variance is $\mathbb{V}\mathrm{ar}\, X = r(1-p)/p^2$.

Since the negative binomial $X \sim \mathcal{NB}(r,p)$ is a convolution (a sum) of r independent geometric random variables, $X = Y_1 + Y_2 + \cdots + Y_r$, $Y_i \sim \mathcal{G}(p)$, the mean and variance of X can be easily derived from the mean and variance of its geometric components Y_i, as in (5.2) and (5.3). Note also that $m_X(t) = (m_Y(t))^r$, where $m_Y(t) = \left(\frac{p}{1-(1-p)e^t}\right)$ is the moment-generating function of the component Y_i in the sum. This is a consequence of the fact that a moment-generating function for a sum of independent random variables is the product of the moment-generating functions of the components; see (5.5).

The distribution remains valid if r is not an integer, although an interpretation involving r successes is lost. For an arbitrary nonnegative r, the distribution is called a Pólya distribution or a generalized negative binomial distribution (although this second term can be ambiguous since several generalizations exist). The constant $\binom{r+x-1}{x} = \frac{(r+x-1)!}{x!(r-1)!}$ is replaced by $\frac{\Gamma(r+x)}{x!\Gamma(r)}$, keeping in mind that $\Gamma(n) = (n-1)!$ when n is an integer. The

Pólya distribution is used in ecology for inference about the abundance of species in nature.

The MATLAB commands for negative binomial CDF, PDF, quantile, and random number are `nbincdf`, `nbinpdf`, `nbininv`, and `nbinrnd`. In WinBUGS the negative binomial distribution is denoted as `dnegbin(p,r)`. Note the opposite order of parameters r and p compared to notation $\mathcal{NB}(r,p)$ and the order adopted by MATLAB.

Example 5.16. **CASES II.** Assume as in Example 5.14 that the prevalence of "CASES" in a large population is $p = 20\%$. Subjects are sampled, one by one, until seven CASES are found and then the sampling is stopped.

(a) What is the probability that the number of CONTROLS among all sampled subjects will be 18?

(b) What is the probability of observing more than the "expected number" of CONTROLS?

The number of CONTROLS X among all sampled subjects is a negative binomial, $X \sim \mathcal{NB}(7, 0.2)$.

$$\mathbb{P}(X = 18) = \binom{25+7-1}{18} 0.2^7 (1-0.2)^1 8 = 0.0310.$$

Also, `nbinpdf(18,7,0.2)=0.0310`.

Thus, with a probability of 0.031 the number of CONTROLS sampled, before seven CASES are observed, is equal to 18.

(b) The expected number of CONTROLS is $\mathbb{E}X = 7\frac{0.8}{0.2} = 28$. The probability of $X > \mathbb{E}X$ is $\mathbb{P}(X > 28) = 1 - \mathbb{P}(X \le 28) = 1 - \sum_{x=0}^{28} \binom{7+x-1}{x} 0.8^x 0.2^7 = 0.4328$. In MATLAB $\mathbb{P}(X > 28)$ is calculated as `1-nbincdf(28,7,0.20)=0.4328`. ✎

The tail probabilities of a negative binomial distribution can be expressed by binomial probabilities. If $X \sim \mathcal{NB}(r,p)$, then

$$\mathbb{P}(X > x) = \mathbb{P}(Y < r),$$

where $Y \sim \mathcal{B}in(x+r, p)$. In words, if we have not seen r successes after seeing x failures, then in $x+r$ experiments the number of successes will be less than r. In part (b) of the previous example, $r = 7, x = 28$, and $p = 0.20$, so

```
1 - nbincdf(28, 7, 0.20)  % 0.4328
binocdf(7-1, 28+7, 0.20)  % 0.4328
```

5.3.7 Multinomial Distribution

The binomial distribution was developed by counting the occurrences two complementary events, A and A^c, in n independent trials. Suppose, instead,

that each trial results in one of $k > 2$ mutually exclusive events, $A_1, \dots A_k$, so that $\mathcal{S} = A_1 \cup \cdots \cup A_k$. One can define the vector of random variables $(X_1, \dots, X_k)$ where a component X_i counts how many times A_i appeared in n trials. The defined random vector is called multinomial.

The probability mass function for $(X_1, \dots, X_k)$ is

$$p_{X_1,\dots,X_k}(x_1,\dots,x_k) = \frac{n!}{x_1! \cdots x_k!} {p_1}^{x_1} \cdots {p_k}^{x_k},$$

where $p_1 + \cdots + p_k = 1$ and $x_1 + \cdots + x_k = n$. Since $p_k = 1 - p_1 - \cdots - p_{k-1}$, there are $k-1$ free parameters to characterize the multinomial distribution, which is denoted by $X = (X_1, \dots, X_k) \sim \mathcal{M}n(n, p_1, \dots, p_k)$.

The mean and variance of the component X_i are the same as in the binomial case. It is easy to see that the marginal distribution for a component X_i is binomial since the events $A_1, \dots, A_k$ can be grouped as A_i, A_i^c. Therefore, $\mathbb{E}(X_i) = np_i$, $\mathbb{V}\text{ar}(X_i) = np_i(1-p_i)$. The components X_i are dependent since they sum up to n. For $i \neq j$, the covariance between X_i and X_j is

$$\mathbb{C}\text{ov}(X_i, X_j) = \mathbb{E}X_iX_j - \mathbb{E}X_i\mathbb{E}X_j = -np_ip_j. \tag{5.8}$$

This is easy to verify if X_i and X_j are represented as the sums of Bernoullis $Y_{i1} + Y_{i2} + \cdots + Y_{ik}$ and $Y_{j1} + Y_{j2} + \cdots + Y_{jk}$, respectively. Since $Y_{im}Y_{jm} = 0$ (in a single trial A_i and A_j cannot occur simultaneously), it follows that

$$\mathbb{E}X_iX_j = (n^2 - n)p_ip_j.$$

Since $\mathbb{E}X_i\mathbb{E}X_j = n^2p_ip_j$, the covariance in (5.8) follows.

If $X = (X_1, X_2, \dots, X_k) \sim \mathcal{M}n(n, p_1, p_2, \dots, p_k)$, then $X' = (X_1 + X_2, \dots, X_k) \sim \mathcal{M}n(n, p_1 + p_2, \dots, p_k)$. This is called the *fusing* property of the multinomial distribution.

If $X_1 \sim \mathcal{P}oi(\lambda_1)$, $X_2 \sim \mathcal{P}oi(\lambda_2)$, …, $X_n \sim \mathcal{P}oi(\lambda_n)$ are n independent Poisson random variables with parameters $\lambda_1, \dots, \lambda_n$, then the conditional distribution of $X_1, X_2, \dots, X_n$, given that $X_1 + X_2 + \cdots + X_n = n$, is $\mathcal{M}n(n, p_1, \dots, p_k)$, where $p_i = \lambda_i/(\lambda_1 + \lambda_2 + \cdots + \lambda_n)$. This fact is used in modeling contingency tables with a fixed total and will be discussed in Chapter 12.

In MATLAB, the multinomial PMF is calculated by `mnpdf(x,p)`, where x is a $1 \times k$ vector of values, such that $\sum_{i=1}^k x_i = n$, and p is a $1 \times k$ vector of probabilities, such that $\sum_{i=1}^k p_i = 1$.

For example,

```
%If n=2, Multinomial is Binomial
mnpdf([5 15],[0.6 0.4])
          %ans = 0.0013
% is the same as
 binopdf(5, 5+15, 0.6)
          %ans = 0.0013
```

In WinBUGS, the multinomial distribution is coded as `dmulti(p[],n)`.

Example 5.17. **ABO Group Distribution.** Suppose that the probabilities of blood groups in a particular population are given as

O	A	B	AB
0.37	0.39	0.18	0.06

If eight subjects are selected at random from this population, what is the probability that

(a) $(O,A,B,AB) = (3,4,1,0)$?

(b) $O = 3$?

In (a), the probability is

```
factorial(8) /(factorial(3) * ...
    factorial(4) * factorial(1) * factorial(0)) * ...
    0.37^3 * 0.39^4 * 0.18^1 * 0.06^0
%ans = 0.0591
%or
 mnpdf([3  4  1  0],[0.37 0.39 0.18 0.06])
%ans = 0.0591.
```

In (b), $O \sim \mathcal{B}in(8,0.37)$ and $\mathbb{P}(O = 3) = 0.2815$.

5.3.8 Quantiles

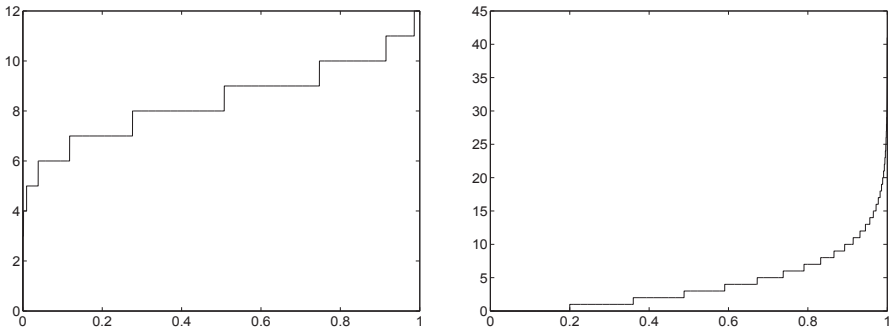
Quantiles of random variables are defined as follows. A p-quantile (or $100 \times p$ percentile) of random variable X is the value x for which $F(x) = p$, if F is a monotone cumulative distribution function for X. For an arbitrary random variable, including discrete, this definition is not unique and modification is needed:

$$F(x) = \mathbb{P}(X \le x) \ge p \quad \text{and} \quad \mathbb{P}(X \ge x) \ge 1 - p.$$

For example, the 0.05 quantile of a binomial distribution with parameters $n = 12$ and $p = 0.7$ is $x = 6$ since $\mathbb{P}(X \le 6) = 0.1178 \ge 0.05$ and $\mathbb{P}(X \ge 6) = 1 - \mathbb{P}(X \le 5) = 1 - 0.0386 = 0.9614 \ge 0.95$. Binomial $\mathcal{B}in(12,0.7)$ and geometric $\mathcal{G}(0.2)$ quantiles are shown in Figure 5.6.

```
quab =[]; quag =[];
for p = 0.00:0.0001:1
    quab = [quab binoinv(p, 12, 0.7)];
    quag = [quag geoinv(p, 0.2)];
end
figure(1)
```

```
plot([0.00:0.0001:1],quab,'k-')
figure(2)
plot([0.00:0.0001:1],quag,'k-')
```



(a) (b)

Fig. 5.6 (a) Binomial $\mathcal{B}in(12,0.7)$ and (b) geometric $\mathcal{G}(0.2)$ quantiles.

5.4 Continuous Random Variables

Continuous random variables take values within an interval (a,b) on a real line **R**. The probability density function (PDF) $f(x)$ fully specifies the variable. The PDF is nonnegative, $f(x) \geq 0$, and integrates to 1, $\int_R f(x)\,dx = 1$. The probability that X takes a value in an interval (a,b) (and for continuous random variables equivalently $[a,b), (a,b]$, or $[a,b]$) is $\mathbb{P}[X \in (a,b)] = \int_a^b f(x)dx$.

The CDF is

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

In terms of the CDF, $\mathbb{P}[X \in (a,b)] = F(b) - F(a)$.

The expectation of X is given by

$$\mathbb{E}X = \int_{\mathbf{R}} xf(x)dx.$$

The expectation of a function of a random variable $g(X)$ is

$$\mathbb{E}g(X) = \int_{\mathbf{R}} g(x)f(x)dx.$$

The kth moment of a continuous random variable X is defined as

$$m_k = \mathbb{E}X^k = \int_{\mathbf{R}} x^k f(x)dx,$$

and the kth central moment is

$$\mu_k = \mathbb{E}(X - \mathbb{E}X)^k = \int_{\mathbf{R}} (x - \mathbb{E}X)^k f(x)dx.$$

As in the discrete case, the first moment is the expectation and the second central moment is the variance, $\mu_2 = \mathbb{V}\mathrm{ar}\,(X) = \mathbb{E}(X - \mathbb{E}X)^2$. The skewness and kurtosis of X are defined via the central moments as in the discrete case (5.1),

$$\gamma = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mathbb{E}(X - \mathbb{E}X)^3}{(\mathbb{V}\mathrm{ar}\,(X))^{3/2}} \quad \text{and} \quad \kappa = \frac{\mu_4}{\mu_2^2} = \frac{\mathbb{E}(X - \mathbb{E}X)^4}{(\mathbb{V}\mathrm{ar}\,(X))^2}.$$

The moment-generating function of a continuous random variable X is

$$m(t) = \mathbb{E}e^{tX} = \int_{\mathbf{R}} e^{tx} f(x)dx.$$

Since $m^{(k)}(t) = \int_{\mathbf{R}} x^k e^{tx} f(x)dx$, $\mathbb{E}X^k = m^{(k)}(0)$. Moment-generating functions are related to Laplace transforms of densities. Since the bilateral Laplace transform of $f(x)$ is defined as

$$\mathcal{L}(f) = \int_{\mathbf{R}} e^{-tx} f(x)dx,$$

it holds that $m(-t) = \mathcal{L}(f)$.

The entropy of a continuous random variable with a density $f(x)$ is defined as

$$\mathcal{H}(X) = -\int_{\mathbb{R}} f(x) \log f(x)dx,$$

whenever this integral exists. Unlike the entropy for discrete random variables, $\mathcal{H}(X)$ can be negative and not necessarily invariant with respect to a transformation of X.

Example 5.18. **Markov's Inequality.** If X is a random variable that takes only nonnegative values, then for any positive constant a,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}X}{a}. \tag{5.9}$$

Indeed,

$$\begin{aligned}\mathbb{E}X &= \int_0^\infty xf(x)dx \geq \int_a^\infty xf(x)dx \\ &\geq \int_a^\infty af(x)dx \\ &= a\int_a^\infty f(x)dx = a\mathbb{P}(X \geq a).\end{aligned}$$

An average mass of a single cell of *E. coli* bacterium is 665 fg (femtogram, fg = 10^{-15}g). If a particular cell of *E. coli* is inspected, what can be said about the probability that its weight will exceed 1000 fg? According to Markov's inequality, this probability does not exceed 665/1000 = 0.665.
✎

Example 5.19. **Durability of the Starr–Edwards Valve.** The Starr–Edwards valve is one of the oldest cardiac valve prostheses in the world. The first aortic valve replacement (AVR) with a Starr–Edwards metal cage and silicone ball valve was performed in 1961. Follow-up studies have documented the excellent durability of the Starr–Edwards valve as an AVR. Suppose that the durability of the Starr–Edwards valve (in years) is a random variable X with density

$$f(x) = \begin{cases} ax^2/100, & 0 < x < 10, \\ a(x-30)^2/400, & 10 \leq x \leq 30, \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the constant a.
(b) Find the CDF $F(x)$ and sketch graphs of f and F.
(c) Find the mean and 60th percentile of X. Which is larger? Find the variance.
Solution: (a) Since $1 = \int_R f(x)dx$,

$$\begin{aligned}1 &= \int_0^{10} ax^2/100dx + \int_{10}^{30} a(x-30)^2/400dx \\ &= ax^3/300\Big|_0^{10} + a(x-30)^3/1200\Big|_{10}^{30}.\end{aligned}$$

This gives $1000a/300 - 0 + 0 - (-20)^3a/1200 = 10a/3 + 20a/3 = 10a = 1$, that is, $a = 1/10$. The density is

$$f(x) = \begin{cases} x^2/1000, & 0 < x < 10, \\ (x-30)^2/4000, & 10 \leq x \leq 30, \\ 0, & \text{otherwise.} \end{cases}$$

(b) The CDF is

$$F(x) = \begin{cases} 0, & x < 0, \\ x^3/3000, & 0 < x < 10, \\ 1 + (x-30)^3/12000, & 10 \le x \le 30, \\ 1, & x \ge 30. \end{cases}$$

(c) The 60th percentile is a solution to the equation $1 + (x-30)^3/12000 = 0.6$ and is $x = 13.131313\ldots$. The mean is $\mathbb{E}X = 25/2$, and the 60th percentile exceeds the mean. $EX^2 = 180$; thus the variance is $\mathbb{V}\text{ar}\, X = 180 - (25/2)^2 = 95/4 = 23.75$.
✎

Example 5.20. **Soliton Waves and Sech Distribution.** Soliton waves were first described by John Scott Russell, a Scottish civil engineer. In August 1834 he was riding beside the Union Canal near Edinburgh, Scotland, and noticed a strange wave building up at the bow of a boat. After the boat stopped, the wave traveled on, "assuming the form of a large solitary elevation, a rounded, smooth and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed." Soliton waves appear within the ocean and the atmosphere, within magnets and super-cooled devices, within the ionized plasma of space, and in optical fibers, to list a few.

The envelope of a soliton wave (Fig. 5.7a), properly scaled, is a probability density as is described next. Let X be a continuous random variable with the density

$$f(x) = \frac{2}{e^{\pi x} + e^{-\pi x}},\ x \in R. \tag{5.10}$$

This function is in fact hyperbolic secant of argument πx, motivating the name "sech,"

$$f(x) = \text{sech}(\pi x),\ x \in R.$$

The density is shown in Figure 5.7b. The odd moments for this distribution are 0, and a few even moments are

$$EX^2 = 1/4, \quad EX^4 = 5/16, \quad EX^6 = 61/64, \quad EX^8 = 1385/256, \ldots .$$

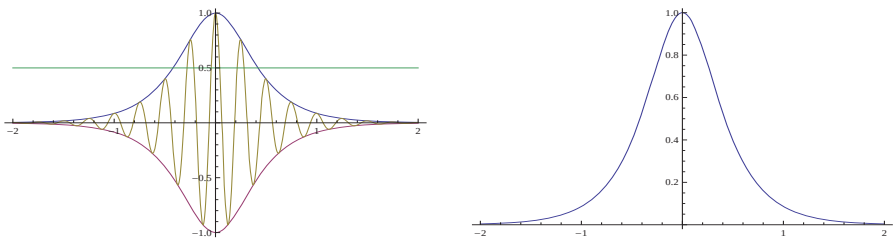
(a) What are the skewness and kurtosis of this distribution?

(b) Calculate the 0.25- and 0.75-quantiles of this distribution.

(c) Find the "width" of the sech envelope, defined as the length of the line segment at height 0.5 that falls inside the envelope, see Figure 5.7a.

(d) What is the probability of random variable X with sech distribution to fall within the "width" range?

Solution. Since this distribution is symmetric about 0, the central moments are equal to raw moments. The skewness $\gamma = 0$, and kurtosis is $\kappa = \frac{5/16}{(1/4)^2} = 5$.



(a) (b)

Fig. 5.7 (a) Soliton waves and (b) density of sech distribution.

(b) By representing (5.10) as

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

and taking the substitution $t = e^{\pi x}$ in the integral $F(x) = \int_{-\infty}^{x} f(t)dt$, we find the CDF,

$$F(x) = \frac{2}{\pi} \arctan(e^{\pi x}), \; x \in R.$$

Since $F(x)$ is monotone and one-to-one, its inverse is unique and represents a quantile function for this distribution. For $F(x) = p$, it is easy to find

$$x = \frac{1}{\pi} \log\left(\tan\left(\frac{\pi p}{2}\right)\right),$$

which for $p = 0.25$ gives $x_{0.25} = -0.2805$. Because of symmetry, $x_{0.75} = 0.2805$.

```
p=0.25;   x25=1/pi * log( tan(pi * p/2))   %-0.2805
```

(c) The solution of $f(x) = 1/2$ can be found in finite form, $x_{1/2} = \frac{1}{\pi}\log(2 \pm \sqrt{3}) = \pm 0.4192$. The length of segment inside the envelope is

$$x_2 - x_1 = \frac{1}{\pi} \log \frac{2 + \sqrt{3}}{2 - \sqrt{3}} = 0.8384.$$

In MATLAB,

```
fzero(@(x) sech(pi * x) - 1/2, 1)   % 0.4192
fzero(@(x) sech(pi * x) - 1/2,-1)   %-0.4192
```

(d) The required probability is $2/3$. Numerically,

```
format long
sechcdf = @(x) 2/pi * atan( exp(pi * x));
```

```
x2=1/pi * log(2 + sqrt(3));
prob =sechcdf(x2)-sechcdf(-x2) %0.666666666666667
```

This probability can be obtained analytically by observing that $\tan\frac{\pi}{12} = 2-\sqrt{3}$ and $\tan\frac{5\pi}{12} = 2+\sqrt{3}$.
✎

5.4.1 Joint Distribution of Two Continuous Random Variables

Two random variables X and Y are jointly continuous if there exists a non-negative function $f(x,y)$ so that for any two-dimensional domain D,

$$\mathbb{P}((X,Y)\in D) = \int\int_D f(x,y)dxdy.$$

When such a two-dimensional density $f(x,y)$ exists, it is a repeated partial derivative of the cumulative distribution function $F(x,y) = \mathbb{P}(X\le x, Y\le y)$,

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x\,\partial y}.$$

The marginal densities for X and Y are, respectively, $f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$ and $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$. The conditional distributions of X when $Y=y$ and of Y when $X=x$ are

$$f(x|y) = f(x,y)/f_Y(y) \quad \text{and} \quad f(y|x) = f(x,y)/f_X(x).$$

The distributional analogy of the multiplication probability rule $\mathbb{P}(AB) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$ is

$$f(x,y) = f(x|y)f_Y(y) = f(y|x)f_X(x). \tag{5.11}$$

When X and Y are independent, the joint density is the product of marginal densities, $f(x,y) = f_X(x)f_Y(y)$. Conversely, if the joint density of (X,Y) can be represented as a product of marginal densities, X and Y are independent.

The definition of covariance and the correlation for X and Y coincides with the discrete case equivalents:

$$\mathbb{C}\text{ov}(X,Y) = \mathbb{E}XY - \mathbb{E}X \cdot \mathbb{E}Y \quad \text{and} \quad \mathbb{C}\text{orr}(X,Y) = \frac{\mathbb{C}\text{ov}(X,Y)}{\sqrt{\mathbb{V}\text{ar}\,(X) \cdot \mathbb{V}\text{ar}\,(Y)}}.$$

Here, $\mathbb{E}XY = \int_{\mathbb{R}^2} xyf(x,y)dxdy$.

Example 5.21. **Probability, Marginals, and Conditional.** A two-dimensional random variable (X,Y) is defined by its density function, $f(x,y) = 2xe^{-x-2y}$, $x \geq 0, y \geq 0$.

(a) Find the probability that random variable (X,Y) falls in the rectangle $0 \leq X \leq 1, 1 \leq Y \leq 2$.

(b) Find the marginal distributions of X and Y.

(c) Find the conditional distribution of $X|\{Y = y\}$ Does it depend on y?

Solution: (a) The joint density separates variables x and y, therefore

$$\mathbb{P}(0 \leq X \leq 1,\ 1 \leq Y \leq 2) = \int_0^1 xe^{-x}dx \times \int_1^2 2e^{-2y}dy.$$

Since

$$\int_0^1 xe^{-x}dx = -xe^{-x}\Big|_0^1 + \int_0^1 e^{-x}dx = -e^{-1} - e^{-1} + 1 = 1 - 2/e,$$

and

$$\int_1^2 2e^{-2y}dy = -e^{-2y}\Big|_1^2 = -e^{-4} + e^{-2} = \frac{e^2 - 1}{e^4}.$$

then

$$\mathbb{P}(0 \leq X \leq 1,\ 1 \leq Y \leq 2) = \frac{e-2}{e} \times \frac{e^2-1}{e^4} \approx 0.0309.$$

(b) Since the joint density separates the variables, it is a product of marginal densities $f(x,y) = f_X(x) \times f_Y(y)$. This is an analytic way to state that components X and Y are independent. Therefore, $f_X(x) = xe^{-x}$, $x \geq 0$ and $f_Y(y) = 2e^{2y}$, $y \geq 0$.

(c) The conditional densities for $X|\{Y = y\}$ and $Y|\{X = x\}$ are defined as

$$f(x|y) = f(x,y)/f_Y(y) \quad \text{and} \quad f(y|x) = f(x,y)/f_X(x).$$

Because of independence of X and Y the conditional densities coincide with the marginal densities. Thus, the conditional density for $X|\{Y = y\}$ does not depend on y.

✎

5.4.2 Conditional Expectation*

Conditional expectation of Y given $\{X = x\}$ is simply the expectation with respect to the conditional distribution,

$$\mathbb{E}(Y|X = x) = \int_{\mathbb{R}} yf(y|x)dy.$$

Since it depends on the value x taken by random variable X, conditional expectation is a function of x. When a particular realization of X is not specified, the conditional expectation of Y given X is denoted by $\mathbb{E}Y|X$ and represents a random variable.

The following properties of conditional expectation and variance are very important and useful in applications:

In general, $\mathbb{E}Y|X$ is a random variable for which

$$\begin{aligned} \mathbb{E}Y &= \mathbb{E}(\mathbb{E}Y|X), \\ \mathbb{V}\text{ar}\, Y &= \mathbb{V}\text{ar}\,(\mathbb{E}Y|X) + \mathbb{E}(\mathbb{V}\text{ar}\, Y|X). \end{aligned} \tag{5.12}$$

These two equations are sometimes called the Iterated Expectation Rule and Total Variance Rule.

Example 5.22. **Conditional Distributions, Expectations, and Variances.** Let a bivariate random variable (X,Y) have a uniform distribution on triangle $x \geq 0$, $y \geq 0$ and $x + y \leq 1$. The density is constant over the triangle, and the constant is a reciprocal of the triangle area,

$$f(x,y) = \begin{cases} 2, 0 \leq x,\ y,\ x + y \leq 1 \\ 0, \text{else} \end{cases}$$

The marginal density for X is obtained by integrating y form the joint density $f(x,y)$. Here variable y ranges from 0 to $1 - x$, and

$$f_X(x) = \int_0^{1-x} 2dy = 2y\Big|_0^{1-x} = 2(1-x),\ 0 \leq x \leq 1.$$

For $f_Y(y)$, the derivation is analogous, $f_Y(y) = 2(1-y),\ 0 \leq y \leq 1$. The means and variances of X (as well as Y) are

$$\mathbb{E}X = \int_0^1 2x(1-x)dx = \left(x^2 - \frac{2x^3}{3}\right)\Big|_0^1 = 1 - \frac{2}{3} = \frac{1}{3},$$

$$\begin{aligned}\mathbb{V}\mathrm{ar}\, X = \mathbb{E}X^2 - (\mathbb{E}X)^2 &= \int_0^1 2x^2(1-x)dx - \frac{1}{9} \\ &= \left(\frac{2x^3}{3} - \frac{x^4}{2}\right)\Big|_0^1 - \frac{1}{9} \\ &= \frac{2}{3} - \frac{1}{2} - \frac{1}{9} = \frac{1}{18}.\end{aligned}$$

The conditional distribution of Y when $X = x$ is

$$f(y|x) = \begin{cases} \frac{1}{1-x}, & 0 \le y \le 1-x \\ 0, & \text{else} \end{cases}$$

The conditional expectation of Y given $\{X = x\}$ is

$$\mathbb{E}(Y|X=x) = \int_0^{1-x} \frac{ydy}{1-x} = \frac{y^2}{2(1-x)}\Big|_0^{1-x} = \frac{1-x}{2}.$$

Since this is true for any x that X takes, the conditional expectation can be expressed in terms of X as

$$\mathbb{E}Y|X = \frac{1-X}{2},$$

and as such represents a random variable. It is straightforward to show (Exercise 5.23) that $\mathbb{V}\mathrm{ar}\,(Y|X=x) = \frac{(1-x)^2}{12}$, that is,

$$\mathbb{V}\mathrm{ar}\, Y|X = \frac{(1-X)^2}{12}.$$

We will check that the Iterated Expectation Rule and Total Variance Rule from (5.12) are satisfied. The iterate expectation is

$$\mathbb{E}(\mathbb{E}Y|X) = \mathbb{E}\frac{1-X}{2} = \frac{1-1/3}{2} = \frac{1}{3},$$

which coincides with $\mathbb{E}Y$. The total variance is

$$\begin{aligned}\mathbb{E}(\mathbb{V}\mathrm{ar}\, Y|X) + \mathbb{V}\mathrm{ar}\,(\mathbb{E}Y|X) &= \mathbb{E}\frac{(1-X)^2}{12} + \mathbb{V}\mathrm{ar}\,\frac{1-X}{2} \\ &= \frac{1}{12}(1 - 2\mathbb{E}X + \mathbb{E}X^2) + \frac{1}{4}\mathbb{V}\mathrm{ar}\, X \\ &= \frac{1}{12}\left(1 - \frac{2}{3} + \frac{1}{6}\right) + \frac{1}{4}\cdot\frac{1}{18} \\ &= \frac{6-4+1}{72} + \frac{1}{72} = \frac{1}{18},\end{aligned}$$

which coincides with $\mathbb{V}\mathrm{ar}\, Y$.
✎

5.5 Some Standard Continuous Distributions

In this section we overview some popular, commonly used continuous distributions: uniform, exponential, gamma, inverse gamma, beta, double exponential, logistic, Weibull, Pareto, and Dirichlet. The normal (Gaussian) distribution will be just briefly mentioned here. Due to its importance, a separate chapter will cover the details of the normal distribution and its close relatives: χ^2, t, Cauchy, F, and lognormal distributions. Some other continuous distributions will be featured in the examples, exercises, and other chapters, such as Maxwell and Rayleigh distributions.

5.5.1 Uniform Distribution

A random variable X has a uniform $\mathcal{U}(a,b)$ distribution if its density is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b, \\ 0, & \text{else}. \end{cases}$$

Sometimes, to simplify notation, the density can be written simply as

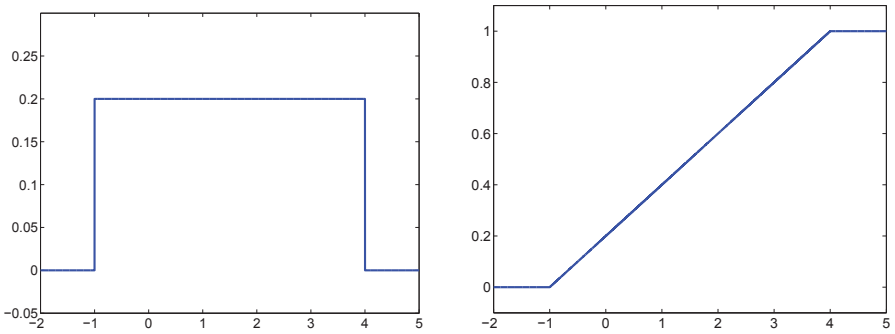
$$f_X(x) = \frac{1}{b-a}\mathbf{1}(a \le x \le b).$$

Here, $\mathbf{1}(A)$ is 1 if A is a true statement and 0 if A is false. Thus, for $x < a$ or $x > b$, $f_X(x) = 0$, since for those values of x the relation $a \le x \le b$ is false and $\mathbf{1}(a \le x \le b) = 0$. For $a = 0$ and $b = 1$, the distribution is called standard uniform.

The CDF of X is given by

$$F_X(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \le x \le b, \\ 1, & x > b. \end{cases}$$

The graphs of the PDF and CDF of a uniform $\mathcal{U}(-1,4)$ random variable are shown in Figure 5.8.



(a) (b)

Fig. 5.8 (a) PDF and (b) CDF for uniform $\mathcal{U}(-1,4)$ distribution. The graphs are plotted as (a) `plot(-2:0.001:5, unifpdf(-2:0.001:5, -1, 4))` and (b) `plot(-2:0.001:5, unifcdf(-2:0.001:5, -1, 4))`.

The expectation of X is $\mathbb{E}X = \frac{a+b}{2}$ and the variance is $\mathbb{V}\mathrm{ar}\,X = (b-a)^2/12$. The nth moment of X is given by $\mathbb{E}X^n = \frac{1}{n+1}\sum_{i=0}^{n} a^i b^{n-i}$. The moment-generating function for the uniform distribution is $m(t) = \frac{e^{tb}-e^{ta}}{t(b-a)}$.

If U is $\mathcal{U}(0,1)$, then $X = -\lambda \log(U)$ is an exponential random variable with scale parameter λ. The sum of two independent standard uniform random variables has triangular distribution,

$$f_X(x) = \begin{cases} x, & 0 \le x \le 1, \\ 2-x, & 1 \le x \le 2, \\ 0, & \text{else}. \end{cases}$$

This is sometimes called a "witch hat" distribution. The distribution of the sum of n independent standard uniforms random variables is known as the Irwing–Hall distribution.

The MATLAB commands for uniform CDF, PDF, quantile, and random number are `unifcdf`, `unifpdf`, `unifinv`, and `unifrnd`. In WinBUGS, the uniform distribution is coded as `dunif(a,b)`.

Example 5.23. **A Gauge That Rounds.** An absolute error E of a measurement read at a particular gauge has uniform $\mathcal{U}(0,1/2)$ distribution. This error is caused by gauge's rounding to the nearest integer. The mean and variance of E are $(0+1/2)/2 = 1/4$ and $(1/2-0)^2/12 = 1/48$. The probability

that in a single measurement the absolute error exceeds 0.3 is `1-unifcdf(0.3, 0, 1/2)` which is equal to 0.4. Since the density is 2 for values between 0 and 1/2, this probability can be easily visualized as an area of a rectangle with basis $0.5 - 0.3 = 0.2$ and height 2.
✎

Example 5.24. **Uniform Inspection Time.** Counts N at a particle counter observed at time $t \geq 0$ are distributed as Poisson $\mathcal{P}oi(\lambda t)$. Suppose the count is inspected at random time $T = t$. If the inspection time T is distributed uniformly between 0 and b, what are the expectation and variance of N?

If the inspection time t was fixed, the expectation and variance would be λt. When inspection time is random, $T \sim \mathcal{U}(0, b)$, then we use iterated expectation and total variance as in (5.12),

$$\begin{aligned} \mathbb{E}N &= \mathbb{E}(\mathbb{E}N|T) = \mathbb{E}(\lambda T) = \lambda b/2, \\ \mathbb{V}\mathrm{ar}\, N &= \mathbb{V}\mathrm{ar}\,(\mathbb{E}N|T) + \mathbb{E}(\mathbb{V}\mathrm{ar}\, N|T) \\ &= \mathbb{V}\mathrm{ar}\,(\lambda T) + \mathbb{E}(\lambda T) = \lambda^2 b^2/12 + \lambda b/2. \end{aligned}$$

Note the overdispersion $\lambda^2 b^2/12$ due to randomness of the inspection time.
✎

5.5.2 Exponential Distribution

The probability density function for an exponential random variable is

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

where $\lambda > 0$ is called the *rate* parameter. An exponentially distributed random variable X is denoted by $X \sim \mathcal{E}(\lambda)$. Its moment-generating function is $m(t) = \lambda/(\lambda - t)$ for $t < \lambda$, and the mean and variance are $1/\lambda$ and $1/\lambda^2$, respectively. The nth moment is $\mathbb{E}X^n = \frac{n!}{\lambda^n}$.

This distribution has several interesting features; for example, its *failure rate*, defined as

$$\lambda_X(t) = \frac{f_X(t)}{1 - F_X(t)},$$

is constant and equal to λ.

The exponential distribution has an important connection to the Poisson distribution. Suppose we observe i.i.d. exponential variates $(X_1, X_2, \dots)$ and define $S_n = X_1 + \cdots + X_n$. For any positive value t, it can be shown that

$\mathbb{P}(S_n < t < S_{n+1}) = p_Y(n)$, where $p_Y(n)$ is the probability mass function for a Poisson random variable Y with parameter λt.

Like a geometric random variable, an exponential random variable has the *memoryless property*, $\mathbb{P}(X \geq u + v | X \geq u) = \mathbb{P}(X \geq v)$ (Exercise 5.25).

The median value, representing a typical observation, is roughly 70% of the mean, showing how extreme values can affect the population mean. This is explicitly shown by the ease in computing the inverse CDF:

$$p = F(x) = 1 - e^{-\lambda x} \iff x = F^{-1}(p) = -\frac{1}{\lambda}\log(1-p).$$

The MATLAB commands for exponential CDF, PDF, quantile, and random number are `expcdf`, `exppdf`, `expinv`, and `exprnd`. MATLAB uses the alternative parametrization with $1/\lambda$ in place of λ. Thus, the CDF of random variable X with $\mathcal{E}(3)$ distribution evaluated at $x = 2$ is calculated in MATLAB as `expcdf(2,1/3)`. In WinBUGS, the exponential distribution is coded as `dexp(lambda)`.

Example 5.25. **Melanoma.** The 5-year cancer survival rate in the case of malignant melanoma of the skin at stage IIIA is 78%. Assume that the survival time T can be modeled by an exponential random variable with unknown rate λ. Given the 5-year survival rate, we will find the probability of a melanoma patient surviving more than 10 years.

Using the given survival rate of 0.78, we first determine the parameter of the exponential distribution – the rate λ. Since $\mathbb{P}(T > t) = \exp(-\lambda t)$, $\mathbb{P}(T > 5) = 0.78$ leads to $\exp\{-5\lambda\} = 0.78$, with solution $\lambda = -\frac{1}{5}\log(0.78)$, which can be rounded to $\lambda = 0.05$.

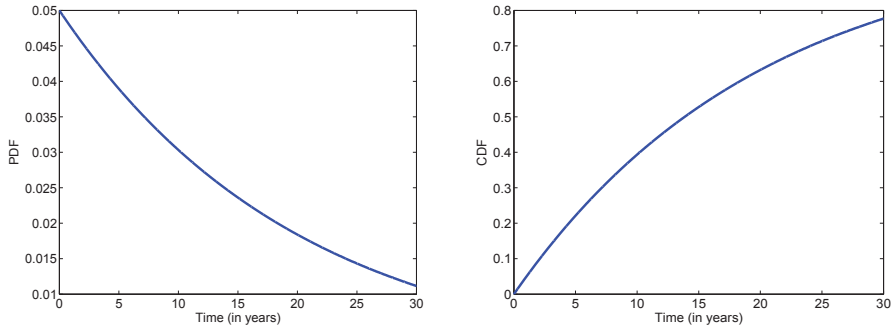
Next, we find the probability that the survival time exceeds 10 years, first directly using the CDF,

$$\mathbb{P}(T > 10) = 1 - F(10) = 1 - \left(1 - e^{-0.05 \cdot 10}\right) = \frac{1}{\sqrt{e}} = 0.6065,$$

and then by MATLAB. One should be careful when parameterizing the exponential distribution in MATLAB. MATLAB uses the scale parameter, a reciprocal of the rate λ.

```
1 - expcdf(10, 1/0.05)
  %ans =  0.6065
  %
  %Figures of PDF and CDF are produced by
time=0:0.001:30;
pdf = exppdf(time, 1/0.05);  plot(time, pdf, 'b-');
cdf = expcdf(time, 1/0.05);  plot(time, cdf, 'b-');
```

This is shown in Figure 5.9.



(a) (b)

Fig. 5.9 Exponential (a) PDF and (b) CDF for rate $\lambda = 0.05$.

Example 5.26. **Minimum of n Exponential Lifetimes.** Let $n = 20$ independent components be connected in a serial system; that is, all components need to be operational for the system to work. The lifetime of each component is exponential $\mathcal{E}(\lambda)$ random variable, where $\lambda = 1/3$ is the rate parameter (in units of 1/year). What is the probability that the system remains operational for more than one month?

If T_i are lifetimes of system components, the system's lifetime is $T = \min\{T_1, T_2, \ldots, T_n\}$ because of the serial connection. When T_i are independent exponentials with rates λ_i, the system's lifetime T is also exponential with rate $\lambda = \sum_{i=1}^{n} \lambda_i$.

This is easy to see; for T to exceed t, each T_i has to exceed t,

$$\mathbb{P}(T > t) = \mathbb{P}(T_1 > t, T_2 > t, \ldots, T_n > t).$$

Due to the independence of T_i's, the probability above is

$$\prod_{i=1}^{n} \mathbb{P}(T_i > t) = \prod_{i=1}^{n} \exp\{-\lambda_i t\} = \exp\left\{-t \sum_{i=1}^{n} \lambda_i\right\}.$$

Thus, $T \sim \mathcal{E}\left(\sum_{i=1}^{n} \lambda_i\right)$.

In this example, all λ_i are equal and $T \sim \mathcal{E}(30 \cdot 1/3)$. We assume that 1 month is 1/12 of a year, and

$$\mathbb{P}(T > 1/12) = \exp\{-10/12\} = 0.4346.$$

Even though each component will work for at least a month with probability of 97.26%, this probability for a serial system of 30 independent components scales down to 43.46%.

5.5.3 Normal Distribution

As we indicated at the start of this section, due to its importance, the normal distribution is covered in a separate chapter. Here we provide a definition and list a few important facts.

The probability density function for a normal (Gaussian) random variable X is given by

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

where μ is the mean and σ^2 is the variance of X. This will be denoted as $X \sim \mathcal{N}(\mu,\sigma^2)$. For $\mu = 0$ and $\sigma = 1$, the distribution is called the standard normal distribution. The CDF of a normal distribution cannot be expressed in terms of elementary functions and so defines a function of its own. For the standard normal distribution, the CDF is

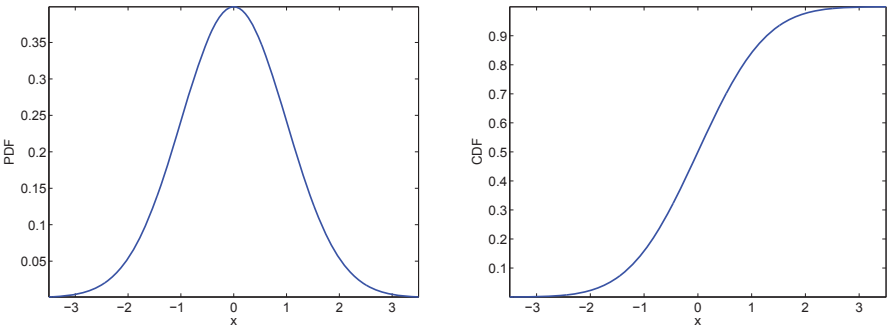
$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\} dt.$$

The standard normal PDF and CDF are shown in Figure 5.10a,b.

The moment-generating function is $m(t) = \exp\{\mu t + \sigma^2 t^2/2\}$. The odd central moments $\mathbb{E}(X-\mu)^{2k+1}$ are 0 because the normal distribution is symmetric about the mean. The even moments are

$$\mathbb{E}(X-\mu)^{2k} = \sigma^{2k}\,(2k-1)!!,$$

where $(2k-1)!! = (2k-1)\cdot(2k-3)\cdots 5\cdot 3\cdot 1$.



(a) (b)

Fig. 5.10 Standard normal (a) PDF and (b) CDF $\Phi(x)$.

The MATLAB commands for normal CDF, PDF, quantile, and random number are `normcdf`, `normpdf`, `norminv`, and `normrnd`. In WinBUGS, the normal distribution is coded as `dnorm(mu,tau)`, where `tau` is a precision parameter, the reciprocal of variance.

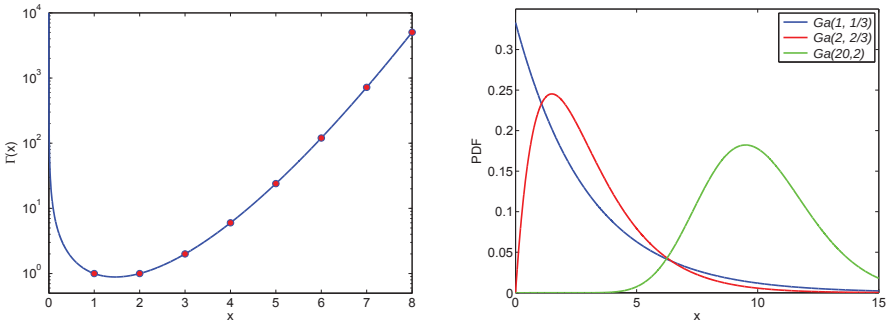
5.5.4 Gamma Distribution

The gamma distribution is an extension of the exponential distribution. Prior to defining its density, we define the gamma function that is critical in normalizing the density. Function $\Gamma(x)$, defined via the integral $\int_0^\infty t^{x-1}e^{-t}dt, \quad x > 0$, is called the gamma function (Fig. 5.11a). If n is a positive integer, then $\Gamma(n) = (n-1)!$. In MATLAB: `gamma(x)`.

Random variable X has a gamma $\mathcal{G}a(r,\lambda)$ distribution if its PDF is given by

$$f_X(x) = \begin{cases} \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, & x \geq 0, \\ 0, & \text{else.} \end{cases}$$

The parameter $r > 0$ is called the *shape* parameter, and $\lambda > 0$ is the *rate* parameter. Figure 5.11b shows gamma densities for $(r,\lambda) = (1,1/3), (2,2/3)$, and $(20,2)$.



(a) (b)

Fig. 5.11 (a) Gamma function, $\Gamma(x)$. The *red dots* are values of the gamma function at integers, $\Gamma(n) = (n-1)!$; (b) Gamma densities: $\mathcal{G}a(1,1/3)$, $\mathcal{G}a(2,2/3)$, and $\mathcal{G}a(20,2)$.

The moment-generating function is $m(t) = (\lambda/(\lambda - t))^r$, so in the case $r = 1$, the gamma distribution becomes the exponential distribution. From $m(t)$ we have $\mathbb{E}X = r/\lambda$ and $\mathbb{V}\text{ar}\, X = r/\lambda^2$.

If $X_1, \ldots, X_n$ are generated from an exponential distribution with (rate) parameter λ, it follows from $m(t)$ that $Y = X_1 + \cdots + X_n$ is distributed as

gamma with parameters λ and n; that is, $Y \sim \mathcal{G}a(n,\lambda)$. A gamma distribution with an integer shape parameter is sometimes called Erlang's distribution. More generally, if $X_i \sim \mathcal{G}a(r_i,\lambda)$ are independent, then $Y = X_1 + \cdots + X_n$ is distributed as gamma with parameters λ and $r = r_1 + r_2 + \cdots + r_n$,; that is, $Y \sim \mathcal{G}a(r,\lambda)$ (Exercise 5.24).

Often, the gamma distribution is parameterized with $1/\lambda$ in place of λ, and this alternative parametrization is used in MATLAB definitions. The CDF in MATLAB is `gamcdf(x,r,1/lambda)`, and the PDF is `gampdf(x,r,1/lambda)`. The function `gaminv(p,r,1/lambda)` computes the pth quantile of the $\mathcal{G}a(r,\lambda)$ random variable. In WinBUGS, $\mathcal{G}a(n,\lambda)$ is coded as `dgamma(n,lambda)`.

Example 5.27. **Corneoretinal Potentials.** Emil du Bois-Reymond (1848) observed that the cornea of the eye is electrically positive relative to the back of the eye. This potential is not affected by the presence or absence of light, and its variability is critical in defining the electro-oculogram (EOG). Eye movements thus produce a moving (rotating) dipole source, and accordingly, signals that are indicative of the movement may be obtained.

Assume that corneoretinal potential is a random variable $X = Y + 0.35$ [mV], where Y is gamma distributed with shape parameter 3 and rate parameter 20 [1/mV] (or equivalently, scale parameter $1/20 = 0.05$ [mV]).

(a) What is the probability to observe corneoretinal potential X exceeding 0.5 [mV].

(b) If an observed corneoretinal potential exceeds x^*, it is recorded as significant. If, in the long run, we wish to label 1% largest potentials as significant, how should the threshold x^* be set?

```
%(a) P(X > 0.5)=P(Y+0.35 > 0.5)=P(Y > 0.15)
1-gamcdf(0.15, 3, 1/20)  %0.4232
%(b) 0.01=P(X > x*)=P(Y-0.35 > x*)=P(Y>x*-0.35).
%x*-35 is 0.99-quantile of gamma distribution with shape=3 and rate=20.
xstar = 0.35 + gaminv(0.99, 3, 1/20)    %0.7703
```

Thus, if modeled as gamma $\mathcal{G}a(3,20)$, the corneoretinal potential will exceed 0.5 with probability 0.4232, and will exceed $x^* = 0.7703$ with probability 0.01.

5.5.5 *Inverse Gamma Distribution*

Random variable X is said to have an inverse gamma $\mathcal{IG}(r,\lambda)$ distribution with parameters $r > 0$ and $\lambda > 0$ if its density is given by

$$f_X(x) = \begin{cases} \frac{\lambda^r}{\Gamma(r)x^{r+1}} e^{-\lambda/x}, & x \geq 0, \\ 0, & \text{else .} \end{cases}$$

The mean and variance of X are $\mathbb{E}X = \lambda/(r-1)$, $r > 1$, and $\mathbb{V}\text{ar}\, X = \lambda^2/[(r-1)^2(r-2)]$, $r > 2$, respectively. If $X \sim \mathcal{G}a(r,\lambda)$, then its reciprocal X^{-1} is $\mathcal{IG}(r,\lambda)$ distributed. We will see that in the Bayesian context, the inverse gamma is a natural prior distribution for a scale parameter.

5.5.6 Beta Distribution

We first define two special functions: beta and incomplete beta. The beta function is defined as $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt = \Gamma(a)\Gamma(b)/\Gamma(a+b)$. In MATLAB, beta function is coded as `beta(a,b)`. An incomplete beta is $B(x,a,b) = \int_0^x t^{a-1}(1-t)^{b-1}dt$, $0 \leq x \leq 1$. In MATLAB, `betainc(x,a,b)` represents the normalized incomplete beta, defined as $I_x(a,b) = B(x,a,b)/B(a,b)$. As we will see in a moment, $B(a,b)$ will be a normalizing constant in PDF, while $B(x,a,b)/B(a,b)$ coincides with CDF of beta distribution.

The density function for a beta random variable is

$$f_X(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, & 0 \leq x \leq 1, \\ 0, & \text{else ,} \end{cases}$$

where B is the beta function and $a,b \geq 0$. Because X is defined only in the interval [0,1], the beta distribution is useful in modeling uncertainty or randomness in proportions or probabilities. A beta-distributed random variable is denoted by $X \sim \mathcal{B}e(a,b)$. The standard uniform distribution $\mathcal{U}(0,1)$ serves as a special case with $(a,b) = (1,1)$. The moments of beta distribution are

$$\mathbb{E}X^k = \frac{\Gamma(a+k)\Gamma(a+b)}{\Gamma(a)\Gamma(a+b+k)} = \frac{a(a+1)\ldots(a+k-1)}{(a+b)(a+b+1)\ldots(a+b+k-1)}$$

so that $\mathbb{E}(X) = a/(a+b)$ and $\mathbb{V}\text{ar}\, X = ab/[(a+b)^2(a+b+1)]$.

In MATLAB, the CDF for a beta random variable (at $x \in (0,1)$) is computed as `betacdf(x,a,b)`, and the PDF is computed as `betapdf(x,a,b)`. The pth percentile is `betainv(p,a,b)`. In WinBUGS, the beta distribution is coded as `dbeta(a,b)`.

To emphasize the modeling diversity of beta distributions, we depict densities for a selection of (a,b), as in Figure 5.12.

If $U_1, U_2, \ldots, U_n$ is a sample from a uniform $\mathcal{U}(0,1)$ distribution, then the distribution of the kth component in the ordered sample is beta, $U_{(k)} \sim$

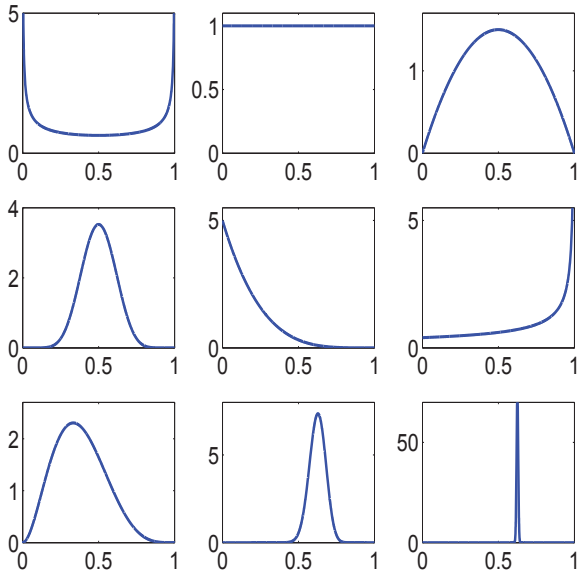


Fig. 5.12 Beta densities for (a,b) as (1/2, 1,2), (1,1), (2,2), (10,10), (1,5), (1, 0.4), (3,5), (50, 30), and (5000, 3000).

$\mathcal{B}e(k,n-k+1)$, for $1 \le k \le n$. Also, if $X \sim \mathcal{G}(m,\lambda)$ and $Y \sim \mathcal{G}(n,\lambda)$, then $X/(X+Y) \sim \mathcal{B}e(m,n)$.

5.5.7 Double Exponential Distribution

A random variable X has double exponential $\mathcal{DE}(\mu,\lambda)$ distribution if its PDF and CDF are given by

$$f_X(x) = \frac{\lambda}{2}\, e^{-\lambda|x-\mu|},$$
$$F_X(x) = \begin{cases} \frac{1}{2}e^{\lambda(x-\mu)}, & x < \mu \\ 1 - \frac{1}{2}e^{-\lambda(x-\mu)}, & x \ge \mu \end{cases}, \quad -\infty < x < \infty, \lambda > 0$$

The expectation of X is $\mathbb{E}X = \mu$, and the variance is $\mathbb{V}\mathrm{ar}\, X = 2/\lambda^2$. The moment-generating function for the double exponential distribution is

$$m(t) = \frac{\lambda^2 e^{\mu t}}{\lambda^2 - t^2}, \quad |t| < \lambda.$$

The double exponential distribution is also known as the *Laplace distribution*. If X_1 and X_2 are independent exponential $\mathcal{E}(\lambda)$, then $X_1 - X_2$ is distributed as $\mathcal{DE}(0,\lambda)$. Also, if $X \sim \mathcal{DE}(0,\lambda)$, then $|X| \sim \mathcal{E}(\lambda)$. In MATLAB the double exponential distribution is not implemented since it can be readily obtained by folding the exponential distribution about y-axis, see Figure 5.13a.

In WinBUGS, $\mathcal{DE}(\mu,\lambda)$ is coded as `ddexp(mu,lambda)`.

Example 5.28. **Neighboring Pixels in Digital Mammograms.** The difference D between two arbitrary neighboring pixels in a digital mammogram image is modeled by a double exponential $\mathcal{DE}(0,\lambda)$ distribution.

(a) It is known that the probability of D being less than -4 is 0.3. Using this information calculate λ.

(b) Find the probability of D falling between -5 and 20.

(c) What are the mean and variance of D?

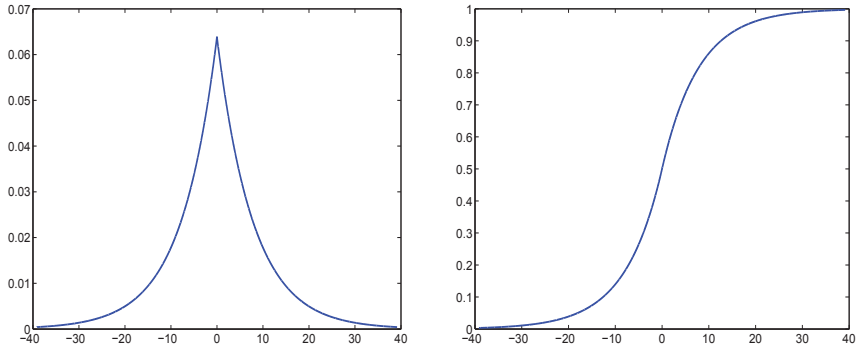
(d) Plot graphs of the PDF and CDF.

```
%mammopixels.m
dexppdf=@(x, mu, lambda) 1/2 * exppdf(abs(x-mu),1./lambda);
dexpcdf=@(x, mu, lambda) 1/2 + sign(x-mu)/2.*expcdf(abs(x-mu),1./lambda);
dexpinv=@(p,  mu, lambda) mu+sign(2*p-1).*expinv(abs(2*p-1),1./lambda);
dexprnd=@(mu,lambda,size) mu+exprnd(1./lambda,size)-exprnd(1./lambda,size);
dexpstat = @(mu, lambda) deal(mu, 2./lambda.^2);

% (a) 0.3=P(D<=-4)=0.5 * exp(- 4*lambda) -> lambda=-1/4*log(2*0.3)=0.1277
% To check:
dexpinv(0.3, 0, 0.1277)  %-4.0002
%(b) P( -5 < D < 20)
dexpcdf(20, 0, 0.1277) - dexpcdf(-5,0,0.1277)  %0.6971
%(c)
[m v]=dexpstat(0, 0.1277)  %m = 0,  v=122.6445
%(d)
mu=0;  lambda=0.1277
x = mu-5/lambda:0.001:mu+5/lambda;
figure;
plot(x, dexppdf(x, mu, lambda));
figure;
plot(x, dexpcdf(x,mu, lambda));
```

5.5.8 Logistic Distribution

The logistic distribution was first defined by Belgian mathematician Pierre Francois Verhulst (1804–1849) who, in 1838, used it in modeling population growth and coined the term *logistic*. Logistic distribution is used for models in pharmacokinetics, regression with binary responses, river discharge and



(a) (b)

Fig. 5.13 (a) PDF and (b) CDF for $D \sim \mathcal{DE}(0, 0.1277)$.

rainfall in hydeology, neural networks, and machine learning, to list just a few modeling applications.

The logistic random variable can be introduced by a property of its CDF expressed by a differential equation. Let $F(x) = \mathbb{P}(X \le x)$ be the CDF for which $F'(x) = F(x) \times (1 - F(x))$. One interpretation of this differential equation is as follows: For a Bernoulli random variable $\mathbf{1}(X \le x) = \begin{cases} 1, X \le x \\ 0, X > x \end{cases}$, the change in $\mathbb{E}\mathbf{1}(X \le x)$ as a function of x, is equal to its variance. The solution in the class of CDFs is

$$F_X(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x},$$

which is called the logistic distribution. Its density is

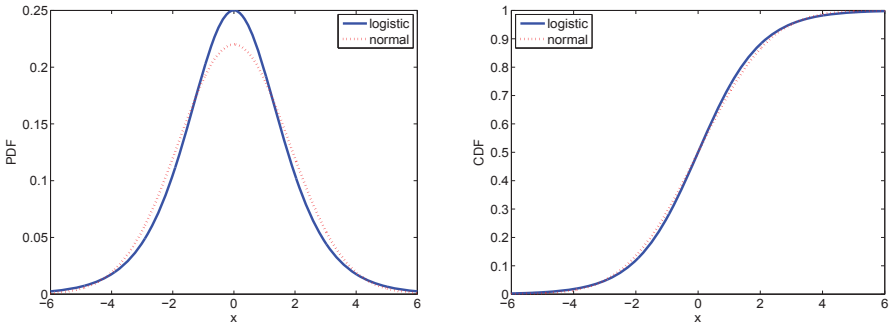
$$f_X(x) = \frac{e^x}{(1 + e^x)^2} = \frac{e^{-x}}{(1 + e^{-x})^2}.$$

Graphs of $f_X(x)$ and $F_X(x)$ are shown in Figure 5.14. The mean of the distribution is 0 and the variance is $\pi^2/3$. For a more general logistic distribution given by the CDF

$$F_X(x) = \frac{1}{1 + e^{-(x-\mu)/\sigma}},$$

the mean is μ, variance $\pi^2\sigma^2/3$, skewness 0, and kurtosis 21/5. For the higher moments, one can use the moment-generating function

$$m(t) = \exp\{\mu t\} B(1 - \sigma t, 1 + \sigma t),$$



(a) (b)

Fig. 5.14 (a) Density and (b) CDF of logistic distribution. Superimposed (*dotted red*) is the normal distribution with matching mean and variance, 0 and $\pi^2/3$, respectively.

where B is the beta function. In WinBUGS the logistic distribution is coded as `dlogis(mu,tau)`, where `tau` is the reciprocal of σ.

If X has a logistic distribution, then $\log(X)$ has a log-logistic distribution (also known as the Fisk distribution). The log-logistic distribution is used in economics (population wealth distribution) and reliability.

The logistic distribution will be revisited in Chapter 15, where we deal with logistic regression.

5.5.9 Weibull Distribution

The Weibull distribution is one of the most important distributions in survival theory and engineering reliability. It is named after Swedish engineer and scientist Waloddi Weibull after his publication in the early 1950s (Weibull, 1951).

The density of the two-parameter Weibull random variable $X \sim \mathcal{W}ei(r,\lambda)$ is given as

$$f_X(x) = \lambda r x^{r-1} e^{-\lambda x^r},\ x > 0. \tag{5.13}$$

The CDF is given as $F_X(x) = 1 - e^{-\lambda x^r}$. Parameter r is the shape parameter, while λ is the rate parameter. Both parameters are strictly positive. In this form, Weibull $X \sim \mathcal{W}ei(r,\lambda)$ is a distribution of $X = Y^{1/r}$ for Y exponential $\mathcal{E}(\lambda)$.

In MATLAB, the Weibull distribution is parameterized by a and r, as in

$$f(x) = a^{-r} r x^{r-1} e^{-(x/a)^r},\ x > 0. \tag{5.14}$$

Note that in this parametrization, a is the scale parameter and relates to λ as $\lambda = a^{-r}$. So when $a = \lambda^{-1/r}$, the CDF in MATLAB is `wblcdf(x,a,r)`, and the PDF is `wblpdf(x,a,r)`. The function `wblinv(p,a,r)` computes the pth quantile of the $\mathcal{W}ei(r,\lambda)$ random variable.

The (r,λ) parametrization of Weibull distribution is not as prevalent as the shape-scale parametrization from (5.14), but the likelihood in (5.13) is more convenient for Bayesian inference. In WinBUGS, $\mathcal{W}ei(r,\lambda)$ is coded as `dweib(r,lambda)`.

The Weibull distribution generalizes the exponential distribution $(r = 1)$ and Rayleigh distribution $(r = 2)$. Figure 5.15 shows the densities of the Weibull distribution for $r = 2$ (blue), $r = 1$ (red), and $r = 1/2$ (black). In all three cases, $\lambda = 1/2$. The values for the scale parameter $a = \lambda^{-1/r}$ are $\sqrt{2}$,2, and 4, respectively.

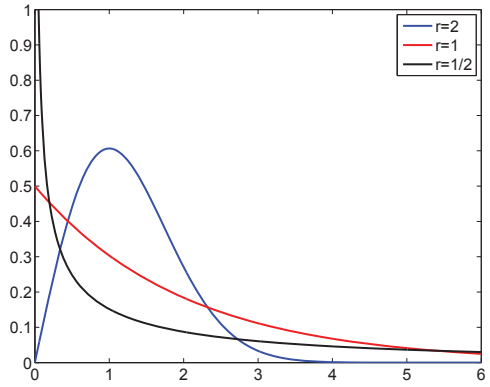


Fig. 5.15 Densities of Weibull distribution with $r = 2$ (*blue*), $r = 1$ (*red*), and $r = 2$ (*black*). In all three cases, $\lambda = 1/2$.

The mean of a Weibull random variable X is $\mathbb{E}X = \frac{\Gamma(1+1/r)}{\lambda^{1/r}} = a\Gamma\left(1+\frac{1}{r}\right)$, and the variance is $\mathbb{V}\text{ar}\,X = \frac{\Gamma(1+2/r)-\Gamma^2(1+1/r)}{\lambda^{2/r}} = a^2(\Gamma\left(1+\frac{2}{r}\right) - \Gamma^2\left(1+\frac{1}{r}\right))$. The kth moment is $\mathbb{E}X^k = \frac{\Gamma(1+k/r)}{\lambda^{k/r}} = a^k\Gamma\left(1+\frac{k}{r}\right)$.

5.5.10 *Pareto Distribution*

The Pareto distribution is named after the Italian economist Vilfredo Pareto (1848-1923). Some examples in which the Pareto distribution provides an exemplary model include wealth distribution in individuals, sizes of human settlements, visits to encyclopedia pages, and file size distribution of

Internet traffic that uses the TCP protocol. A random variable X has a Pareto $\mathcal{P}a(c,\alpha)$ distribution with parameters $0 < c < \infty$ and $\alpha > 0$ if its density is given by

$$f_X(x) = \begin{cases} \frac{\alpha}{c}\ \left(\frac{c}{x}\right)^{\alpha+1}, & x \geq c, \\ 0, & \text{else}\ . \end{cases}$$

The CDF is

$$F_X(x) = \begin{cases} 0, & x < c, \\ 1 - \left(\frac{c}{x}\right)^{\alpha}, & x \geq c. \end{cases}$$

The mean and variance of X are $\mathbb{E}X = \alpha c/(\alpha - 1)$, $\alpha > 1$, and $\mathbb{V}\mathrm{ar}\, X = \alpha c^2/[(\alpha - 1)^2(\alpha - 2)]$, $\alpha > 2$. The median is $m = c \cdot 2^{1/\alpha}$. If $X_1, \ldots, X_n$ are independent $\mathcal{P}a(c,\alpha)$, then $Y = 2c\sum_{i=1}^{n} \ln(X_i) \sim \chi^2$ with $2n$ degrees of freedom.

In MATLAB one can specify the generalized Pareto distribution, which for some selection of its parameters is equivalent to the aforementioned Pareto distribution. In WinBUGS, the code is `dpar(alpha,c)` (note the permuted order of parameters).

5.5.11 Dirichlet Distribution

The Dirichlet distribution is a multivariate version of the beta distribution in the same way that the multinomial distribution is a multivariate extension of the binomial. A random variable $X = (X_1, \ldots, X_k)$ with a Dirichlet distribution of $(X \sim \mathcal{D}ir(a_1, \ldots, a_k))$ has a PDF of

$$f(x_1, \ldots, x_k) = \frac{\Gamma(A)}{\prod_{i=1}^{k} \Gamma(a_i)} \prod_{i=1}^{k} x_i^{a_i - 1},$$

where $A = \sum a_i$, and $x = (x_1, \ldots, x_k) \geq 0$ is defined on the simplex $x_1 + \cdots + x_k = 1$. Then

$$\mathbb{E}(X_i) = \frac{a_i}{A}, \quad \mathbb{V}\mathrm{ar}\,(X_i) = \frac{a_i(A - a_i)}{A^2(A+1)}, \quad \text{and} \quad \mathbb{C}\mathrm{ov}(X_i, X_j) = -\frac{a_i a_j}{A^2(A+1)}.$$

The Dirichlet random variable can be generated from gamma random variables $Y_1, \ldots, Y_k \sim \mathcal{G}a(a,b)$ as $X_i = Y_i/S_Y$, $i = 1, \ldots, k$, where $S_Y = \sum_i Y_i$. The marginal distribution of a component X_i is $\mathcal{B}e(a_i, A - a_i)$. This is illustrated in the following MATLAB m-file that generates random Dirichlet vectors:

```
function drand = dirichletrnd(a,n)
```

```
% function drand = dirichletrnd(a,n)
% a - vector of parameters 1 x m
% n - number of random realizations
% drand - matrix m x n, each column one realization.
%--------------------------------------------------
a=a(:);
m=size(a,1);
a1=zeros(m,n);
for i = 1:m
    a1(i,:) = gamrnd(a(i,1),1,1,n);
end
for i=1:m
drand(i, 1:n )= a1(i, 1:n ) ./ sum(a1);
end
```

5.6 Random Numbers and Probability Tables

In older introductory statistics texts, many back-end pages have been devoted to various statistical tables. Several decades ago, many books of statistical tables were published. Also, the most respected of statistical journals occasionally published articles providing statistical tables.

In 1947 the RAND Corporation published the monograph *A Million Random Digits with 100,000 Normal Deviates*, which at the time was a state-of-the-art resource for simulation and Monte Carlo methods. The book can be found at `http://www.rand.org/pubs/monograph_reports/MR1418.html`.

These days, much larger tables of random numbers can be produced by a single line of code, resulting in a set of random numbers that can pass a battery of stringent randomness tests. With MATLAB and many other widely available software packages, statistical tables and tables of random numbers are now obsolete. For example, tables of binomial CDF and PDF for a specific n and p can be reproduced by

```
%n=12, p=0.7
disp('binocdf(0:12, 12, 0.7)');
binocdf(0:12, 12, 0.7)
disp('binopdf(0:12, 12, 0.7)');
binopdf(0:12, 12, 0.7)
```

We will show how to sample and simulate from a few distributions in MATLAB and compare empirical means and variances with their theoretical counterparts. The following annotated MATLAB code simulates from binomial, Poisson, and geometric distributions and compares theoretical and empirical means and variances:

```
%various_simulations.m
simu = binornd(12, 0.7, [1,100000]);
```

```
  % simu is 10000 observations from Bin(12,0.7)
disp('simu = binornd(12, 0.7, [1,100000]); 12*0.7 - mean(simu)');

12*0.7 - mean(simu)  %0.001069
  %should be small since the theoretical mean is n*p
disp('simu = binornd(12, 0.7, [1,100000]); ...
                                    12 * 0.7 * 0.3 - var(simu)');

12 * 0.7 * 0.3 - var(simu)  %-0.008350
  %should be small since the theoretical variance is n*p*(1-p)

%% Simulations from Poisson(2)
poi = poissrnd(2, [1, 100000]);
disp('poi = poissrnd(2, [1, 100000]); mean(poi)');
mean(poi)   %1.9976
disp('poi = poissrnd(2, [1, 100000]); var(poi)');
var(poi)    %2.01501

%%% Simulations from Geometric(0.2)
geo = geornd(0.2, [1, 100000]);
disp('geo = geornd(0.2, [1, 100000]); mean(geo)');
mean(geo)  %4.00281
disp('geo = geornd(0.2, [1, 100000]); var(geo)');
var(geo)   %20.11996
```

5.7 Transformations of Random Variables*

When a random variable with known density is transformed, the result is a random variable as well. The question is how to find its distribution. The general theory for distributions of functions of random variables is beyond the scope of this text, and the reader can find comprehensive coverage in Ross (2010a, b).

We have already seen that, for a discrete random variable X, the PMF of a function $Y = g(X)$ is simply the table

$g(X)$	$g(x_1)$	$g(x_2)$	$\cdots$	$g(x_n)$	$\cdots$
Prob	p_1	p_2	$\cdots$	p_n	$\cdots$

in which only realizations of X are transformed while the probabilities are kept unchanged.

For continuous random variables the distribution of a function is more complex. In some cases, however, looking at the CDF is sufficient.

In this section we will discuss two topics: (i) how to find the distribution for a transformation of a single continuous random variable and (ii) how to approximate moments, in particular means and variances, of complex functions of many random variables.

Suppose that a continuous random variable has a density $f_X(x)$ and that a function g is monotone on the domain of f, with the inverse function h, $h = g^{-1}$. Then the random variable $Y = g(X)$ has a density

$$f_Y(y) = f(h(y))|h'(y)|. \tag{5.15}$$

If g is not one-to-one, but has k one-to-one inverse branches, $h_1, h_2, \ldots, h_k$, then

$$f_Y(y) = \sum_{i=1}^{k} f(h_i(y))|h_i'(y)|. \tag{5.16}$$

An example of a function which is not one-to-one is $g(x) = x^2$, for which inverse branches $h_1(y) = \sqrt{y}$ and $h_2(y) = -\sqrt{y}$ are one-to-one.

Example 5.29. Square Root of Exponential. Let X be a random variable with an exponential $\mathcal{E}(\lambda)$ distribution, where $\lambda > 0$ is the rate parameter. Find the distribution of the random variable $Y = \sqrt{X}$.

Here $g(x) = \sqrt{x}$ and $g^{-1}(y) = y^2$. The Jacobian is $|g^{-1}(y)'| = 2y$, $y \geq 0$. Thus,

$$f_Y(y) = \lambda e^{-\lambda y^2} \cdot 2y,\ y \geq 0, \lambda > 0,$$

which is known as the Rayleigh distribution.

An alternative approach to finding the distribution of Y is to consider the CDF:

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(\sqrt{X} \leq y) = \mathbb{P}(X \leq y^2) = 1 - e^{-\lambda y^2}$$

since X has the exponential distribution. The density is now obtained by taking the derivative of $F_Y(y)$,

$$f_Y(y) = (F_Y(y))' = 2\lambda y e^{-\lambda y^2},\ \ y \geq 0, \lambda > 0.$$

✎

The distribution of a function of one or many random variables is an ultimate summary. However, the result could be quite messy and sometimes the distribution lacks a closed form. Moreover, not all facets of the resulting distribution may be of interest to researchers; sometimes only the mean and variance are needed.

If X is a random variable with $\mathbb{E}X = \mu$ and $\mathbb{V}\mathrm{ar}\, X = \sigma^2$, then for a function $Y = g(X)$ the following approximation holds:

$$\mathbb{E}Y \approx g(\mu) + \frac{1}{2}g''(\mu)\sigma^2,$$
$$\mathbb{V}\mathrm{ar}\ Y \approx \big(g'(\mu)\big)^2\sigma^2. \tag{5.17}$$

If n independent random variables are transformed as $Y = g(X_1, X_2, \ldots, X_n)$, then

$$\mathbb{E}Y \approx g(\mu_1, \mu_2, \ldots, \mu_n) + \frac{1}{2}\sum_{i=1}^{n}\frac{\partial^2 g}{\partial x^2}(\mu_1, \mu_2, \ldots, \mu_n)\sigma_i^2,$$
$$\mathbb{V}\mathrm{ar}\ Y \approx \sum_{i=1}^{n}\left(\frac{\partial g}{\partial x_i}(\mu_1, \mu_2, \ldots, \mu_n)\right)^2\sigma_i^2, \tag{5.18}$$

where $\mathbb{E}X_i = \mu_i$ and $\mathbb{V}\mathrm{ar}\, X_i = \sigma_i^2$.

The approximation for the mean $\mathbb{E}Y$ is obtained by the second-order Taylor expansion and is more precise than the approximation for the variance $\mathbb{V}\mathrm{ar}\, Y$, which is of the first order ("linearization"). The second-order approximation for $\mathbb{V}\mathrm{ar}\, Y$ is straightforward but involves third and fourth moments of Xs. Also, when the variables $X_1, \ldots, X_n$ are correlated, the factor $2\sum_{1\le i<j\le n}\frac{\partial^2 g}{\partial x_i \partial x_j}(\mu_1, \ldots, \mu_n)\mathbb{C}\mathrm{ov}(X_i, X_j)$ should be added to the expression for $\mathbb{V}\mathrm{ar}\, Y$ in (5.18).

If g is a complicated function, the mean $\mathbb{E}Y$ is often approximated by a first-order approximation, $\mathbb{E}Y \approx g(\mu_1, \mu_2, \ldots, \mu_n)$, that involves no derivatives.

Example 5.30. **String Vibrations.** In string vibration, the frequency of the fundamental harmonic is often of interest. The fundamental harmonic is produced by the vibration with nodes at the two ends of the string. In this case, the length of the string L is half of the wavelength of the fundamental harmonic. The frequency ω (in Hz) depends also on the tension of the string T, and the string mass M,

$$\omega = \frac{1}{2}\sqrt{\frac{T}{ML}}.$$

Quantities L, T, and M are measured imperfectly and are considered independent random variables. The means and variances are estimated as follows:

Variable (unit)	Mean	Variance
L (m)	0.5	0.0001
T (N)	70	0.16
M (kg/m)	0.001	10^{-8}

Approximate the mean μ_ω and variance σ_ω^2 of the resulting frequency ω.

The partial derivatives

$$\begin{aligned}
\frac{\partial\omega}{\partial T} &= \frac{1}{4}\sqrt{\frac{1}{TML}}, & \frac{\partial^2\omega}{\partial T^2} &= -\frac{1}{8}\sqrt{\frac{1}{T^3ML}},\\
\frac{\partial\omega}{\partial M} &= -\frac{1}{4}\sqrt{\frac{T}{M^3L}}, & \frac{\partial^2\omega}{\partial M^2} &= \frac{3}{8}\sqrt{\frac{T}{M^5L}},\\
\frac{\partial\omega}{\partial L} &= -\frac{1}{4}\sqrt{\frac{T}{ML^3}}, & \frac{\partial^2\omega}{\partial L^2} &= \frac{3}{8}\sqrt{\frac{T}{ML^5}},
\end{aligned}$$

evaluated at the means $\mu_L = 0.5$, $\mu_T = 70$, and $\mu_M = 0.001$, are

$$\begin{aligned}
&\frac{\partial\omega}{\partial T}(\mu_L,\mu_T,\mu_M) = 1.3363, && \frac{\partial^2\omega}{\partial T^2}(\mu_L,\mu_T,\mu_M) = -0.0095,\\
&\frac{\partial\omega}{\partial M}(\mu_L,\mu_T,\mu_M) = -9.3541\cdot 10^4, && \frac{\partial^2\omega}{\partial M^2}(\mu_L,\mu_T,\mu_M) = 1.4031\cdot 10^8,\\
&\frac{\partial\omega}{\partial L}(\mu_L,\mu_T,\mu_M) = -187.0829, && \frac{\partial^2\omega}{\partial L^2}(\mu_L,\mu_T,\mu_M) = 561.2486,
\end{aligned}$$

and the mean and variance of ω are

$$\boxed{\mu_\omega \approx 187.8117} \quad \text{and} \quad \boxed{\sigma_\omega^2 \approx 91.2857}.$$

The first-order approximation for μ_ω is $\frac{1}{2}\sqrt{\frac{\mu_T}{\mu_M\mu_L}} = 187.0829$.

5.8 Mixtures*

In modeling tasks it is sometimes necessary to combine two or more random variables in order to get a satisfactory model. There are two ways of combining random variables: by taking the linear combination $a_1X_1 + a_2X_2 + \ldots$ for which a density in the general case is often convoluted and difficult to express in a finite form, or by combining densities and PMFs directly.

For example, for two densities f_1 and f_2, the density $g(x) = \varepsilon f_1(x) + (1-\varepsilon) f_2(x)$ is a mixture of f_1 and f_2 with weights ε and $1-\varepsilon$. It is important for the weights to be nonnegative and add up to 1 so that $g(x)$ remains a density.

Very popular mixtures are point mass mixture distributions that combine a density function $f(x)$ with a point mass (Dirac) function δ_{x_0} at a value x_0. The Dirac functions belong to a class of special functions. Informally, one may think of δ_{x_0} as a limiting function for a sequence of functions

$$f_{n,x_0} = \begin{cases} n, \; x_0 - \frac{1}{2n} < x < x_0 + \frac{1}{2n}, \\ 0, \text{ else,} \end{cases}$$

when $n \to \infty$. It is easy to see that for any finite n, f_{n,x_0} is a density since it integrates to 1; however, the function domain shrinks to a singleton x_0, while its value at x_0 goes to infinity.

For example, $f(x) = 0.3\delta_0 + 0.7 \times \frac{1}{\sqrt{2\pi}} \exp\{-\frac{x^2}{2}\}$ is a normal distribution contaminated by a point mass at zero with a weight 0.3.

5.9 Markov Chains*

You may have encountered statistical jargon containing the term "Markov chain." In Bayesian calculations the acronym MCMC stands for Markov chain Monte Carlo simulations, while in statistical models of genomes, hidden Markov chain models are popular. Here we give a basic definition and a few examples of Markov chains.

A sequence of random variables $X_0, X_1, \ldots, X_n, \ldots$, with values in the set of "states" $\mathcal{S} = \{1, 2, \ldots\}$, constitutes a Markov chain if the probability of transition to a future state, $X_{n+1} = j$, depends only on the value at the current state, $X_n = i$, and not on any previous values X_{n-1}, $X_{n-2}, \ldots$, X_0. A popular way of putting this is to say that in Markov chains the future depends on the present and not on the past. Formally,

$$\mathbb{P}(X_{n+1} = j | X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i) = \mathbb{P}(X_{n+1} = j | X_n = i) = p_{ij},$$

where $i_0, i_1, \ldots, i_{n-1}, i, j$ are the states from $\mathcal{S}$. The probability p_{ij} is independent of n and represents the transition probability from state i to state j. In our brief coverage of Markov chains, we will consider chains with a finite number of states, N.

For states $\mathcal{S} = \{1, 2, \ldots, N\}$, the transition probabilities form an $N \times N$ matrix $\mathbf{P} = (p_{ij})$. Each row of this matrix sums up to 1 since the probabilities of all possible moves from a particular state, including the probability of remaining in the same state, sum up to 1:

$$p_{i1} + p_{i2} + \cdots + p_{ii} + \cdots + p_{iN} = 1.$$

The matrix $\mathbf{P}$ describes the evolution and long-time behavior of the Markov chain it represents. In fact, if the distribution $\pi^{(0)}$ for the initial variable X_0 is specified, the pair $\pi^{(0)}, \mathbf{P}$ fully describes the Markov chain.

Matrix $\mathbf{P}^2$ gives the probabilities of transition in two steps. Its element $p_{ij}^{(2)}$ is $\mathbb{P}(X_{n+2} = j | X_n = i)$.

Likewise, the elements of matrix $\mathbf{P}^m$ are the probabilities of transition in m steps,

$$p_{ij}^{(m)} = \mathbb{P}(X_{n+m} = j | X_n = i),$$

for any $n \geq 0$ and any $i, j \in \mathcal{S}$.

If the distribution for X_0 is $\pi^{(0)} = \left(\pi_1^{(0)}, \pi_2^{(0)}, \ldots, \pi_N^{(0)}\right)$, then the distribution for X_n is

$$\pi^{(n)} = \pi^{(0)} \mathbf{P}^n. \tag{5.19}$$

Of course, if the state X_0 is known, $X_0 = i_0$, then $\pi^{(0)}$ is a vector of 0s except at position i_0, where the value is 1.

For n large, the probability $\pi^{(n)}$ "forgets" the initial distribution at state X_0 and converges to $\pi = \lim_{n \to \infty} \pi^{(n)}$. This distribution is called the stationary distribution of a chain and satisfies

$$\pi = \pi \mathbf{P}.$$

Operationally, to find stationary distribution, one solves the system

$$\begin{cases} (\mathbf{I} - \mathbf{P})\pi' = 0 \\ \mathbf{1}'\pi = 1. \end{cases}$$

Result. If for a finite state Markov chain one can find an integer k so that all entries in $\mathbf{P}^k$ are strictly positive, then stationary distribution π exists.

Example 5.31. **Ehrenfest Model.** Ehrenfest model (Ehrenfest, 1907) illustrates the diffusion in gasses by considering random transition of molecules between two compartments.

Consider N balls numbered from 1 to N, distributed in two boxes, A and B. The system is in state i if i balls are in the box A (and $N - i$ balls in the box B). A number between 1 and N is randomly selected, and the ball with the selected number switches the boxes. The system constitutes a Markov chain, since the future state of the system depends on the present and not on the past states. We will analyze the case of $N = 4$.

Possible states of the system are $\{0,1,2,3,4\}$, so the MC has 5 states. The transition probabilities among the states are given as follows:

```
N=4;     %total number of particles
```

```
ns=N+1;  %number of MC states
%forming transition matrix
P=zeros(ns);
P(1,2)=1;    P(ns,ns-1)=1; %states 0 and N are ''reflective''
for j=2:ns-1
    i=j-1;                %number of particles in box A
P(j, j-1)=i/N             %A -> B
P(j, j+1)=(N-i)/N         %B -> A
end
```

Therefore, the transition matrix is

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/4 & 0 & 3/4 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 3/4 & 0 & 1/4 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

What is the most likely state of the system after $M = 11$ steps if all balls originally were in A?

```
pi0 =[0 0 0 0 1]  %probability of initial state i=4 is 1.
pi0 * P^11
%0    0.4995         0    0.5005         0
```

The most likely state is $i = 3$. For any even number of transitions, the most likely state is $i = 2$ with constant probability of $3/4$.

The stationary probabilities are found by solving the following system:

```
linsolve([(eye(ns)-P)'; ones(1,ns)],[ zeros(ns,1);1])
```

The stationary probabilities coincide with binomial $\mathcal{B}in(4+1,1/2)$ PDF.

```
st=[0.0625 0.25 0.375 0.25 0.0625]
st * P
%0.0625    0.2500    0.3750    0.2500    0.0625
```

MATLAB script ehrenfestsim.m simulates dynamic change of states in Ehrenfest model with $N = 20 \times 20 = 400$ particles, that are initially all in box A. Figure 5.16 summarizes the calculations. It shows the content of boxes A and B after 10,000 transitions, as well as the proportion of balls in each of the boxes.

Example 5.32. **Point-Accepted Mutation.** Point-accepted mutation (PAM) implements a simple theoretical model for scoring the alignment of protein sequences. Specifically, at a fixed position, the rate of mutation at each moment is assumed to be independent of previous events. Then the evolution of this fixed position in time can be treated as a Markov chain, where the

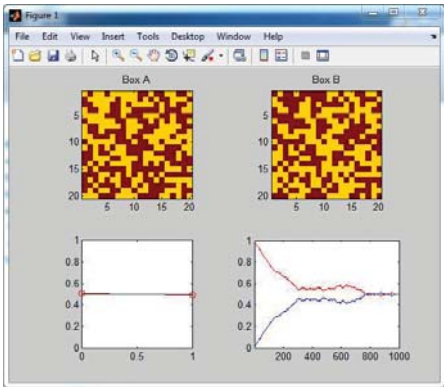


Fig. 5.16 Ehrenfest model simulation by `ehrenfestsim.m`. The top two panels show the contents of two boxes A and B after 10,000 transitions. The lower left panel shows the proportion of balls in the boxes (0 for A and 1 for B), and the lower right panel shows how the proportions changed over 10,000 transitions. The red curve is the proportion for box A.

PAM matrix represents its transition matrix. The original PAMs are 20×20 matrices describing the evolution of 20 standard amino acids (Dayhoff et al. 1978). As a simplified illustration, consider the case of a nucleotide sequence with only four states (A, T, G, and C). Assume that in a given time interval ΔT the probabilities that a given nucleotide mutates to each of the other three bases or remains unchanged can be represented by a 4×4 mutation matrix M:

$$M = \left(\begin{array}{c|cccc} & \text{A} & \text{T} & \text{G} & \text{C} \\ \hline \text{A} & 0.98 & 0.01 & 0.005 & 0.005 \\ \text{T} & 0.01 & 0.96 & 0.02 & 0.01 \\ \text{G} & 0.01 & 0.01 & 0.97 & 0.01 \\ \text{C} & 0.02 & 0.03 & 0.01 & 0.94 \end{array}\right)$$

Consider the fixed position with the letter T at $t = 0$:

$$s_0 = (0 \;\; 1 \;\; 0 \;\; 0).$$

Then, at times Δ, 2Δ, 10Δ, 100Δ, 1000Δ, and 10000Δ, by (5.19), the probabilities of the nucleotides (A, T, G, C) are $s_1 = s_0 M$,$s_2 = s_0 M^2$, $s_{10} = s_0 M^{10}$, $s_{100} = s_0 M^{100}$,$s_{1000} = s_0 M^{1000}$, and $s_{10000} = s_0 M^{10000}$, as given in the following table:

	Δ	2Δ	10Δ	100Δ	1000Δ	10000Δ
A	0.0100	0.0198	0.0909	0.3548	0.3721	0.3721
T	0.9600	0.9222	0.6854	0.2521	0.2465	0.2465
G	0.0200	0.0388	0.1517	0.2747	0.2651	0.2651
C	0.0100	0.0193	0.0719	0.1184	0.1163	0.1163

5.10 Exercises

5.1. **Phase I Clinical Trials and CTCAE Terminology.** In Phase I clinical trials, a safe dosage of a drug is assessed. In administering the drug, doctors are grading subjects' toxicity responses on a scale from 0 to 5. In CTCAE (Common Terminology Criteria for Adverse Events, National Institute of Health), Grade refers to the severity of adverse events. Generally, Grade 0 represents no measurable adverse events (sometimes omitted as a grade); Grade 1 events are mild; Grade 2 are moderate; Grade 3 are severe; Grade 4 are life-threatening or disabling; Grade 5 are fatal. This grading system inherently places a value on the importance of an event, although there is not necessarily "proportionality" among grades (a "2" is not necessarily twice as bad as a "1"). Some adverse events are difficult to "fit" into this point schema, but altering the general guidelines of severity scaling would render the system useless for comparing results between trials, which is an important purpose of the system.

Assume that based on a large number of trials (administrations to patients with renal cell carcinoma), the toxicity of the drug PNU (a murine Fab fragment of the monoclonal antibody 5T4 fused to a mutated superantigen staphylococcal enterotoxin A) at a particular fixed dosage is modeled by discrete random variable X,

X	0	1	2	3	4	5
Prob	0.620	0.190	0.098	0.067	0.024	0.001

.

Plot the PMF and CDF and find $\mathbb{E}X$ and $\mathbb{V}\mathrm{ar}(X)$.

5.2. **Mendel and Dominance.** Suppose that a specific trait, such as eye color or left-handedness, in a person is dependent on a pair of genes, and suppose that D represents a dominant and d a recessive gene. Thus, a person having DD is pure dominant and dd is pure recessive while Dd is a hybrid. The pure dominants and hybrids are alike in outward appearance. A child receives one gene from each parent.
Suppose two hybrid parents have 4 children. What is the probability that 3 out of 4 children have outward appearance of the dominant gene.

5.3. **Chronic Kidney Disease.** Chronic kidney disease (CKD) is a serious condition associated with premature mortality, decreased quality of life, and increased healthcare expenditures. Untreated CKD can result in end-stage renal disease and necessitate dialysis or kidney transplantation. Risk factors for CKD include cardiovascular disease, diabetes, hypertension, and obesity. To estimate the prevalence of CKD in the United States (overall and by health risk factors and other characteristics), the CDC (CDC's MMWR Weekly, 2007; Coresh et al., 2003) analyzed the most recent data from the National Health and Nutrition Examination Survey (NHANES). The total crude (i.e., not age-standardized) CKD prevalence estimate for adults aged > 20 years in the United States was 17%. By age group, CKD was more prevalent among persons aged > 60 years (40%) than among persons aged 40–59 years (13%) or 20–39 years (8%).
(a) From the population of adults aged > 20 years, 10 subjects are selected at random. Find the probability that 3 of the selected subjects have CKD.
(b) From the population of adults aged > 60, 5 subjects are selected at random. Find the probability that at least one of the selected have CKD.
(c) From the population of adults aged > 60, 16 subjects are selected at random and it was found that 6 of them had CKD. From this sample of 16, subjects are selected at random, *one-by-one with replacement,* and inspected. Find the probability that among 5 inspected (i) exactly 3 had CKD; (ii) at least one of the selected have CKD.
(d) From the population of adults aged > 60 subjects are selected at random until a subject is found to have CKD. What is the probability that exactly 3 subjects are sampled.
(e) Suppose that persons aged > 60 constitute 23% of the population of adults older than 20. For the other two age groups, 20–39, and 40–59, the percentages are 42% and 35%. Ten people are selected at random. What is the probability that 5 are from the > 60 group, 3 from the 20–39 group, and 2 from the 40-59 group.

5.4. **Experimenting to See All Possible Outcomes.** In a chemical experiment two outcomes are possible, A and A^c, with probabilities p and $q = 1 - p$. A student is repeating the experiment until both A and A^c are observed.
(a) Find the distribution of random variable X, the number of experiments necessary to observe A and A^c.
(b) What is the expected number of experiments?
(c) If the expected number of experiments is 3, what can you say about p?
Hint: Use $P(X = k) = P(X > k - 1) - P(X > k),\ k = 2,3,\ldots$. Argue that $P(X > k) = p^k + q^k$.

5.5. **Ternary Channel.** Refer to Exercise 3.40 in which a communication system was transmitting three signals, s_1, s_2, and s_3.
(a) If s_1 is sent $n = 1000$ times, find an approximation to the probability of the event that it was correctly received between 730 and 770 times, inclusive.
(b) If s_2 is sent $n = 1000$ times, find an approximation to the probability of the event that the channel did not switch to s_3 at all, that is, if 1,000 s_2 signals are sent and not a single s_3 was received. Can you use the same approximation as in (a)?

5.6. **Random Circular Sector with Cells.** On a circular plate, there are 400 randomly located cells. A part of the plate in the shape of a circular sector with central angle $\varphi = \frac{\pi}{100}$ (in radians) is selected at random.
Find an approximation to the probability that the number of cells in the selected sector is
(a) zero;
(b) 4 or more.
Hint: Argue that the number of cells in the selected area is Poisson with $\lambda = 2$.

5.7. **Conditioning a Poisson.** If $X_1 \sim \mathcal{P}oi(\lambda_1)$ and $X_2 \sim \mathcal{P}oi(\lambda_2)$ are independent, show that the distribution of X_1, given $X_1 + X_2 = n$, is binomial $\mathcal{B}in\left(n, \lambda_1/(\lambda_1 + \lambda_2)\right)$.

5.8. **Rh+ Plates.** Assume that there are 6 plates with red blood cells, three are Rh+ and three are Rh–.
Two plates are selected (a) with, (b) without replacement. Find the probability that one plate out of the 2 selected/inspected is of Rh+ type.
Now, increase the number of plates keeping the proportion of Rh+ fixed to 1/2. For example, if the total number of plates is 10000, 5000 of each type, what are the probabilities from (a) and (b)?

5.9. **Your Teammate's Misconceptions about Density and CDF.** Your teammate thinks that if f is a probability density function for the continuous random variable X, then $f(10)$ is the probability that $X = 10$. (a) Explain to your teammate why his/her reasoning is false.
Your teammate is not satisfied with your explanation and challenges you by asking, "If $f(10)$ is not the probability that $X = 10$, then just what does $f(10)$ signify?" (b) How would you respond?
Your teammate now thinks that if F is a cumulative probability density function for the continuous random variable X, then $F(5)$ is the probability that $X = 5$. (c) Explain why your teammate is wrong.
Your teammate then asks you, "If $F(5)$ is not the probability of $X = 5$, then just what does $F(5)$ represent?" (d) How would you respond?

5.10. **Falls among Elderly.** Falls are the second leading cause of unintentional injury-related death for people of all ages and the leading cause for people 60 years and older in the United States. Falls are also the most costly injury among older persons in the United States.
One in three adults aged 65 years and older falls annually.
(a) Find the probability that 3 among 11 adults aged 65 years and older will fall in the following year.
(b) Find the probability that among 110,000 adults aged 65 years and older the number of falls will be between 36,100 and 36,700, inclusive. Find the exact probability by assuming a binomial distribution for the number of falls, and an approximation to this probability via de Moivre's theorem; see page 252.

5.11. **Cell Clusters in 3D Petri Dishes.** The number of cell clusters in a 3D Petri dish has a Poisson distribution with mean $\lambda = 5$. Find the percentage of Petri dishes that have (a) 0 clusters, (b) at least one cluster, (c) more than 8 clusters, and (d) between 4 and 6 clusters. Use MATLAB and `poisspdf`, `poisscdf` functions.

5.12. **Left-Handed Twins.** The identical twin of a left-handed person has a 76 % chance of being left-handed, implying that left-handedness has partly genetic and partly environmental causes. Ten identical twins of ten left-handed persons are inspected for left-handedness. Let X be the number of left-handed among the inspected. What is the probability that X
(a) falls anywhere between 5 and 8, inclusive;
(b) is at most 6;
(c) is not less than 6.
(d) Would you be surprised if the number of left-handed among the 10 inspected was 3? Why or why not?

5.13. **Pot Smoking Is Not Cool!** A nationwide survey of seniors by the University of Michigan reveals that almost 70% disapprove of daily pot smoking, according to a report in Parade, September 14, 1980. If 12 seniors are selected at random and asked their opinion, find the probability that the number who disapprove of smoking pot daily is
(a) anywhere from 7 to 9;
(b) at most 5;
(c) not less than 8.

5.14. **Power Supply.** A power supply is connected to 20 independent loads. Each load is ON 30% of the time and draws a current of 0.75 amps. Let X be a current in the power supply at a particular moment.
(a) If X exceeds 13 amps, the power supply is declared to be in a critical regime. What is the probability of this happening?
(b) Find the probability that X is below 5 amps.

(c) Find the expectation and variance of X.

5.15. **Emergency Help by Phone.** The emergency hotline in a hospital tries to answer questions to its patient support within 3 minutes. The probability is 0.9 that a given call is answered within 3 minutes and the calls are independent.
(a) What is the expected total number of calls that occur until the first call is answered late?
(b) What is the probability that exactly one of the next 10 calls is answered late?

5.16. **Min of Three.** Let X_1, X_2, and X_3 be three mutually independent random variables, with a discrete uniform distribution on $\{1,2,3\}$, given as $P(X_i = k) = 1/3$ for $k = 1,2$ and 3.
(a) Let $M = \min\{X_1, X_2, X_3\}$. What is the distribution (probability mass function) and cumulative distribution function of M?
(b) What is the distribution (probability mass function) and cumulative distribution function of random variable $R = \max\{X_1, X_2, X_3\} - \min\{X_1, X_2, X_3\}$.

5.17. **Cystic Fibrosis in Japan.** Some rare diseases, including those of genetic origin, are life-threatening or chronically debilitating diseases that are of such low prevalence that special combined efforts are needed to address them. An accepted definition of low prevalence is a prevalence of less than 5 in a population of 10,000. A rare disease has such a low prevalence in a population that a doctor in a busy general practice would not expect to see more than one case in a given year.
Assume that cystic fibrosis, which is a rare genetic disease in most parts of Asia, has a prevalence of 2 per 10,000 in Japan. What is the probability that in a Japanese city of 15,000 there are
(a) exactly 3 incidences,
(b) at least one incidence,
of cystic fibrosis.

5.18. **Random Variables as Models.** Tubert-Bitter et al. (1996) found that the number of serious gastrointestinal reactions reported to the British Committee on Safety of Medicines was 538 out of 9,160,000 prescriptions of the anti-inflammatory drug *Piroxicam.*
(a) What is the rate of gastrointestinal reactions per 10,000 prescriptions?
(b) Using the Poisson model with the rate λ as in (a), find the probability of exactly two gastrointestinal reactions per 10,000 prescriptions.
(c) Find the probability of finding at least two gastrointestinal reactions per 10,000 prescriptions.

5.19. **Jack and Jill, Poisson, and Bayes' Rule.** Jack and Jill are partners in a typing service. Jill handles 60% of the typing work in their partnership. She makes errors (uncorrected errors) at an average rate of one

per 4 pages while Jack makes errors at a rate of one per page. Assume that for each typist these errors occur independently and at a constant rate throughout the paper. Assume, in addition, that for both typists the number of errors per page is well approximated by a Poisson distribution.
You submit a 5-page paper to the partnership for typing without knowing whether Jack or Jill will type it.
(a) It comes back error-free. What is the probability that Jack typed it?
(b) What is the probability that Jack typed the paper if 3 errors are found.

5.20. **Variance of Difference of Two Multinomial Components.** Let $(X_1, X_2, \ldots, X_k)$ be a discrete random vector with multinomial $\mathcal{M}n(n, p_1, \ldots, p_k)$ distribution. Show that the variance of $X_i - X_j$ is $n(p_i + p_j - (p_i - p_j)^2)$.

5.21. **A 2D PDF.** Let

$$f(x,y) = \begin{cases} \frac{3}{8}(x^2 + 2xy), & 0 \le x \le 1,\ 0 \le y \le 2 \\ 0, & \text{else} \end{cases}$$

be a bivariate PDF of a random vector (X, Y).
(a) Show that $f(x,y)$ is a density.
(b) Show that marginal distributions are $f_X(x) = \frac{3}{2}x + \frac{3}{4}x^2,\ 0 \le x \le 1$, and $f_Y(y) = \frac{3+8y}{4+12y},\ 0 \le y \le 2$.
(c) Show $\mathbb{E}X = 11/16$ and $\mathbb{E}Y = 5/4$.
(d) Show that conditional distributions are

$$f(x|y) = \frac{3x(x+2y)}{1+3y},\ 0 \le x \le 1, \quad \text{for any fixed } y \in [0,2],$$

$$f(y|x) = \frac{2y+x}{4+2x},\ 0 \le y \le 2, \quad \text{for any fixed } x \in [0,1].$$

(e) Show that

$$\mathbb{E}X|Y = \frac{3+8Y}{4+12Y} \quad \text{and} \quad \mathbb{E}Y|X = \frac{8+3X}{6+3X}.$$

(f) Demonstrate that iterated expectation rule (5.12) is satisfied,

$$\mathbb{E}(\mathbb{E}X|Y) = 11/16 \quad \text{and} \quad \mathbb{E}(\mathbb{E}Y|X) = 5/4.$$

5.22. **2-D Density Tasks.** If

$$f(x,y) = \begin{cases} \frac{1}{4}xy(x+y)\exp\{-x-y\}, & 0 \le x < \infty,\ 0 \le y < \infty \\ 0, & \text{else} \end{cases}$$

Find
(a) marginal distribution $f_X(x)$,

(b) conditional distribution $f(x|y)$,
(c) expectation $\mathbb{E}X$, and
(d) conditional expectation $\mathbb{E}X|Y$.
(f) Are X and Y independent? Explain.

5.23. **Conditional Variance.** In the context of Example 5.22 show that

$$\mathbb{V}\text{ar}\,(Y|X=x) = \frac{(1-x)^2}{12}.$$

5.24. **Additivity of Gammas.** If $X_i \sim \mathcal{G}a(r_i,\lambda)$ are independent, prove that $Y = X_1 + \cdots + X_n$ is distributed as gamma with parameters $r = r_1 + r_2 + \cdots + r_n$ and λ; that is, $Y \sim \mathcal{G}a(r,\lambda)$.

5.25. **Memoryless Property.** Prove that the geometric $\mathcal{G}e(p)$ distribution ($\mathbb{P}(X=x) = (1-p)^x p, x = 0,1,2,\dots$) and the exponential distribution ($\mathbb{P}(X \le x) = 1 - e^{-\lambda x},\ x \ge 0, \lambda \ge 0$) both possess the *Memoryless Property*; that is, they satisfy

$$\mathbb{P}(X \ge v | X \ge u) = \mathbb{P}(X \ge v-u),\ v \ge u.$$

5.26. **Rh System.** Rh antigens are transmembrane proteins with loops exposed at the surface of red blood cells. They appear to be used for the transport of carbon dioxide and/or ammonia across the plasma membrane. They are named for the rhesus monkey in which they were first discovered. There are a number of different Rh antigens. Red blood cells that are *Rh positive* express the antigen designated as D. About 15% of the population do not have RhD antigens and thus are *Rh negative*. The major importance of the Rh system for human health is to avoid the danger of RhD incompatibility between a mother and her fetus.
(a) From the general population 8 people are randomly selected and checked for their Rh factor. Let X be the number of Rh negative among the eight selected. Find $\mathbb{P}(X=2)$.
(b) In a group of 16 patients, three members are Rh negative. Eight patients are selected at random. Let Y be the number of Rh negative among the eight selected. Find $\mathbb{P}(Y=2)$.
(c) From the general population subjects are randomly selected and checked for their Rh factor. Let Z be the number of Rh positive subjects before the first Rh negative subject is selected. Find $\mathbb{P}(Z=2)$.
(d) Identify the distributions of the random variables in (a), (b), and (c).
(e) What are the expectations and variances for the random variables in (a), (b), and (c)?

5.27. **Blood Types.** The prevalence of blood types in the US population is O+: 37.4%, A+: 35.7%, B+: 8.5%, AB+: 3.4%, O–: 6.6%, A–: 6.3%, B–: 1.5%, and AB–: 0.6%.

(a) A sample of 24 subjects is randomly selected from the US population. What is the probability that 8 subjects are O+? Random variable X describes the number of O+ subjects among 24 selected. Find $\mathbb{E}X$ and $\mathbb{V}\text{ar}\,X$.
(b) Among 16 subjects, eight are O+. From these 16 subjects, five are selected at random as a group. What is the probability that among the five selected at most two are O+?
(c) Use Poisson approximation to find the probability that among 500 randomly selected subjects the number of AB– subjects is at least 1.
(d) Random sampling from the population is performed until the first subject with B+ blood type is found. What is the expected number of subjects sampled?

5.28. **Variance of the Exponential.** Show that for an exponential random variable X with density $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$, the variance is $1/\lambda^2$.
Hint: You can use the fact that $\mathbb{E}X = 1/\lambda$. To find $\mathbb{E}X^2$ you need to repeat the integration-by-parts twice.

5.29. **Equipment Aging.** Suppose that the lifetime T of a particular piece of laboratory equipment (in 1000 hour units) is an exponentially distributed random variable such that $\mathbb{P}(T > 10) = 0.8$.
(a) Find the "rate" parameter, λ.
(b) What are the mean and standard deviation of the random variable T?
(c) Find the median, the first and third quartiles, and the inter-quartile range of the lifetime T. Recall that for an exponential distribution, you can find any percentile exactly.

5.30. **A Simple Continuous Random Variable.** Assume that the measured responses in an experiment can be modeled as a continuous random variable with density

$$f(x) = \begin{cases} c - x, & 0 \leq x \leq c \\ 0, & \text{else} \end{cases}$$

(a) Find the constant c and sketch the graph of the density $f(x)$.
(b) Find the CDF $F(x) = \mathbb{P}(X \leq x)$, and sketch its graph.
(c) Find $\mathbb{E}(X)$ and $\mathbb{V}\text{ar}\,(X)$.
(d) What is $\mathbb{P}(X \leq 1/2)$?

5.31. **2D Continuous Random Variable Question.** A two-dimensional random variable (X, Y) is defined by its density function, $f(x,y) = Cxe^{-xy}$, $0 \leq x \leq 1$; $0 \leq y \leq 1$.
(a) Find the constant C.
(b) Find the marginal distributions of X and Y.

5.32. **Insulin Sensitivity.** The insulin sensitivity (SI), obtained in a glucose tolerance test is one of the patient responses used to diagnose type II diabetes. Leading a sedative lifestyle and being overweight are well-established risk factors for type II diabetes. Hence, body mass index (BMI) and hip to waist ratio (HWR = HIP/WAIST) may also predict an impaired insulin sensitivity. In an experiment, 106 males (coded 1) and 126 females (coded 2) had their SI measured and their BMI and HWR registered. Data ( `diabetes.xls`) are available on the text web page. For this exercise you will need only the 8th column of the data set, which corresponds to the SI measurements.
(a) Find the sample mean and sample variance of SI.
(b) A gamma distribution with parameters α and β seems to be an appropriate model for SI. What α, β should be chosen so that the $\mathbb{E}X$ matches the sample mean of SI and $\mathbb{V}\text{ar}\,X$ matches the sample variance of SI.
(c) With α and β selected as in (b), simulate a random sample from gamma distribution with a size equal to that of SI ($n = 232$). Use `gamrnd`. Compare two histograms, one with the simulated values from the gamma model and the second from the measurements of SI. Use 20 bins for the histograms. Comment on their similarities/differences.
(d) Produce a Q–Q plot to compare the measured SI values with the model. Suppose that you selected $\alpha = 3$ and $\beta = 3.3$, and that `dia` is your data set. Take $n = 232$ equally spaced points between [0,1] and find their gamma quantiles using `gaminv(points,alpha,beta)`. If the model fits the data, these theoretical quantiles should match the ordered sample.
Hint: (i) Here MATLAB's parametrization of gamma density is used, $\alpha = r$ and $\beta = 1/\lambda$. In terms of α and β, $\mathbb{E}X = \alpha\beta$ and $\mathbb{V}\text{ar}\,X = \alpha\beta^2$.
(ii) The plot of theoretical quantiles against the ordered sample is called a Q–Q plot. An example of producing a Q–Q plot in MATLAB is as follows:

```
xx = 0.5/232: 1/232: 1;
yy=gaminv(xx, 3, 3.3);
plot(yy, sort(dia(:,8)),'*')
hold on
plot(yy, yy,'r-')
```

5.33. **Correlation between a Uniform and Its Power.** Suppose that X has uniform $\mathcal{U}(-1,1)$ distribution and that $Y = X^k$.
(a) Show that for k even, $\mathbb{C}\text{orr}(X,Y) = 0$.
(b) Show that for arbitrary k, $\mathbb{C}\text{orr}(X,Y) \to 0$, when $k \to \infty$.

5.34. **Precision of Lab Measurements.** The error X in measuring the weight of a chemical sample is a random variable with PDF.

$$f(x) = \begin{cases} \frac{3x^2}{16}, & -2 < x < 2 \\ 0, & \text{otherwise} \end{cases}$$

(a) A measurement is considered to be *accurate* if $|X| < 0.5$. Find the probability that a randomly chosen measurement can be classified as accurate.
(b) Find and sketch the graph of the cumulative distribution function $F(x)$.
(c) The loss in dollars, which is caused by measurement error, is $Y = X^2$. Find the mean of Y (expected loss).
(d) Compute the probability that the loss is less than \$3.
(e) Find the median of Y.

5.35. **Lifetime of Cells.** Cells in the human body have a wide variety of life spans. One cell may last a day; another a lifetime. Red blood cells (RBC) have a lifespan of several months and cannot replicate, which is the price RBCs pay for being specialized cells. The lifetime of a RBC can be modeled by an exponential distribution with density $f(t) = \frac{1}{\beta} e^{-t/\beta}$, where $\beta = 4$ (in units of months). For simplicity, assume that when a particular RBC dies, it is instantly replaced by a newborn RBC of the same type. For example, a replacement RBC could be defined as any new cell born approximately at the time when the original cell died.
(a) Find the expected lifetime of a single RBC. Find the probability that the cell's life exceeds 150 days. *Hint:* Days have to be expressed in units of β.
(b) A single RBC and its replacements are monitored over the period of 1 year. How many deaths/replacements are observed on average? What is the probability that the number of deaths/replacements exceeds 5. *Hint:* Utilize a link between exponential and Poisson distributions. In simple terms, if lifetimes are exponential with parameter β, then the number of deaths/replacements in the time interval $[0,t]$ is Poisson with parameter $\lambda = t/\beta$. Time units for t and β have to be the same.
(c) Suppose that a single RBC and two of its replacements are monitored. What is the distribution of their total lifetime? Find the probability that their total lifetime exceeds 1 year. *Hint:* Consult the gamma distribution. If n random variables are exponential with parameter β, then their sum is gamma distributed with parameters $\alpha = n$ and β.
(d) A particular RBC is observed $t = 2.2$ months after its birth and is found to still be alive. What is the probability that the total lifetime of this cell will exceed 7.2 months?

5.36. ***k*-out-of-*n* and Weibull Lifetime.** Engineering systems of type k-out-of-n are described in Exercise 3.10. Suppose that a k-out-of-n system consists of n identical and independent elements for which the lifetime has Weibull distribution with parameters r and λ. More precisely, if T is a lifetime of a component,

$$P(T \geq t) = e^{-\lambda t^r}.$$

Time t is in units of months, and consequently, rate parameter λ is in units $(\text{month})^{-1}$. Parameter r is dimensionless.
Assume that $n = 20, k = 7$, $r = 3/2$ and $\lambda = 1/4$.
(a) Find the probability that the k-out-of-n system is working at time $t = 3$.
(b) Plot this probability as a function of time.
(c) At time $t = 3$ the system is found operational. What is the distribution of the number of failed components? What is the expected number of failed components?
Hint: For each component the probability of the system working at time t is $p = e^{-1/2t^{3/2}}$. The probability that a k-out-of-n system is operational corresponds to the tail probability of binomial distribution: $\mathbb{P}(X \geq k)$, where X is the number of components working. Use `binocdf` and be careful about the discrete nature of the binomial distribution.
In part (c), first find the probability that a component fails in the time interval $[0,3]$. Denote this probability with f. Then, the number of failed components Y cannot exceed $n - k$, and given the independence of components, it is binomial. That is, $Y \sim \mathcal{B}in(n-k, f)$.

5.37. **Silver-Coated Nylon Fiber.** Silver-coated nylon fiber is used in hospitals for its anti-static electricity properties, as well as for antibacterial and antimycotic effects. In the production of silver-coated nylon fibers, the extrusion process is interrupted from time to time by blockages occurring in the extrusion dyes. The time in hours between blockages, T, has an exponential $\mathcal{E}(1/10)$ distribution, where $1/10$ is the rate parameter.
Find the probabilities that
(a) a run continues for at least 10 hours,
(b) a run lasts less than 15 hours, and
(c) a run continues for at least 20 hours, given that it has lasted 10 hours.
Use MATLAB and `expcdf` function. Be careful about the parametrization of exponentials in MATLAB.

5.38. **Xeroderma Pigmentosum.** Xeroderma pigmentosum (XP) was first described in 1874 by Hebra et al. XP is the condition characterized as dry, pigmented skin. It is a hereditary condition with an incidence of 1:250,000 live births (Robbin et al., 1974). In a city with a population of 1,000,000, find the distribution of the number of people with XP. What is the expected number? What is the probability that there are no XP-affected subjects?

5.39. **Failure Time.** Let X model the time to failure (in years) of a Beckman Coulter TJ-6 laboratory centrifuge. Suppose that the PDF of X is $f(x) = c/(3+x)^3$ for $x \geq 0$.

(a) Find the value of c such that f is a legitimate PDF.
(b) Compute the mean and median time to failure of the centrifuge.

5.40. **Resistors.** If n resistors with resistances $R_1, R_2, \ldots, R_n$ are connected in-line, the total resistance R is

$$R = R_1 + R_2 + \cdots + R_n.$$

If the connection is parallel (resistors branch out from a single node, and join up again somewhere else in the circuit), then

$$1/R = 1/R_1 + 1/R_2 + \cdots + 1/R_n.$$

Suppose that resistances of two resistors are independent random variables with means $\mu_1 = 2\ \Omega$ and $\mu_2 = 3\ \Omega$ and variances $\sigma_1^2 = 0.02^2\ \Omega^2$ and $\sigma_2^2 = 0.01^2\ \Omega^2$.
Estimate the mean and variance of the total resistance if the resistors are connected
(a) in line;
(b) parallel.
(c) In the case of parallel connection, assume that $R_1 \sim \mathcal{IG}(r_1, \lambda)$ and $R_2 \sim \mathcal{IG}(r_2, \lambda)$, where $r_1, r_2 > 1$ and λ is the rate parameter. Thus, R is $\mathcal{IG}(r_1 + r_2, \lambda)$ and $ER = \lambda/(r_1 + r_2 - 1)$. How does this exact value compare to the first-order approximation

$$ER = g(ER_1, ER_2)$$

for $g(x_1, x_2) = 1/(1/x_1 + 1/x_2)$?

5.41. **Beta Fit.** Assume that the fraction of impurities in a certain chemical solution is modeled by a Beta $\mathcal{B}e(\alpha, \beta)$ distribution with known parameter $\alpha = 1$. The average fraction of impurities is 0.1.
(a) Find the parameter β.
(b) What is the standard deviation of the fraction of impurities?
(c) Find the probability that the fraction of impurities exceeds 0.25.

5.42. **Uncorrelated but Possibly Dependent.** Show that for any two random variables X and Y with equal second moments, the variables $Z = X + Y$ and $W = X - Y$ are uncorrelated. Note, that Z and W could be dependent.

5.43. **Nights of Mr. Jones.** If Mr. Jones had insomnia one night, the probability that he would sleep well the following night is 0.6; otherwise, he would have insomnia. If he slept well one night, the probabilities of sleeping well or having insomnia the following night would be 0.5 each.
On Monday night Mr. Jones had insomnia. What is the probability that he had insomnia on the following Friday night?

5.44. **Stationary Distribution of MC.** Consider a Markov chain with transition matrix

$$\mathbf{P} = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{pmatrix}.$$

(a) Show that all entries of $\mathbf{P}^2$ are strictly positive.
(b) Using MATLAB, find $\mathbf{P}^{100}$ and guess what the stationary distribution $\pi = (\pi_1, \pi_2, \pi_3)$ would be. Confirm your guess by solving the equation $\pi = \pi\mathbf{P}$, which gives the exact stationary distribution. *Hint:* The system $\pi = \pi\mathbf{P}$ needs a closure equation $\pi_1 + \pi_2 + \pi_3 = 1$.

5.45. **Influence of Two Previous Trials.** In a potentially infinite sequence of trials, the probability of success is $1/2$, unless the previous two trials resulted in a success. In this case, the probability of success is $2/3$. Code successes as 1 and failures as 0. Such binary sequence defines a MC where the states are 00, 01, 10, and 11; see Figure 5.17.
(a) Write down the transition matrix $\boldsymbol{P}$.

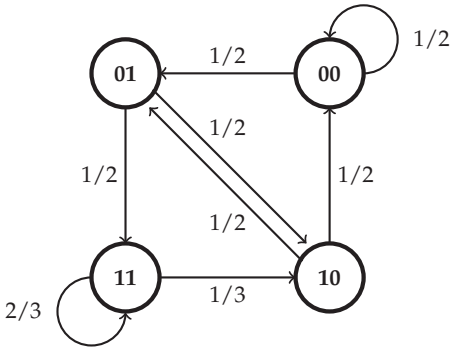


Fig. 5.17 Markov chain schematic graph

(b) Using MATLAB find $\boldsymbol{P}^{100}$ and argue that the stationary probabilities for 00, 01, 10, and 11 are $2/9$, $2/9$, $2/9$, and $1/3$, respectively. Confirm this numerical result by solving the system

$$\pi = \pi\boldsymbol{P},$$

where $\boldsymbol{P}$ is the transition matrix and $\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$ is the row vector of stationary probabilities. Since $\boldsymbol{P}$ is not of full rank, the equation $\sum_i^4 \pi_i = 1$ completes the system.

HINT: `> linsolve([(eye(4)-P)'; ones(1,4)],[zeros(4,1); 1])`

(c) Argue that the proportion of successes in a long run is 5/9.

5.46. **Heat Production by a Resistor.** Joule's Law states that the amount of heat produced by a resistor is

$$Q = I^2 \, R \, T,$$

where
Q is heat energy (in Joules),
I is current (in Amperes),
R is resistance (in Ohms), and
T is duration of time (in seconds).
Suppose that in an experiment, I, R, and T are independent random variables with means $\mu_I = 10$ A, $\mu_R = 30$ Ω, and $\mu_T = 120$ s. Suppose that the variances are $\sigma_I^2 = 0.01$ A^2, $\sigma_R^2 = 0.02$ Ω^2, and $\sigma_T^2 = 0.001$ s^2. Estimate the mean μ_Q and the variance σ_Q^2 of the produced energy Q.

MATLAB AND WINBUGS FILES AND DATA SETS USED IN THIS CHAPTER

http://statbook.gatech.edu/Ch5.RanVar/

apgar.m, bookplots.m, circuitgenbin.m, corneoretinal.m, covcord2d.m, dexp.m, Discrete.m, empiricalcdf.m, histp.m, hyper.m, lefthanded.m, lifetimecells.m, mamopixels.m, markovchain.m, MCEhrenfest.m, melanoma.m, mingling.m, plotbino.m, plotsdistributions.m, plotuniformdist.m, randdirichlet.m, stringerror.m

hearttransplant1.odc, hearttransplant2.odc, lifetimecells.odc, simulationc.odc, simulationd.odc

diabetes.xls

CHAPTER REFERENCES

Apgar, V. (1953). A proposal for a new method of evaluation of the newborn infant. *Curr. Res. Anesth. Analg.*, **32**, (4), 260–267. PMID 13083014.

CDC (2007). *Morbidity and Mortality Weekly Report.* **56**, 8, 161–165.
Coresh, J., Astor, B. C., Greene, T., Eknoyan, G., and Levey, A. S. (2003). Prevalence of chronic kidney disease and decreased kidney function in the adult US population: 3rd national health and nutrition examination survey. *Am. J. Kidney Dis.*, **41**, 1–12.
Dayhoff, M. O., Schwartz, R., and Orcutt, B. C. (1978). A model of evolutionary change in proteins. Atlas of protein sequence and structure. *Nat. Biomed. Res. Found.*, **5**, Suppl. 3, 345–358.
du Bois-Reymond, E. H. (1848). *Untersuchungen Ueber Thierische Elektricität*, Vol. 1, G. Reimer, Berlin.
Ehrenfest, P. und T. (1907). Uber zwei bekannte Einwände gegen das Boltzmannsche H-Theorem. âĂŐPhysik. Z., **8**, 311–331.
Gjini, A., Stuart, J. M., George, R. C., Nichols, T., and Heyderman, R. S. (2004). Capture-recapture analysis and pneumococcal meningitis estimates in England. *Emerg. Infect. Dis.*, **10**, 1, 87–93.
Hebra F. and Kaposi M. (1874). On diseases of the skin including exanthemata. *New Sydenham Soc.*, **61**, 252–258.
Montmort, P. R. (1714). *Essai d'Analyse sur les Jeux de Hazards*, 2ed. Jombert, Paris.
Pielou, E.C. (1961). Segregation and symmetry in two-species populations as studied by nearest-neighbor relationships. *J. Ecol.*, **49**, 2, 255–269.
Robbin, J. H., Kraemer, K. H., Lutzner, M. A., Festoff, B. W., and Coon, H. P. (1974). Xeroderma pigmentosum: An inherited disease with sun sensitivity, multiple cutaneous neoplasms and abnormal DNA repair. *Ann. Intern. Med.*, **80**, 221–248.
Ross, M. S. (2010a). *A First Course in Probability*, Pearson Prentice-Hall.
Ross, M. S. (2010b) *Introduction to Probability Models*, 10th ed. Academic Press, Burlington.
Tubert-Bitter, P., Begaud, B., Moride, Y., Chaslerie, A., and Haramburu, F. (1996). Comparing the toxicity of two drugs in the framework of spontaneous reporting: A confidence interval approach. *J. Clin. Epidemiol.*, **49**, 121–123.
Weibull, W. (1951). A statistical distribution function of wide applicability. *J. Appl. Mech.*, **18**, 293–297.

Chapter 6
Normal Distribution

The adjuration to be normal seems shockingly repellent to me.

– Karl Menninger

WHAT IS COVERED IN THIS CHAPTER

- Definition of Normal Distribution, Bivariate Case
- Standardization, Quantiles of Normal Distribution, Sigma Rules
- Linear Combinations of Normal Random Variables
- Central Limit Theorem, de Moivre's Approximation
- Distributions Related to Normal: Chi-Square, Wishart, t, F, Log-normal, and Some Noncentral Distributions
- Transformations to Normality

6.1 Introduction

In Chapters 2 and 5 we occasionally referred to a normal distribution either informally (bell-shaped distributions/histograms) or formally, as in Section 5.5.3, where the normal density and its moments were briefly introduced. This chapter is devoted to the normal distribution due to its importance in statistics. What makes the normal distribution so important? The normal distribution is the proper statistical model for many natural and social phenomena. But even if some measurements cannot be modeled by the

normal distribution (it could be skewed, discrete, multimodal, etc.), their sample means would closely follow the normal law, under very mild conditions. The central limit theorem covered in this chapter makes it possible to use probabilities associated with the normal curve to answer questions about the sums and averages in sufficiently large samples. This translates to the ubiquity of normality – many estimators, test statistics, and nonparametric tests covered in later chapters of this text are approximately normal, when sample sizes are not small (typically larger than 20 to 30), and this asymptotic normality is used in a substantial way. Several other important distributions can be defined through a normal distribution. Also, normality is a quite stable property – an arbitrary linear combination of normal random variables remains normal. The property of linear combinations of random variables preserving the distribution of their components is not shared by any other probability law and is a characterizing property of a normal distribution.

6.2 Normal Distribution

In 1738, Abraham de Moivre developed the normal distribution as an approximation to the binomial distribution, and it was subsequently used by Laplace in 1783 to study measurement errors and by Gauss in 1809 in the analysis of astronomical data. The name *normal* came from Quetelet, who demonstrated that many human characteristics distributed themselves in a bell-shaped manner (centered about the "average man," *l'homme moyen*), including such measurements as chest girths of 5,738 Scottish soldiers, the heights of 100,000 French conscripts, and the body weight and height of people he measured. From his initial research on height and weight has evolved the internationally recognized measure of obesity called the Quetelet index (QI), or body mass index (BMI), QI = (weight in kilograms)/(squared height in meters).

Table 6.1 provides frequencies of chest sizes of 5,738 Scottish soldiers as well as the relative frequencies. Using this now famous data set, Quetelet argued that many human measurements distribute as normal. Figure 6.1 gives a normalized histogram of Quetelet's data set with superimposed normal density in which the mean and the variance are taken as the sample mean (39.8318) and sample variance (2.0496^2).

The PDF for a normal random variable with mean μ and variance σ^2 is

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

The distribution function is computed using integral approximation because no closed form exists for the antiderivative of $f(x)$; this is generally

Table 6.1 Chest sizes of 5738 Scottish soldiers, data compiled from the 13th edition of the *Edinburgh Medical Journal* (1817).

Size	Frequency	Relative frequency (in %)
33	3	0.05
34	18	0.31
35	81	1.41
36	185	3.22
37	420	7.32
38	749	13.05
39	1073	18.70
40	1079	18.80
41	934	16.28
42	658	11.47
43	370	6.45
44	92	1.60
45	50	0.87
46	21	0.37
47	4	0.07
48	1	0.02
Total	5738	99.99

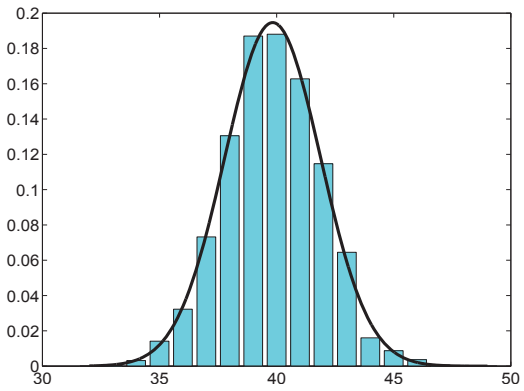


Fig. 6.1 Normalized bar plot of Quetelet's data set. Superimposed is the normal density with mean $\mu = 39.8318$ and variance $\sigma^2 = 2.0496^2$.

not a problem for practitioners because most software packages will compute interval probabilities numerically. In MATLAB, `normcdf(x, mu, sigma)` and `normpdf(x, mu, sigma)` calculate the CDF and PDF at x, and `norminv(p, mu, sigma)` computes the inverse CDF at given probability p, that is, the p-quantile. Equivalently, a normal CDF can be expressed in terms of a special function called the *error integral*:

$$\texttt{erf(x)} = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

It holds that `normcdf(x)= 1/2+1/2*erf(x/sqrt(2))`. A random variable X with a normal distribution will be denoted $X \sim \mathcal{N}(\mu,\sigma^2)$.

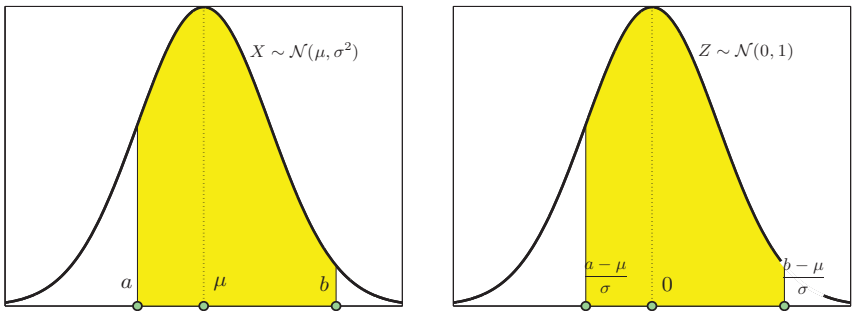
In addition to software, CDF values are often given in tables. Such tables contain only quantiles and CDF values for the *standard* normal distribution, $Z \sim \mathcal{N}(0,1)$, for which $\mu = 0$ and $\sigma^2 = 1$. Such tables are sufficient since an arbitrary normal random variable X can be *standardized* to Z if its mean and variance are known:

$$X \sim \mathcal{N}(\mu,\sigma^2) \quad \longrightarrow \quad Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1).$$

For a standard normal random variable, Z the PDF is denoted by ϕ, and CDF by Φ,

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)\, dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt. \quad [\texttt{normcdf(x)}]$$

Suppose we are interested in the probability that a random variable X distributed as $\mathcal{N}(\mu,\sigma^2)$ falls between two bounds a and b, $\mathbb{P}(a < X < b)$. It is irrelevant whether the bounds are included or not since the normal distribution is continuous and $\mathbb{P}(a < X < b) = \mathbb{P}(a \le X \le b)$. Also, any of the bounds can be infinite.



(a) (b)

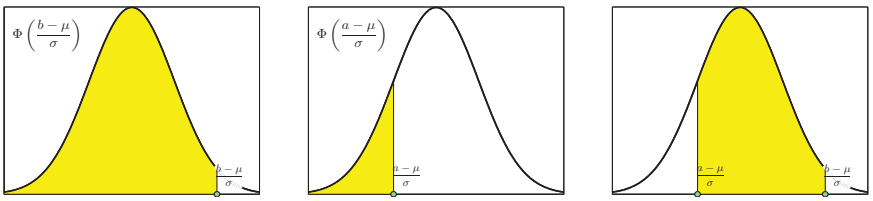
Fig. 6.2 Illustration of the relation $\mathbb{P}(a \le X \le b) = \mathbb{P}\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right)$.

In terms of Φ,

$$X \sim \mathcal{N}(\mu,\sigma^2):$$
$$\mathbb{P}(a \le X \le b) = \mathbb{P}\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Figures 6.2 and 6.3 provide the illustration. In MATLAB:

```
normcdf((b-mu)/sigma) - normcdf((a - mu)/sigma)
       %or equivalently
normcdf(b, mu, sigma) - normcdf(a, mu, sigma)
```



(a) (b) (c)

Fig. 6.3 Calculation of $\mathbb{P}(a \le X \le b)$ for $X \sim \mathcal{N}(\mu,\sigma^2)$. (a) $\mathbb{P}(X \le b) = \mathbb{P}(Z \le \frac{b-\mu}{\sigma}) = \Phi\left(\frac{b-\mu}{\sigma}\right)$; (b) $\mathbb{P}(X \le a) = \mathbb{P}(Z \le \frac{a-\mu}{\sigma}) = \Phi\left(\frac{a-\mu}{\sigma}\right)$; (c) $\mathbb{P}(a \le X \le b)$ as the difference of the two probabilities in (a) and (b).

Note that when the bounds are infinite, since Φ is a CDF,

$$\Phi(-\infty) = 0, \text{ and } \Phi(\infty) = 1.$$

Traditional statistics textbooks provide tables of cumulative probabilities for the standard normal distribution, $p = \Phi(x)$, for values of x typically between –3 and 3 with an increment of 0.01. The tables have been used in two ways: (i) directly, that is, for a given x the user finds $p = \Phi(x)$; and (ii) inversely, given p, one finds approximately what x gives $\Phi(x) = p$, which is of course a p-quantile of the standard normal. Given the limited precision of the tables, the results in direct and inverse uses have been approximate.

In MATLAB, the tables can be reproduced by a single line of code:

```
x=(-3:0.01:3)'; tables=[x normcdf(x)]
```

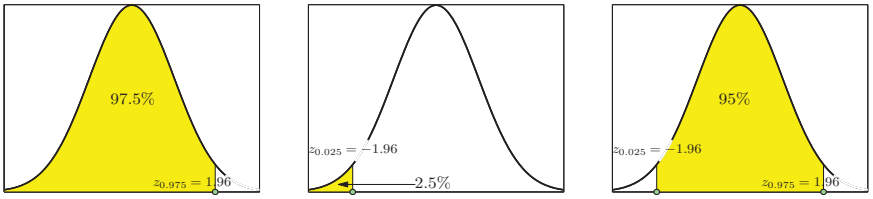
Similarly, the normal p-quantiles z_p defined as $p = \Phi(x_p)$ can be tabulated as

```
probs=(0.005:0.005:0.995)'; tables=[probs norminv(probs)]
```

There are several normal quantiles that are frequently used in the construction of confidence intervals and tests; these are the 0.9, 0.95, 0.975, 0.99, 0.995, and 0.9975 quantiles,

$z_{0.9} = 1.28155 \approx 1.28$ $z_{0.95} = 1.64485 \approx 1.64$ $z_{0.975} = 1.95996 \approx 1.96$
$z_{0.99} = 2.32635 \approx 2.33$ $z_{0.995} = 2.57583 \approx 2.58$ $z_{0.9975} = 2.80703 \approx 2.81$

For example, the 0.975 quantile of the normal is $z_{0.975} = 1.96$. This is equivalent to saying that 95% of the area below the standard normal density $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\}$ lies between –1.96 and 1.96. Note that the shortest interval containing $1-\alpha$ probability is defined by quantiles $z_{\alpha/2}$ and $z_{1-\alpha/2}$ (see Figure 6.4 as an illustration for $\alpha = 0.05$). Since the standard normal density is symmetric about 0, $z_p = -z_{1-p}$.



(a) (b) (c)

Fig. 6.4 (a) Normal quantiles (a) $z_{0.975} = 1.96$, (b) $z_{0.025} = -1.96$, and (c) 95% area between quantiles –1.96 and 1.96.

6.2.1 Sigma Rules

Sigma rules state that for any normal distribution, the probability that an observation will fall in the interval $\mu \pm k\sigma$ for $k = 1, 2$, and 3 is 68.27%, 95.45%, and 99.73%, respectively. More precisely,

$$\begin{array}{ll} \mathbb{P}(\mu-\sigma < X < \mu+\sigma) = \mathbb{P}(-1 < Z < 1) = \Phi(1) - \Phi(-1) = 0.682689 & \approx 68.27\% \\ \mathbb{P}(\mu-2\sigma < X < \mu+2\sigma) = \mathbb{P}(-2 < Z < 2) = \Phi(2) - \Phi(-2) = 0.954500 & \approx 95.45\% \\ \mathbb{P}(\mu-3\sigma < X < \mu+3\sigma) = \mathbb{P}(-3 < Z < 3) = \Phi(3) - \Phi(-3) = 0.997300 & \approx 99.73\% \end{array}$$

Have you ever wonder about the origin of the term *Six Sigma*? It does not involve $\mathbb{P}(\mu - 6\sigma < X < \mu + 6\sigma)$ as one may expect.

The Six Sigma doctrine is a standard according to which an item with measurement $X \sim \mathcal{N}(\mu, \sigma^2)$ should satisfy $X < 6\sigma$ to be conforming if μ is allowed to vary between -1.5σ and 1.5σ.

Thus, effectively, accounting for the variability in the mean, the Six Sigma constraint becomes

$$\mathbb{P}(X < \mu + 4.5\sigma) = P(Z < 4.5) = \Phi(4.5) = 0.99999660.$$

This means that only 3.4 items per million produced are allowed to exceed $\mu + 4.5\sigma$ (be defective). Such standard of quality was set by the Motorola Company in the 1980s, and it evolved into a doctrine for improving efficiency and quality in management.

6.2.2 *Bivariate Normal Distribution**

When the components of a random vector have a normal distribution, we say that the vector has a multivariate normal distribution. For independent components, the density of a multivariate distribution is simply the product of the univariate densities. When components are correlated, the distribution involves the covariance matrix that describes the correlation. Next we discuss the bivariate normal distribution, which will be important later on, in the context of correlation and regression.

The pair (X,Y) is distributed as bivariate normal $\mathcal{N}_2(\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho)$ if the joint density is

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_y)^2}{\sigma_Y^2}\right]\right\}. \tag{6.1}$$

The parameters $\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2$, and ρ are

$$\mu_X = \mathbb{E}(X),\ \mu_Y = \mathbb{E}(Y),\ \sigma_X^2 = \mathbb{V}\mathrm{ar}\,(X),\ \sigma_Y^2 = \mathbb{V}\mathrm{ar}\,(Y),\ \text{and}\ \ \rho = \mathbb{C}\mathrm{orr}(X,Y).$$

One can define bivariate normal distribution with a density as in (6.1) by transforming two independent, standard normal random variables Z_1 and Z_2,

$$\begin{aligned} X &= \mu_1 + \sigma_X Z_1, \\ Y &= \mu_2 + \rho\sigma_Y Z_1 + \sqrt{1-\rho^2}\sigma_Y Z_2. \end{aligned}$$

The marginal distributions in (6.1) are $X \sim \mathcal{N}(\mu_X,\sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y,\sigma_Y^2)$. The bivariate normal vector (X,Y) has a covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_X\sigma_Y\rho \\ \sigma_X\sigma_Y\rho & \sigma_Y^2 \end{pmatrix}. \tag{6.2}$$

The covariance matrix Σ is nonnegative definite. A sufficient condition for nonnegative definiteness in this case is $|\Sigma| \geq 0$ (see also Exercise 6.2).

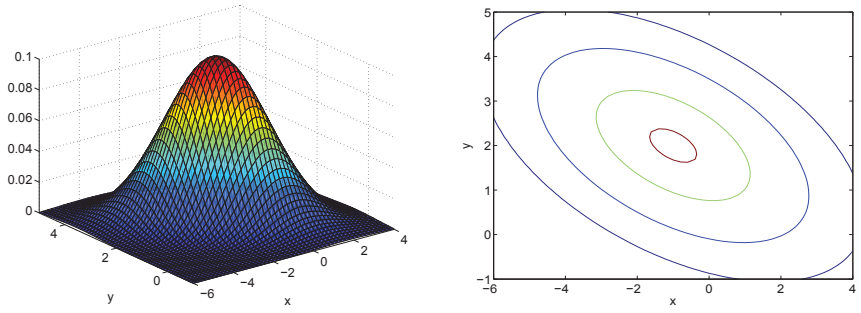
Figure 6.5a shows the density of a bivariate normal distribution with mean

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$

and covariance matrix

$$\Sigma = \begin{pmatrix} 3 & -0.9 \\ -0.9 & 1 \end{pmatrix}.$$

Figure 6.5b shows contours of equal probability.



(a) (b)

Fig. 6.5 (a) Density of bivariate normal distribution with mean `mu=[-1 2]` and covariance matrix `Sigma=[3 -0.9; -0.9 1]`. (b) Contour plots of a density at levels `[0.001 0.01 0.05 0.1]`

Several properties of bivariate normal are listed below:

(i) If (X,Y) is bivariate normal, then $aX+bY$ has a univariate normal distribution.

(ii) If (X,Y) is bivariate normal, then $(aX+bY, cX+dY)$ is also bivariate normal.

(iii) If the components in (X,Y) are such that $\mathrm{Cov}(X,Y) = \sigma_X\sigma_Y\rho = 0$, then X and Y are independent.

(iv) Any bivariate normal pair (X,Y) can be transformed into a pair $(U,V) = (aX+bY, cX+dY)$ such that U and V are independent. If $\sigma_X^2 = \sigma_Y^2$, then one such transformation is $U = X+Y,\ \ V = X-Y$. For an arbitrary bivariate normal distribution, the rotation

$$\begin{aligned} U &= X\cos\varphi - Y\sin\varphi \\ V &= X\sin\varphi + Y\cos\varphi \end{aligned}$$

makes components (U,V) independent if the rotation angle φ satisfies

$$\cot 2\varphi = \frac{\sigma_X^2 - \sigma_Y^2}{2\sigma_X \sigma_Y \rho}.$$

(v) If (X, Y) is bivariate normal, then the conditional distribution of Y when $X = x$ is normal with expectation and variance

$$\mu_X + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X), \quad \text{and} \quad \sigma_Y^2(1 - \rho^2),$$

respectively. The linearity in x of the conditional expectation of Y will be the basis for linear regression, covered in Chapter 14. Also, the fact that $X = x$ is known decreases the variance of Y; indeed, $\sigma_Y^2(1 - \rho^2) \leq \sigma_Y^2$.

More generally, when the components of a p-dimensional random vector all have a normal distribution, we say that the vector has a multivariate normal distribution. For independent components, the density of a multivariate distribution is simply the product of the univariate normal densities. When the components are correlated, the distribution involves the covariance matrix that describes the correlation.

A random vector $\boldsymbol{X} = (X_1, \ldots, X_p)'$ has a multivariate normal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, denoted as $\boldsymbol{X} \sim \mathcal{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if its density is

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(1/2)(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})},$$

where $\boldsymbol{x} \in \mathbb{R}^p$, and $\boldsymbol{\Sigma}$ is a non-negative definite $p \times p$ matrix. Here $|\boldsymbol{\Sigma}|$ is the determinant and $\boldsymbol{\Sigma}^{-1}$ the inverse of the covariance matrix $\boldsymbol{\Sigma}$.

6.3 Examples with a Normal Distribution

We provide two examples with typical calculations involving normal distributions, with solutions in MATLAB and WinBUGS.

Example 6.1. **IgE Concentration.** Total serum IgE (immunoglobulin E) concentration allergy tests allow for the measurement of the total IgE level in a serum sample. Elevated levels of IgE are associated with the presence of an allergy. An example of testing for total serum IgE is the PRIST (paper radioimmunosorbent test). This test involves serum samples reacting with IgE that has been tagged with radioactive iodine. The bound radioactive iodine, calculated upon completion of the test procedure, is proportional to the amount of total IgE in the serum sample. The determination of normal IgE levels in a population of healthy, nonallergic individuals varies by the fact that some individuals may have subclinical allergies and therefore have abnormal serum IgE levels. The log concentration of IgE (in IU/ml) in a cohort of healthy subjects is distributed as a normal $\mathcal{N}(9, (0.9)^2)$ random

variable. What is the probability that in a randomly selected subject from the same cohort the log concentration will

(a) Exceed 10 IU/ml?

(b) Be between 8.1 and 9.9 IU/ml?

(c) Differ from the mean by no more than 1.8 IU/ml?

(d) Find the number x_0 such that the IgE log concentration in 90% of the subjects from the same cohort exceeds x_0.

(e) In what bounds (symmetric about the mean) does the IgE log concentration fall with a probability of 0.95?

(f) If the IgE log concentration is $\mathcal{N}(9,\sigma^2)$, find σ so that

$$\mathbb{P}(8 \leq X \leq 10) = 0.64.$$

Let X be the IgE log concentration in a randomly selected subject. Then $X \sim \mathcal{N}(9, 0.9^2)$. The solution is given by the following MATLAB code (ige.m):

```
%(a)
%P(X>10)= 1-P(X <= 10)
1-normcdf(10,9,0.9) %or   1-normcdf((10-9)/0.9)
%ans = 0.1333
%(b)
%P(8.1 <= X <= 9.9)
%P((8.1-9)/0.9 <= Z <= (9.9-9)/0.9)
%P(-1 <= Z <= 1) ::: Note the 1-sigma rule.
normcdf(9.9, 9, 0.9) - normcdf(8.1, 9, 0.9)
%or, normcdf((9.9-9)/0.9)-normcdf((8.1-9)/0.9)
%ans = 0.6827
%(c)
%P(9-1.8 <= X <= 9+1.8) = P(-2 <= Z <= 2)
%Note the 2-sigma rule.
normcdf(9+1.8, 9, 0.9) - normcdf(9-1.8, 9, 0.9)
% ans = 0.9545
%(d)
%0.90 = P(X > x0)=1-P(X <= x0)
%that is P(Z <= (x0-9)/0.9)=0.1
norminv(1-0.9,   9, 0.9)
%ans = 7.8466
%(e)
%P(9-delta <= X <= 9+delta)=0.95
[9-0.9*norminv(1-0.05/2), 9+0.9*norminv(1-0.05/2)]
%ans =  7.2360   10.7640
%(f)
%P(-1/sigma) <= Z <= 1/sigma)=0.64
%note that 0.36/2 + 0.64 + 0.36/2 = 1
1/norminv( 1 - 0.36/2 )
%ans = 1.0925
```

Example 6.2. **Aplysia Nerves.** In this example, easily solved analytically and using MATLAB, we will show how to use WinBUGS and obtain an approximate solution. The analysis is not Bayesian; WinBUGS will simply serve as a random number generator and the required probability and quantile will be found approximately by simulation.

Characteristics of Aplysia nerves in response to extension were examined by Koike (1987). Only the Aplysia nerve was easily elongated up to about five times its resting or relaxing length without impairing propagation of the action potential along the axon in the nerve. The conduction velocity along the elongated nerve increased linearly in proportion to the nerve length in a range from the relaxing length to about 1 to 1.5 times extension. For an expansion factor of 1.5, the conducting velocity factors are normally distributed with a mean of 1.4 and a standard deviation of 0.1. Using WinBUGS, we are interested in finding

(a) the proportion of Aplysia nerves elongated by a factor of 1.5 for which the conduction velocity factor exceeds 1.5;

(b) the proportion of Aplysia nerves elongated by a factor of 1.5 for which the conduction velocity factor falls in the interval $[1.35, 1.61]$; and

(c) the velocity factor x that is exceeded by 5% of Aplysia nerves elongated by a factor of 1.5.

```
#aplysia.odc
model{
mu <- 1.4
stdev <- 0.1
prec<- 1/(stdev * stdev)
y ~ dnorm(mu, prec)
#a
propexceed <- step(y - 1.5)
#b
propbetween <-  step(y-1.35)*step(1.61-y)
#c
#done in Sample Monitor Tool by
#selecting 95th percentile
}
```

There are no data to load; after the `check model` in `Model>Specification` go directly to `compile`, and then to `gen inits`. Update 10,000 iterations, and set in `Sample Monitor Tool` from `Inference>Samples` the nodes `y`, `propexceed`, and `propbetween`. For part (c) select the 95th percentile in `Sample Monitor Tool` under `percentiles`. Finally, run the `Update Tool` for 1,000,000 updates and check the results in `Sample Monitor Tool` by setting a star (`*`) in the `node` window and looking at `stats`.

	mean	sd	MC error	val2.5pc	median	val97.5pc	start	sample
propbetween	0.6729	0.4691	4.831E-4	0.0	1.0	1.0	10001	1000000
propexceed	0.1587	0.3654	3.575E-4	0.0	0.0	1.0	10001	1000000
y	1.4	0.1001	1.005E-4	1.204	1.4	1.565	10001	1000000

Here is the same computation in MATLAB.

```
1-normcdf(1.5, 1.4, 0.1)   %0.1587
normcdf(1.61, 1.4, 0.1)-normcdf(1.35, 1.4, 0.1) %0.6736
norminv(1-0.05, 1.4, 0.1)  %1.5645
```

6.4 Combining Normal Random Variables

Any linear combination of independent normal random variables is also normally distributed. Thus, we need only keep track of the mean and variance of the variables involved in the linear combination, since these two parameters completely characterize the distribution. Let $X_1, X_2, \ldots, X_n$ be independent normal random variables such that $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$; then for any selection of constants $a_1, a_2, \ldots, a_n$,

$$a_1 X_1 + a_2 X_2 + \cdots + a_n X_n = \sum_{i=1}^{n} a_i X_i \sim \mathcal{N}(\mu, \sigma^2),$$

where

$$\mu = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n = \sum_{i=1}^{n} a_i\mu_i,$$
$$\sigma^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ldots a_n^2\sigma_n^2 = \sum_{i=1}^{n} a_i^2\sigma_i^2.$$

Two special cases are important: (i) $a_1 = 1, a_2 = -1$ and (ii) $a_1 = \cdots = a_n = 1/n$. In case (i) we have a difference of two normals; its mean is the difference of the corresponding means and variance is a *sum* of two variances. Case (ii) corresponds to the arithmetic mean of normals, $\overline{X}$. For example, if $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, then the sample mean $\overline{X} = (X_1 + \cdots + X_n)/n$ has a normal $\mathcal{N}(\mu, \sigma^2/n)$ distribution. Thus, variances for X_is and $\overline{X}$ are related as

$$\sigma_{\overline{X}}^2 = \frac{\sigma^2}{n}$$

or, equivalently, for standard deviations

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

Example 6.3. **The Piston Production Error.** The profile of a piston comprises a ring in which inner and outer radii X and Y are normal random variables, $\mathcal{N}(88, 0.01^2)$ and $\mathcal{N}(90, 0.02^2)$, respectively. The thickness $D = Y - X$ is the random variable of interest.

(a) Find the distribution of D.

(b) For a randomly selected piston, what is the probability that D will exceed 2.04?

(c) If D is averaged over a batch of $n = 64$ pistons, what is the probability that $\overline{D}$ will exceed 2.04? Exceed 2.004?

```
sqrt(0.01^2 + 0.02^2)                    %0.0224
1-normcdf((2.04  - 2)/0.0224)            %0.0371
1-normcdf((2.04  - 2)/(0.0224/sqrt(64))) %0
1-normcdf((2.004 - 2)/(0.0224/sqrt(64))) %0.0766
```

Compare the probabilities of events $\{D > 2.04\}$ and $\{\overline{D} > 2.04\}$. Why is the probability of $\{\overline{D} > 2.04\}$ essentially 0, when the analogous probability for an individual measure D is 3.71%?

Example 6.4. **Diluting Acid.** In a laboratory, students are told to mix 100 ml of distilled water with 50 ml of sulfuric acid and 30 ml of C_2H_5OH. Of course, the measurements are not exact. The water is measured with a mean of 100 ml and a standard deviation of 4 ml, the acid with a mean of 50 ml and a standard deviation of 2 ml, and C_2H_5OH with a mean of 30 ml and a standard deviation of 3 ml. The three measurements are normally distributed and independent.

(a) What is the probability of a given student measuring out at least 103 ml of water?

(b) What is the probability of a given student measuring out between 148 and 157 ml of water plus acid?

(c) What is the probability of a given student measuring out a total of between 175 and 180 ml of liquid?

```
1 - normcdf(103, 100, 4)                        %0.2266
normcdf(157, 150, sqrt(4^2 + 2^2)) ...
    - normcdf(148, 150, sqrt(4^2 + 2^2))        %0.6139
normcdf(180, 180, sqrt(4^2 + 2^2 + 3^2 )) ...
    - normcdf(175, 180, sqrt(4^2 + 2^2 + 3^2))  %0.3234
```

Example 6.5. **Two Plate Assembly Simulation.** The following example is adapted from Banks et al. (1984). In assembly of two square 4×4 steel plates, comprising a part of a medical device, each plate has a hole drilled

in its center. The plates are to be joined by a pin. Assembling machine adjusts the plates with respect to the lower left corner denoted as (0,0) in coordinate system xOy.

The coordinates of hole centers X_i and Y_i for ith plate ($i = 1,2$) are independent normally distributed random variables with mean 2 and standard deviation 0.001.

The hole diameters D_1 and D_2 are normally distributed with mean of 0.2 and standard deviation 0.0012, for both plates. The pin diameter, R, is also normally distributed with mean 0.195 and standard deviation of 0.0005.

(a) What proportion of pins will go through assembled plates? We will approximate this proportion by 1,000,000 simulated assembles using MATLAB. The clearance

$$\min\{D_1, D_2\} - \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2} - R,$$

has to be positive for a successful assembly. Why is the second term needed?

(b) In an assembled pair of plates the pin will wobble if it is too loose. This wobbling will occur if

$$\min\{D_1, D_2\} - R \geq 0.006.$$

What fraction of assembled plates would not wobble? This is conditional probability, since we restrict attention on the assembled plates only. Thus in simulating this proportion we ignore the cases when the assembly was not possible.

The following MATLAB script estimates the desired proportions:

```
%Normal Probabilities by Simulation
rng(10,'twister')
M=1000000 ; %number of simulations
  clear = 0;
  clearnowobb=0;
for i = 1:M
    X1 = 2 + 0.001  * randn;
    Y1 = 2 + 0.001  * randn;
    X2 = 2 + 0.001  * randn;
    Y2 = 2 + 0.001  * randn;
    C=sqrt((X1-X2)^2 + (Y1-Y2)^2);
    D1=0.2+0.0012*randn;
    D2=0.2+0.0012*randn;
    D = min(D1, D2);
    R = 0.195 + 0.0005*randn;
    clear = clear + (D-C-R > 0);
    clearnowobb = clearnowobb + (D-C-R > 0)*(D-R<0.006);
end
  p1=clear/M             %(a) 0.9553
  p2 = clearnowobb/clear %(b) 0.9346
```

Thus, by simulation, we estimated that 95.53% of assembles are possible and that among the assembled plates 93.46% would not wobble.
✎

6.5 Central Limit Theorem

The central limit theorem (CLT) elevates the status of the normal distribution above other distributions. We have already seen that a linear combination of independent normals is a normal random variable itself. That is, if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2), \quad \text{and} \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

The CLT states that $X_1, \ldots, X_n$ need not be normal in order for $\sum_{i=1}^{n} X_i$ or, equivalently, for $\overline{X}$ to be *approximately* normal. This approximation is quite good for n as low as 30. As we said, variables $X_1, X_2, \ldots, X_n$ need not be normal but must satisfy some conditions. For CLT to hold, it is sufficient for X_is to be independent, equally distributed, and have finite variances and, consequently, means. Other than that, the X_is can be arbitrary – skewed, discrete, etc. The conditions of i.i.d. and finiteness of variances are sufficient – more precise formulations of the CLT are beyond the scope of this text. Dasgupta (2008) provides comprehensive coverage.

CLT. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with a mean μ and finite variance σ^2. Then,

$$\sum_{i=1}^{n} X_i \overset{\text{approx}}{\sim} \mathcal{N}(n\mu, n\sigma^2) \quad \text{and} \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{\text{approx}}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

A special case of CLT involving Bernoulli random variables results in a normal approximation to binomials because the sum of many i.i.d. Bernoullis is at the same time exactly binomial and approximately normal. This approximation is handy when n is very large.

de Moivre (1738). Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli $\mathcal{B}er(p)$ random variables with parameter p.

Then,

$$Y = \sum_{i=1}^{n} X_i \overset{\text{approx}}{\sim} \mathcal{N}(np, npq)$$

and

$$\mathbb{P}(k_1 \le Y \le k_2) = \Phi\left(\frac{k_2 + 1/2 - np}{\sqrt{npq}}\right) - \Phi\left(\frac{k_1 - 1/2 - np}{\sqrt{npq}}\right),$$

where Φ is the CDF of standard normal random variable.

De Moivre's approximation is good if both np and nq exceed 10 and n exceeds 30. If that is not the case, a Poisson approximation to binomial (page 179) could be better.

The factors $1/2$ in de Moivre's formula are continuity corrections. For example, Y, which is discrete, is approximated with a continuous distribution. $\mathbb{P}(Y \le k_2 + 1)$ and $\mathbb{P}(Y < k_2 + 1)$ are the same for a normal but not for a binomial distribution for which $\mathbb{P}(Y < k_2 + 1) = \mathbb{P}(Y \le k_2)$. Likewise, $\mathbb{P}(Y \ge k_1 - 1)$ and $\mathbb{P}(Y > k_1 - 1)$ are the same for a normal but not for a binomial distribution for which $\mathbb{P}(Y > k_1 - 1) = \mathbb{P}(Y \ge k_1)$. Thus, $\mathbb{P}(k_1 \le Y \le k_2)$ for a binomial distribution is better approximated by $\mathbb{P}(k_1 - 1/2 \le Y \le k_2 + 1/2)$.

All approximations used to be much more important in the era before modern computing power was available. MATLAB is capable of calculating exact binomial probabilities for huge values of n, and for practical reasons de Moivre's approximation is obsolete. For example,

```
format long
binocdf(1999988765, 4000000000, 1/2)
%ans = 0.361195130797824
format short
```

However, the theoretical value of de Moivre's approximation is significant since many estimators and tests based on a binomial distribution can use well-developed normal distribution machinery for an analysis beyond the computation.

The following MATLAB program exemplifies the CLT by averages of simulated uniform random variables:

```
% Central Limit Theorem Demo
figure;
subplot(3,2,1)
hist(rand(1, 10000),40)        %histogram of 10000 uniforms
subplot(3,2,2)
hist(mean(rand(2, 10000)),40)  %histogtam of 10000
                               %averages of 2 uniforms
subplot(3,2,3)
```

```
hist(mean(rand(3, 10000)),40)  %histogtam of 10000
                               %averages of 3 uniforms
subplot(3,2,4)
hist(mean(rand(5, 10000)),40)  %histogtam of 10000
                               %averages of 5 uniforms
subplot(3,2,5)
hist(mean(rand(10, 10000)),40) %histogtam of 10000
                               %averages of 10 uniforms
subplot(3,2,6)
hist(mean(rand(100, 10000)),40)%histogtam of 10000
                               %averages of 100 uniforms
```

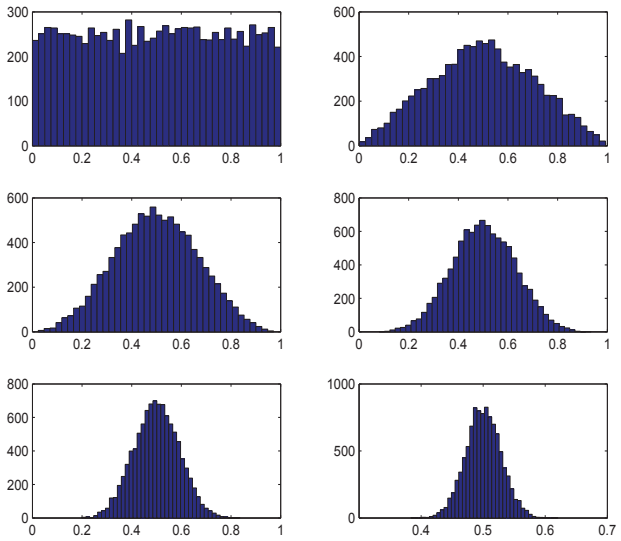


Fig. 6.6 Convergence to normal distribution shown via averages of 1, 2, 3, 5, 10, and 100 independent uniform (0,1) random variables.

Figure 6.6 shows the histograms of 10,000 simulations of averages of $k = 1,2,3,5,10$, and 100 uniform random variables. It is interesting to see the metamorphosis of a flat single uniform ($k = 1$), via a "witch hat distribution" ($k = 2$), into bell-shaped distributions close to the normal. For additional simulation experiments, see the script cltdemo.m.

Example 6.6. **Is Grandpa's Genetic Theory Valid?** The domestic cat's wild appearance is increasingly overshadowed by color mutations, such as black, white spotting, maltesing (diluting), red and tortoiseshell, shading, and Siamese pointing. By favoring the odd or unusually colored and marked cats over the "plain" tabby, people have consciously and unconsciously enhanced these color mutations over the course of domestication. Today, "colored" cats outnumber the wild looking tabby cats, and pure tabbies are

becoming rare. Some may not be quite as taken by the coat of our domestic feline friends as Jason's grandpa is. He has a genetic theory that asserts that three-fourths of cats with more than three colors in their fur are female. A total of $n = 300$ three-color cats (TCCs) are observed and 86 are found to be male. If Jason's grandpa's genetic theory is true, then the number of male TCCs is binomial $\mathcal{B}(300, 0.25)$, with an expectation of 75 and variance of $56.25 = 7.5^2$.

(a) What is the probability that, assuming Jason's grandpa's theory, one will observe 86 or more male cats? How does this finding support the theory?

(b) What is the probability that, assuming the independence of a cat's fur and gender, one will observe 86 or more male cats?

(c) What is the probability that one will observe exactly 75 male TCCs?

We will find exact solutions using binomial distribution and compare results with normal approximations.

```
format long  %for precise comparisons
%(a)
1 - binocdf(85, 300, 0.25) %0.08221654140000, exact
1 - normcdf(85, 75, 7.5)   %0.09121121972587
1 - normcdf(86, 75, 7.5)   %0.07123337741399
1 - normcdf(85.5, 75, 7.5) %0.08075665923377, approx
  %85.5 is taken as continuity-corrected argument
%(b)
1 - binocdf(85, 300, 0.5)  %0.99999999999998
 %virtually a sure event
%(c)
binopdf(75, 300, 0.25)     %0.05312831515720, exact
normcdf(75.5, 75, 7.5)-normcdf(74.5, 75, 7.5)
                           %0.05315292860073, approx
```

Example 6.7. **Avio Company.** The Avio Company sells 410 plane tickets for a 400-seater flight. Find the probability that the company overbooked the flight if a person who bought a ticket shows up at the gate with a probability of 0.96.

Each sold ticket can be thought of as an "experiment" where "success" means showing up at the gate for the flight. The number of people that show up X is binomial $\mathcal{B}in(410, 0.96)$. The following MATLAB script calculates the normal approximation:

```
410*0.96                        %393.6000
sqrt(410*0.96*0.04)             %3.9679
1-normcdf((400.5-393.6)/3.9679) %0.0410
```

Notice that in this case the normal approximation is not very good since the exact binomial probability is 0.0329:

```
1-binocdf(400, 410, 0.96)        %0.0329
```

The reason is that the normal approximation works well when the probabilities are not close to 0 or 1, and here 0.96 is quite close to 1 for a given sample size of 410.

The Poisson approximation to the binomial performs better. The probability of missing the flight is $1 - 0.96 = 0.04$, and overbooking will happen if 9 or fewer passengers miss the flight:

```
   %prob that 9 or less fail to show
poisscdf(9, 0.04*410)            %0.0355
```

6.6 Distributions Related to Normal

Four distributions – chi-square χ^2, t, F, and lognormal – are specially related to the normal distribution. This relationship is described in terms of functions of independent standard normal variables. Let $Z_1, Z_2, \ldots, Z_n$ be n independent standard normal (mean 0, variance 1) random variables. Then:

• The sum of squares $Z_1^2 + \cdots + Z_n^2$ is chi-square distributed with n degrees of freedom, χ_n^2:

$$\chi_n^2 \sim Z_1^2 + Z_2^2 + \cdots + Z_n^2.$$

• The ratio of a standard normal Z and the square root of an independent chi-square χ^2 random variable normalized by its number of degrees of freedom, has a t-distribution with n degrees of freedom, t_n:

$$t_n \sim \frac{Z}{\sqrt{\frac{\chi_n^2}{n}}}.$$

• The ratio of two independent chi-squares normalized by their respective numbers of degrees of freedom is distributed as an F:

$$F_{m,n} \sim \frac{\chi_m^2/m}{\chi_n^2/n}.$$

The degrees of freedom for F are *m – numerator df* and *n – denominator df*.

• As the name indicates, the lognormal ("log-is-normal") distribution is connected to a normal distribution via a logarithm function. If X has a lognormal disrtibution, then the distribution of $Y = \log X$ is normal.

A more detailed description of these four distributions follows next.

6.6.1 Chi-square Distribution

The probability density function for a chi-square random variable with parameter k, called the *degrees of freedom*, is

$$f_X(x) = \frac{(1/2)^{k/2}\ x^{k/2-1}}{\Gamma(k/2)} e^{-x/2}, \quad 0 \le x < \infty.$$

The chi-square distribution (χ^2) is a special case of the gamma distribution with parameters $r = k/2$ and $\lambda = 1/2$. Its mean and variance are $\mu = k$ and $\sigma^2 = 2k$, respectively.

If $Z \sim \mathcal{N}(0,1)$, then $Z^2 \sim \chi^2_1$, that is, a chi-square random variable with one degree of freedom. Furthermore, if $U \sim \chi^2_m$ and $V \sim \chi^2_n$ are independent, then $U + V \sim \chi^2_{m+n}$.

From these results it can be shown that if $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ and $\overline{X}$ is the sample mean, then the *sample variance* $s^2 = \sum_i (X_i - \overline{X})^2/(n-1)$ is proportional to a chi-square random variable with $n-1$ degrees of freedom:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}. \tag{6.3}$$

This result was proved first by German geodesist Helmert (1876). The χ^2-distribution was previously defined by Abbe and Bienaymé in the mid-1800s.

The formal proof of (6.3) is beyond the scope of this text, but an intuition can be obtained by inspecting

$$\begin{aligned}\frac{(n-1)s^2}{\sigma^2} &= \left(\frac{X_1 - \overline{X}}{\sigma}\right)^2 + \left(\frac{X_2 - \overline{X}}{\sigma}\right)^2 + \cdots + \left(\frac{X_n - \overline{X}}{\sigma}\right)^2 \\ &= (Y_1 - \overline{Y})^2 + (Y_2 - \overline{Y})^2 + \cdots + (Y_n - \overline{Y})^2,\end{aligned}$$

where Y_i are independent normal $\mathcal{N}(\mu/\sigma, 1)$.

$$(Y_1-\overline{Y})^2+(Y_2-\overline{Y})^2=\left(\frac{Y_1-Y_2}{\sqrt{2}}\right)^2=Z_1^2, \quad \text{for } \overline{Y}=\frac{Y_1+Y_2}{2},$$

$$(Y_1-\overline{Y})^2+(Y_2-\overline{Y})^2+(Y_3-\overline{Y})^2=\left(\frac{Y_1-Y_2}{\sqrt{2}}\right)^2+\left(\frac{Y_1+Y_2-2Y_3}{\sqrt{6}}\right)^2=Z_1^2+Z_2^2,$$

$$\text{for } \overline{Y}=\frac{Y_1+Y_2+Y_3}{3},$$

etc.

Note that the right-hand sides are sums of squares of uncorrelated standard normal variables.

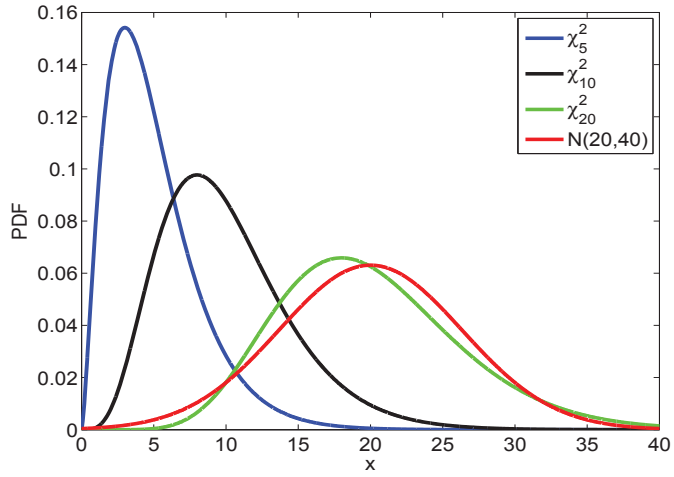


Fig. 6.7 χ^2-distribution with 5, 10, and 20 degrees of freedom. A normal $\mathcal{N}(20,40)$ distribution is superimposed to illustrate a good approximation to χ^2_n by $\mathcal{N}(n,2n)$ for n large

In MATLAB, the CDF and PDF for a χ^2_k are `chi2cdf(x,k)` and `chi2pdf(x,k)`, respectively. The pth quantile of the χ^2_k distribution is `chi2inv(p,k)`.

Example 6.8. χ^2_{10} **as a Sum of Squares of Ten Standard Normals.** In this example we demonstrate by simulation that the sum of squares of standard normal random variates follows the χ^2-distribution. In particular we compare $Z_1^2+Z_2^2+\cdots+Z_{10}^2$ with χ^2_{10}.

Figure 6.8, produced by the code in `nor2chi2.m`, shows a normalized histogram of the sums of squares of ten standard normals with a superimposed χ^2_{10} density (above) and a Q–Q plot comparing the sorted generated sample with χ^2_{10} quantiles (below). As expected, the simulated empirical distribution is very close to the theoretical chi-square distribution.

```
figure;
subplot(2,1,1)
  %form a matrix of standard normals 10 x 10000
  %square the entries, sum up columnwise, to
  % get a vector of 10000 chi2 with 10 df.
  histn(sum(normrnd(0,1,[10, 10000]).^2),0, 1,30)
    hold on
  plot((0.1:0.1:30), chi2pdf((0.1:0.1:30),10),'r-','LineWidth',2)
  axis tight
subplot(2,1,2)
  %check the Q-Q plot
  xx = sum(normrnd(0,1,[10, 10000]).^2);
  tt = 0.5/10000:1/10000:1;
  yy = chi2inv(tt,10);
  plot(sort(xx), yy,'*')
    hold on
  plot(yy, yy,'r-')
```

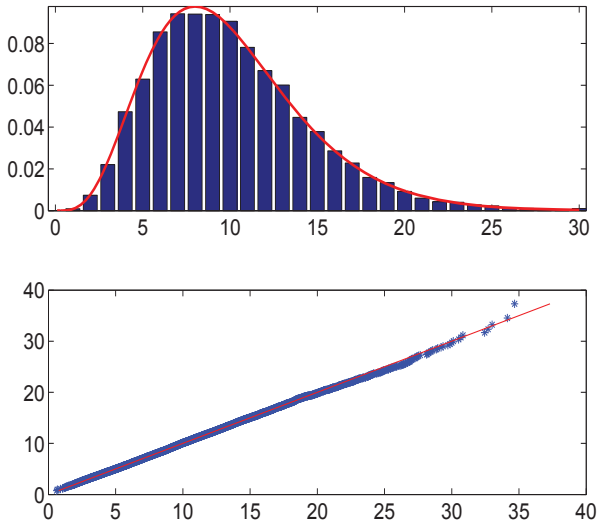


Fig. 6.8 Sum of 10 squared standard normals compared to χ^2_{10} distribution. Above: Normalized histogram with superimposed χ^2_{10} density (*red*); Below: Q–Q-plot of sorted sums against χ^2_{10} quantiles.

Example 6.9. **Targeting Meristem Cells.** A gene transfer system for meristem cells can be developed on the basis of a ballistic approach (Sautter, 1993). Instead of a macroprojectile, microtargeting uses the law of Bernoulli for acceleration of highly uniform-sized gold particles. The particle is aimed at an area as small as 150 μm in diameter, which corresponds to the size of

a meristem. Suppose that a particle is fired at a meristem at the origin of a plane coordinate system, with units in microns. The particle lands at (X,Y), where X and Y are independent and each has a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 10^2$. The particle is successively delivered if it lands within $\sqrt{738}\,\mu$m of the target (origin). What is the probability of this event? The particle is successively delivered if $X^2 + Y^2 \le 738$, or $(X/10)^2 + (Y/10)^2 \le 7.38$. Since both $X/10$ and $Y/10$ have a standard normal distribution, random variable $(X/10)^2 + (Y/10)^2$ is χ^2_2-distributed. Since `chi2cdf(7.38,2)=0.975`, we conclude that the particle is successfully delivered with a probability of 0.975.
✎

A square root of chi-square random variable χ^2_k with k degrees of freedom is called chi (χ_k) random variable. The density of χ_k random variable X is

$$f_X(x) = \frac{2^{1-k/2} x^{k-1} e^{-x/2}}{\Gamma\left(\frac{k}{2}\right)}, \quad 0 \le x < \infty.$$

The mean and variance of X are

$$\mathbb{E}X = \frac{\sqrt{2}\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \quad \text{and} \quad \mathbb{V}\text{ar}\, X = k - (\mathbb{E}X)^2.$$

Special cases of χ-distribution are Rayleigh and Maxwell distributions. In their standard form (scale/rate = 1), these two distributions are χ_2 and χ_3 respectively. Absolute value of a standard normal random variable is χ_1 distributed.

A multivariate version of the χ^2-distribution is called a Wishart distribution. It is a distribution of random matrices that are symmetric and positive definite. As such, it is a proper model for normal covariance matrices, and we will see later its use in Bayesian inference involving bivariate normal distributions.

A $p \times p$ random matrix X has a Wishart distribution if its density is given by

$$f(X) = \frac{|X|^{(n-p-1)/2} \exp\{-\frac{1}{2} tr(\Sigma^{-1} X)\}}{2^{np/2} \pi^{p(p-1)/4} |\Sigma|^{n/2} \prod_{i=1}^{p} \Gamma\left(\frac{n+1-i}{2}\right)},$$

where Σ is the scale matrix and n is the number of degrees of freedom. Operator tr is the trace of a matrix, that is, the sum of its diagonal elements, and $|\Sigma|$ and $|X|$ are determinants of Σ and X, respectively.

For $p = 1$ and $\Sigma = 1$, the Wishart distribution is χ^2_n. In MATLAB, it is possible to simulate from the Wishart distribution as `wishrnd(Sigma,n)`. In WinBUGS, the Wishart distribution is coded as `dwish(R[,],n)`, where the precision matrix R is defined as Σ^{-1}.

6.6.2 *t*-Distribution

Random variable X has t-distribution with k degrees of freedom, $X \sim t_k$, if its PDF is

$$f_X(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma(k/2)}\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}}, \quad -\infty < x < \infty.$$

The t-distribution is similar in shape to the standard normal distribution except for having fatter tails. If $X \sim t_k$, then $\mathbb{E}X = 0$, $k > 1$ and $\mathbb{V}\mathrm{ar}\,X = k/(k-2)$, $k > 2$. For $k = 1$, the t-distribution coincides with the Cauchy distribution.

The t-distribution has an important role to play in statistical inference. With a set of i.i.d. $X_1,\ldots,X_n \sim \mathcal{N}(\mu,\sigma^2)$, we can standardize the sample mean using the simple transformation of $Z = (\overline{X}-\mu)/\sigma_{\overline{X}} = \sqrt{n}(\overline{X}-\mu)/\sigma$. However, if the variance is unknown, by using the same transformation, except for substituting the sample standard deviation s for σ, we arrive at a t-distribution with $n-1$ degrees of freedom:

$$t = \frac{\overline{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}.$$

More technically, if $Z \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_k$ are independent, then $t = Z/\sqrt{Y/k} \sim t_k$. In MATLAB, the CDF at x for a t-distribution with k degrees of freedom is calculated as `tcdf(x,k)`, and the PDF is computed as `tpdf(x,k)`. The pth percentile is computed with `tinv(p,k)`. In WinBUGS, the t-distribution is coded as `dt(mu,tau,k)`, where `tau` is a precision parameter and `k` is the number of degrees of freedom.

The t-distribution was originally found by German mathematician and astronomer Jacob Lüroth in 1876 (Lüroth, 1876). William Sealy Gosset rediscovered the t-distribution in 1908 and published the results under the pen name "Student."

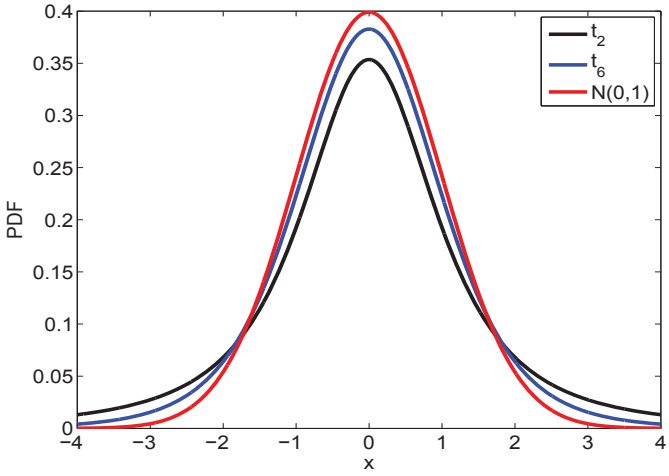


Fig. 6.9 t-distribution with 2 and 6 degrees of freedom. A standard normal distribution is superimposed as the solid red line.

6.6.3 Cauchy Distribution

The Cauchy distribution is a special case of the t-distribution; it is symmetric and bell-shaped like the normal distribution, but with much fatter tails. In fact, it is a popular distribution to use in nonparametric robust procedures and simulations because the distribution is so spread out; it has no mean and variance (none of the Cauchy moments exist). Physicists know this distribution as the *Lorentz distribution*. If $X \sim \mathcal{C}a(a,b)$, then X has a density

$$f_X(x) = \frac{1}{\pi} \frac{b}{b^2 + (x-a)^2}, \quad -\infty < x < \infty.$$

The standard Cauchy $\mathcal{C}a(0,1)$ distribution coincides with the t-distribution with 1 degree of freedom.

The Cauchy distribution is also related to the normal distribution. If Z_1 and Z_2 are two independent $\mathcal{N}(0,1)$ random variables, then their ratio $C = Z_1/Z_2$ is Cauchy, $\mathcal{C}a(0,1)$. Finally, if $C_i \sim \mathcal{C}a(a_i,b_i)$ for $i = 1,\ldots,n$, then $S_n = C_1 + \cdots + C_n$ is Cauchy distributed with parameters $a_S = \sum_i a_i$ and $b_S = \sum_i b_i$. The consequence of this additivity is interesting. If one observes n Cauchy $\mathcal{C}a(0,1)$ random variables $X_i, i = 1,\ldots,n$, and takes the average $\overline{X}$, the average is also Cauchy $\mathcal{C}a(0,1)$. This means that for Cauchy CLT does not hold; a single measurement is as precise as the average of any finite number of measurements.

Here is a simple geometric example that leads to a Cauchy distribution:

Example 6.10. **Geometric Interpretation of Cauchy.** A ray passing through the point $(-1,0)$ in R^2 intersects the y-axis at the coordinate $(0,Y)$. If the angle α between the ray and the positive direction of the x-axis is uniform $\mathcal{U}(-\pi/2,\pi/2)$, what is the distribution for Y?

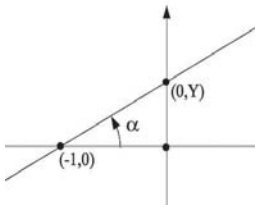


Fig. 6.10 If the angle α between the ray and x-axis is uniform $\mathcal{U}(-\pi/2,\pi/2)$, Y is Cauchy $\mathcal{C}a(0,1)$.

Here $Y=\tan\alpha$, $\alpha=h(Y)=\arctan(Y)$ and $h'(y)=\frac{1}{1+y^2}$. The density for uniform $\mathcal{U}(-\pi/2,\pi/2)$ is constant $1/\pi$ if $\alpha\in(-\pi/2,\pi/2)$, and 0 else. From (5.15),

$$f_Y(y)=\frac{1}{\pi}|h'(y)|=\frac{1}{\pi}\frac{1}{1+y^2},$$

which is the density of the Cauchy $\mathcal{C}a(0,1)$ distribution.
✎

6.6.4 *F*-Distribution

Random variable X has an F-distribution with m and n degrees of freedom, denoted as $F_{m,n}$, if its density is given by

$$f_X(x)=\frac{m^{m/2}n^{n/2}}{B(m/2,n/2)}\;\; x^{m/2-1}(n+mx)^{-(m+n)/2},\quad x>0.$$

The CDF of an F-distribution is not of closed form, but it can be expressed in terms of an incomplete beta function (page 206) as

$$F(x)=1-I_\nu(n/2,m/2),\;\; \nu=n/(n+mx),\quad x>0.$$

The mean is given by $\mathbb{E}X = n/(n-2), n > 2$, and the variance by $\mathbb{V}\mathrm{ar}\, X = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, n > 4$.

If $X \sim \chi^2_m$ and $Y \sim \chi^2_n$ are independent, then $(X/m)/(Y/n) \sim F_{m,n}$. Because of this representation, m and n are often called, respectively, the *numerator* and *denominator* degrees of freedom. F and beta distributions are related. If $X \sim \mathcal{B}e(a,b)$, then $bX/[a(1-X)] \sim F_{2a,2b}$. Also, if $X \sim F_{m,n}$, then $mX/(n+mX) \sim \mathcal{B}e(m/2, n/2)$.

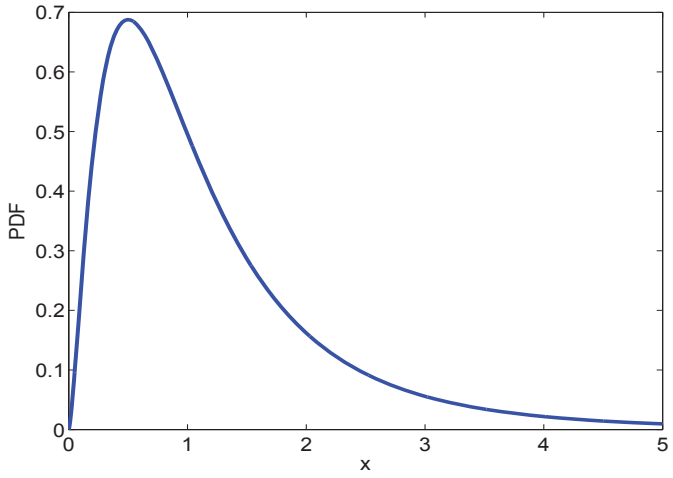


Fig. 6.11 $F_{5,10}$ PDF. `t2 = 0:0.005:5; plot(t2, fpdf(t2, 5, 10))`

The F-distribution is one of the most important distributions for statistical inference; in introductory statistical courses, the test for equality of variances, ANOVA, and multivariate regression are based on the F-distribution. For example, if s_1^2 and s_2^2 are sample variances of two independent normal samples with variances σ_1^2 and σ_2^2 and sizes m and n respectively, the ratio $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ is distributed as $F_{m-1,n-1}$. The F-distribution is named after Sir Ronald Fisher, who in fact tabulated not F but $z = \frac{1}{2}\log F$. The F-distribution in its current form was first tabulated and used by George W. Snedecor, and the distribution is sometimes called Snedecor's F, or the Fisher–Snedecor F.

In MATLAB, the CDF at x for an F-distribution with m, n degrees of freedom is calculated as `fcdf(x,m,n)`, and the PDF is computed as `fpdf(x,m,n)`. The pth percentile is computed with `finv(p,m,n)`. Figure 6.11 provides a plot of a $F_{5,10}$ PDF.

6.6.5 Noncentral χ^2, t, and F Distributions

Noncentral χ^2, t, and F distributions are generalizations of standard χ^2, t, and F distributions. They are used mainly in power analysis of tests and sample size designs. For example, we will use noncentral t for power analysis of one-sample and two-sample t tests later in the text.

Random variable $\chi^2_n(\delta)$ has a *noncentral χ^2-distribution* with n degrees of freedom and parameter of noncentrality δ if it can be represented as

$$\chi^2_n(\delta) = Z_1 + Z_2 + \cdots + Z_{n-1} + X_n,$$

where $Z_1, Z_2, \ldots Z_{n-1}, X_n$ are independent random variables. Random variables $Z_1, \ldots, Z_{n-1}$ have a standard normal $\mathcal{N}(0,1)$ distribution while X_n is distributed as $\mathcal{N}(\delta,1)$. In MATLAB the noncentral χ^2 is denoted as `ncx2pdf`, `ncx2cdf`, `ncx2inv`, `ncx2stat`, and `ncx2rnd` for PDF, CDF, quantile, descriptive statistics, and random number generator.

Random variable $t_n(\delta)$ has a *noncentral t-distribution* with n degrees of freedom and noncentrality parameter δ if it can be represented as

$$t_n(\delta) = \frac{X}{\sqrt{\chi^2_n/n}},$$

where X and χ^2_n are independent, $X \sim \mathcal{N}(\delta,1)$, and χ^2_n has a (central) χ^2 distribution with n degrees of freedom. In MATLAB, functions `nctpdf`, `nctcdf`, `nctinv`, `nctstat`, and `nctrnd`, stand for PDF, CDF, quantile, descriptive statistics, and random number generator of the noncentral t.

Figure 6.12 plots the densities of noncentral t for values of the noncentrality parameter $-1, 0$, and 2. Noncentral t for $\delta = 0$ is a standard t-distribution.

Random variable $F_{m,n}(\delta)$ has a *noncentral F-distribution* with m, n degrees of freedom and parameter of noncentrality δ if it can be represented as

$$F_{m,n}(\delta) = \frac{\chi^2_m(\delta)/m}{\chi^2_n/n},$$

where $\chi^2_m(\delta)$ and χ^2_n are independent, with noncentral (δ) and standard χ^2 distributions with m and n degrees of freedom, respectively. In MATLAB, functions `ncfpdf`, `ncfcdf`, `ncfinv`, `ncfstat`, and `ncfrnd`, stand for the PDF, CDF, quantile, descriptive statistics, and random number generator of the noncentral F.

The noncentral F will be used in Chapter 11 for power calculations in several ANOVA designs.

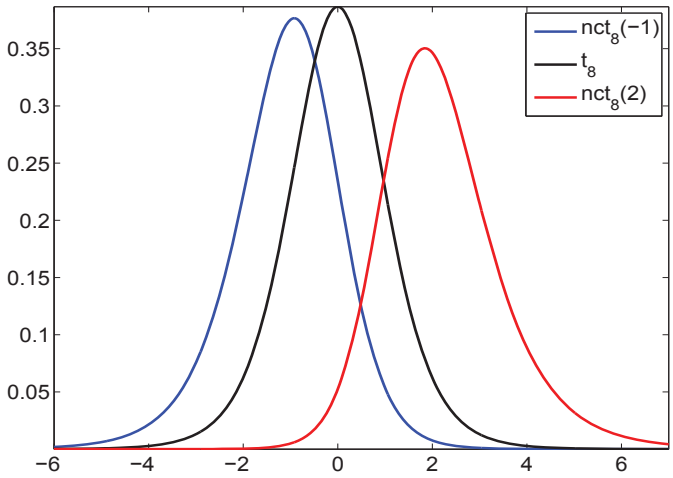


Fig. 6.12 Densities of noncentral $t_8(\delta)$ distribution for $\delta = -1, 0$, and 2.

6.6.6 Lognormal Distribution

A random variable X has a lognormal distribution with parameters μ and σ^2, $X \sim \mathcal{LN}(\mu, \sigma^2)$, if its density function is given by

$$f(x) = \frac{1}{x\sqrt{2\pi}\sigma} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\}, \quad x > 0.$$

If Y has a normal distribution, then $X = e^Y$ is lognormal.

Parameter μ is the mean and σ is the standard deviation of the distribution for the normal random variable $Y = \log X$, not the lognormal random variable X, and this can sometimes be confusing.

The moments of the lognormal distribution can be computed from the moment-generating function of the normal distribution. The nth moment is $\mathbb{E}(X^n) = \exp\{n\mu + n^2\sigma^2/2\}$, from which the mean and variance of X are

$$\mathbb{E}(X) = \exp\{\mu + \sigma^2/2\}, \quad \text{and} \quad \mathbb{V}\text{ar}\,(X) = \exp\{2(\mu + \sigma^2)\} - \exp\{2\mu + \sigma^2\}.$$

The median is $\exp\{\mu\}$ and the mode is $\exp\{\mu - \sigma^2\}$.

The lognormality is preserved under multiplication and division, i.e., the products and quotients of lognormal random variables remain lognormally distributed. If $X_i \sim \mathcal{LN}(\mu_i, \sigma_i^2)$, then $\prod_{i=1}^n X_i \sim \mathcal{LN}(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$.

Several biomedical phenomena are well modeled by a lognormal distribution, such as the age at onset of Alzheimer's disease, latent periods

of infectious diseases, or survival time after diagnosis of cancer. For measurement errors that are multiplicative, the lognormal distribution is the convenient model. More applications and properties can be found in Crow and Shimizu (1988).

In MATLAB, the CDF of a lognormal distribution with parameters m and s is evaluated at x as `logncdf(x,m,s)`, and the PDF is computed as `lognpdf(x,m,s)`. The pth percentile is computed with `logninv(p,m,s)`. Here the parameter `s` stands for σ, not σ^2. In WinBUGS, the lognormal distribution is coded as `dlnorm(mu,tau)`, where `tau` stands for the precision parameter $\frac{1}{\sigma^2}$.

Example 6.11. **Renner's Honey Data.** The content of hydroxymethylfurfurol (HMF, $\frac{mg}{kg}$) in 1573 honey samples (Renner, 1970) is well conforming to the lognormal distribution. The data set `renner.mat|dat` contains the interval midpoints (first column) and interval frequencies (second column). The parameter μ was estimated as -0.6084 and σ as 1.0040. The histogram and fitting density are shown in Figure 6.13 and the code is given in `renner.m`.

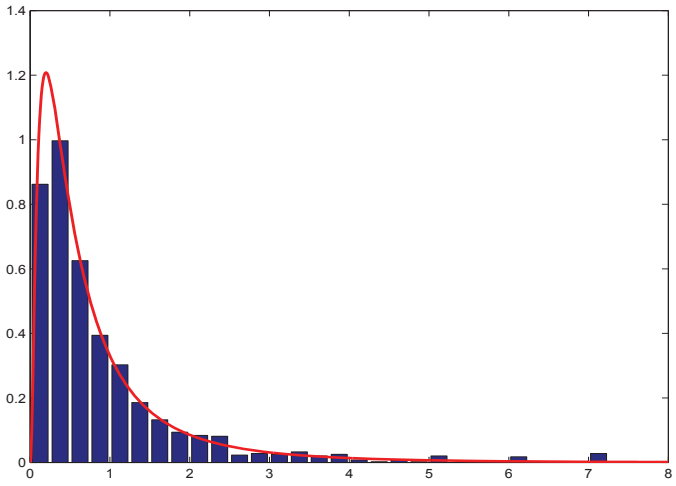


Fig. 6.13 Normalized histogram of Renner's honey data and lognormal distribution with parameters $\mu = -0.6083$ and $\sigma^2 = 1.0040^2$ that fits data well.

The goodness of such fitting procedures will be discussed in Chapter 17 more formally. Note that μ and σ are the mean and standard deviation of the logarithms of observations, not the observations themselves.

```
load 'renner.dat'
% mid-intervals, int. length = 0.25
rennerx = renner(:,1);
% frequencies in the interval
```

```
rennerf = renner(:,2);
n = sum(renner(:,2)); % sample size (n=1573)
bar(rennerx,   rennerf./(0.25 * n))
hold on
   m = sum(log(rennerx) .* rennerf)/n   %m =-0.6083
   s = sqrt(   sum( rennerf .*(log(rennerx) - m).^2 )/n   )
   %s=1.0040
xx = 0:0.01:8;
yy = lognpdf(xx, m, s);
plot(xx, yy, 'r-','linewidth',2)
```

6.7 Delta Method and Variance-Stabilizing Transformations*

The CLT states that for independent identically distributed random variables $X_1,\ldots,X_n$ with mean μ and finite variance σ^2,

$$\sqrt{n}(\overline{X}-\mu) \overset{approx}{\sim} \mathcal{N}(0,\sigma^2),$$

where the symbol $\overset{approx}{\sim}$ means *distributed approximately as*. Other than for a finite variance, there are no restrictions on the type, distribution, or any other feature of random variables X_i.

For a function g,

$$\sqrt{n}\left(g(\overline{X})-g(\mu)\right) \overset{approx}{\sim} \mathcal{N}(0,g'(\mu)^2\sigma^2).$$

The only restriction on g is that the derivative evaluated at μ must be finite and nonzero.

This result is called the *delta method* and the proof, which uses a simple Taylor expansion argument, will be omitted since it also uses facts concerning the convergence of random variables not covered in the text.

Example 6.12. Reciprocal and Square of Sample Mean. For n large

$$1/\overline{X} \overset{approx}{\sim} \mathcal{N}\left(\frac{1}{\mu},\ \frac{\sigma^2}{\mu^4}\right),$$
$$\left(\overline{X}\right)^2 \overset{approx}{\sim} \mathcal{N}\left(\mu^2,\ 4\mu^2\sigma^2\right).$$

The delta method is useful for many asymptotic arguments. Now we focus on the selection of the transformation g that stabilizes the variance.

Important statistical methodologies often assume that observations have variances that are constant for all possible values of the mean. Observations coming from a normal $\mathcal{N}(\mu,\sigma^2)$ distribution would satisfy this requirement since σ^2 does not depend on the mean μ. However, constancy of variances with respect to the mean is rather an exception than the rule. For example, if random variates from the exponential $\mathcal{E}(\lambda)$ distribution are generated, then the variance $\sigma^2 = 1/\lambda^2$ depends on the mean $\mu = 1/\lambda$, as $\sigma^2 = \mu^2$.

For some important distributions we will find a transformation that will make the variance constant and thus uninfluenced by the mean. This will prove beneficial for a range of inferential statistical procedures covered later in the text (confidence intervals, testing hypotheses).

Suppose that the variance $\mathbb{V}\text{ar } X = \sigma_X^2(\mu)$ can be expressed as a function of the mean $\mu = \mathbb{E}X$. For $Y = g(X)$, $\mathbb{V}\text{ar } Y \approx [g'(\mu)]^2\sigma_X^2(\mu)$, see (5.18). The condition that the variance of Y is constant leads to a simple differential equation

$$[g'(\mu)]^2\sigma_X^2(\mu) = c^2$$

with the following solution:

$$g(x) = c\int \frac{dx}{\sigma_X(x)}\,dx. \tag{6.4}$$

This is the theoretical basis for many proposed variance-stabilizing transformations. Note that $\sigma_X(x)$ in (6.4) is a function expressing the variance as a function of the mean.

Example 6.13. **Stabilizing Variance.** Suppose data are sampled from (a) Poisson $\mathcal{P}oi(\lambda)$, (b) exponential $\mathcal{E}(\lambda)$, and (c) binomial $\mathcal{B}in(n,p)$ distributions.

In (a), the mean and variance are equal, $\sigma^2(\mu) = \mu\ (= \lambda)$, and (6.4) becomes

$$g(x) = c\int \frac{dx}{\sqrt{x}}\,dx = 2c\sqrt{x} + d$$

for some constants c and d. Thus, as the variance-stabilizing transformation for Poisson observations we can take $g(x) = \sqrt{x}$.

In (b) and (c), $\sigma^2(\mu) = \mu^2$ and $\sigma^2(\mu) = \mu - \mu^2/n$, and, after solving the integral in (6.4), we find that the transformations are $g(x) = \log(x)$ and $g(x) = \arcsin\sqrt{x/n}$ (Exercise 6.19).

Example 6.14. **Box–Cox Transformation.** Box and Cox (1964) introduced a family of transformations, indexed by a parameter λ, applicable to positive data $X_1, \ldots, X_n$:

$$Y_i = \begin{cases} \frac{X_i^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log X_i, & \lambda = 0. \end{cases} \tag{6.5}$$

This transformation is mostly applied to responses in linear models exhibiting nonnormality or heterogeneity of variances (heteroscedasticity). For a properly selected λ, transformed data $Y_1, \ldots, Y_n$ may look "more normal" and amenable to standard modeling techniques. The parameter λ is selected by maximizing,

$$(\lambda - 1) \sum_{i=1}^{n} \log X_i - \frac{n}{2} \log \left[\frac{1}{n} \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \right], \tag{6.6}$$

where Y_i are as given in (6.5) and $\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$. As an illustration, we apply the Box–Cox transformation to apparently skewed data of pyruvate kinase concentrations.

Exercise 2.19 featured a multivariate data set dmd.dat in which the fourth column gives pyruvate kinase concentrations in 194 female relatives of boys with Duchenne muscular dystrophy (DMD). The distribution of this measurement is skewed to the right (Fig. 6.14a). We will find the Box–Cox transformation to symmetrize the data (make it approximately normal). Panel (b) gives the values of likelihood in (6.6) for different values of λ. Note that (6.6) is maximized for λ approximately equal to –0.15. Figure 6.14c gives the histogram for data transformed by the Box–Cox transformation with $\lambda = -0.15$. The histogram is notably symmetrized. For details see boxcox.m.

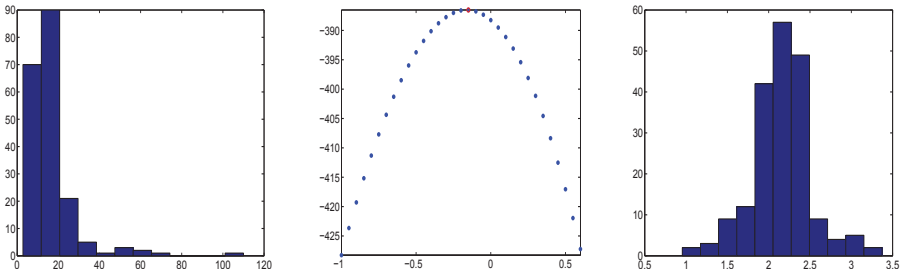


Fig. 6.14 (a) Histogram of row data of pyruvate kinase concentrations; (b) log-likelihood is maximized at $\lambda = -0.15$; and (c) histogram of Box–Cox-transformed data.

6.8 Exercises

6.1. **Standard Normal Calculations.** Random variable X has a standard normal distribution. What is larger, $\mathbb{P}(|X| \leq 0.7)$ or $\mathbb{P}(|X| \geq 0.7)$?

6.2. **Nonnegative Definiteness of Σ Constrains ρ.** A symmetric 2×2-matrix $A = [a \;\; b; b \;\; d]$ is nonnegative definite if $a \geq 0$ and $\det(A) = ad - b^2 \geq 0$. Show that condition $\det(\Sigma) \geq 0$ for Σ in (6.2), implies $-1 \leq \rho \leq 1$.

6.3. **Herrings.** The alewife (*Pomolobus pseudoharengus*, Wilson 1811) grows to maximum length of about 15 in., but adults average only about 10.5 in. long and about 8 oz. in weight; 16,400,000 fish taken in New England in 1898 weighed about 8,800,000 lbs.

Fig. 6.15 Alewife fish.

Assume that the length of an individual fish (Fig. 6.15) is normally distributed with mean 10.5 in. and standard deviation 1.6 in. and that the weight is distributed as χ^2 with 8 degrees of freedom.
(a) What percentage of fish are between 10.5 and 13 in. long?
(b) What percentage of fish weigh more than 10 oz.?
(c) Ten percent of fish are longer than x. Find x.

6.4. **Sea Urchins.** In a laboratory experiment, researchers at Barry University, (Miami Shores, FL) studied the rate at which sea urchins ingested turtle grass (*Florida Scientist*, Summer/Autumn 1991). The urchins were starved for 48 h, then fed 5-cm blades of green turtle grass. The mean ingestion time was found to be 2.83 h and the standard deviation 0.79 h. Assume that green turtle grass ingestion time for the sea urchins has an approximately normal distribution.
(a) Find the probability that a sea urchin will require between 2.3 and 4 h to ingest a 5-cm blade of green turtle grass.
(b) Find the time t^* (hours) so that 95% of sea urchins take more than t^* hours to ingest a 5-cm blade of green turtle grass.

6.5. **Pyruvate Kinase for Controls Is Normal.** Refer to Exercise 2.19. The histogram for PK response for controls, X, is fairly bell-shaped (as much

as 142 observations show), so you decided to fit it with a normal distribution, $\mathcal{N}(12, 4^2)$.
(a) How would you defend the choice of a normal model that allows for negative values when the measured level is always positive?
(b) Find the probability that X falls between 4 and 20.
(c) Find the probability that X exceeds 20.
(d) Find the value x_0 so that 93% of all PK measurements exceed x_0.

6.6. **Leptin.** Leptin (from the Greek word *leptos*, meaning thin) is a 16-kDa hormone that plays a key role in regulating energy intake and energy expenditure, including the regulation (decrease) of appetite and (increase) of metabolism. Serum leptin concentrations can be measured in several ways. One approach is by using a radioimmunoassay in venous blood samples (Linco Research Inc., St Charles, MO). Several studies have consistently found women to have higher serum leptin concentrations than do men. For example, among US adults across a broad age range, the mean serum leptin concentration in women is approximately normal $\mathcal{N}(12.7\ \mu\text{g/L}, (1.3\ \mu\text{g/L})^2)$ and in men approximately normal $\mathcal{N}(4.6\ \mu\text{g/L}, (0.5\ \mu\text{g/L})^2)$.
(a) What is the probability that the concentration of leptin in a randomly selected US adult male exceeds 6 μg/L?
(b) What proportion of US women have concentration of leptin in the interval $12.7 \pm 2\ \mu$g/L?
(c) What interval, symmetric about the mean 12.7 μg/L, contains leptin concentrations of 95% of adult US women?

6.7. **Pulse Rate.** The pulse rate of 1-month-old infants has a mean of 115 beats per minute and a standard deviation of 16 beats per minute.
(a) Explain why the average pulse rate in a sample of 64 1-month-old infants is approximately normally distributed.
(b) Find the mean and the variance of the normal distribution in (a).
(c) Find the probability that the average pulse rate of a sample of 64 will exceed 120.

6.8. **Side Effects.** One of the side effects of flooding a lake in northern boreal forest areas[1] (e.g., for a hydroelectric project) is that mercury is leached from the soil, enters the food chain, and eventually contaminates the fish. The concentration of mercury in fish will vary among individual fish because of differences in eating patterns, movements around the lake, etc. Suppose that the concentrations of mercury in individual fish follows an approximately normal distribution with a mean of 0.25 ppm and a standard deviation of 0.08 ppm. Fish are safe to eat if the mercury level is below 0.30 ppm. What proportion of fish are safe to eat?

[1] The northern boreal forest, sometimes also called the taiga or northern coniferous forest, stretches unbroken from eastern Canada westward throughout the majority of Canada to the central region of Alaska.

6.9. **Macrolepiota Procera.** The size of mushroom caps varies. While many species of *Marasmius* and *Collybia* are only 12 to 20 mm (1/2 to 3/4 in.) in diameter, some fungi are nearly 200 mm (8 in.) across. The cap diameter of parasol mushroom (*Macrolepiota procera*, Fig. 6.16) is a normal random variable with parameters $\mu = 230$ mm and $\sigma = 25$ mm.

Fig. 6.16 Parasol mushroom *Macrolepiota procera.*

(a) What proportion of parasol caps has a diameter between 200 and 250 mm?
(b) Five percent of parasol caps are larger than x_0 in diameter. Find x_0.

6.10. **Duration of Gestation in Humans.** Altman (1980) quotes the following incident from the UK: "In 1949 a divorce case was heard in which the sole evidence of adultery was that a baby was born 349 days after the husband had gone abroad on military service. The appeal judges agreed that medical evidence was unlikely but scientifically possible." So the appeal failed. "Most people think that the husband was hard done by," Altman adds.
So let us judge the judges. The reported mean duration of an uncomplicated human gestation is between 266 and 288 days, depending on many factors but mainly on the method of calculation. Assume that population mean and standard deviations are $\mu = 280$ and $\sigma = 10$ days, respectively. In fact, smaller standard deviations have been reported, so 10 days is a conservative choice. The normal model fits the data reasonably well if the samples are large.
Under the normal $\mathcal{N}(\mu,\sigma^2)$ model, find the probability that a gestation period will be equal to or greater than 349 days.

6.11. **Tolerance Design.** Eggert (2005) provides the following engineering design question. A 5-in. diameter pin will be assembled into a 5.005-in. journal bearing. The pin manufacturing tolerance is specified to $t_{pin} = 0.003$ inch. A minimum clearance fit of 0.001 in. is needed.
Determine tolerance required of the hole, t_{hole}, such that 99.9% of the mates will exceed the minimum clearance. Assume that manufacturing

variations are normally distributed. The tolerance is defined as 3 standard deviations.

6.12. **Ulnar Variance.** The lower arm is made up of two bones – the ulna and the radius. The length of these bones can lead to an ulnar variance, which can cause wrist pain, degenerative ailments, improper hand and wrist functioning.
This exercise uses data reported in Jung et al. (2001), who studied radiographs of the wrists of 120 healthy volunteers in order to determine the normal range of ulnar variance. The radiographs had been taken in various positions under both unloaded (static) and loaded (dynamic) conditions.
The ulnar variance in neutral rotation was modeled by normal distribution with a mean of $\mu = 0.74$ mm and standard deviation of $\sigma = 1.46$ mm.
(a) What is the probability that a radiogram of a normal person will show negative ulnar variance in neutral rotation (ulnar variance, unlike the statistical variance, can be negative)?
The researchers modeled the maximum ulnar variance (UV_{max}) as normal $\mathcal{N}(1.52, 1.56^2)$ when gripping in pronation and minimum ulnar variance (UV_{min}) as normal $\mathcal{N}(0.19, 1.43^2)$ when relaxed in supination.
(b) Find the probability that the mean dynamic range in ulnar variance, $C = UV_{max} - UV_{min}$, will exceed 1 mm.

6.13. **Independence of Sample Mean and Standard Deviation in Normal Samples.** Simulate 1000 samples from the standard normal distribution, each of size 100, and find their sample mean and standard deviation.
(a) Plot a scatterplot of sample means vs. the corresponding sample standard deviations. Are there any trends?
(b) Find the coefficient of correlation between sample means and standard deviations from (a) arranged as two vectors. Is the coefficient close to zero?

6.14. **Sonny and Multiple Choice Exam.** An instructor gives a 100-question multiple-choice final exam. Each question has 4 choices. In order to pass, a student has to have at least 35 correct answers. Sonny decides to guess at random on each question. What is the probability that Sonny will pass the exam?

6.15. **Amount of Liquid in a Bottle.** Suppose that the volume of liquid in a bottle of a certain chemical solution is normally distributed with a mean of 0.5 L and standard deviation of 0.01 L.
(a) Find the probability that a bottle will contain at least 0.48 L of liquid.
(b) Find the volume that corresponds to the 95th percentile.

6.16. **Marginals and Conditionals of a 2D Normal.** Find marginal and conditional densities $f_X(x)$, $f_Y(y)$, $f(x|y)$ and $f(y|x)$, if (X, Y) has density

$$f(x,y) = \frac{3\sqrt{3}}{\pi} \exp\left\{-4x^2 - 6xy - 9y^2\right\}.$$

6.17. **Meristem Cells in 3D.** Suppose that a particle is fired at a cell sitting at the origin of a spatial coordinate system, with units in microns. The particle lands at (X,Y,Z), where X,Y, and Z are independent, and each has a normal distribution with a mean of $\mu = 0$ and variance of $\sigma^2 = 250$. The particle is successfully delivered if it lands within $70\,\mu$m of the origin. Find the probability that the particle was not successfully delivered.

6.18. **Glossina morsitans.** *Glossina morsitans* (tsetse fly) is a large biting fly that inhabits most of midcontinental Africa. This fly is infamous as the primary biological vector (the meaning of vector here is epidemiological, not mathematical. A vector is any living carrier that transmits an infectious agent) of trypanosomes, which cause human sleeping sickness. The data in the table below are reported in Pearson (1914) and represent the frequencies of length in microns of trypanosomes found in *Glossina morsitans*.

Microns	15	16	17	18	19	20	21	22	23	24	25
Frequency	7	31	148	230	326	252	237	184	143	115	130
Microns	26	27	28	29	30	31	32	33	34	35	Total
Frequency	110	127	133	113	96	54	44	11	7	2	2500

The original data distinguished five different strains of trypanosomes, but it seems that the summary data set, as shown in the table, can be well approximated by a mixture of two normal distributions, $p_1\mathcal{N}\mu_1,\sigma_1^2) + p_2\mathcal{N}\mu_2,\sigma_2^2)$.
Using MATLAB's `gmdistribution.fit` identify the means of the two normal components, as well as their weights in the mixture, p_1 and p_2. Plot the normalized histogram and superimpose the density of the mixture. Data can be found in `glossina.mat`.

6.19. **Stabilizing the Variance.** In Example 6.13 it was stated that the variance stabilizing transformations for exponential $\mathcal{E}(\lambda)$ and binomial $\mathcal{B}in(n,p)$ distributions are $g(x) = \log(x)$ and $g(x) = \arcsin\sqrt{\frac{x}{n}}$, respectively. Prove these statements.

6.20. **From Normal to Lognormal.** Derive the density of a lognormal distribution by transforming $X \sim \mathcal{N}(0,1)$ into $Y = \exp\{X\}$.

6.21. **Changing the Threshold for FPG.** Woolf and Rothmich (1998) report that a change of the diagnostic threshold for fasting plasma glucose (FPG) from 140 to 126 mg per dL, drastically increased the number of people diagnosed as diabetics:

> Lowering the diagnostic threshold shifts the definition of diabetes into the central bulge of the distribution curve where the glucose level of most Americans falls. Among U.S. adults 40 to 74 years of age who have not been diagnosed with diabetes, 1.9 million have FPG levels of 126 to 140 mg per dL, which is almost as many as the number of people who have levels over 140 mg per dL. Under the new guidelines (ADA 1997), many Americans with FPG levels of 126 to 140 mg per dL, who previously would have been told that they had normal (or impaired) glucose tolerance, will now be informed that they harbor a disease.

Assume that the FPG of a randomly selected adult of age 40 to 74 from the US state of Georgia, can be modeled as lognormal $\mathcal{LN}(\mu,\sigma^2)$, where $\mu = 4.46$ and $\sigma^2 = 0.22^2$.
(a) Estimate how many people will fall in the range 126–140 if the population of adults of age 40 to 74 in Georgia is approximately 4 million.
(b) Find the FPG* level so that 95% of the population falls below FPG*.
(c) The lognormal model is not symmetric (lognormal distribution is positively skewed), so the mean is larger than the median. Find the median. In one sentence explain what this median represents in the terms of FPG.
Hint: In (a) you need first to estimate proportion of the population in 126–140 FPG range. MATLAB parametrizes lognormal distributions with μ and σ. Be careful about the mean and variance of FPG. They are **not** $\mu = 4.46$ and $\sigma^2 = 0.22^2$.

6.22. **The Square of a Standard Normal.** If $X \sim \mathcal{N}(0,1)$, show that $Y = X^2$ has a density of

$$f_Y(y) = \frac{1}{\sqrt{2}\Gamma\left(\frac{1}{2}\right)} y^{1/2-1} e^{-y/2}, \quad y \geq 0,$$

which is χ^2 with 1 degree of freedom.

MATLAB AND WINBUGS FILES AND DATA SETS USED IN THIS CHAPTER

http://statbook.gatech.edu/Ch6.Norm/

acid.m, aviocompany.m, boxcox.m, ch2itf.m, cltdemo.m, glossina.m, histn.m, ige.m, meanvarind.m, nor2chi2.m, piston.m, plot2dnormal.m, plotnct.m, quetelet.m, renner.m, simulplates.m, tsetse.m

aplysia.odc

glossina.mat, renner.dat|mat

CHAPTER REFERENCES

Altman, D. G. (1980). Statistics and ethics in medical research: misuse of statistics is unethical. *Br. Med. J.*, **281**, 1182–1184.

Banks, J., Carson, J. S. II, and Nelson, B. (1984). *Discrete Event System Simulation*, 2nd ed., Prentice Hall, Upper Saddle River, NJ.

Casella, G. and Berger, R. (2002). *Statistical Inference.* Duxbury Press, Belmont, CA.

Crow E. L. and Shimizu K., eds. (1988). *Lognormal Distributions: Theory and Application*. Dekker, New York.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer Texts in Statistics, Springer, New York.

Eggert, R. J. (2005). *Engineering Design*. Pearson Prentice Hall, Boston.

Helmert, F. R. (1876). Die Genauigkeit der Formel von Peters zur Berechnung des wahrscheinlichen Fehlers directer Beobachtungen gleicher Genauigkeit. *Astronom. Nachr.*, **88**, 113–132.

Jung, J. M., Baek, G. H., Kim, J. H., Lee, Y. H., and Chung, M. S. (2001). Changes in ulnar variance in relation to forearm rotation and grip. *J. Bone Joint Surg. Br.*, **83**, 7, 1029–1033. PubMed PMID: 11603517.

Koike, H. (1987). The extensibility of Aplysia nerve and the determination of true axon length. *J. Physiol.*, **390**, 469–487.

Lüroth, J. (1876). Vergleichung von zwei Werten des wahrscheinlichen Fehlers. *Astron. Nachr.*, **87**, 14, 209–220.

Pearson, K. (1914). On the probability that two independent distributions of frequency are really samples of the same population, with special reference to recent work on the identity of trypanosome strains. *Biometrika*, **10**, 1, 85–143.

Renner E. (1970). *Mathematisch-statistische Methoden in der praktischen Anwendung*. Parey, Hamburg.

Sautter, C. (1993). Development of a microtargeting device for particle bombardment of plant meristems. *Plant Cell Tiss. Org.*, **33**, 251–257.

Woolf, S. H. and Rothemich, S. F. (1998). New diabetes guidelines: A closer look at the evidence. *Am. Fam. Physician*, **58**, 6, 1287–1290.

Chapter 7
Point and Interval Estimators

A grade *is an inadequate report of an inaccurate judgment by a biased and variable judge of the extent to which a student has attained an undefined level of mastery of an unknown proportion of an indefinite amount of material.*

– Paul Dressel

WHAT IS COVERED IN THIS CHAPTER

- Moment-Matching and Maximum Likelihood Estimators
- Unbiased and Consistent Estimators
- Estimation of Mean and Variance
- Confidence Intervals
- Estimation of Population Proportions
- Sample Size Design by Length of Confidence Intervals
- Prediction and Tolerance Intervals
- Intervals for the Poisson Rate

7.1 Introduction

One of the primary objectives of inferential statistics is estimation of population characteristics, or descriptors, on the basis of limited information contained in a sample. The population descriptors are formalized by a statistical model, which can be postulated at various levels of specificity: a broad class of models, a parametric family, or a fully specific unique model.

Often, a functional or distributional form is fully specified but dependent on one or more parameters. Such a model is called parametric. When the model is parametric, the task of estimation is to find the best possible sample counterparts as estimators for the parameters and to assess the accuracy of the estimators.

The estimation procedure follows standard rules. Usually, a sample is taken and a *statistic*, as a function of observations, is calculated. The value of the statistic serves as a point estimator for the unknown population parameter. For example, responses in political pools observed as sample proportions are used to estimate the population proportion of voters in favor of a particular candidate. The associated model is binomial and the parameter of interest is the binomial proportion in the population.

The estimators for a parameter can be given as a single value – *point estimators* or as a range of values – *interval estimators*. For example, the sample mean is a point estimator of the population mean. Confidence intervals and credible sets in a Bayesian context are examples of interval estimators.

In this chapter, we first discuss general methods for finding estimators and then focus on estimation of specific population parameters: means, variances, proportions, rates, etc. Some estimators are universal; that is, they are not connected with any specific distribution. Universal estimators are a sample mean for the population mean and a sample variance for the population variance. However, for interval estimators and for Bayesian estimators, a knowledge of sampling distribution is critical.

In Chapter 2 we learned about many sample summaries that are good estimators for their population counterparts; these will be discussed further in this chapter. We have also seen some robust competitors based on order statistics and ranks; these will be discussed further in Chapter 18.

The methods for how to propose an estimator for a population parameter are discussed next. The methods will use knowledge of the form of population distribution or, equivalently, distribution of sample summaries treated as random variables.

7.2 Moment-Matching and Maximum Likelihood Estimators

We describe two approaches for devising point estimators: moment matching and maximum likelihood.

Matching Estimation. Matching theoretical descriptors, most often moments, with their empirical counterparts, is a natural way to propose an estimator. The theoretical moments of a random variable X with a density specified up to a parameter, $f(x|\theta)$, are functions of that parameter:

$$\mathbb{E}X^k = h(\theta).$$

For example, if the measurements have a Poisson distribution $\mathcal{P}oi(\lambda)$, the second moment $\mathbb{E}X^2$ is $\lambda + \lambda^2$, which is a function of λ. Here, $h(x) = x + x^2$.

Suppose we obtained a sample $X_1, X_2, \ldots, X_n$ from $f(x|\theta)$. The empirical counterparts for theoretical moments $\mathbb{E}X^k$ are sample moments

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

By matching the theoretical and empirical moments, an estimator $\hat{\theta}$ is found as a solution of the equation

$$\overline{X^k} = h(\theta).$$

For example, for the exponential distribution $\mathcal{E}(\lambda)$, the first theoretical moment is $\mathbb{E}X = 1/\lambda$. An estimator for rate parameter λ is obtained by solving the moment-matching equation $\overline{X} = 1/\lambda$, resulting in $\hat{\lambda}_{mm} = 1/\overline{X}$. Moment-matching estimators are not unique; different theoretical and sample moments can be matched. In the context of an exponential model, the second theoretical moment is $\mathbb{E}X^2 = 2/\lambda^2$, leading to an alternative matching equation,

$$\overline{X^2} = 2/\lambda^2,$$

with the solution

$$\hat{\lambda}_{mm} = \sqrt{\frac{2}{\overline{X^2}}} = \sqrt{\frac{2n}{\sum_{i=1}^{n} X_i^2}}.$$

The following simple MATLAB code simulates a sample of size 10^6 from an exponential distribution with rate parameter $\lambda = 3$, then calculates moment-matching estimators based on the first two moments.

```
Y = exprnd(1/3, 10e6, 1);
%parametrization in MATLAB is 1/lambda
1/mean(Y) %matching the first moment
 ans = 2.9981
sqrt(2/mean(Y.^2)) %matching the second moment
 ans = 2.9984
```

Example 7.1. Moment Matching for Gamma. Consider a sample from a gamma distribution with parameters r and λ. It is known that for $X \sim \mathcal{G}a(r,\lambda)$, $\mathbb{E}(X) = \frac{r}{\lambda}$, and $\mathbb{V}\text{ar}\, X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{r}{\lambda^2}$. It is easy to see that

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

Thus, the moment-matching estimators are

$$\hat{r}_{mm} = \frac{(\overline{X})^2}{\overline{X^2} - (\overline{X})^2} \qquad \text{and} \qquad \hat{\lambda}_{mm} = \frac{\overline{X}}{\overline{X^2} - (\overline{X})^2}.$$

✎

Matching estimation uses mostly moments, but any other statistic that is (i) easily calculated from a sample and (ii) whose population counterpart depends on parameter(s) of interest can be used in matching. For example, the sample/population quantiles can be used.

Example 7.2. **Melanoma Survival Rate.** In one study on cancer, the highest 5-year survival rate (90%) for women was for malignant melanoma of the skin. Assume that survival time T has an exponential distribution with an unknown rate parameter λ. Using quantiles, estimate λ.

From

$$P(T > 5) = 0.90 \quad \Rightarrow \quad \exp\{-5 \cdot \lambda\} = 0.90$$

it follows that $\hat{\lambda} = 0.0211$.
✎

Maximum Likelihood. An alternative method, which uses a functional form for distributions of measurements, is maximum likelihood estimation (MLE).

The MLE was first proposed and used by R. A. Fisher in the 1920s and remains one of the most popular tools in estimation theory and broader statistical inference. The method can be formulated as an optimization problem involving the search for extrema when the model is considered as a function of parameters.

Suppose that the sample $X_1, \ldots, X_n$ comes from a population with distribution $f(x|\theta)$ indexed by θ, which could be a scalar or a vector of parameters. Elements of the sample are independent, thus the joint distribution of $X_1, \ldots, X_n$ is a product of individual densities:

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

When the sample is observed, the joint distribution remains dependent upon the parameter,

$$L(\theta|X_1,\ldots,X_n) = \prod_{i=1}^{n} f(X_i|\theta), \tag{7.1}$$

and, as a function of the parameter, L is called the *likelihood*. The value of the parameter θ that maximizes the likelihood $L(\theta|X_1,\ldots,X_n)$ is the MLE, $\hat{\theta}_{mle}$.

The problem of finding the maximum of L and the value $\hat{\theta}_{mle}$ at which L is maximized is an optimization problem. In some cases, the maximum can be found directly or with the help of the log transformation of L. Other times, the procedure must be iterative and the solution is an approximation. In some cases, depending on the model and sample size, the maximum is not unique or does not exist.

In the most common cases, maximizing the logarithm of likelihood, *log-likelihood*, is simpler than maximizing the likelihood directly. This is because the product in L becomes the sum when a logarithm is applied:

$$\ell(\theta|X_1,\ldots,X_n) = \log L(\theta|X_1,\ldots,X_n) = \sum_{i=1}^{n} \log f(X_i|\theta),$$

and finding an extremum of a sum is simpler. Since the logarithm is a monotonically increasing function, the maxima of L and ℓ are achieved at the same value $\hat{\theta}_{mle}$ (see Figure 7.1 for an illustration).

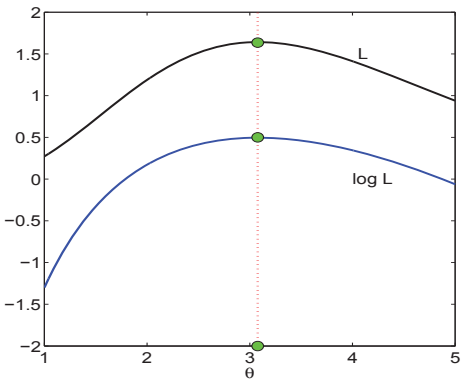


Fig. 7.1 Likelihood and log-likelihood of exponential distribution with rate parameter λ when the sample $X = [0.4, 0.3, 0.1, 0.5]$ is observed. The MLE is $1/\overline{X} = 3.077$.

Analytically,

$$\hat{\theta}_{mle} = \text{argmax}_\theta \ell(\theta|X_1,\ldots,X_n),$$

and it can be found as a solution of

$$\frac{\partial \ell(\theta|X_1,\ldots,X_n)}{\partial \theta} = 0 \quad \text{subject to} \quad \frac{\partial^2 \ell(\theta|X_1,\ldots,X_n)}{\partial \theta^2} < 0.$$

In simple terms, the MLE makes the first derivative (with respect to θ) of the log-likelihood equal to 0 and the second derivative negative, which is a condition for a maximum.

As an illustration, consider the MLE of λ in the exponential model, $\mathcal{E}(\lambda)$. After $X_1,\ldots,X_n$ is observed, the likelihood becomes

$$L(\lambda|X_1,\ldots,X_n) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} = \lambda^n \exp\left\{-\lambda \sum_{i=1}^{n} X_i\right\}.$$

The likelihood L is obtained as a product of densities $f(x_i|\lambda)$ where the arguments x_is are fixed observations X_i. The product is taken over all observations, as in (7.1). We can search for the maximum of L directly, but since it is a product of two terms involving λ, it is beneficial to look at the log-likelihood instead.

The log-likelihood is

$$\ell(\lambda|X_1,\ldots,X_n) = n\log\lambda - \lambda \sum_{i=1}^{n} X_i.$$

The equation to be solved is

$$\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0,$$

and the solution is $\hat{\lambda}_{mle} = \frac{n}{\sum_{i=1}^{n} X_i} = 1/\overline{X}$. The second derivative of the log-likelihood, $\frac{\partial^2 \ell}{\partial \lambda^2} = -\frac{n}{\lambda^2}$, is always negative; thus, the solution $\hat{\lambda}_{mle}$ maximizes ℓ, and consequently L. Figure 7.1 shows the likelihood and log-likelihood as functions of λ. For sample $X = [0.4, 0.3, 0.1, 0.5]$, the maximizing λ is $1/\overline{X} = 3.0769$. Note that both the likelihood and log-likelihood are maximized at the same value.

For the alternative parametrization of exponentials via a scale parameter, as in MATLAB, $f(x|\lambda) = \frac{1}{\lambda} e^{-x/\lambda}$, the estimator is, of course, $\hat{\lambda}_{mle} = \overline{X}$.

An important property of MLE is their *invariance property*.

Invariance Property of MLEs. Let $\hat{\theta}_{mle}$ be an MLE of θ and let $\eta = g(\theta)$, where g is an arbitrary function. Then $\hat{\eta}_{mle} = g(\hat{\theta}_{mle})$ is an MLE of η.

For example, if the MLE for λ in the exponential distribution was $1/\overline{X}$, then for a function of the parameter $\eta = \lambda^2 - \sin(\lambda)$ the MLE is $(1/\overline{X})^2 - \sin\left(1/\overline{X}\right)$.

In MATLAB, the function `mle` finds the MLE when inputs are data and the name of a distribution with a list of options. The normal distribution is the default. For example, `parhat = mle(data)` calculates the MLE for μ and σ of a normal distribution, evaluated at vector `data`. One of the outputs is the confidence interval. For example, `[parhat, parci] = mle(data)` returns MLEs and 95% confidence intervals for the parameters. The confidence intervals, as interval estimators, will be discussed later in this chapter. The command `[...] = mle(data,'distribution',dist)` computes parameter estimations for the distribution specified by `dist`. Acceptable strings for `dist` are as follows:

```
'beta'             'bernoulli'                'binomial'
'discrete uniform' 'exponential'              'extreme value'
'gamma'            'generalized extreme value' 'generalized pareto'
'geometric'        'lognormal'                'negative binomial'
'normal'           'poisson'                  'rayleigh'
'uniform'          'weibull'
```

Example 7.3. **MLE of Beta in MATLAB.** The following MATLAB commands show how to estimate parameters a and b in a beta distribution. We will simulate a sample of size 1,000 from a beta $\mathcal{B}e(2,3)$ distribution and then find MLEs of a and b from the sample.

```
a = betarnd( 2, 3,[1, 1000]);
thetahat = mle(a,'distribution', 'beta')
%thetahat = 1.9991    3.0267
```

It is possible to find the MLE using MATLAB's `mle` command for distributions that are not on the list. The code is given at the end of Example 7.4in which moment-matching estimators and MLEs for parameters in a Maxwell distribution are compared.

Example 7.4. **Moment-Matching Estimators and MLEs in a Maxwell Distribution.** The Maxwell distribution models random speeds of molecules in thermal equilibrium as given by statistical mechanics. A random variable X with a Maxwell distribution is given by the probability density function

$$f(x|\theta) = \sqrt{\frac{2}{\pi}}\,\theta^{3/2}\,x^2\,e^{-\theta x^2/2}, \qquad \theta > 0, x > 0.$$

Assume that we observed velocities $X_1, \ldots, X_n$ and want to estimate the unknown parameter θ.

The following theoretical moments for the Maxwell distribution are available: the expectation $\mathbb{E}X = 2\sqrt{\frac{2}{\pi\theta}}$, the second moment $\mathbb{E}X^2 = 3/\theta$, and the fourth moment $\mathbb{E}X^4 = 15/\theta^2$. To find moment-matching estimators for θ, the theoretical moments are "matched" with their empirical counterparts $\overline{X}$, $\overline{X^2} = \frac{1}{n}\sum_{i=1}^n X_i^2$, and $\overline{X^4} = \frac{1}{n}\sum_{i=1}^n X_i^4$, and the resulting equations are solved with respect to θ :

$$\begin{aligned} \overline{X} = 2\sqrt{\frac{2}{\pi\theta}} \quad &\Rightarrow \quad \hat{\theta}_1 = \frac{8}{\pi(\overline{X})^2}, \\ \frac{1}{n}\sum_{i=1}^n X_i^2 = \frac{3}{\theta} \quad &\Rightarrow \quad \hat{\theta}_2 = \frac{3n}{\sum_{i=1}^n X_i^2}, \\ \frac{1}{n}\sum_{i=1}^n X_i^4 = \frac{15}{\theta^2} \quad &\Rightarrow \quad \hat{\theta}_3 = \sqrt{\frac{15n}{\sum_{i=1}^n X_i^4}}. \end{aligned}$$

To find the MLE of θ, we show that the log-likelihood has the form $\frac{3n}{2}\log\theta - \frac{\theta}{2}\sum_{i=1}^n X_i^2 +$ factor free of θ. The maximum of the log-likelihood is achieved at $\hat{\theta}_{\text{MLE}} = \frac{3n}{\sum_{i=1}^n X_i^2}$, which is the same as the moment-matching estimator $\hat{\theta}_2$.

Specifically, if $X_1 = 1.4$, $X_2 = 3.1$, and $X_3 = 2.5$ are observed, the MLE of θ is $\hat{\theta}_{\text{MLE}} = \frac{9}{17.82} = 0.5051$. The other two moment-matching estimators are $\hat{\theta}_1 = 0.4677$ and $\hat{\theta}_3 = 0.5768$.

In MATLAB, the Maxwell distribution can be custom-defined using a 'handle' to an anonymous function `@`:

```
maxwell = @(x,theta)  sqrt(2/pi)  *  ...
  theta.^(3/2) * x.^2   .* exp( - theta * x.^2/2);
mle([1.4 3.1 2.5], 'pdf', maxwell, 'start', rand)
   %ans = 0.5051
```

In most cases, taking the log of likelihood simplifies finding the MLE. Here is an example in which the maximization of likelihood was done without the use of derivatives.

Example 7.5. Suppose the observations $X_1 = 2$, $X_2 = 5$, $X_3 = 0.5$, and $X_4 = 3$ come from the uniform $\mathcal{U}(0,\theta)$ distribution. We are interested in estimating θ. The density for the single observation X is $f(x|\theta) = \frac{1}{\theta}\mathbf{1}(0 \le x \le \theta)$, and the likelihood, based on n observations $X_1, \ldots, X_n$, is

$$L(\theta|X_1, \ldots, X_n) = \frac{1}{\theta^n} \cdot \mathbf{1}(0 \le X_1 \le \theta) \cdot \mathbf{1}(0 \le X_2 \le \theta) \cdot \ \ldots \ \cdot \mathbf{1}(0 \le X_n \le \theta).$$

The product in the expression above can be simplified: if all Xs are less than or equal to θ, then their maximum $X_{(n)}$ is less than θ as well. Thus,

$$\mathbf{1}(0 \le X_1 \le \theta) \cdot \mathbf{1}(0 \le X_2 \le \theta) \cdot \ \ldots \ \cdot \mathbf{1}(0 \le X_n \le \theta) \ = \ \mathbf{1}(X_{(n)} \le \theta).$$

Maximizing the likelihood now can be performed by inspection. In order to maximize $\frac{1}{\theta^n}$, subject to $X_{(n)} \le \theta$, we should take the smallest θ possible, and that θ is $X_{(n)} = \max X_i$. Therefore, $\hat{\theta}_{mle} = X_{(n)}$, and in this problem, the estimator is $X_{(4)} = X_2 = 5$.

An alternative estimator can be found by moment matching. It can be shown (the arguments are beyond the scope of this book) that in estimating θ in $\mathcal{U}(0,\theta)$, only $\max X_i$ should be used. What is the distribution of $\max X_i$?

We will find this distribution for general i.i.d. $X_i, i = 1,\ldots,n$, with CDF $F(x)$ and PDF $f(x) = F'(x)$.

The CDF is, by definition,

$$\begin{aligned} G(x) &= \mathbb{P}(\max X_i \le x) = \mathbb{P}(X_1 \le x, X_2 \le x, \ldots, X_n \le x) \\ &= \prod_{i=1}^{n} \mathbb{P}(X_i \le x) = (F(x))^n. \end{aligned}$$

The reasoning in the equation above is as follows: If the maximum is $\le x$, then all X_i are $\le x$, and vice versa. The density for $\max X_i$ is $g(x) = G'(x) = n\,F^{n-1}(x)\,f(x)$, and the first moment is

$$\mathbb{E} \max X_i = \int_{\mathbb{R}} x\,g(x)\,dx = \int_{\mathbb{R}} x\,n\,F^{n-1}(x)\,f(x)\,dx.$$

For the uniform distribution $\mathcal{U}(0,\theta)$,

$$\mathbb{E} \max X_i = \int_0^\theta x \cdot n\,(x/\theta)^{n-1} \cdot 1/\theta\,dx = \frac{n}{\theta^n} \int_0^\theta x^n\,dx = \frac{n}{n+1}\theta.$$

The expectation of the maximum $\mathbb{E} \max X_i$ is matched with the largest order statistic in the sample, $X_{(n)}$. Thus, in solving the moment-matching equation, we obtain an alternative estimator for θ, $\hat{\theta}_{mm} = \frac{n+1}{n} X_{(n)}$. In this problem, $\hat{\theta}_{mm} = 25/4 = 6.25$. For a Bayesian estimator, see Example 8.6.

7.3 Unbiasedness and Consistency of Estimators

Based on a sample $X_1,\ldots,X_n$ from a population with distribution $f(x|\theta)$, let $\hat{\theta}_n = g(X_1,\ldots,X_n)$ be a statistic that estimates the parameter θ. The statistic,

or estimator, $\hat{\theta}_n$ as a function of the sample is a random variable. As a random variable, the estimator has an expectation of $\mathbb{E}\hat{\theta}_n$, a variance of $\mathbb{V}\mathrm{ar}\,\hat{\theta}_n$, and its own distribution called a *sampling distribution*.

Example 7.6. **AB Blood-Group Proportion.** Suppose we are interested in finding the proportion of AB blood-group subjects in a particular geographic region. This proportion, θ, is to be estimated on the basis of the sample $Y_1, Y_2, \ldots, Y_n$, each having a Bernoulli $\mathcal{B}er(\theta)$ distribution taking values 1 and 0 with probabilities θ and $1-\theta$, respectively. The realization $Y_i = 1$ indicates the presence of the AB group in observation i. The sum $X = \sum_{i=1}^{n} Y_i$ is, by definition, binomial $\mathcal{B}in(n,\theta)$.

The estimator for θ is $\hat{\theta}_n = \overline{Y} = \frac{X}{n}$. It is easy to check that this estimator is both moment-matching ($\mathbb{E}Y_i = \theta$) and MLE (the likelihood is $\theta^{\sum Y_i}(1-\theta)^{n-\sum Y_i}$). Thus, $\hat{\theta}_n$ has a binomial distribution with rescaled realizations $\{0, 1/n, 2/n, \ldots, (n-1)/n, 1\}$, that is,

$$\mathbb{P}\left(\hat{\theta}_n = \frac{k}{n}\right) = \binom{n}{k}\theta^k(1-\theta)^{n-k}, \quad k = 0,1,\ldots,n,$$

which is the estimator's sampling distribution.

It can be shown, by referring to a binomial distribution, that the expectation of $\hat{\theta}_n$ is the expectation of the binomial, $n\theta$, multiplied by $1/n$,

$$\mathbb{E}\hat{\theta}_n = \frac{1}{n} \times n\theta = \theta,$$

and that the variance is

$$\mathbb{V}\mathrm{ar}\,\hat{\theta}_n = \left(\frac{1}{n}\right)^2 \times n\theta(1-\theta) = \frac{\theta(1-\theta)}{n}.$$

If $\mathbb{E}\hat{\theta}_n = \theta$, then the estimator $\hat{\theta}$ is called *unbiased*. The expectation is taken with respect to the sampling distribution. The quantity

$$b(\theta) = \mathbb{E}\hat{\theta}_n - \theta$$

is called the *bias* of $\hat{\theta}$.

The error in estimation can be assessed by various measures. The usual measure is the *mean squared error* (MSE).

The MSE is defined as

$$\text{MSE}(\hat{\theta},\theta) = \mathbb{E}(\hat{\theta}_n - \theta)^2.$$

The MSE represents the expected squared deviation of the estimator from the parameter it estimates. This expectation is taken with respect to the sampling distribution of $\hat{\theta}_n$.

From the definition of MSE,

$$\begin{aligned}\mathbb{E}(\hat{\theta}_n - \theta)^2 &= \mathbb{E}(\hat{\theta}_n - \mathbb{E}\hat{\theta}_n + \mathbb{E}\hat{\theta}_n - \theta)^2 \\ &= \mathbb{E}(\hat{\theta}_n - \mathbb{E}\hat{\theta}_n)^2 - 2\mathbb{E}(\hat{\theta}_n - \mathbb{E}\hat{\theta}_n)(\mathbb{E}\hat{\theta}_n - \theta) + (\mathbb{E}\hat{\theta}_n - \theta)^2 \\ &= \mathbb{E}(\hat{\theta}_n - \mathbb{E}\hat{\theta}_n)^2 + (\mathbb{E}\hat{\theta}_n - \theta)^2.\end{aligned}$$

Consequently, the MSE can be represented as a sum of the variance of the estimator and its bias squared:

$$\text{MSE}(\hat{\theta},\theta) = \mathbb{V}\text{ar}\,\hat{\theta} + b(\theta)^2.$$

The square root of the MSE is sometimes used; it is called the *root mean squared error* (RMSE). For example, in estimating the population proportion, the estimator $\hat{p} = X/n$, for the $X \sim \mathcal{B}in(n,p)$ model, is unbiased, $\mathbb{E}(\hat{p}) = p$. In this case, the MSE is $\mathbb{V}\text{ar}\,(\hat{p}) = pq/n$, and the RMSE is $\sqrt{pq/n}$. Note that the RMSE is a function of the parameter. If parameter p is replaced by its estimator $\hat{p}$, then the RMSE becomes the *standard error*, *s.e.*, of the estimator. For binomial p, the standard error of $\hat{p}$ is $s.e.(\hat{p}) = \sqrt{\hat{p}\hat{q}/n}$.

Remark. The *standard error* (*s.e.*) of any estimator usually refers to a sample counterpart of its RMSE, which is a sample counterpart of standard deviation for unbiased estimators. For example, if $X_1, X_2, \ldots, X_n$ are $\mathcal{N}(\mu,\sigma^2)$, then $s.e.(\overline{X}) = s/\sqrt{n}$.

Inspecting the variance of an unbiased estimator, when the sample size increases, allows for checking estimator's consistency. The consistency is a desirable property of estimators. Informally, it is defined as the convergence of an estimator, in a stochastic sense, to the parameter it estimates.

If, for an unbiased estimator $\hat{\theta}_n$, $\mathbb{V}\text{ar}\,\hat{\theta}_n \to 0$ when the sample size $n \to \infty$, the estimator is called *consistent*.

More advanced definitions of convergences of random variables, which are beyond the scope of this text, are required in order to deduce more pre-

cise definitions of asymptotic unbiasedness, weak and strong consistency. These definitions will not be discussed here.

Example 7.7. **Estimating Normal Variance.** Suppose that we are interested in estimating the parameter θ in a population with a distribution of $\mathcal{N}(0,\theta), \theta > 0$, and that the proposed estimator, when the sample $X_1, X_2, \ldots, X_n$ is observed, is $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$.

It can be demonstrated that, when $X \sim \mathcal{N}(0,\theta)$, $\mathbb{E}X^2 = \theta$ and $\mathbb{E}X^4 = 3\theta^2$, by representing X as $\sqrt{\theta}Z$ for $Z \sim \mathcal{N}(0,1)$ and using the fact that $\mathbb{E}Z^2 = 1$ and $\mathbb{E}Z^4 = 3$.

The estimator $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \overline{X^2}$ is unbiased and consistent. Since $\mathbb{E}\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i^2 = \frac{1}{n}n\theta = \theta$, the estimator is unbiased. To show consistency, it is sufficient to demonstrate that the variance tends to 0 as the sample size increases. This is evident from

$$\mathbb{V}\mathrm{ar}\,\hat{\theta} = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}\,X_i^2 = \frac{1}{n^2}3n\theta^2 = \frac{3\theta^2}{n} \to 0, \text{ when } n \to \infty.$$

Alternatively, we can use the fact that $\frac{1}{\theta}\sum_{i=1}^{n} X_i^2$ has a χ_n^2-distribution, therefore the sampling distribution of $\hat{\theta}$ is a scaled χ_n^2, where the scaling factor is $\frac{1}{n\theta}$. The unbiasedness and consistency follow from $\mathbb{E}\chi_n^2 = n$ and $\mathbb{V}\mathrm{ar}\,\chi_n^2 = 2n$ by accounting for the scaling factor.
✎

Some important examples of unbiased and consistent estimators are provided next.

7.4 Estimation of a Mean, Variance, and Proportion

7.4.1 Point Estimation of Mean

For a sample $X_1, \ldots, X_n$ of size n we have already discussed the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ as an estimator of location. A natural estimator of the population mean μ is the sample mean $\hat{\mu} = \overline{X}$. The estimator $\overline{X}$ is an "optimal" estimator of a mean in many different models/distributions and for many different definitions of optimality.

The estimator $\overline{X}$ varies from sample to sample. More precisely, $\overline{X}$ is a random variable with a fixed distribution depending on the common distribution of observations, X_i.

The following is true for *any* distribution in the population as long as $\mathbb{E}X_i = \mu$ and $\mathbb{V}\text{ar}(X_i) = \sigma^2$ exist:

$$\mathbb{E}\overline{X} = \mu, \quad \mathbb{V}\text{ar}(\overline{X}) = \frac{\sigma^2}{n}. \tag{7.2}$$

The preceding equations are a direct consequence of independence in a sample and imply that $\overline{X}$ is an unbiased and consistent estimator of μ. If, in addition, we assume normality $X_i \sim \mathcal{N}(\mu, \sigma^2)$, then the sampling distribution of $\overline{X}$ is known exactly (page 248),

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

and the relations in (7.2) are apparent.

Chebyshev's Inequality and Strong Law of Large Numbers*. There are two general results in probability that theoretically justify the use of the sample mean $\overline{X}$ to estimate the population mean, μ. These are Chebyshev's inequality and strong law of large numbers (SLLN). We will briefly overview these results.

The Chebyshev inequality states that when $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with mean μ and finite variance σ^2, the probability that $\overline{X}$ will deviate from μ is small,

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2},$$

for any $\epsilon > 0$. The inequality is a direct consequence of (5.9) with $(\overline{X}_n - \mu)^2$ in place of X and ϵ^2 in place of a.

To translate this to specific numbers, we can choose ϵ small, say 0.000001. Assume that the X_is have a variance of 1. The Chebyshev inequality states that with n larger than the solution of $1/(n \times 0.0000001^2) = 0.9999$, the distance between $\overline{X}_n$ and μ will be smaller than 0.000001 with a probability of 99.99%. Admittedly, n here is an experimentally unfeasible number; however, for any small ϵ, finite σ^2, and "confidence" $1 - \frac{\sigma^2}{n\epsilon^2}$ close to 1, such n is finite.

The laws of large numbers state that, as a numerical sequence, $\overline{X}_n$ converges to μ. Care is nevertheless needed. The sequence $\overline{X}_n$ is not a sequence of numbers, but a sequence of random variables, which are functions defined on sample spaces $\mathcal{S}$. Thus, direct application of a calculus-type of convergence is not appropriate. However, for any fixed realization from the sample space $\mathcal{S}$, the sequence $\overline{X}_n$ becomes numerical and a traditional con-

vergence can be stated. Thus, a correct statement for the so-called SLLN is

$$\mathbb{P}(\overline{X}_n \to \mu) = 1,$$

that is, viewed as an event, $\{\overline{X}_n \to \mu\}$ is a sure event – it happens with a probability of 1.

7.4.2 Point Estimation of Variance

To obtain some intuition, we start, once again, with a finite population: $y_1, \ldots, y_N$. The population variance is $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2$, where $\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$ is the population mean.

For a sample $X_1, X_2, \ldots, X_n$ that is observed, an estimator of variance σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

for μ known, and

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

for μ not known, which is estimated by $\overline{X}$.

In the expression for s^2 we divide the sum by $n-1$ instead of the "expected" n in order to ensure the unbiasedness of s^2, $\mathbb{E}s^2 = \sigma^2$. The proof of this fact is straightforward and does not require any distributional assumptions, except that the population variance σ^2 is finite.

Note that by the definition of variance, $\mathbb{E}(X_i - \mu)^2 = \sigma^2$ and $\mathbb{E}(\overline{X} - \mu)^2 = \sigma^2/n$.

$$\begin{aligned}
(n-1)s^2 &= \sum_{i=1}^{n}(X_i - \overline{X})^2 \\
&= \sum_{i=1}^{n}[(X_i - \mu) - (\overline{X} - \mu)]^2 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\overline{X} - \mu)\sum_{i-1}^{n}(X_i - \mu) + n(\overline{X} - \mu)^2 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2, \quad \text{since } \sum_{i=1}^{n}(X_i - \mu) = n(\overline{X} - \mu).
\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{E}(s^2) &= \frac{1}{n-1}\mathbb{E}(n-1)s^2 \\ &= \frac{1}{n-1}\mathbb{E}[\sum_{i=1}^{n}(X_i-\mu)^2 - n(\overline{X}-\mu)^2] \\ &= \frac{1}{n-1}(n\sigma^2 - n\frac{\sigma^2}{n}) \\ &= \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2.\end{aligned}$$

When, in addition, the population is normal $\mathcal{N}(\mu,\sigma^2)$, then

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1},$$

meaning that the statistic $\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2$ has a χ^2-distribution with $n-1$ degrees of freedom (see equation 6.3 and the related discussion).

For a sample from a normal distribution, the unbiasedness of s^2 is a consequence of the following two facts: $s^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$ and $\mathbb{E}\chi^2_{n-1} = (n-1)$. The variance of s^2 is

$$\mathbb{V}\text{ar}\, s^2 = \left(\frac{\sigma^2}{n-1}\right)^2 \times \mathbb{V}\text{ar}\, \chi^2_{n-1} = \frac{2\sigma^4}{n-1}, \tag{7.3}$$

since $\mathbb{V}\text{ar}\, \chi^2_{n-1} = 2(n-1)$. Unlike the unbiasedness result, $\mathbb{E}s^2 = \sigma^2$, which does not require a normality assumption, the result in (7.3) is valid only when observations come from a normal distribution. In the general case,

$$\mathbb{V}\text{ar}\, s^2 = \frac{\mu_4 - \mu_2^2}{n} - \frac{2(\mu_4 - 2\mu_2^2)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}, \tag{7.4}$$

where $\mu_k = \mathbb{E}(X - \mathbb{E}X)^k$ is kth central moment. It is easy to see how for a normal distribution, (7.4) becomes (7.3), since in this case $\mu_4 = 3\mu_2$ and $\mu_2 = \sigma^2$.

Although s^2 is an unbiased estimator for σ^2, s is not an unbiased estimator for σ, a fact that is often overlooked. If the population is normal, then $\sqrt{(n-1)/2}\frac{\Gamma((n-1)/2)}{\Gamma(n/2)}\, s$ is an unbiased estimator of σ. This bias correction for s is important when n is small; for n large the correction is negligible. For example, if $n = 50$, the unbiased estimator of σ is 1.0051 s.

As Figure 7.2 shows, the empirical distribution of normalized sample variances is close to a χ^2-distribution. We generated $M = 100{,}000$ samples of size $n = 8$ from a normal $\mathcal{N}(0, 5^2)$ distribution and found sample variances s^2 for each sample. The sample variances were multiplied by $n - 1 = 7$ and divided by $\sigma^2 = 25$. The histogram of these rescaled sample variances is plotted and the density of a χ^2-distribution with 7 degrees of freedom is superimposed in red. The code generating Figure 7.2 is given next.

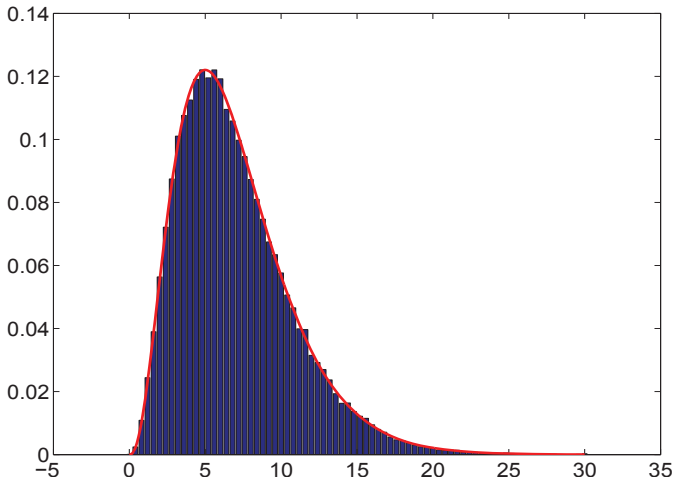


Fig. 7.2 Histogram of normalized sample variances $(n-1)s^2/\sigma^2$ obtained from $M = 100{,}000$ independent samples from $\mathcal{N}(0, 5^2)$, each of size $n = 8$. The density of a χ^2-distribution with 7 degrees of freedom is superimposed in *red*.

```
M=100000;  n = 8;
   X = 5 * randn([n, M]);
   ch2 = (n-1) * var(X)/25;
histn(ch2,0,0.4,30)
hold on
plot( (0:0.1:30), chi2pdf((0:0.1:30), n-1),'r-')
```

The code is efficient since a `for-end` loop is avoided. The simulated object `X` is an $n \times M$ matrix consisting of M columns (samples) of length n. The operator `var(X)` acts on columns of `X` producing M sample variances.

Several Robust Estimators of the Standard Deviation*. Suppose that a sample $X_1, \ldots, X_n$ is observed but its normality is not assumed. We discuss two estimators of the standard deviation that are calibrated by the normal distribution and are robust with respect to outliers and deviations from normality.

Gini's mean difference is defined as

$$G = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} |X_i - X_j|.$$

The statistic $G\frac{\sqrt{\pi}}{2}$ is an estimator of the standard deviation and is more robust to outliers than the standard statistic s.

A proposal by Croux and Rousseeuw (1992) involves absolute differences, as in Gini's mean difference estimator, but uses a kth-order statistic rather than the average. The estimator of σ is

$$Q = 2.2219 \{|X_i - X_j|, i < j\}_{(k)}, \quad \text{where } k = \binom{\lfloor n/2 \rfloor + 1}{2}.$$

The constant 2.2219 is used to calibrate the estimator, so that if the sample is a standard normal, then $Q = 1$. In calculating Q, all $\binom{n}{2}$ differences $|X_i - X_j|$ are ordered, and the kth in rank is selected and multiplied by 2.2219. This choice of k requires an additional multiplicative correction factor $n/(n + 1.4)$ for n odd, or $n/(n + 3.8)$ for n even.

MATLAB scripts `ginimd.m` and `crouxrouss.m` can be used to evaluate the estimators. The algorithm is naïve and uses a double loop to evaluate G and Q. The evaluation breaks down for sample sizes exceeding a few hundreds because of memory problems. A smarter algorithm that avoids looping is implemented in versions `ginimd2.m` and `crouxrouss2.m`. In these versions, the sample size can go up to 6,000.

In the next MATLAB session, we show how the robust estimators of the standard deviation perform. If 1,000 standard normal random variates are generated and one value is replaced with a clear outlier, say $X_{1000} = 20$, we will explore the influence of this outlier to both standard and robust estimators of the standard deviation. Note that s is quite sensitive, the outlier will inflate the estimator by almost 20%. The robust estimators are affected as well, but not as much as s.

```
x =randn(1, 1000);
x(1000)=20;
std(x)
% ans = 1.1999
s1 = ginimd2(x)
%s1 =1.0555
s2 = crouxrouss2(x)
%s2 =1.0287
iqr(x)/1.349
%ans = 1.0172
```

There are many other robust estimators of the variance/standard deviation. Good references containing extensive material on robust estimation are Wilcox (2005) and Staudte and Sheater (1990).

Estimation of Covariance. If $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent realizations of a bivariate random variable (X, Y), then an unbiased estimator of covariance

$$\sigma_{XY} = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)$$

is the sample covariance (page 32)

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}).$$

In the case of normal distribution, the variance of this estimator is

$$\mathbb{V}\text{ar}\,(s_{XY}) = \frac{\sigma_X^2 \sigma_Y^2 + \sigma_{XY}^2}{n-1}.$$

7.4.3 Point Estimation of Population Proportion

It is natural to estimate the population proportion p by a sample proportion. The sample proportion is the MLE and moment-matching estimator for p.

For sample proportions a binomial distribution is used as the theoretical model. Let $X \sim \mathcal{B}in(n, p)$, where parameter p is unknown. The MLE of p based on a single observation X is obtained by maximizing the likelihood

$$\binom{n}{X} p^X (1-p)^{n-X}$$

or the log-likelihood

$$\text{factor free of } p + X \log(p) + (n - X) \log(1 - p).$$

The maximum is obtained by solving

$$\begin{aligned} &(\text{factor free of } p + X \log(p) + (n - X) \log(1 - p))' = 0 \\ &\frac{X}{p} - \frac{n - X}{1 - p} = 0, \end{aligned}$$

which after some algebra gives the solution $\hat{p}_{mle} = \frac{X}{n}$.

In Example 7.6, we argued that the exact distribution for X/n is a rescaled binomial and that the statistic is unbiased, with the variance converging to 0 when the sample size increases. These two properties define a consistent estimator.

7.5 Confidence Intervals

Whenever the sampling distribution of a point estimator $\hat{\theta}_n$ is continuous, then necessarily $\mathbb{P}(\hat{\theta}_n = \theta) = 0$. In other words, the probability that the estimator exactly matches the parameter it estimates is 0.

Instead of the point estimator, one may report two estimators, $L = L(X_1, \ldots, X_n)$ and $U = U(X_1, \ldots, X_n)$, so that the interval $[L, U]$ covers θ with a probability of $1 - \alpha$, for small α. In this case, the interval $[L, U]$ will be called a $(1-\alpha)100\%$ confidence interval for θ.

For the construction of a confidence interval for a parameter, one needs to know the sampling distribution of the associated point estimator. The lower and upper interval bounds L and U depend on the quantiles of this distribution. We will derive the confidence interval for the normal mean, normal variance, population proportion, and Poisson rate. Many other confidence intervals, including differences, ratios, and some functions of statistics, are tightly connected to testing methodology and will be discussed in subsequent chapters.

Note that when the population is normal and $X_1, \ldots, X_n$ is observed, the exact sampling distributions of

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \quad \text{and}$$

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \times \frac{1}{\sqrt{\frac{(n-1)s^2}{\sigma^2}/(n-1)}}$$

are standard normal and t_{n-1} distributions, respectively.

The expression for t is shown as a product to emphasize the construction of a t-distribution from a standard normal (in *blue*) and χ^2 (in *red*), as in page 255. When the population is not normal but n is large, both statistics Z and t have an approximate standard normal distribution, due to the CLT.

We saw that the point estimator for the population probability of a success is the sample proportion $\hat{p} = X/n$, where X is the number of successes in n trials. The statistic X/n is based on a binomial sampling scheme in which X has exactly a binomial $\mathcal{B}in(n, p)$ distribution. Using this exact dis-

tribution would lead to confidence intervals in which the bounds and confidence levels are discretized. The normal approximation to the binomial (CLT in the form of de Moivre's approximation) leads to

$$\hat{p} \overset{approx}{\sim} \mathcal{N}\left(p, \frac{p(1-p)}{n}\right), \tag{7.5}$$

and the confidence intervals for the population proportion p would be based on normal quantiles.

7.5.1 Confidence Intervals for the Normal Mean

Let $X_1, \ldots, X_n$ be a sample from a normal $\mathcal{N}(\mu, \sigma^2)$ distribution where the parameter μ is to be estimated and σ^2 is known.

Starting from the identity

$$\mathbb{P}(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha$$

and the fact that $\overline{X}$ has a $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ distribution, we can write

$$\mathbb{P}\left(-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} + \mu \leq \overline{X} \leq z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} + \mu\right) = 1 - \alpha;$$

see Figure 7.3a for an illustration. Simple algebra gives

$$\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \tag{7.6}$$

which is a $(1-\alpha)100\%$ confidence interval.

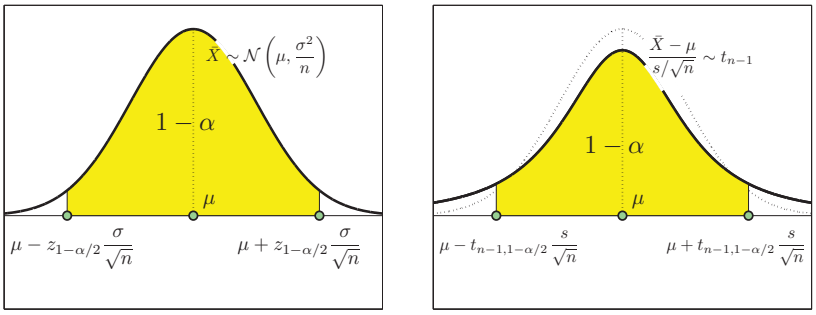
If σ^2 is not known, then a confidence interval with the sample standard deviation s in place of σ can be used. The z quantiles are valid for large n, but for small n ($n < 40$) we use t_{n-1} quantiles, since the sampling distribution for $\frac{\overline{X}-\mu}{s/\sqrt{n}}$ is t_{n-1}. Thus, for σ^2 unknown,

$$\overline{X} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} \tag{7.7}$$

is the confidence interval for μ of level $1-\alpha$.

Below is a summary of the above-stated intervals:

The $(1-\alpha)$ 100% confidence interval for an unknown normal mean μ on the basis of a sample of size n is



(a) (b)

Fig. 7.3 (a) When σ^2 is known, $\overline{X}$ has a normal $\mathcal{N}(\mu, \sigma^2/n)$ distribution and $\mathbb{P}(\mu - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$, leading to the confidence interval in (7.6). (b) If σ^2 is not known and s^2 is used instead, then $\frac{\overline{X}-\mu}{s/\sqrt{n}}$ is t_{n-1}, leading to the confidence interval in (7.7).

$$\left[\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

when the variance σ^2 is known, and

$$\left[\overline{X} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}},\ \overline{X} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}\right]$$

when the variance σ^2 is not known and s^2 is used instead.

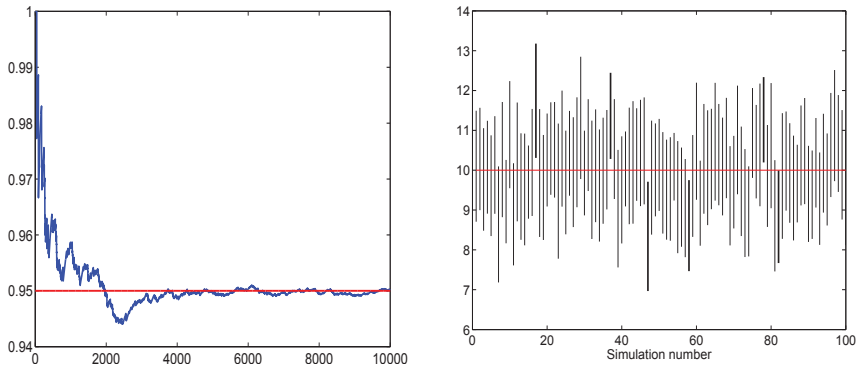
Interpretation of Confidence Intervals. What does a "confidence of 95%" mean? A common misconception is that this means that the unknown mean falls in the calculated interval with a probability of 0.95. Such a probability statement is valid for credible sets in the Bayesian context, which will be discussed in Chapter 8.

The interpretation of the $(1-\alpha)$ 100% confidence interval is as follows. If a random sample from a normal population is selected a large number of times and the confidence interval for the population mean μ is calculated, the proportion of such intervals covering μ approaches $1-\alpha$.

The following MATLAB code illustrates this. The code generates $M = 10{,}000$ random samples of size $n = 40$ from a normal population with a mean of $\mu = 10$ and a variance of $\sigma^2 = 4^2$; then it calculates a 95% confidence interval from each sample. It then counts how many of the intervals cover the mean μ, `cover = 1`, and finally finds their proportion, `covers/M`. The code was run consecutively several times and the following empirical confidences were obtained: 0.9461, 0.9484, 0.9469, 0.9487, 0.9502, 0.9482, 0.9502,

0.9482, 0.9530, 0.9517, 0.9503, 0.9514, 0.9496, 0.9515, etc., all clearly scattering around 0.95. Figure 7.4a plots the behavior of the coverage proportion when simulations range from 1 to 10,000. Figure 7.4b plots the first 100 intervals in the simulation and their position with respect to $\mu = 10$. The confidence intervals in simulations 17, 37, 47, 58, 78, and 82 fail to cover μ.

```
M=10000;          %simulate M times
n = 40;           % sample size
alpha = 0.05;     %1-alpha = confidence
tquantile = tinv(1-alpha/2, n-1);
covers =[];
for i = 1:M
    X = 10 + 4*randn(1,n); %sample, mean=10, var =16
    xbar = mean(X); s = std(X);
    LB = xbar - tquantile * s/sqrt(n);
    UB = xbar + tquantile * s/sqrt(n);
    % cover=1 if the interval covers population mean 10
    if UB < 10 | LB > 10
            cover = 0;
    else
            cover = 1;
    end
        covers =[covers cover]; %saves cover history
end
sum(covers)/M %proportion of intervals covering the mean
```



(a) (b)

Fig. 7.4 (a) Proportion of intervals covering the mean plotted against the iteration number, as in `plot(cumsum(covers)./(1:length(covers)) )`. (b) First 100 simulated intervals. The intervals 17, 37, 47, 58, 78, and 82 fail to cover the true mean.

7.5.2 *Confidence Interval for the Normal Variance*

Earlier (page 256) we argued that the sampling distribution of $\frac{(n-1)s^2}{\sigma^2}$ was χ^2 with $n-1$ degrees of freedom. From the definition of χ^2_{n-1} quantiles,

$$1-\alpha = \mathbb{P}(\chi^2_{n-1,\alpha/2} \le \chi^2_{n-1} \le \chi^2_{n-1,1-\alpha/2}),$$

as in Figure 7.5. Replacing χ^2_{n-1} with $\frac{(n-1)s^2}{\sigma^2}$, we get

$$1-\alpha = \mathbb{P}\left(\chi^2_{n-1,\alpha/2} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{n-1,1-\alpha/2}\right).$$

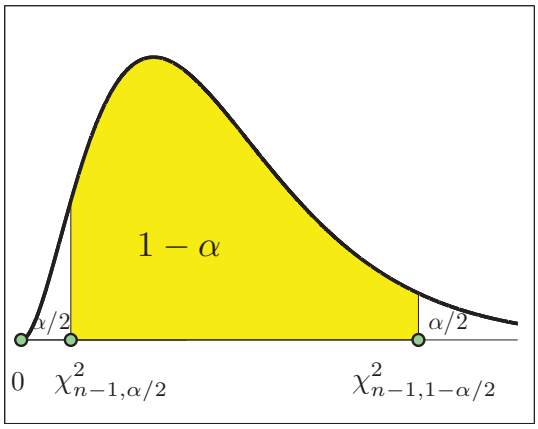


Fig. 7.5 Confidence interval for normal variance σ^2 is derived from $\mathbb{P}(\chi^2_{n-1,\alpha/2} \le (n-1)s^2/\sigma^2 \le \chi^2_{n-1,1-\alpha/2}) = 1-\alpha$.

Simple algebra with the inequalities above (taking the reciprocal of all three parts, being careful about the direction of the inequalities, and multiplying everything by $(n-1)s^2$) gives

$$\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}.$$

The $(1-\alpha)$ 100% confidence interval for an unknown normal variance is

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}},\ \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right]. \tag{7.8}$$

Remark. If the population mean μ is known, then s^2 is calculated as $\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$, and the χ^2 quantiles gain one degree of freedom (n instead of $n-1$). This makes the confidence interval a bit tighter.

Example 7.8. ***Amanita muscaria.*** With its bright red, sometimes dinner-plate-sized caps, the fly agaric (*Amanita muscaria*) is one of the most striking of all mushrooms. The white warts that adorn the cap, the white gills, a well-developed ring, and the distinctive volva of concentric rings distinguish the fly agaric from all other red mushrooms. The spores of the mushroom print white, are elliptical, and have a larger axis in the range of 7 to 13 μm (Fig. 7.6).

Fig. 7.6 Spores of *Amanita muscaria*.

Measurements of the diameter X of spores for $n = 51$ mushrooms are given in the following table:

10	11	12	9	10	11	13	12	10	11
11	13	9	10	9	10	8	12	10	11
9	10	7	11	8	9	11	11	10	12
10	8	7	11	12	10	9	10	11	10
8	10	10	8	9	10	13	9	12	9
9									

Assume that the measurements are normally distributed with mean μ and variance σ^2, but both parameters are unknown. The sample mean and variances are $\overline{X} = 10.098$, $s^2 = 2.1702$, and $s = 1.4732$. Also, the confidence interval would use an appropriate t-quantile, in this case `tinv(1-0.05/2, 51-1) = 2.0086`.

The 95% confidence interval for the population mean, μ, is

$$\left[10.098 - 2.0086 \times \frac{1.4732}{\sqrt{51}},\ 10.098 + 2.0086 \times \frac{1.4732}{\sqrt{51}}\right] = [9.6836, 10.5124].$$

Thus, the unknown mean μ belongs to the interval $[9.6836, 10.5124]$ with confidence 95%. That means that if the sample is obtained many times and for each sample the confidence interval is calculated, 95% of the intervals would contain μ.

To find, say, the 90% confidence interval for the population variance, σ^2, we need χ^2 quantiles, `chi2inv(1-0.10/2, 51-1) = 67.5048`, and `chi2inv(0.10/2, 51-1) = 34.7643`. According to (7.8), the interval is

$$[(51-1) \times 2.1702/67.5048,\ \ (51-1) \times 2.1702/34.7643] = [1.6074, 3.1213].$$

Thus, the interval $[1.6074, 3.1213]$ covers the population variance σ^2 with a confidence of 90%.
✎

Example 7.9. A Confidence Interval for σ^2 by CLT. An alternative confidence interval for the normal variance is possible. Since by the CLT $s^2 \overset{approx}{\sim} \mathcal{N}\left(\sigma^2, \frac{2\sigma^4}{n-1}\right)$ (Can you explain why?), when n is not small, an approximate $(1-\alpha)100\%$ confidence interval for σ^2 is

$$\left[s^2 - z_{1-\alpha/2} \cdot \frac{\sqrt{2}\, s^2}{\sqrt{n-1}},\ s^2 + z_{1-\alpha/2} \cdot \frac{\sqrt{2}\, s^2}{\sqrt{n-1}}\right].$$

In Example 7.8, $s^2 = 2.1702$ and $n = 51$. A 90% confidence interval for the variance was $[1.6074, 3.1213]$. By normal approximation,

```
 s2 = 2.1702; n=51; alpha = 0.1;
[s2 - norminv(1-alpha/2)*sqrt(2)* s2/sqrt(n-1), ...
 s2 + norminv(1-alpha/2)*sqrt(2)* s2/sqrt(n-1)]
%ans = 1.4563    2.8841
```

The interval $[1.4563, 2.8841]$ is shorter, compared to the standard confidence interval $[1.6074, 3.1213]$ obtained using χ^2 quantiles, as $1.4278 < 1.5139$. Insisting on equal-probability tails does not lead to the shortest interval since the χ^2-distribution is asymmetric. In addition, the approximate interval is centered at s^2. Why, then, is this interval not used? The coverage probability of a CLT-based interval is smaller than the nominal $1-\alpha$, and unless n is large (>100, say), this discrepancy can be significant (Exercise 7.28).
✎

7.5.3 Confidence Intervals for the Population Proportion

The sample proportion $\hat{p} = \frac{X}{n}$ has a range of optimality properties (unbiasedness, consistency); however, its realizations are discrete. For this reason, confidence intervals for p are obtained using the normal approximation, or connections of binomial with other continuous distributions, such as F.

Recall that for n large and np or nq not small (>10), the binomial X can be approximated by a $\mathcal{N}(np, npq)$ distribution. This approximation leads to $\frac{X}{n} \overset{approx}{\sim} \mathcal{N}\left(p, \frac{pq}{n}\right)$.

Note, however, that the standard deviation of $\hat{p}$, $\sqrt{\frac{pq}{n}}$, is not known, as it depends on p, and for the confidence interval we can use a plug-in estimator $\sqrt{\frac{\hat{p}\hat{q}}{n}}$ instead.

Let p be the population proportion and $\hat{p}$ the observed sample proportion. Assume that the smaller of $\frac{np}{q}$ and $\frac{nq}{p}$ is larger than 10. Then the $(1-\alpha)100\%$ confidence interval for unknown p is

$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}\,\hat{q}}{n}},\ \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}\,\hat{q}}{n}}\right].$$

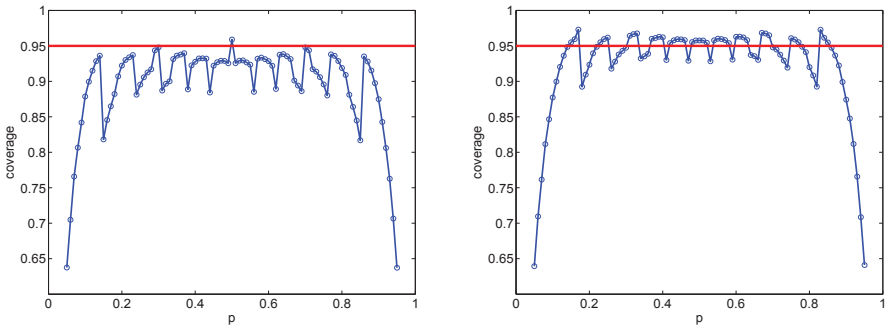
This interval is known as the Wald interval (Wald and Wolfowitz, 1939).

The Wald interval is used quite frequently, but its performance is suboptimal and even poor when p is close to 0 or 1. Figure 7.7a demonstrates the performance of Wald's 95% confidence interval for $n = 20$ and p ranging from 0.05 to 0.95 with a step of 0.01. The plot is obtained by simulation (`waldsimulation.m`). For each p, 100,000 binomial proportions are simulated, the Wald confidence intervals calculated, and the proportion of those intervals containing p is plotted. Notice that for nominal 95% confidence, the actual coverage probability may be much smaller, depending on p.

Unless the sample size n is very large, the Wald interval should not be used. The performance of Wald's interval can be improved by continuity corrections:

$$\left[\hat{p} - \frac{1}{2n} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}\,\hat{q}}{n}},\ \hat{p} + \frac{1}{2n} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}\,\hat{q}}{n}}\right].$$

Figure 7.7b shows the coverage probability of Wald's corrected interval.



(a) (b)

Fig. 7.7 (a) Simulated coverage probability for Wald's confidence interval for the true binomial proportion p ranging from 0.05 to 0.95, and $n = 20$. For each p, 100,000 binomial proportions are simulated, the Wald confidence intervals calculated, and the proportion of those containing p plotted. (b) The same as (a), but for the corrected Wald interval.

There is a range of intervals that have a performance superior to Wald's interval. An overview of several alternatives is provided next.

Adjusted Wald Interval. The adjusted Wald interval (Agresti and Coull, 1998) uses $p^* = \frac{X+2}{n+4}$ as an estimator of the proportion. Adding "two successes and two failures" was proposed by Wilson (1927):

$$\left[p^* - z_{1-\alpha/2}\sqrt{\frac{p^*q^*}{n+4}},\ p^* + z_{1-\alpha/2}\sqrt{\frac{p^*q^*}{n+4}} \right].$$

We will see in the next chapter that Wilson's proposal p^* has a Bayesian justification (page 343).

Wilson Score Interval. The Wilson score interval is another adjustment to the Wald interval based on the so-called Wilson-score test (Wilson, 1927; Hogg and Tanis, 2001):

$$\left[\frac{1}{1+z^2/n}\left(\hat{p} + \frac{z^2}{2n} - z\sqrt{\frac{\hat{p}\,\hat{q}}{n} + \frac{z^2}{4n^2}}\right),\ \frac{1}{1+z^2/n}\left(\hat{p} + \frac{z^2}{2n} + z\sqrt{\frac{\hat{p}\,\hat{q}}{n} + \frac{z^2}{4n^2}}\right) \right],$$

where z is $z_{1-\alpha/2}$. This interval can be obtained by solving the inequality

$$|\hat{p} - p| \ \leq\ z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

with respect to p. After squaring the left- and right-hand sides and some algebra, we get the quadratic inequality

$$p^2\left(1+\frac{z^2_{1-\alpha/2}}{n}\right)-p\left(2\hat{p}+\frac{z^2_{1-\alpha/2}}{n}\right)+\hat{p}^2\le 0,$$

for which the solution coincides with Wilson's score interval.

Clopper–Pearson Interval. The Clopper–Pearson confidence interval (Clopper and Pearson, 1934) does not use a normal approximation but, rather, exact links among binomial, beta, and F distributions. For $0<X<n$, the $(1-\alpha)\cdot 100\%$ Clopper–Pearson confidence interval is

$$\left[\frac{X}{X+(n-X+1)F^*},\ \frac{(X+1)F^{**}}{n-X+(X+1)F^{**}}\right],$$

where F^* is the $(1-\alpha/2)$-quantile of the F_{ν_1,ν_2}-distribution with $\nu_1=2(n-X+1)$ and $\nu_2=2X$ and F^{**} is the $(1-\alpha/2)$-quantile of the F_{ν_1,ν_2}-distribution with $\nu_1=2(X+1)$ and $\nu_2=2(n-X)$. In terms of beta distribution, Clopper–Pearson interval takes a very simple form, its MATLAB code is `[betainv(alpha/2, X, n-X+1), betainv(1-alpha/2, X+1, n-X)].`

When $X=0$, the interval is $[0,1-(\alpha/2)^{1/n}]$ and for $X=n$, $[(\alpha/2)^{1/n},1]$.

Anscombe's ArcSin Interval. For $X\sim \mathcal{B}in(n,p)$ Anscombe (1948) showed that if $p^*=\frac{X+3/8}{n+3/4}$, then the quantity

$$2\sqrt{n}(\arcsin\sqrt{p^*}-\arcsin\sqrt{p})$$

has an approximately standard normal distribution. From this result it follows that

$$\left[\sin^2\left(\arcsin\sqrt{p^*}-\frac{z_{1-\alpha/2}}{2\sqrt{n}}\right),\ \sin^2\left(\arcsin\sqrt{p^*}+\frac{z_{1-\alpha/2}}{2\sqrt{n}}\right)\right]$$

is the $(1-\alpha)100\%$ confidence interval for p.

The next example shows the comparative performance of different confidence intervals for the population proportion.

Example 7.10. **Cyclosporine Reversal Study.** An interesting case study involved research on the therapeutic benefits of cyclosporine on patients with chronic inflammatory bowel disease (Crohn's disease). In a double-blind clinical trial, researchers reported (Brynskov et al., 1989) that out of 37 patients with Crohn's disease resistant to standard therapies, 22 improved after a 3-month period. This proportion was significantly higher than that for the placebo group (11/34). The study was published in the *New England Journal of Medicine.*

However, at the 6-month follow-up, no significant differences were found between the treatment group and the control. In the cyclosporine

group, 30 patients *did not* improve, compared to 23 out of 34 in the placebo group (Brynskov et al., 1991). Thus, the proportion of patients who benefited in the cyclosporine group dropped from $\hat{p}_1 = 22/37 = 59.46\%$ at the 3-month to $\hat{p}_2 = 7/37 = 18.92\%$ at the 6-month follow-up. The researchers state: "We conclude that a short course of cyclosporin treatment does not result in long-term improvement in active chronic Crohn's disease."

To illustrate the performance of several introduced confidence intervals for the population proportion, we will find Wald's, Wilson's, Wilson score, Clopper–Pearson's, and Arcsin 95% confidence intervals for the proportion of patients who benefited in the cyclosporine group at the 3-month and 6-month follow-ups. Calculations are performed in MATLAB.

```
%Cyclosporine Clinical Trials
%
n = 37; %number of subjects in cyclosporine group
%  three months
X1 = 22;     p1hat = X1/n;   q1hat = 1-p1hat;
%  six months
X2 = 7;       p2hat = X2/n;   q2hat = 1- p2hat;
%================================
%Wald Intervals
W3 = [p1hat  - norminv(0.975) * sqrt( p1hat * q1hat / n), ...
    p1hat  + norminv(0.975) * sqrt( p1hat * q1hat / n)]
W6 = [p2hat  - norminv(0.975) * sqrt( p2hat * q2hat / n), ...
    p2hat  + norminv(0.975) * sqrt( p2hat * q2hat / n)]
%W3 = 0.4364       0.75279
%W6 = 0.06299      0.31539
%================================
% Wilson Intervals
     p1hats = (X1+2)/(n+4);   q1hats = 1-p1hats;
     p2hats = (X2+2)/(n+4);   q2hats = 1- p2hats;
Wi3 = [p1hats - norminv(0.975)*sqrt( p1hats * q1hats/(n+4)), ...
    p1hats + norminv(0.975) * sqrt( p1hats * q1hats/(n+4))];
Wi6 = [p2hats - norminv(0.975)*sqrt( p2hats * q2hats/(n+4)), ...
    p2hats + norminv(0.975) * sqrt( p2hats * q2hats/(n+4))];
% Wi3 =     0.43457       0.73617
% Wi6 =     0.092815      0.34621
%==========================
%Wilson Score Intervals
z=norminv(0.975);
Wis3 = [  1/(1 + z^2/n) *  (p1hat  + z^2/(2 * n)  - ...
    z * sqrt( p1hat * q1hat / n  + z^2/(4 * n^2))), ...
    1/(1 + z^2/n) * (p1hat  + z^2/(2 * n)  +  ...
    z * sqrt( p1hat * q1hat / n  + z^2/(4 * n^2)))];
Wis6 = [  1/(1 + z^2/n) *  (p2hat  + z^2/(2 * n)  - ...
    z * sqrt( p2hat * q2hat / n  + z^2/(4 * n^2))), ...
    1/(1 + z^2/n) * (p2hat  + z^2/(2 * n)  +  ...
    z * sqrt( p2hat * q2hat / n  + z^2/(4 * n^2)))];
%Wis3 =   0.43486     0.73653
%Wis6 =   0.0948      0.34205
%========================
```

```
% Clopper-Pearson Intervals
       Fs = finv(0.975, 2*(n-X1 + 1), 2*X1);
       Fss = finv(0.975, 2*(X1+1), 2*(n-X1));
      CP3 = [ X1/(X1 + (n-X1+1).*Fs),  ...
          (X1+1).*Fss./(n - X1 + (X1+1).*Fss)];

       Fs = finv(0.975, 2*(n-X2 + 1), 2*X2);
       Fss = finv(0.975, 2*(X2+1), 2*(n-X2));
      CP6 = [ X2/(X2 + (n-X2+1).*Fs), ...
          (X2+1).*Fss./(n - X2 + (X2+1).*Fss)];
%CP3 = 0.421         0.75246
%CP6 = 0.079621      0.35155
%=========================================
% Anscombe ARCSIN intervals
%
p1h = (X1 + 3/8)/(n + 3/4);  p2h = (X2 + 3/8)/(n + 3/4);

AA3 = [(sin(asin(sqrt(p1h))-norminv(0.975)/(2*sqrt(n))))^2, ...
       (sin(asin(sqrt(p1h))+norminv(0.975)/(2*sqrt(n))))^2];
AA6 = [(sin(asin(sqrt(p2h))-norminv(0.975)/(2*sqrt(n))))^2, ...
       (sin(asin(sqrt(p2h))+norminv(0.975)/(2*sqrt(n))))^2];

%AA3 = 0.43235        0.74353
%AA6 = 0.085489       0.3366
```

Figure 7.8 shows the pairs of confidence intervals at the 3- and 6-month follow-ups. Wald's intervals are in black, Wilson's in red, the Wilson score in green, Clopper–Pearson's in magenta, and ArcSin in blue. Notice that for all methods, the confidence intervals at the 3- and 6-month follow-ups are well separated, suggesting a significant change in the proportions. There are differences among the intervals, in their centers and lengths, for a particular time of follow-up. However, as Figure 7.8 indicates, these differences are not large.

Next, we discuss the confidence interval for the probability of success when in n trials no successes have been observed.

7.5.4 *Confidence Intervals for Proportions When* $X = 0$

When the binomial probability is small, it is not unusual that out of n trials no successes are observed. How do we find a $(1-\alpha)100\%$ confidence interval in such a case? The Clopper–Pearson interval is possible for $X = 0$, and it is given by $[0, 1-(\alpha/2)^{1/n}]$.

Yet it is possible to establish an alternative interval based on the following consideration. First, we have $(1-p)^n$ is as the probability of no success in n trials, and this probability is at least α:

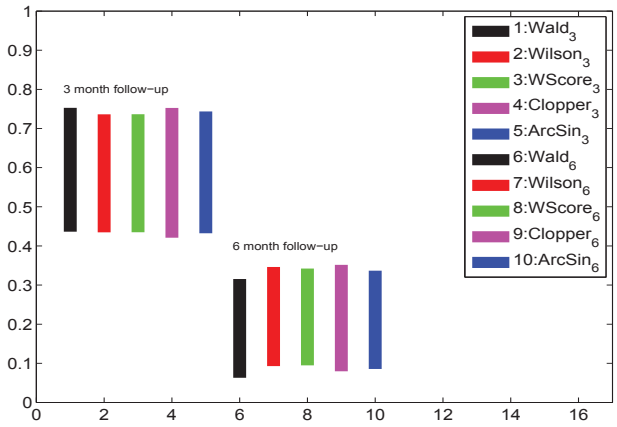


Fig. 7.8 Confidence intervals at 3- and 6-month follow-ups. Wald's intervals are in *black*, Wilson's in *red*, the Wilson Score in *green*, Clopper–Pearson's in *magenta*, and ArcSin in *blue*.

$$(1-p)^n \geq \alpha.$$

Since $n\log(1-p) \geq \log(\alpha)$ and $\log(1-p) \approx -p$, then

$$p \leq -\log(\alpha)/n.$$

This is a basis for the so-called $3/n$ rule: the 95% confidence interval for p is $[0,3/n]$ if no successes have been observed since $-\log(0.05) = 2.9957 \approx 3$. By symmetry, the 95% confidence interval for p when n successes are observed in n experiments is $[1-3/n,1]$. When n is small, this rule leads to intervals that are too wide to be useful. See Exercise 7.31 for a comparison of the Clopper–Pearson and $3/n$-rule intervals. We will argue in the next chapter that in the case where no successes are observed, one should approach the inference in a Bayesian manner.

7.5.5 Designing the Sample Size with Confidence Intervals

In all previous examples it was assumed that we had data in hand. Thus, we looked at the data after the sampling procedure had been completed. It is often the case that we have control over what sample size to adopt before the sampling. How large should the sample be? On one hand, a sample that is too small may affect the validity of our statistical conclusions. On the other hand, an unnecessarily large sample wastes money, time, and resources.

The length L of the $(1-\alpha)100\%$ confidence interval is $L = 2z_{1-\alpha/2}\sigma/\sqrt{n}$ for the normal mean and $L = 2z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$ for the population proportion. The sample size n is determined by solving the preceeding equations when L is fixed.

(i) Sample size for estimating the mean: σ^2 is known:

$$n \geq \frac{4z_{1-\alpha/2}^2\sigma^2}{L^2}, \tag{7.9}$$

where L is the length of the interval.

(ii) Sample size for estimating the proportion:

$$n \geq \frac{4z_{1-\alpha/2}^2\hat{p}(1-\hat{p})}{L^2}, \tag{7.10}$$

where $\hat{p}$ is estimated or elicited.

Designing the sample size usually precedes the sampling. In the absence of data, $\hat{p}$ is elicited from experts or inferred from prior studies. In the absence of any information, the most conservative choice is $\hat{p} = 0.5$.

It is possible to express L^2 in the units of variance of observations, σ^2, for the normal and $p(1-p)$ for the Bernoulli distribution. Therefore, it is sufficient to state that L/σ is $1/2$, for example, or that $L/\sqrt{p(1-p)}$ is $1/4$, and the required sample size can be calculated.

Example 7.11. **Cholesterol Level.** You are asked to design a cholesterol study experiment and you would like to estimate the mean cholesterol level of all students on a large metropolitan campus. You plan to take a random sample of n students and measure their cholesterol levels. Previous studies have shown that the standard deviation is 25, and you intend to use this value in planning your study. If a 99% confidence interval with a total length not exceeding 12 is desired, how many students should you include in your sample?

For a 99% confidence level, the normal 0.995 quantile is needed, $z_{0.995} = 2.58$. Then, $n \geq \frac{4 \cdot 2.5758^2 \cdot 25^2}{12^2} = 115.1892$, and desired sample size is 116 since 115.1892 should be rounded to the closest larger integer.

The *margin of error* is defined as half of the length of a 95% confidence interval for unknown proportion, location, scale, or some other population parameter of interest.

In popular use, the margin of error is usually connected with public opinion polls and represents the quantifiable sampling error built into well-designed sampling schemes. For estimating the true proportion of voters favoring a particular candidate, an approximate 95% confidence interval is

$$\left[\ \hat{p} - 1.96\sqrt{\hat{p}\ \hat{q}/n},\ \hat{p} + 1.96\sqrt{\hat{p}\ \hat{q}/n}\ \right],$$

where $\hat{p}$ is the sample proportion of voters favoring the candidate, $\hat{q} = 1 - \hat{p}$, 1.96 is the normal 97.5 percentile, and n is the sample size. Since $\hat{p}\ \hat{q} \le 1/4$, the margin of error, $1.96\sqrt{\hat{p}\ \hat{q}/n}$, is usually conservatively rounded to $\frac{1}{\sqrt{n}}$.

For example, if a survey of $n = 1600$ voters yields that 52% favor a particular candidate, then the margin of error can be estimated as $1/\sqrt{1600} = 1/40 = 0.025 = 2.5\%$ and is independent of the realized proportion of 52%.

However, if the true proportion is not close to 1/2, the precision of the margin of error can be improved by selecting a less conservative upper bound on $\hat{p}\ \hat{q}$. For example, if a survey of $n = 1600$ citizens yields that 16% of them favor policy P, the margin of error can be estimated as $1.96 \cdot \sqrt{0.2 \cdot 0.8/1600} \approx 1/50 = 0.02 = 2\%$ provided that we are certain that the true proportion of citizens supporting policy P does not exceed 20%.

7.6 Prediction and Tolerance Intervals*

In addition to confidence intervals for parameters, a researcher may be interested in predicting future observations. This leads to prediction intervals.

We will focus on the prediction interval for predicting future observations from a normal population $\mathcal{N}(\mu,\sigma^2)$ once $X_1,\ldots,X_n$ have been observed. Any future observation will be denoted by X_{n+1}.

Consider $\overline{X}$ and X_{n+1}. These two random variables are independent and their difference has a normal distribution,

$$\overline{X} - X_{n+1} \sim \mathcal{N}(0,\sigma^2/n + \sigma^2),$$

thus, $Z = \frac{\overline{X} - X_{n+1}}{\sigma\sqrt{1+1/n}}$ has a standard normal distribution. This leads to $(1-\alpha)100\%$ prediction intervals for X_{n+1}:

$$\left[\overline{X} - t_{n-1,1-\alpha/2}\ s\ \sqrt{1+\frac{1}{n}},\ \overline{X} + t_{n-1,1-\alpha/2}\ s\ \sqrt{1+\frac{1}{n}}\right].$$

When σ^2 is known, s is replaced by σ and $t_{n-1,1-\alpha/2}$ by $z_{1-\alpha/2}$.

Note that prediction intervals contain the factor $\sqrt{1+\frac{1}{n}}$ in place of $\sqrt{\frac{1}{n}}$ in matching confidence intervals for the normal population mean. When n is large, the prediction interval can be substantially larger than the confidence interval. This is because the uncertainty about a future observation consists of (1) uncertainty about its mean and (2) uncertainty about the individual response.

Prediction intervals based on a random sample were used to predict the value of a future observation from the sampled population. In practice, the interest may be in the characteristics of a majority of the units in the population rather than a single unit or the overall mean. For example, a manufacturer of medical devices might want to learn the proportion of production for which a particular dimension falls within a given range.

Tolerance intervals (TI) are used for this purpose. A tolerance interval is constructed so that it would contain at least a specified proportion, say, $1-\gamma$ of the population with a specified confidence, say, $1-\alpha$. Such an interval is usually referred to as the $1-\gamma$ content – $1-\alpha$ coverage TI, or simply a $(1-\gamma, 1-\alpha)$ TI. The ends of a tolerance interval are called tolerance limits.

For normal populations, the two-sided interval is defined as

$$[\overline{X}-ks, \overline{X}+ks], \qquad k=\sqrt{\frac{(n^2-1)\, z^2_{1-\gamma/2}}{n\, \chi^2_{n-1,\alpha}}} \tag{7.11}$$

and interpreted as follows: With a confidence of $1-\alpha$, the proportion $1-\gamma$ of population measurements will fall between the lower and upper bounds in (7.11). The interval in (7.11) is called a $(1-\gamma, 1-\alpha)$-tolerance interval.

Example 7.12. **(0.95, 0.99)-Tolerance Interval.** For sample size $n=20$, $\overline{X}=12$, $s=0.1$, confidence $1-\alpha=99\%$, and proportion $1-\gamma=95\%$, the tolerance factor k is calculated using the following MATLAB script:

```
n=20;
z = norminv(1-0.05/2)  %proportion of 1-0.05=0.95
  %z =  1.9600
xi = chi2inv(0.01, n-1)  %confidence 1-0.01=0.99
  %xi = 7.6327
k = sqrt( (n^2-1) * z^2/(n * xi) )
  %k = 3.1687
[12-k*0.1   12+k*0.1]
  %11.6831   12.3169
```

and the $(0.95, 0.99)$-tolerance interval is $[11.6831, 12.3169]$.
✎

For an example of a tolerance interval for a binomial X, see Exercise 7.37.

Example 7.13. **Distribution-Free Tolerance Intervals.** When the distribution of observations $X_1, \ldots, X_n$ is arbitrary, but continuous, and n is the smallest integer satisfying

$$(1-\gamma/2)^n - \frac{1}{2}(1-\gamma)^n \leq \frac{\alpha}{2},$$

then the full range $(X_{(1)}, X_{(n)})$ is a $(1-\gamma, 1-\alpha)$ tolerance interval. For example, for a $(0.95, 0.95)$ tolerance interval in MATLAB, n can be found as

```
beta=0.05; alpha=0.05;
fzero(@(n) (1-beta/2)^n - 1/2*(1-beta)^n - alpha/2, 100)
%145.2464
```

This means that $(X_{(1)}, X_{(146)})$ is a $(0.95, 0.95)$ distribution-free tolerance interval.
✎

7.7 Confidence Intervals for Quantiles*

The confidence interval for a normal quantile is based on a noncentral t-distribution. Let $X_1, \ldots, X_n$ be a sample of size n with the sample mean $\overline{X}$ and sample standard deviation s.

we want to find a confidence interval on the population's pth quantile, $\mu + z_p \times \sigma$, with a confidence level of $1-\alpha$.

The confidence interval is given by $[L, U]$, where

$$L = \overline{X} + s \cdot nct(\alpha/2, n-1, \sqrt{n} \cdot z_p)/\sqrt{n},$$
$$U = \overline{X} + s \cdot nct(1-\alpha/2, n-1\sqrt{n} \cdot z_p)/\sqrt{n},$$

and $nct(q, df, nc)$ is the q-quantile of the noncentral t-distribution (page 264) with df degrees of freedom and noncentrality parameter nc.

The confidence intervals for quantiles can be based on order statistics when normality is not assumed. For example, instead of declaring the confidence interval for the mean, we should report the confidence interval on the median if the normality of the data is a concern. Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the order statistics of the sample. Then a $(1-\alpha)100\%$ confidence interval for the median Me is

$$X_{(h)} \le Me \le X_{(n-h+1)}.$$

The value of h is usually tabulated. For large n ($n > 40$), a good approximation for h is an integer part of

$$\frac{n - z_{1-\alpha/2}\sqrt{n} - 1}{2}.$$

For example, if $n = 300$, the 95% confidence interval for the median is $[X_{(132)}, X_{(169)}]$ as demonstrated below:

```
n = 300;
h = floor( (n - 1.96 * sqrt(n) - 1)/2 )
% h = 132
n - h + 1
% ans =  169
```

7.8 Confidence Intervals for the Poisson Rate*

Recall that an observation X coming from $\mathcal{P}oi(\lambda)$ has both a mean and a variance equal to the rate parameter, $\mathbb{E}X = \mathbb{V}\mathrm{ar}\, X = \lambda$. Also, Poisson random variables are additive in the rate parameter:

$$X_1, X_2, \ldots, X_n \sim \mathcal{P}oi(\lambda) \qquad \Rightarrow \qquad n\overline{X} = \sum_{i=1}^{n} X_i \sim \mathcal{P}oi(n\lambda). \tag{7.12}$$

The asymptotically shortest Wald-type $(1-\alpha)100\%$ interval for λ is obtained using the fact that $Z = \sqrt{\frac{n}{\lambda}}\,(\overline{X} - \lambda)$ is approximately the standard normal. The inequality

$$\sqrt{\frac{n}{\lambda}}\,|\overline{X} - \lambda| \;\le\; z_{1-\alpha/2}$$

leads to

$$\lambda^2 - \lambda\left(2\overline{X} + \frac{z^2_{1-\alpha/2}}{n}\right) + (\overline{X})^2 \;\le\; 0,$$

which solves to

$$\left[\overline{X}+\frac{z_{1-\alpha/2}^2}{2n}-z_{1-\alpha/2}\sqrt{\frac{\overline{X}}{n}+\frac{z_{1-\alpha/2}^2}{4n^2}},\right.$$
$$\left.\overline{X}+\frac{z_{1-\alpha/2}^2}{2n}+z_{1-\alpha/2}\sqrt{\frac{\overline{X}}{n}+\frac{z_{1-\alpha/2}^2}{4n^2}}\right]. \quad (7.13)$$

Other Wald-type intervals are derived from the fact that $\frac{\sqrt{\overline{X}}-\sqrt{\lambda}}{\sqrt{1/(4n)}}$ is approximately the standard normal. Then, the variance-stabilizing, modified variance-stabilizing, and recentered variance-stabilizing $(1-\alpha)100\%$ confidence intervals are given as

$$\left[\overline{X}-z_{1-\alpha/2}\sqrt{\frac{\overline{X}}{n}},\quad \overline{X}+z_{1-\alpha/2}\sqrt{\frac{\overline{X}}{n}}\right],$$
$$\left[\overline{X}+\frac{z_{1-\alpha/2}^2}{4n}-z_{1-\alpha/2}\sqrt{\frac{\overline{X}}{n}},\quad \overline{X}+\frac{z_{1-\alpha/2}^2}{4n}+z_{1-\alpha/2}\sqrt{\frac{\overline{X}}{n}}\right], \text{ and}$$
$$\left[\overline{X}+\frac{z_{1-\alpha/2}^2}{4n}-z_{1-\alpha/2}\sqrt{\frac{\overline{X}+3/8}{n}},\quad \overline{X}+\frac{z_{1-\alpha/2}^2}{4n}+z_{1-\alpha/2}\sqrt{\frac{\overline{X}+3/8}{n}}\right].$$

Details can be found in Barker (2002).

An alternative approach is based on the link between Poisson and chi-square distributions. Namely, if $X\sim\mathcal{P}oi(\lambda)$, then

$$\mathbb{P}(X>x)=\mathbb{P}(Y<2\lambda),\ \text{for } Y\sim\chi_{2x}^2$$

and the $(1-\alpha)100\%$ confidence interval for λ when X is observed is

$$\left[\frac{1}{2}\,\chi_{2X,\alpha/2}^2,\ \frac{1}{2}\,\chi_{2(X+1),1-\alpha/2}^2\right],$$

where $\chi_{2X,\alpha/2}^2$ and $\chi_{2(X+1),1-\alpha/2}^2$ are $\alpha/2$ and $1-\alpha/2$ quantiles of the χ^2-distribution with $2X$ and $2(X+1)$ degrees of freedom, respectively. By convention, $\chi_{0,\alpha}^2=0$. Due to the additivity property (7.12), the confidence interval changes slightly for the case of an observed sample of size n, $X_1,X_2,\ldots,X_n$. One finds $S=\sum_{i=1}^n X_i$, which is a Poisson with parameter $n\lambda$ and proceeds as in the single-observation case. Because the interval obtained is for $n\lambda$, the bounds should be divided by n to get the interval for λ:

$$\left[\frac{1}{2n}\,\chi^2_{2S,\alpha/2},\ \frac{1}{2n}\,\chi^2_{2(S+1),1-\alpha/2}\right]. \tag{7.14}$$

The interval in (7.14) is sometimes referred to as Garwood's interval (Garwood, 1936).

Example 7.14. **Counts of α-Particles.** Rutherford et al. (1930, pp. 171–172) provide descriptions and data on an experiment by Rutherford and Geiger (1910) on the collision of α-particles emitted from a small bar of polonium with a small screen placed at a short distance from the bar. The number of such collisions in each of 2,608 eight-minute intervals was recorded. The distance between the bar and screen was gradually decreased so as to compensate for the decay of radioactive substance.

X	0	1	2	3	4	5	6	7	8	9	10	11	≥ 12
Freq	57	203	383	525	532	408	273	139	45	27	10	4	2

It is postulated that because of the large number of atoms in the bar and the small probability of any of them emitting a particle, the observed frequencies should be well modeled by a Poisson distribution.

```
%Rutherford.m
X=[ 0  1  2  3  4  5  6  7  8  9  10  11  12 ];
fr=[ 57 203  383  525  532  408  273  139  45  27  10  4  2];
n = sum(fr); %number of experiments//time intervals
rfr = fr./n;       %relative frequencies %n=2608
xbar = X * rfr' ;  %lambdahat = xbar =  3.8704
tc = X * fr';      %total number of counts  tc = 10094
 %Recentered Variance Stabilizing
[xbar + (norminv(0.975))^2/(4*n) - ...
         norminv(0.975) * sqrt(( xbar + 3/8)/n )...
 xbar + (norminv(0.975))^2/(4*n)  + ...
         norminv(0.975) * sqrt( (xbar+ 3/8)/n )]
                  %   3.7917    3.9498
 % Garwood's interval
[1/(2 *n) * chi2inv(0.025, 2 * tc)  ...
       1/(2 * n) * chi2inv(0.975, 2*(tc + 1))]
       %  3.7953    3.9467
```

The estimator for λ is $\hat{\lambda} = \overline{X} = 3.8704$, the Wald-type recentered variance stabilizing interval is $[3.7917, 3.9498]$, and the Garwood confidence interval is $[3.7953, 3.9467]$. The intervals are very close to each other and quite tight due to the large sample size.

7.9 Exercises

7.1. **Tricky Estimation.** A publisher gives the proofs of a new book to two different proofreaders, who read it separately and independently. The first proofreader found 60 misprints, the second proofreader found 70 misprints, and 50 misprints were found by both. Estimate how many misprints remain undetected in the book? *Hint:* Refer to Example 5.10.

7.2. **Laplace's Rule of Succession.** Laplace's Rule of Succession states that if an event appeared X times out of n trials, the probability that it will appear in a future trial is $\frac{X+1}{n+2}$.
(a) If $\frac{X+1}{n+2}$ is taken as an estimator for binomial p, compare the MSE of this estimator with the MSE of the traditional estimator, $\hat{p} = \frac{X}{n}$.
(b) Represent MSE from (a) as the sum of the estimator's variance and the bias squared.

7.3. **Neurons Fire in Potter's Lab.** The data set  `neuronfires.mat` was compiled in Professor Steve Potter's lab at Georgia Tech. It consists of 989 firing times of a cell culture of neurons. The recorded firing times are time instances when a neuron sent a signal to another linked neuron (a spike). The cells from the cortex of an embryonic rat brain were cultured for 18 days on multielectrode arrays. The measurements were taken while the culture was stimulated at a rate of 1 Hz. It was postulated that firing times form a Poisson process; thus, the interspike intervals should have an exponential distribution.
(a) Calculate the interspike intervals T using MATLAB's `diff` command. Check the histogram for T and discuss its resemblance to the exponential density. By the moment-matching estimator, argue that exponential parameter λ is close to 1.
(b) According to (a), the model for interspike intervals is $T \sim \mathcal{E}(1)$. You are interested in the proportion of intervals that are shorter than 3, $T \leq 3$. Find this proportion from the theoretical model $\mathcal{E}(1)$ and compare it to the estimate from the data. For the theoretical model, use `expcdf` and for empirical data use `sum(T <= 3)/length(T)`.

7.4. **Moment Matching Uniform.** Let $X_1, X_2, \ldots, X_n$ be a sample from uniform $\mathcal{U}(\mu - \delta, \mu + \delta)$ distribution. Show that the moment-matching estimators of μ and δ are

$$\hat{\mu} = \overline{X} \qquad \text{and} \qquad \hat{\delta} = \sqrt{3\left(\overline{X^2} - (\overline{X})^2\right)}\,,$$

where $\overline{X}$ and $\overline{X^2}$ are first and second sample moments.

7.5. **The MLE in a Discrete Case.** A sample $-1,1,1,0,-1,1,1,1,0,1,1,0,-1,1,1$ was observed from a population with a probability mass function of

X	–1	0	1
Prob	θ	2θ	$1-3\theta$

(a) What is the possible range for θ?
(b) What is the MLE for θ?
(c) How would the MLE look for a sample of size n?

7.6. **Two Thetas.** (a) A sample of size $n = 10$,

$$0,0,1,1,1,2,1,1,0, \text{ and } 2,$$

is obtained from a partially specified discrete distribution

X	0	1	2
Prob	θ	$1/2$	$1/2-\theta$

How would you estimate θ given the sample?

(b) A sample of size $n = 4$,

$$1.1,\ 0.7,\ 0.5, \text{ and } 1.7,$$

is obtained from normal $\mathcal{N}(\theta, 0.6^2)$ distribution. As a confidence interval (CI) for θ, the interval $[0,2]$ is proposed. With what confidence does this interval contain θ?

7.7. **MLE for Two Continuous Distributions.** Find the MLE for parameter θ if the model for observations $X_1, X_2, \ldots, X_n$, is

$$\text{(a)} \quad f(x|\theta) = \frac{\theta}{x^2},\ 0 < \theta \le x;$$
$$\text{(b)} \quad f(x|\theta) = \frac{\theta - 1}{x^\theta},\ x \ge 1,\ \theta > 1.$$

7.8. **Match the Moment.** The geometric distribution (X is the number of failures before the first success) has a probability mass function

$$f(x|p) = (1-p)^x\ p,\ \ x = 0,1,2,\ldots.$$

Suppose $X_1, X_2, \ldots, X_n$ are observations from this distribution. It is known that $\mathbb{E}X_i = \frac{1-p}{p}$.
(a) What would you report as the moment-matching estimator if the sample $X_1 = 2, X_2 = 6, X_3 = 1$ were observed?
(b) What is the MLE for p?

7.9. **Weibull Distribution.** The two-parameter Weibull distribution is given by the density

$$f(x) = a\lambda^a x^{a-1} e^{-(\lambda x)^a}, \qquad a > 0, \lambda > 0, x \geq 0,$$

with mean and variance

$$\mathbb{E}X = \frac{\Gamma(1+1/a)}{\lambda}, \quad \text{and} \quad \mathbb{V}\text{ar}\, X = \frac{1}{\lambda^2}\left[\Gamma(1+2/a) - \Gamma(1+1/a)^2\right].$$

Assume that the "shape" parameter a is known and equal to $1/2$.
(a) Propose two moment-matching estimators for λ.
(b) If $X_1 = 1, X_2 = 3, X_3 = 2$, what are the values of the estimator?

Hint: Recall that $\Gamma(n) = (n-1)!$

7.10. **Rate Parameter of Gamma.** Let $X_1, \ldots, X_n$ be a sample from a gamma distribution given by the density

$$f(x) = \frac{\lambda^a x^{a-1}}{\Gamma(a)}\, e^{-\lambda x}, \qquad a > 0, \lambda > 0, x \geq 0,$$

where shape parameter a is known and rate parameter λ is unknown and of interest.
(a) Find the MLE of λ.
(b) Using the fact that $X_1 + X_2 + \cdots + X_n$ is also gamma distributed with parameters na and λ, find the expected value of the MLE from (a) and show that it is a biased estimator of λ.
(c) Modify the MLE so that it is unbiased. Compare MSEs for the MLE and the modified estimator.

7.11. **Estimating the Parameter of a Rayleigh Distribution.** If two random variables X and Y are independent of each other and normally distributed with variances equal to σ^2, then the variable $R = \sqrt{X^2 + Y^2}$ follows the Rayleigh distribution with scale parameter σ. An example of such a variable would be the distance of darts from the target in a dart-throwing game where the deviations in the two dimensions of the target plane are independent and normally distributed. The Rayleigh random variable R has a density

$$f(r) = \frac{r}{\sigma^2} \exp\left\{-\frac{r^2}{2\sigma^2}\right\}, \quad r \geq 0,$$

$$\mathbb{E}R = \sigma\sqrt{\frac{\pi}{2}} \qquad \mathbb{E}R^2 = 2\sigma^2.$$

(a) Find the two moment-matching estimators of σ.
(b) Find the MLE of σ.
(c) Assume that $R_1 = 3, R_2 = 4, R_3 = 2$, and $R_4 = 5$ are Rayleigh-distributed random observations representing the distance of a dart from the center. Estimate the variance of the horizontal error, which is theoretically a zero-mean normal.
(d) In Example 5.29, the distribution of a square root of an exponential random variable with a rate parameter λ was Rayleigh with the following density:

$$f(r) = 2\lambda r \exp\{-\lambda r^2\}.$$

To find the MLE for λ, can you use the MLE for σ from (b)?

7.12. **Monocytes among Blood Cells.** Eisenhart and Wilson (1943) report the number of monocytes in 100 blood cells of a cow in 113 successive weeks.

Monocytes	Frequency	Monocytes	Frequency
0	0	7	12
1	3	8	10
2	5	9	11
3	13	10	7
4	19	11	3
5	13	12	2
6	15	13+	0

(a) If the underlying model is Poisson, what is the estimator of λ?
(b) If the underlying model is binomial $\mathcal{B}in(100, p)$, what is the estimator of p?
(c) For the models specified in (a) and (b) find theoretical or "expected" frequencies.
Hint: Suppose the model predicts $\mathbb{P}(X = k) = p_k,\ k = 0,1,\ldots,13$. The expected frequency of $X = k$ is $113 \times p_k$. For a follow-up see Exercise 17.7.

7.13. **Estimation of θ in $\mathcal{U}(0,\theta)$.** Which of the two estimators in Example 7.5 is unbiased? Find the MSE of both estimators. Which one has a smaller MSE?

7.14. **Estimating the Rate Parameter in a Double Exponential Distribution.** Let $X_1, \ldots, X_n$ follow double exponential distribution with density

$$f(x|\theta) = \frac{\theta}{2} e^{-\theta|x|},\ -\infty < x < \infty,\ \theta > 0.$$

For this distribution, $\mathbb{E}X = 0$ and $\mathbb{V}\mathrm{ar}\,(X) = \mathbb{E}X^2 = 2/\theta^2$. The double exponential distribution, also known as Laplace's distribution, is a model frequently encountered in statistics, see page 207.

(a) Find a moment-matching estimator for θ.
(b) Find the MLE of θ.
(c) Evaluate the two estimators from (a) and (b) for a sample $X_1 = -2, X_2 = 3, X_3 = 2$, and $X_4 = -1$.

7.15. **Reaction Times I.** A sample of 20 students is randomly selected and given a test to determine their reaction time in response to a given stimulus. Assume that individual reaction times are normally distributed. If the mean reaction time is determined to be $\overline{X} = 0.9$ (in seconds) and the standard deviation is $s = 0.12$, find the confidence intervals:
(a) 95% CI for the unknown population mean μ.
(b) 98.5% CI interval for the unknown population mean μ.
(c) 95% CI for the unknown population variance σ^2.

7.16. **Reaction Times II.** Under the conditions in the previous problem, assume that the population standard deviation was known to be $\sigma = 0.12$.
(a) Find the 98.5% CI for the unknown mean μ;
(b) Find the sample size necessary to produce a 95% CI for μ of length 0.07.

7.17. **Toxins.** An investigation on toxins produced by molds that infect corn crops was performed. A biochemist prepared extracts of the mold culture with organic solvents and then measured the amount of toxic substance per gram of solution. From 11 preparations of the mold culture, the following measurements of the toxic substance (in milligrams) were obtained: 3, 2, 5, 3, 2, 6, 5, 4.5, 3, 3, and 4.
Compute a 99% confidence interval for the mean weight of toxic substance per gram of mold culture. State the assumption you make about the population.

7.18. **Bias of s_*^2.** For a sample $X_1, \ldots, X_n$ from a $\mathcal{N}(\mu, \sigma^2)$ population, find the bias of $s_*^2 = \frac{1}{n}\sum_i (X_i - \overline{X})^2$ as an estimator of variance σ^2.
Using (7.3), show that the variance of s_*^2 is smaller than the variance of unbiased estimator s^2.

7.19. **COPD Patients.** Acute exacerbations of disease symptoms in patients with chronic obstructive pulmonary disease (COPD) often lead to hospitalizations and impose a great financial burdens on healthcare systems. A study by Ghanei et al. (2007) aimed to determine factors that may predict rehospitalization in COPD patients.
A total of 157 COPD patients were randomly selected from all COPD patients admitted to the chest clinic of Baqiyatallah Hospital during the year 2006. Subjects were followed for 12 months to observe the occurrence of any disease exacerbation that might lead to hospitalization. Over the 12-month period, 87 patients experienced disease exacerbation. The authors found significant associations between COPD exacerbation and monthly income, comorbidity score, and depression using logistic

regression tools. We are not interested in these associations in this exercise, but we are interested in the population proportion of all COPD patients that experienced disease exacerbation over a 12-month period, p.
(a) Find an estimator of p based on the data available. What is an approximate distribution of this estimator?
(b) Find the 90% confidence interval for the unknown proportion p.
(c) How many patients should be sampled and monitored so that the 90% confidence interval as in (b) does not exceed 0.03 in length.
(d) The hospital challenges the claim by the local health system authorities that half of the COPD patients experience disease exacerbation in a 1-year period, claiming that the proportion is significantly higher. Can the hospital support their claim based on the data available? Use $\alpha = 0.05$. Would you reverse the decision if α were changed to 10%?

7.20. **Right to Die.** A Gallup Poll estimated the support among Americans for "right to die" laws. In the survey, 1528 adults were asked whether they favor voluntary withholding of life-support systems from the terminally ill. The results: 1238 said yes.
(a) Find the 99% confidence interval for the percentage of all adult Americans who are in favor of "right to die" laws.
(b) If the margin of error (half of the length of a 95% confidence interval, see page 310) is to be smaller than 0.01, what sample size is needed to achieve this requirement? Assume $\hat{p} = 0.8$.

7.21. **Exponentials Parameterized by the Scale.** A sample $X_1, \ldots, X_n$ was selected from a population that has an exponential $\mathcal{E}(\lambda)$ distribution with a density of $f(x|\lambda) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}}$, $x \geq 0, \lambda > 0$. We are interested in estimating the parameter λ.
(a) What are the moment-matching and MLE estimators of λ based on $X_1, \ldots, X_n$?
(b) Two independent observations $Y_1 \sim \mathcal{E}(\lambda/2)$ and $Y_2 \sim \mathcal{E}(2\lambda)$ are available. Combine them (make a specific linear combination) to obtain an unbiased estimator of λ. What is the variance of the proposed estimator?
(c) Two independent observations $Z_1 \sim \mathcal{E}(1.1\lambda)$ and $Z_2 \sim \mathcal{E}(0.9\lambda)$ are available. An estimator of λ in the form $\hat{\lambda} = pZ_1 + (1-p)Z_2$, $0 \leq p \leq 1$ is proposed. What p minimizes the magnitude of bias of $\hat{\lambda}$? What p minimizes the variance of $\hat{\lambda}$?

7.22. **Bias in Estimator for Exponential λ Distribution.** If the exponential distribution is parameterized with λ as the scale parameter, $f(x|\lambda) = \frac{1}{\lambda} \exp\{-x/\lambda\}$, $x \geq 0, \lambda > 0$, (as in MATLAB), then $\hat{\lambda} = \overline{X}$ is an unbiased estimator of λ. However, if it is parameterized with λ as a rate parameter, $f(x|\lambda) = \lambda \exp\{-\lambda x\}$, $x \geq 0, \lambda > 0$, then $\hat{\lambda} = 1/\overline{X}$ is biased. Find the

bias of this estimator. *Hint:* Argue that $1/\sum_{i=1}^{n} X_i$ has an inverse gamma distribution with parameters n and λ and take the expectation.

7.23. **Yucatan Miniature Pigs.** Ten adult male Yucatan miniature pigs were exposed to various durations of constant light ("Lighting"), then sacrificed after experimentally controlled time delay ("Survival"), as described in Dureau et al. (1996). Following the experimental protocol, entire eyes were fixed in Bouin's fixative for 3 days. The anterior segment (cornea, iris, lens, ciliary body) was then removed and the posterior segment divided into five regions: posterior pole (including optic nerve head and macula) ("P"), nasal ("N"), temporal ("T"), superior ("S"), and inferior ("I"). Specimens were washed for 2 days, embedded in paraffin, and subjected to microtomy perpendicular to the retinal surface. Every 200 μm, a 10-μm-thick section was selected, and 20 sections were kept for each retinal region. Sections were stained with hematoxylin. The outer nuclear layer (ONL) thickness was measured by an image-analyzing system (Biocom, Les Ulis, France), and three measures were performed for each section at regularly spaced intervals so that 60 measures were made for each retinal region. The experimental protocol for 11 animals was as follows (Lighting and Survival times are in weeks):

Animal	Lighting duration	Survival time
Control	0	0
1	1	12
2	2	10
3	4	0
4	4	4
5	4	6
6	8	0
7	8	4
8	8	8
9	12	0
10	12	4

The data set  `pigs.mat` contains the data structure pigs with `pigs.pc, pigs.p1,...,pigs.p10`, representing the posterior pole measurements for the 11 animals. This data set and complete data `yucatanpigs.dat` can be found on the book's website page.

Observe the data `pigs.pc` and argue that it deviates from normality by using MATLAB's `qqplot`. Transform pigs.pc as `x = (pigs.pc - 14)/(33 -14)`, to confine x between 0 and 1 and assume a beta $\mathcal{B}e(a,a)$ distribution. The MLE for a is complex (involves a numerical solution of equations with digamma functions), but the moment-matching estimator is straightforward.

Find a moment-matching estimator for a.

7.24. **Computer Games.** According to Hamilton (1990), certain computer games are thought to improve spatial skills. A mental rotations test, measuring spatial skills, was administered to a sample of school children after they had played one of two types of computer game. Construct 95% confidence intervals based on the following mean scores, assuming that the children were selected randomly and that the mental rotations test scores had a normal distribution in the population.
(a) After playing the "Factory" computer game: $\overline{X} = 22.47, s = 9.44, n = 19$.
(b) After playing the "Stellar" computer game: $\overline{X} = 22.68, s = 8.37, n = 19$.
(c) After playing no computer game (control group): $\overline{X} = 18.63, s = 11.13, n = 19$.

7.25. **Effectiveness in Treating Cerebral Vasospasm.** In a study on the effectiveness of hyperdynamic therapy in treating cerebral vasospasm, Pritz et al. (1996) reported on the therapy where success was defined as clinical improvement in terms of neurological deficits. The study reported 16 successes in 17 patients.
(a) Using the methods discussed in the text, find 95% confidence intervals for the success rate.
(b) Does any of the methods produce an upper bound larger than 1?
(c) How would you find the 95% confidence interval if the study reported 17 successes in 17 patients?

7.26. **Alcoholism and the Blyth–Still Confidence Interval.** Genetic markers were observed for a group of 50 Caucasian alcoholics in a study that aimed at determining whether alcoholism has (in part) a genetic basis. The antigen (marker) B15 was present in 5 alcoholics. Find the Blyth–Still 99% confidence interval for the proportion of Caucasian alcoholics having this antigen.

If either p or q is close to 0, then a precise $(1-\alpha)100\%$ confidence interval for the unknown proportion p was proposed by Blyth and Still (1983). For $X \sim \mathcal{B}in(n,p)$,

$$\left[\frac{(X-0.5) + \frac{z^2_{1-\alpha/2}}{2} - z_{1-\alpha/2}\sqrt{(X-0.5) - \frac{(X-0.5)^2}{n} + \frac{z^2_{1-\alpha/2}}{4}}}{n + z^2_{1-\alpha/2}}, \right.$$
$$\left. \frac{(X+0.5) + \frac{z^2_{1-\alpha/2}}{2} + z_{1-\alpha/2}\sqrt{(X+0.5) - \frac{(X+0.5)^2}{n} + \frac{z^2_{1-\alpha/2}}{4}}}{n + z^2_{1-\alpha/2}} \right].$$

7.27. **Spores of *Amanita Phalloides.*** Exercise 2.4 provides measurements in μm of 28 spores of the mushroom *Amanita phalloides*.
Assuming normality of measurements, find the following:
(a) A point estimator for the unknown population variance σ^2. What is the sampling distribution of the point estimator?
(b) A 90% confidence interval for the population variance.
(c) (By MATLAB) the minimal sample size that ensures that the upper bound U of the 90% confidence interval for the variance is at most 30% larger than the lower bound L, that is, $U/L \leq 1.3$.
(d) Miller (1991) showed that the coefficient of variation in a normal sample of size n has an approximately normal distribution:

$$s/\overline{X} \overset{approx}{\sim} \mathcal{N}\left(\frac{\sigma}{\mu}, \frac{1}{n-1}\left(\frac{\sigma}{\mu}\right)^2\left[\frac{1}{2}+\left(\frac{\sigma}{\mu}\right)^2\right]\right).$$

Based on this asymptotic distribution, a $(1-\alpha)100\%$ confidence interval for the population coefficient of variation $\frac{\sigma}{\mu}$ is approximately

$$\left(\frac{s}{\overline{X}} - z_{1-\alpha/2}\frac{s}{\overline{X}}\sqrt{\frac{1}{n-1}\left[\frac{1}{2}+\left(\frac{s}{\overline{X}}\right)^2\right]},\ \frac{s}{\overline{X}} + z_{1-\alpha/2}\frac{s}{\overline{X}}\sqrt{\frac{1}{n-1}\left[\frac{1}{2}+\left(\frac{s}{\overline{X}}\right)^2\right]}\right).$$

This approximation works well if n exceeds 10 and the coefficient of variation is less than 0.7. Find the 95% confidence interval for the population coefficient of variation σ/μ based on 28 spores measurements.
(e) Standardly used $(1-\alpha)100\%$ confidence interval for σ/μ is McKay's interval (McKay. 1932),

$$\left[\frac{s}{\overline{X}}\left[\left(\frac{u_1}{n}-1\right)\left(\frac{s}{\overline{X}}\right)^2+\frac{u_1}{n-1}\right]^{-1/2},\ \frac{s}{\overline{X}}\left[\left(\frac{u_2}{n}-1\right)\left(\frac{s}{\overline{X}}\right)^2+\frac{u_2}{n-1}\right]^{-1/2}\right],$$

where $u_1 = \chi^2_{n-1,1-\alpha/2}$, and $u_2 = \chi^2_{n-1,\alpha/2}$.
Find the McKay's 95% confidence interval for σ/μ.

7.28. **CLT-Based Confidence Interval for Normal Variance.** Refer to Example 7.9. Using MATLAB, simulate a normal sample with mean 0 and variance 1 of size $n = 50$ and find if a 95% confidence interval for the population variance contains a 1 (the true population variance). Check this coverage for a standard confidence interval in (7.8) and for a CLT-based interval from Example 7.9. Repeat this simulation $M = 10000$ times, keeping track of the number of successful coverages. Show that the interval (7.8) achieves the nominal coverage, while the CLT-based interval has a smaller coverage of about 2%. Repeat the simulation for sample sizes of $n = 30$ and $n = 200$.

7.29. **Stent Quality Control.** A stent is a tube or mechanical scaffold used to counteract significant decreases in vessel or duct diameter by acutely propping open the conduit. Stents are often used to alleviate diminished blood flow to organs and extremities beyond an obstruction in order to maintain an adequate delivery of oxygenated blood.
In the production of stents, the quality control procedure aims to identify defects in composition and coating. Precision z-axis measurements (10 nm and greater) are obtained, along with surface roughness and topographic surface finish details, using a laser confocal imaging system (an example is the Olympus LEXT OLS3000). Samples of 50 stents from a production process are selected every hour. Typically, 1% of stents are nonconforming. Let X be the number of stents in the sample of 50 that are nonconforming. A production problem is suspected if X exceeds its mean by more than three standard deviations.
(a) Find the critical value for X that will implicate a production problem.
(b) Find an approximation for the probability that in the next-hour batch of 50 stents, the number X of nonconforming stents will be critical, i.e., will raise suspicion that the process has gone awry.
(c) Suppose now that the population proportion of nonconforming stents, p, is unknown. How would one estimate p by taking a 50-stent sample? Is the proposed estimator unbiased?
(d) Suppose now that a batch of 50 stents produced $X = 1$. Find the 95% confidence interval for p.

7.30. **Clopper–Pearson and $3/n$-Rule Confidence Intervals.** Using MATLAB, compare the performance of Clopper–Pearson and $3/n$-rule confidence intervals when $X = 0$. Use $\alpha = 0.001, 0.005, 0.01, 0.05, 0.1$ and $n = 10 : 10 : 200$. Which interval is superior and under what conditions?

7.31. **Fluid Overload in Hemodialysis.** The overload of fluid volume and hypertension are known to contribute to high cardiovascular morbidity and mortality seen in dialysis patients. The correct assessment of volume status is especially important as only a small increase in extracellular volume over prolonged periods of time can lead to a considerable cardiac strain and, as a consequence, to left ventricular hypertrophy. In clinical practice, volume overload is most often judged by a battery of clinical signs such as edema, dyspnea, hypertension, and coughing. A study by Ribitsch et al. (2012) compares volume overload in stable hemodialysis (HD) patients assessed by standard clinical judgment with data obtained from bioimpedance analysis.
Data set  `hemodialysis.dat|mat|xlsx` provides measurements on 28 HD patients (17 males and 11 females) from the dialysis unit of the University Medical Center Graz. The variables are described in the following table:

Column	Variable	Unit	Description
1	M_0	(kg)	Pre-dialytic body mass
2	BMI	(kg/m^2)	Body mass index
3	P_0	(mmHg)	Pre-dialytic mean arterial pressure
4	P_1	(mmHg)	Post-dialytic mean arterial pressure
5	V_E	(L)	Extracelular volume
6	V_O	(L)	Volume overload
7	V_U	(L)	Delivered ultrafiltration volume
8	B_0	(pg/ml)	Pre-dialytic NT-pro-BNP
9	B_1	(pg/ml)	Post-dialytic NT-pro-BNP
10	S_W		Wizemann's clinical score

(a) Find the 95% CI for the population mean of the difference $D = P_1 - P_0$ in stable hemodialysis patients. Assume that this difference is normally distributed.
(b) Find the 90% CI for the population variance of V_0. Assume normality of V_0.
(c) Find the 99% CI for the population proportion of patients for which $B_1 > B_0$.

7.32. **Sensor Agreement.** A company producing an approved medical sensor A is applying to FDA for the approval of a new sensor B. Both sensors are prone to errors, and a gold standard is absent. The FDA is requesting that the new sensor is comparable to the one currently in use.
Data are

		Sensor A	
		Result +	Result −
Sensor B	Result +	208	22
	Result −	11	5819

Find agreement rate $\hat{p}$, that is, the proportion of cases where the sensors agreed (both positive or both negative).
Calculate the 95% Clopper–Pearson CI for the population agreement rate p, and report the lower bound. To establish equivalence, the FDA requires for this lower bound to be at least 0.98. Is this the case?

7.33. **Seventeen Pairs of Rats, Carbon Tetrachloride, and Vitamin B.** In a widely cited experiment by Sampford and Taylor (1959), 17 pairs of rats were formed by selecting pairs from the same litter. All rats were given carbon tetrachloride, and one rat from each pair was treated with vitamin B_{12}, while the other served as a control. In 7 of 17 pairs, the treated rat outlived the control rat.
(a) Based on this experiment, estimate the population proportion p of pairs in which the treated rat would outlive the control rat.
(b) If the estimated proportion in (a) is the "true" population probability, what is the chance that in an independent replication of this experiment

one will get exactly 7 pairs (out of 17) in which the treated rat outlives the control?
(c) Find the 95% confidence interval for the unknown p. Does the interval contain 1/2? What does $p = 1/2$ mean in the context of this experiment, and what do you conclude from the confidence interval?
Would the conclusion be the same if in 140 out of 340 pairs the treated rat outlived the control?
(d) The length of the 95% confidence interval based on $n = 17$ in (c) may be too large. What sample size (number of rat pairs) is needed so that the 95% confidence interval has a length not exceeding $\ell = 0.2$?

7.34. **Hemocytometer Counts.** A set of 1,600 squares on a hemocytometer is inspected, and the number of cells is counted in each square. The number of squares with a particular count is given in the table below:

Count	0	1	2	3	4	5	6	7
# Squares	5	24	77	139	217	262	251	210
Count	8	9	10	11	12	13	14	15
# Squares	175	108	63	36	20	9	2	1

Assume that the count has a Poisson $\mathcal{P}oi(\lambda)$ distribution.
(a) Find an estimator of λ using method of moments.
(b) Find the 95% CI for λ. Compare solutions obtained by alternative intervals in (7.13) and (7.14).

7.35. **Predicting Alkaline Phosphatase.** Refer to BUPA liver disorder data,  `BUPA.dat|mat|xlsx`. The second column gives measurements of alkaline phosphatase among 345 male individuals affected by liver disorder. If variable X represents the logarithm of this measurement, its distribution is symmetric and bell-shaped, so it can be assumed normal. From the data, $\overline{X} = 4.21$ and $s^2 = 0.0676$.
Suppose that a new patient with liver disorder just checked in. Find the 95% prediction interval for his log-level of alkaline phosphatase in the following cases:
(a) The population variance is known and equal to 1/15.
(b) The population variance is not known.
(c) Compare the interval in (b) with a 95% confidence interval for the population mean. Why is the interval in (b) larger?

7.36. **CNFL for DSP.** Corneal nerve fiber length (CNFL), as measured using corneal confocal microscopy (CCM), can be used to reliably rule diabetic sensorimotor polyneuropathy (DSP) in or out, according to research published online on February 8, 2012, in *Diabetes Care*, `doi:10.2337/dc11-1396`. Part of the reported results can be summarized as follows:

	DSP	No DSP	Total
CNFL ≤ 140 (Positive)	28	20	48
CNFL > 140 (Negative)	5	100	105
Total	33	120	153

Find the 95% CI's for population sensitivity and specificity using:
(a) Wald's interval;
(b) Clopper–Pearson's interval; and
(c) Anscombe's ArcSin interval.
Which one is the shortest?
(d) It is desired to repeat the study and design sample sizes of DSP and control subjects that would lead to Wald-type 95% confidence intervals on sensitivity and specificity not exceeding 0.16 in length each.
Hint. Assume that data from the table can be used in assessing the sensitivity/specificity needed in the expression for sample size. Use the sample size formula in (7.10). Since the sensitivity gives sample size for cases and specificity gives sample size for controls, the total sample size for the new study should be the sum of the two sample sizes found.

7.37. **Tolerance Interval for Binomial X.** A $(1-\gamma, 1-\alpha)$ tolerance interval for binomial $X \sim \mathcal{B}in(n,p)$ is determined in two stages. In stage one, a $(1-\alpha)100\%$ confidence interval on p is found, (p_L, p_U). In stage two, the tolerance bounds are determined via the quantiles of binomial distribution,

$$\left[F^{-1}\left(\frac{\gamma}{2}, n, p_L\right),\ F^{-1}\left(1-\frac{\gamma}{2}, n, p_U\right)\right].$$

Here $F^{-1}(\alpha, n, p)$ is α-quantile of binomial $\mathcal{B}in(n,p)$ distribution, `binoinv(alpha,n,p)`.

In a previous experiment, the number of "successes" was 46 out of 100 trials. What is the (0.95, 0.95) tolerance interval for number of successes in a future experiment with 100 trials?

MATLAB FILES AND DATA SETS USED IN THIS CHAPTER

http://statbook.gatech.edu/Ch7.Estim/

AmanitaCI.m, arcsinint.m, bickellehmann.m, clopperint.m, CLTvarCI.m, confintscatterpillar.m, crouxrouss.m, crouxrouss2.m, cyclosporine.m, dists2.m, estimweibull.m, ginimd.m, ginimd2.m, lfev.m, MaxwellMLE.m, MixtureModelExample.m, muscaria.m, plotlike.m, Rutherford.m, simuCI.m, tolerance.m, waldsimulation.m

amanita28.dat, hemodialysis.dat|xlsx, hypertension.dat, neuronfires.dat|mat|xlsx

CHAPTER REFERENCES

Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *Am. Stat.*, **52**, 119–126.

Barker, L. (2002). A comparison of nine confidence intervals for a Poisson parameter when the expected number of events is ≤ 5. *Am. Stat.*, **56**, 85–89.

Blyth, C. and Still, H. (1983). Binomial confidence intervals. *J. Am. Stat. Assoc.*, **78**, 108–116.

Brynskov, J., Freund, L., Rasmussen, S. N., et al. (1989). A placebo-controlled, double-blind, randomized trial of cyclosporine therapy in active chronic Crohn's disease. *New Engl. J. Med.*, **321**, 13, 845–850.

Brynskov, J., Freund, L., Rasmussen, S. N., et al. (1991). Final report on a placebo-controlled, double-blind, randomized, multicentre trial of cyclosporin treatment in active chronic Crohn's disease. *Scand. J. Gastroenterol.*, **26**, 7, 689–695.

Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, **26**, 404–413.

Croux, C. and Rousseeuw, P. J. (1992). Time-efficient algorithms for two highly robust estimators of scale. *Comput. Stat.*, **1**, 411–428.

Dressel, P. L. (1957). Facts and fancy in assigning grades. *Basic College Quarterly*, **2**, 6–12.

Dureau, P., Jeanny, J.-C., Clerc, B., Dufier, J.-L., and Courtois, Y. (1996). Long term light-induced retinal degeneration in the miniature pig. *Mol. Vis.* **2**, 7. http://www.emory.edu/molvis/v2/dureau.

Eisenhart, C. and Wilson, P. W. (1943). Statistical method and control in bacteriology. *Bact. Rev.*, **7**, 57–137,

Garwood, F. (1936). Fiducial limits for the Poisson distribution. *Biometrika*, **28**, 437–442.

Hamilton, L. C. (1990). *Modern Data Analysis: A First Course in Applied Statistics*. Brooks/Cole, Pacific Grove.

Hogg, R. V. and Tanis, E. A. (2001). *Probability and Statistical Inference*, 6th edn. Prentice-Hall, Upper Saddle River.

McKay, A. T. (1932). Distribution of the coefficient of variation and the extended *t*-distribution. *J. Roy. Statist. Soc.*, **95**, 695–698.

Miller, E. G. (1991). Asymptotic test statistics for coefficient of variation. *Comm. Stat. Theory Meth.*, **20**, 10, 3351–3363.

Pritz, M. B., Zhou, X. H., and Brizendine, E. J. (1996). Hyperdynamic therapy for cerebral vasospasm: a meta-analysis of 14 studies. *J. Neurovasc. Dis.*, **1**, 6–8.

Ribitsch, W., Stockinger, J., and Schneditz, D. (2012). Bioimpedance-based volume at clinical target weight is contracted in hemodialysis patients with a high body mass index. *Clinical Nephrology*, **77**, 5, 376–382.

Rutherford, E., Chadwick, J., and Ellis, C. D. (1930). *Radiations from Radioactive Substances*. Macmillan, London, pp. 171–172.

Rutherford, E. and Geiger, H. (1910). The probability variations in the distribution of α-particles (with a note by H. Bateman). *Philos. Mag.*, **6**, 20, 697–707.

Sampford, M. R. and Taylor, J. (1959). Censored observations in randomized block experiments. *J. Roy. Stat. Soc. Ser. B*, **21**, 214–237.

Staudte, R. G. and Sheater, S. J. (1990). *Robust Estimation and Testing*. Wiley, New York.

Wald, A. and Wolfowitz, J. (1939). Confidence limits for continuous distribution functions. *Ann. Math. Stat.*, **10**, 105–118.

Wilcox, R. R. (2005). *Introduction to Robust Estimation and Hypothesis Testing*, 2nd ed. Academic Press, San Diego.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.*, **22**, 209–212.

Chapter 8
Bayesian Approach to Inference

In 1954 I proved that the only sound methods were Bayesian; yet you continue to use non-Bayesian ideas without pointing out a flaw in either my premise or my proof, why?

– Leonard Jimmie Savage

WHAT IS COVERED IN THIS CHAPTER

- Bayesian Paradigm
- Likelihood, Prior, Marginal, Posterior, Predictive Distributions
- Conjugate Priors, Prior Elicitation
- Bayesian Computation
- Estimation, Credible Sets, Testing, Bayes Factor, Prediction

8.1 Introduction

Several paradigms provide a basis for statistical inference; the two most dominant are the *frequentist* (sometimes called classical, traditional, or Neyman–Pearsonian) and *Bayesian*. The term Bayesian refers to Reverend Thomas Bayes, a nonconformist minister interested in mathematics whose posthumously published essay (Bayes, 1763) is fundamental for this kind of inference. According to the Bayesian paradigm, the unobservable parameters in a statistical model are treated as random. Before data are collected, *prior distributions* are elicited to quantify our knowledge about the param-

eters. This knowledge comes from expert opinion, theoretical considerations, or previous similar experiments. When data are available, the prior distributions are updated to the *posterior distributions.* These are conditional distributions that incorporate the observed data. The transition from the prior to the posterior is possible via Bayes' theorem.

The Bayesian approach is relatively modern in statistics; it became influential with advances in Bayesian computational methods in the 1980s and 1990s.

Before launching into a formal exposition of Bayes' theorem, we revisit Bayes' rule for events (page 100). Prior to observing whether an event A has appeared or not, we set the probabilities of n hypotheses, $H_1, H_2, \ldots, H_n$, under which event A may appear. We called them *prior* probabilities of the hypotheses, $\mathbb{P}(H_1), \ldots, \mathbb{P}(H_n)$. Bayes' rule showed us how to update these prior probabilities to the posterior probabilities once we obtained information about event A. Recall that the posterior probability of the hypothesis H_i, given the evidence about A, was

$$\mathbb{P}(H_i|A) = \frac{\mathbb{P}(A|H_i)\mathbb{P}(H_i)}{\mathbb{P}(A)}.$$

Therefore, Bayes' rule gives a recipe for updating the prior probabilities of events to their posterior probabilities once additional information from the experiment becomes available. The focus of this chapter is on how to update prior knowledge about a model; however, this knowledge, or lack of it, is expressed in terms of probability distributions rather than by events.

Suppose that before the data are observed, a description of the population parameter θ is given by a probability density $\pi(\theta)$. The process of specifying the prior distribution is called *prior elicitation.* The data are modeled via the likelihood, which depends on θ and is denoted by $f(x|\theta)$. Bayes' theorem updates the prior $\pi(\theta)$ to the posterior $\pi(\theta|x)$ by incorporating observations x summarized via the likelihood:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)}. \tag{8.1}$$

Here, $m(x)$ normalizes the product $f(x|\theta)\pi(\theta)$ to be a density and is a constant once the prior is specified and the data are observed. Given the data x and the prior distribution, the posterior distribution $\pi(\theta|x)$ summarizes all available information about θ.

Although the equation in (8.1) is referred to as a theorem, there is nothing to prove there. Recall that the probability of intersection of two events A and B was calculated as $\mathbb{P}(AB) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(B)$ [multiplication rule in (3.6)]. By analogy, the joint distribution of X and θ, $h(x,\theta)$,

would have two representations, as in (5.11), depending on the order of conditioning:

$$h(x,\theta) = f(x|\theta)\pi(\theta) = \pi(\theta|x)m(x),$$

and Bayes' theorem solves this equation with respect to the posterior $\pi(\theta|x)$.

To summarize, Bayes' rule updates the probabilities of events when new evidence becomes available, while Bayes' theorem provides the recipe for updating prior distributions of model's parameters once experimental observations become available.

$\mathbb{P}(\text{hypothesis})$	$\xrightarrow{\text{BAYES' RULE}}$	$\mathbb{P}(\text{hypothesis}\|\text{evidence})$
$\pi(\theta)$	$\xrightarrow{\text{BAYES' THEOREM}}$	$\pi(\theta\|\text{data})$

The Bayesian paradigm has many advantages, but the two most important are: (i) the uncertainty is expressed via the probability distribution and the statistical inference can be automated; thus, it follows a conceptually simple recipe embodied in Bayes' theorem, and (ii) available prior information is coherently incorporated into the statistical model describing the data.

The FDA guidelines document (FDA, 2010) recommends the use of a Bayesian methodology in the design and analysis of clinical trials for medical devices. This document eloquently outlines the reasons why a Bayesian methodology is recommended.

- Valuable prior information is often available for medical devices because of their mechanism of action and evolutionary development.
- The Bayesian approach, when correctly employed, may be less burdensome than a frequentist approach.
- In some instances, the use of prior information may alleviate the need for a larger sized trial. In some scenarios, when an adaptive Bayesian model is applicable, the size of a trial can be reduced by stopping the trial early when conditions warrant.
- The Bayesian approach can sometimes be used to obtain an exact analysis when the corresponding frequentist analysis is only approximate or is too difficult to implement.
- Bayesian approaches to multiplicity problems are different from frequentist ones and may be advantageous. Inferences on multiple endpoints and testing of multiple subgroups (e.g., race or sex) are examples of multiplicity.
- Bayesian methods allow for great flexibility in dealing with missing data.

In the context of clinical trials, an unlimited look at the accumulated data, when sampling is sequential in nature, will not affect the inference. In the

frequentist approach, interim data analyses affect type I errors. The ability to stop a clinical trial early is important from the moral and economic viewpoints. Trials should be stopped early due to both futility, to save resources or stop an ineffective treatment, and superiority, to provide patients with the best possible treatments as fast as possible.

Bayesian models facilitate meta-analysis. Meta-analysis is a methodology for the fusion of results of related experiments performed by different researchers, labs, etc. An example of a rudimentary meta-analysis is discussed in Section 8.10.

8.2 Ingredients for Bayesian Inference

A density function for a typical observation X that depends on an unknown, possibly multivariate, parameter θ is called a model and denoted by $f(x|\theta)$. As a function of θ, $f(x|\theta) = L(\theta)$ is called the *likelihood*. If a sample $x = (x_1, x_2, \ldots, x_n)$ is observed, the likelihood takes a familiar form, $L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta)$. This form was used in Chapter 7 to produce MLEs for θ.

Thus both terms model and likelihood are used to describe the distribution of observations. In the standard Bayesian inference the functional form of f is given in the same manner as in the classical parametric approach; the functional form is fully specified up to a parameter θ. According to the generally accepted *likelihood principle,* all information from the experimental data is summarized in the likelihood function, $f(x|\theta) = L(\theta|x_1, \ldots, x_n)$.

For example, if each datum $X|\theta$ were assumed to be exponential with the rate parameter θ and $X_1 = 2, X_2 = 3$, and $X_3 = 1$ were observed, then full information about the experiment would be given by the likelihood

$$\theta e^{-2\theta} \times \theta e^{-3\theta} \times \theta e^{-\theta} = \theta^3 e^{-6\theta}.$$

This model is $\theta^3 \exp\left\{-\theta \sum_{i=1}^{3} X_i\right\}$ if the data are kept unspecified, but in the likelihood function the expression $\sum_{i=1}^{3} X_i$ is treated as a constant term, as was done in the maximum likelihood estimation (page 283).

The parameter θ, with values in the parameter space Θ, is not directly observable and is considered a random variable. This is the key difference between Bayesian and classical approaches. Classical statistics consider the parameter to be a fixed number or vector of numbers, while Bayesians express the uncertainty about θ by considering it as a random variable. This random variable has a distribution $\pi(\theta)$ called the prior distribution. The prior distribution not only quantifies available knowledge, it also describes the uncertainty about a parameter before data are observed. If the prior distribution for θ is specified up to a parameter τ, $\pi(\theta|\tau)$, then τ is called a *hyperparameter.* Hyperparameters are parameters of a prior distribution,

and they are either specified or may have their own priors. This may lead to a hierarchical structure of the model where the priors are arranged in a hierarchy.

The previous discussion can be summarized as follows:

The goal in Bayesian inference is to start with prior information on the parameter of interest, θ, and update it using the observed data. This is achieved via Bayes' theorem, which gives a simple recipe for incorporating observations x in the distribution of θ, $\pi(\theta|x)$, called the *posterior* distribution. All information about θ coming from the prior distribution and the observations are contained in the posterior distribution. The posterior distribution is the ultimate summary of the parameter and serves as the basis for all Bayesian inferences.

According to Bayes' theorem, to find $\pi(\theta|x)$, we divide the *joint* distribution of X and θ ($h(x,\theta) = f(x|\theta)\pi(\theta)$) by the *marginal* distribution for X, $m(x)$, which is obtained by integrating out θ from the joint distribution $h(x,\theta)$:

$$m(x) = \int_{\Theta} h(x,\theta)d\theta = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta.$$

The marginal distribution is also called the *prior predictive* distribution. Thus, in terms of the likelihood and the prior distribution only, the Bayes theorem can be restated as

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_{\Theta} f(x|\theta)\pi(\theta)d\theta}.$$

The integral in the denominator is a major hurdle in Bayesian computation, since for complex likelihoods and priors it could be intractable.

The following table summarizes the notation:

Likelihood, model	$f(x\|\theta)$
Prior distribution	$\pi(\theta)$
Joint distribution	$h(x,\theta) = f(x\|\theta)\pi(\theta)$
Marginal distribution	$m(x) = \int_{\Theta} f(x\|\theta)\pi(\theta)d\theta$
Posterior distribution	$\pi(\theta\|x) = f(x\|\theta)\pi(\theta)/m(x)$

We illustrate these concepts by discussing a few examples in which the posterior distribution can be explicitly obtained. Note that the marginal

distribution has the form of an integral, and in many cases these integrals cannot be found in a finite form. It is fair to say that the number of likelihood/prior combinations that lead to an explicit posterior is rather limited. However, in the general case, the posterior can be evaluated numerically or, as we will see later, a sample can be simulated from the posterior distribution. All of the, admittedly abstract, concepts listed above will be exemplified by several worked-out models. We start with the most important model in which both the likelihood and prior are normal.

Example 8.1. **Normal Likelihood with Normal Prior.** The normal likelihood and normal prior combination is important because it is frequently used in practice. Assume that an observation X is normally distributed with mean θ and known variance σ^2. The parameter of interest, θ, is normally distributed as well, with its parameters μ and τ^2. Parameters μ and τ^2 are hyperparameters, and we will consider them given. Starting with our Bayesian model of $X|\theta \sim \mathcal{N}(\theta,\sigma^2)$ and $\theta \sim \mathcal{N}(\mu,\tau^2)$, we will find the marginal and posterior distributions. Before we start with a derivation of the posterior and marginal, we need a simple algebraic identity:

$$A(x-a)^2 + B(x-b)^2 = (A+B)(x-c)^2 + \frac{AB}{A+B}(a-b)^2, \tag{8.2}$$

for $c = \frac{Aa+Bb}{A+B}$.

We start with the joint distribution of (X,θ), which is the product of two distributions:

$$h(x,\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\theta)^2\right\} \times \frac{1}{\sqrt{2\pi\tau^2}}\exp\left\{-\frac{1}{2\tau^2}(\theta-\mu)^2\right\}.$$

The exponent in the joint distribution $h(x,\theta)$ is

$$-\frac{1}{2\sigma^2}(x-\theta)^2 - \frac{1}{2\tau^2}(\theta-\mu)^2,$$

which, after applying the identity in (8.2), can be expressed as

$$-\frac{\sigma^2+\tau^2}{2\sigma^2\tau^2}\left(\theta - \left(\frac{\tau^2}{\sigma^2+\tau^2}x + \frac{\sigma^2}{\sigma^2+\tau^2}\mu\right)\right)^2 - \frac{1}{2(\sigma^2+\tau^2)}(x-\mu)^2. \tag{8.3}$$

Note that the exponent in (8.3) splits into two parts, one containing θ and the other θ-free. Accordingly the joint distribution $h(x,\theta)$ splits into the product of two densities. Since $h(x,\theta)$ can be represented in two ways, as $f(x|\theta)\pi(\theta)$ and as $\pi(\theta|x)m(x)$, and since we started with $f(x|\theta)\pi(\theta)$, the exponent in (8.3) corresponds to $\pi(\theta|x)m(x)$. Thus, the marginal distribution simply resolves to $X \sim \mathcal{N}(\mu,\sigma^2+\tau^2)$ and the posterior distribution of θ comes out to be

$$\theta|X \sim \mathcal{N}\left(\frac{\tau^2}{\sigma^2+\tau^2}X + \frac{\sigma^2}{\sigma^2+\tau^2}\mu,\ \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right).$$

✎

Below is a specific example of our first Bayesian inference.

Example 8.2. **Jeremy's IQ.** Jeremy, an enthusiastic bioengineering student, posed a statistical model for his scores on a standard IQ test. He thinks that, in general, his scores are normally distributed with unknown mean θ (true IQ) and a variance of $\sigma^2 = 80$. Prior (and expert) opinion is that the IQ of bioengineering students in Jeremy's school, θ, is a normal random variable, with mean $\mu = 110$ and variance $\tau^2 = 120$. Jeremy took the test and scored $X = 98$. The traditional estimator of θ would be $\hat{\theta} = X = 98$. The posterior is normal with a mean of $\frac{120}{80+120} \times 98 + \frac{80}{80+120} \times 110 = 102.8$ and a variance of $\frac{80\times120}{80+120} = 48$. We will see later that the mean of the posterior is Bayes' estimator of θ, and a Bayesian would estimate Jeremy's IQ as 102.8. ✎

If n normal variates, $X_1, X_2, \ldots, X_n$, are observed instead of a single observation X, then the sample is summarized as $\overline{X}$ and the Bayesian model for θ is essentially the same as that for a single X, but with σ^2/n in place of σ^2. In this case, the likelihood and the prior are

$$\overline{X}|\theta \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{n}\right) \text{ and } \theta \sim \mathcal{N}(\mu, \tau^2),$$

producing

$$\theta|\overline{X} \sim \mathcal{N}\left(\frac{\tau^2}{\frac{\sigma^2}{n}+\tau^2}\overline{X} + \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+\tau^2}\mu, \frac{\frac{\sigma^2}{n}\ \tau^2}{\frac{\sigma^2}{n}+\tau^2}\right).$$

Notice that the posterior mean

$$\frac{\tau^2}{\frac{\sigma^2}{n}+\tau^2}\overline{X} + \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+\tau^2}\mu$$

is a weighted average of the MLE $\overline{X}$ and the prior mean μ with weights $w = n\tau^2/(\sigma^2 + n\tau^2)$ and $1 - w = \sigma^2/(\sigma^2 + n\tau^2)$. When the sample size n increases, the contribution of the prior mean to the estimator diminishes as $w \to 1$. In contrast, when n is small and our prior opinion about μ is strong (i.e., τ^2 is small), the posterior mean remains close to the prior mean μ. Later, we will explore several more cases in which the posterior mean has a form of a weighted average of the MLE for the parameter and the prior mean.

Example 8.3. **Likelihood, Prior, and Posterior.** Suppose that $n = 10$ observations are coming from $\mathcal{N}(\theta, 10^2)$. Assume that the prior on θ is $\mathcal{N}(20, 20)$. For the observations
$\{2.944, -13.361, 7.143, 16.235, -6.917, 8.580, 12.540, -15.937, -14.409, 5.711\}$
the posterior is $\mathcal{N}(6.835, 6.667)$. The three densities: likelihood, prior, and posterior, are shown in Figure 8.1.

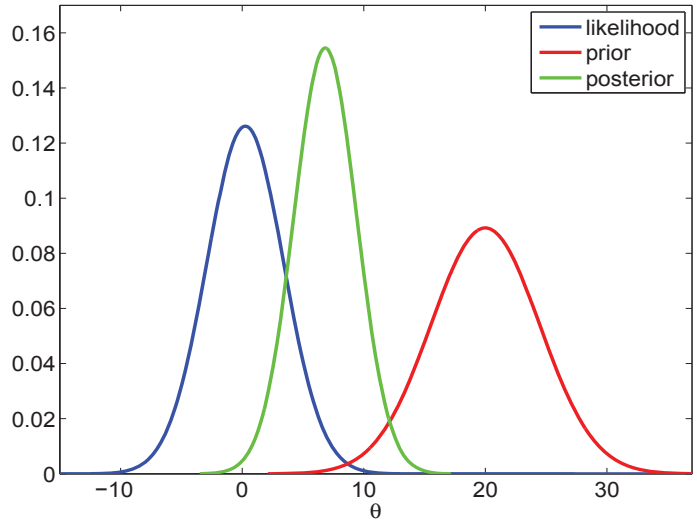


Fig. 8.1 The likelihood centered at MLE $\overline{X} = 0.2529$, $\mathcal{N}(0.2529, 10^2/10)$ (*blue*), $\mathcal{N}(20, 20)$ prior (*red*), and posterior for data $\{2.9441, -13.3618, \ldots, 5.7115\}$ (*green*).

8.3 Conjugate Priors

A major technical difficulty in Bayesian analysis is finding an explicit posterior distribution, given the likelihood and prior. The posterior is proportional to the product of the likelihood and prior, but the normalizing constant, marginal $m(x)$, is often difficult to find since it involves integration.

In Examples 8.1 and 8.3, where the prior is normal, the posterior distribution remains normal. In such cases, the effect of likelihood is only to "update" the prior parameters and not to change the prior's functional form. We say that such priors are *conjugate* with the likelihood. Conjugacy is popular because of its mathematical convenience; once the conjugate likelihood/prior pair is identified, the posterior is found without integration.

The normalizing marginal $m(x)$ is selected such that $f(x|\theta)\pi(\theta)$ is a density from the same class to which the prior belongs. Operationally, one multiplies "kernels" of likelihood and priors, ignoring all multiplicative terms that do not involve the parameter. For example, a kernel of gamma $\mathcal{G}a(r,\lambda)$ density $f(\theta|r,\lambda) = \frac{\lambda^r\theta^{r-1}}{\Gamma(r)}e^{-\lambda\theta}$ would be $\theta^{r-1}e^{-\lambda\theta}$. We would write: $f(\theta|r,\lambda) \propto \theta^{r-1}e^{-\lambda\theta}$, where the symbol $\propto$ stands for "proportional to." Several examples in this chapter involve conjugate pairs (Examples 8.4 and 8.6).

In the pre-Markov chain Monte Carlo era, conjugate priors were extensively used (and overused and misused) precisely because of this computational convenience. Today, the general agreement is that simple conjugate analysis is of limited practical value since, given the likelihood, the conjugate prior has limited modeling capability.

There are quite a few instances of conjugacy. Table 8.1 gives several important cases. As a practice, you may want to derive the posteriors listed in the third column of the table. It is recommended that you consult Chapter 5 on functional forms of densities involved in the Bayesian model.

Table 8.1 Some conjugate pairs. Here $\mathbf{X}$ stands for a sample of size n, $X_1,\ldots,X_n$. For functional expressions of the densities and their moments refer to Chapter 5

Likelihood	Prior	Posterior
$X_i\|\theta \sim \mathcal{N}(\theta,\sigma^2)$	$\theta \sim \mathcal{N}(\mu,\tau^2)$	$\theta\|\mathbf{X} \sim \mathcal{N}\left(\frac{\tau^2}{\tau^2+\sigma^2/n}\overline{X} + \frac{\sigma^2/n}{\tau^2+\sigma^2/n}\mu, \frac{\tau^2\sigma^2/n}{\tau^2+\sigma^2/n}\right)$
$X_i\|\theta \sim \mathcal{B}in(m,\theta)$	$\theta \sim \mathcal{B}e(\alpha,\beta)$	$\theta\|\mathbf{X} \sim \mathcal{B}e(\alpha+\sum_{i=1}^n X_i, \beta+mn-\sum_{i=1}^n X_i)$
$X_i\|\theta \sim \mathcal{P}oi(\theta)$	$\theta \sim \mathcal{G}a(\alpha,\beta)$	$\theta\|\mathbf{X} \sim \mathcal{G}a(\alpha+\sum_{i=1}^n X_i, \beta+n)$
$X_i\|\theta \sim \mathcal{NB}(m,\theta)$	$\theta \sim \mathcal{B}e(\alpha,\beta)$	$\theta\|\mathbf{X} \sim \mathcal{B}e(\alpha+mn, \beta+\sum_{i=1}^n X_i)$
$X_i\|\theta \sim \mathcal{G}a(1/2, 1/(2\theta))$	$\theta \sim \mathcal{IG}(\alpha,\beta)$	$\theta\|\mathbf{X} \sim \mathcal{IG}(\alpha+n/2, \beta+\frac{1}{2}\sum_{i=1}^n X_i)$
$X_i\|\theta \sim \mathcal{U}(0,\theta)$	$\theta \sim \mathcal{P}a(\theta_0,\alpha)$	$\theta\|\mathbf{X} \sim \mathcal{P}a(\max\{\theta_0,X_1,\ldots,X_n\}, \alpha+n)$
$X_i\|\theta \sim \mathcal{N}(\mu,\theta)$	$\theta \sim \mathcal{IG}(\alpha,\beta)$	$\theta\|\mathbf{X} \sim \mathcal{IG}(\alpha+n/2, \beta+\frac{1}{2}\sum_{i=1}^n (X_i-\mu)^2)$
$X_i\|\theta \sim \mathcal{G}a(\nu,\theta)$	$\theta \sim \mathcal{G}a(\alpha,\beta)$	$\theta\|\mathbf{X} \sim \mathcal{G}a(\alpha+n\nu, \beta+\sum_{i=1}^n X_i)$
$X_i\|\theta \sim \mathcal{P}a(c,\theta)$	$\theta \sim \mathcal{G}a(\alpha,\beta)$	$\theta\|\mathbf{X} \sim \mathcal{G}a(\alpha+n, \beta+\sum_{i=1}^n \log(X_i/c))$

Example 8.4. **Binomial Likelihood with Beta Prior.** An easy, yet important, example of a conjugate structure is the binomial likelihood and beta prior. Suppose that we observed $X = x$ from a binomial $\mathcal{B}in(n,p)$ distribution,

$$f(x|\theta) = \binom{n}{x} p^x (1-p)^{n-x},$$

and that the population proportion p is the parameter of interest. If the prior on p is beta $\mathcal{B}e(\alpha,\beta)$ with hyperparameters α and β and density

$$\pi(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1},$$

the posterior is proportional to the product of the likelihood and the prior

$$\pi(p|x) = C \cdot p^x (1-p)^{n-x} \cdot p^{\alpha-1}(1-p)^{\beta-1} = C \cdot p^{x+\alpha-1}(1-p)^{n-x+\beta-1}$$

for some constant C. The normalizing constant C is free of p and is equal to $\frac{\binom{n}{x}}{m(x)B(\alpha,\beta)}$, where $m(x)$ is the marginal distribution.

By inspecting the expression $p^{x+\alpha-1}(1-p)^{n-x+\beta-1}$, it can be seen that the posterior density remains beta; it is $\mathcal{B}e(x+\alpha, n-x+\beta)$, and that the normalizing constant resolves to $C = 1/B(x+\alpha, n-x+\beta)$. From the equality of constants, it follows that

$$\frac{\binom{n}{x}}{m(x)B(\alpha,\beta)} = \frac{1}{B(x+\alpha, n-x+\beta)},$$

and one can express the marginal density as

$$m(x) = \frac{\binom{n}{x} B(x+\alpha, n-x+\beta)}{B(\alpha,\beta)},$$

which is known as a *beta-binomial distribution*.
✎

8.4 Point Estimation

The posterior is the ultimate experimental summary for a Bayesian. The posterior location measures, especially the mean, are of great importance. The posterior mean is the most frequently used Bayes' estimator for a parameter. The posterior mode and median are alternative Bayes' estimators.

The posterior mode maximizes the posterior density in the same way that the MLE maximizes the likelihood. When the posterior mode is used as an estimator, it is called the maximum posterior (MAP) estimator. The MAP estimator is popular in some Bayesian analyses in part because it is computationally less demanding than the posterior mean or median. The reason for this is simple: to find a MAP, the posterior does not need to be fully specified because $\operatorname{argmax}_\theta \pi(\theta|x) = \operatorname{argmax}_\theta f(x|\theta)\pi(\theta)$, that is, the

product of the likelihood and the prior as well as the posterior are maximized at the same point.

Example 8.5. **Binomial-Beta Conjugate Pair.** In Example 8.4 we argued that for the likelihood $X|\theta \sim \mathcal{B}in(n,\theta)$ and the prior $\theta \sim \mathcal{B}e(\alpha,\beta)$, the posterior distribution is $\mathcal{B}e(x+\alpha, n-x+\beta)$. The Bayes estimator of θ is the expected value of the posterior

$$\hat{\theta}_B = \frac{\alpha + x}{(\alpha + x) + (\beta + n - x)} = \frac{\alpha + x}{\alpha + \beta + n}.$$

This is actually a weighted average of the MLE, X/n, and the prior mean $\alpha/(\alpha+\beta)$,

$$\hat{\theta}_B = \frac{n}{\alpha + \beta + n} \cdot \frac{X}{n} + \frac{\alpha + \beta}{\alpha + \beta + n} \cdot \frac{\alpha}{\alpha + \beta}.$$

Notice that, as n becomes large, the posterior mean approaches the MLE because the weight $\frac{n}{n+\alpha+\beta}$ tends to 1. In contrast, when α or β or both are large compared to n, the posterior mean is close to the prior mean. Due to this interplay between n and prior parameters, the sum $\alpha+\beta$ is called the prior sample size, and it measures the influence of the prior as if additional experimentation was performed and $\alpha+\beta$ trials have been added. This is in the spirit of Wilson's proposal to "add two failures and two successes" to an estimator of proportion (page 305). Wilson's estimator can be seen as a Bayes estimator with a beta $\mathcal{B}e(2,2)$ prior.

Large α indicates a small prior variance, since for fixed β, the variance of $\mathcal{B}e(\alpha,\beta)$ is proportional to $1/\alpha^2$, and the prior is concentrated about its mean.
✎

In general, the posterior mean will fall between the MLE and the prior mean. This was demonstrated in Example 8.1. As another example, suppose we flipped a coin four times and tails showed up on all four occasions. We are interested in estimating the probability of showing heads, θ, in a Bayesian fashion. If the prior is $\mathcal{U}(0,1)$, the posterior is proportional to $\theta^0(1-\theta)^4$, which is a beta $\mathcal{B}e(1,5)$. The posterior mean *shifts* the MLE (0) toward the expected value of the prior (1/2) to get $\hat{\theta}_B = 1/(1+5) = 1/6$, which is a more reasonable estimator of θ than the MLE. Note that the $3/n$ rule produces a confidence interval for p of $[0,3/4]$, which is too wide to be useful (Section 7.5.4).

Example 8.6. **Uniform/Pareto Model.** In Example 7.5 we had the observations $X_1 = 2$, $X_2 = 5$, $X_3 = 0.5$, and $X_4 = 3$ from a uniform $\mathcal{U}(0,\theta)$ distribution. We are interested in estimating θ in Bayesian fashion. Let the prior on θ be Pareto $\mathcal{P}a(\theta_0,\alpha)$ for $\theta_0 = 6$ and $\alpha = 2$. Then the posterior

is also Pareto $\mathcal{P}a(\theta^*,\alpha^*)$ with $\theta^* = \max\{\theta_0,\ X_{(n)}\} = \max\{6,5\} = 6$, and $\alpha^* = \alpha + n = 2 + 4 = 6$. The posterior mean is $\frac{\alpha^*\theta^*}{\alpha^*-1} = 36/5 = 7.2$, and the median is $\theta^* \cdot 2^{1/\alpha^*} = 6 \cdot 2^{1/6} = 6.7348$.

Figure 8.2 shows the prior (dashed red line) with the prior mean as a red dot. After observing $X_1,\ldots,X_4$, the posterior mode did not change since the elicited $\theta_0 = 6$ was larger than $\max X_i = 5$. However, the posterior has a smaller variance than the prior. The posterior mean is shown as a green dot, the posterior median as a black dot, and the posterior (and prior) mode as a blue dot.

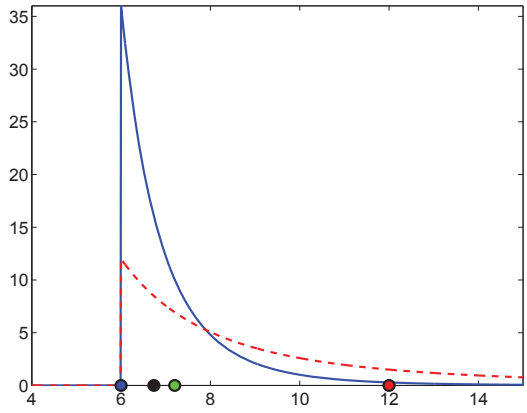


Fig. 8.2 Pareto $\mathcal{P}a(6,2)$ prior (*dashed red line*) and $\mathcal{P}a(6,6)$ posterior (*solid blue line*). The *red dot* is the prior mean, the *green dot* is the posterior mean, the *black dot* is the posterior median, and the *blue dot* is the posterior (and prior) mode.

Another widely used conjugate pair is Poisson–gamma pair.

Example 8.7. **Poisson–Gamma Conjugate Pair.** Let $X_1,\ldots,X_n$, given θ are Poisson $\mathcal{P}oi(\theta)$ with probability mass function

$$f(x_i|\theta) = \frac{\theta^{x_i}}{x_i!}e^{-\theta},$$

and $\theta \sim \mathcal{G}a(\alpha,\beta)$ is given by $\pi(\theta) \propto \theta^{\alpha-1}e^{-\beta\theta}$. Then

$$\pi(\theta|X_1,\ldots,X_n) = \pi(\theta|\sum X_i) \propto \theta^{\sum X_i+\alpha-1}e^{-(n+\beta)\theta},$$

which is $\mathcal{G}a(\sum_i X_i + \alpha, n+\beta)$. The mean is $\mathbb{E}(\theta|X) = (\sum X_i + \alpha)/(n+\beta)$, and it can be represented as a weighted average of the MLE and the prior mean:

$$\mathbb{E}\theta|X = \frac{n}{n+\beta}\,\frac{\sum X_i}{n} + \frac{\beta}{n+\beta}\,\frac{\alpha}{\beta}.$$

Let us apply the this equation in a specific example. Let a rare disease have an incidence of X cases per 100,000 people, where X is modeled as Poisson, $X|\lambda \sim \mathcal{P}oi(\lambda)$, where λ is the rate parameter. Assume that for different cohorts of 100,000 subjects, the following incidences are observed: $X_1 = 2$, $X_2 = 0$, $X_3 = 0$, $X_4 = 4$, $X_5 = 0$, $X_6 = 1$, $X_7 = 3$, and $X_8 = 2$. The experts indicate that λ should be close to 2 and our prior is $\lambda \sim \mathcal{G}a(0.2, 0.1)$. We matched the mean, since for a gamma distribution the mean is $0.2/0.1 = 2$ but the variance $0.2/0.1^2 = 20$ is quite large, thereby expressing our uncertainty. By setting the hyperparameters to 0.02 and 0.01, for example, we would have variance of the gamma prior that is even larger. The MLE of λ is $\hat{\lambda}_{mle} = \overline{X} = 3/2$. The Bayes estimator is

$$\hat{\lambda}_B = \frac{8}{8+0.1}3/2 + \frac{0.1}{8+0.1}2 = 1.5062.$$

Note that since the prior was not informative, the Bayes estimator is quite close to the MLE.
✎

Normal-Inverse Gamma Conjugate Analysis. Let $y_1, y_2, \ldots, y_n$ be the observations from normal $\mathcal{N}(\mu, \sigma^2)$ distribution where both μ and σ^2 are of interest. For this problem there is a conjugate joint prior for (μ, σ^2), normal-inverse gamma $\mathcal{NIG}(\mu_0, c, a, b)$,

$$\pi(\mu, \sigma^2) = \pi(\mu|\sigma^2)\pi(\sigma^2) = \mathcal{N}(\mu_0, \sigma^2/c) \times \mathcal{IG}(a, b).$$

Note that apriori μ and σ^2 are not independent, their joint prior is not a product of densities that fully separates the variables.

Instead of variance σ^2, often the precision parameter $\tau = 1/\sigma^2$ is modeled and estimated. In many cases the estimation of τ is more stable than that of σ^2. From the definition of inverse-gamma it follows that if $\sigma^2 \sim \mathcal{IG}(a, b)$ then $\tau \sim \mathcal{G}a(a, b)$. Thus,

$$\begin{aligned}\pi(\mu, \tau) &= \mathcal{N}\left(\mu_0, \frac{1}{c\tau}\right) \times \mathcal{G}a(a, b) \\ &= \sqrt{\frac{c\tau}{2\pi}} \exp\left\{\frac{c\tau}{2}(\mu - \mu_0)^2\right\} \times \frac{b^a \tau^{a-1}}{\Gamma(a)} \exp\{-b\tau\}.\end{aligned}$$

After observing $\boldsymbol{y} = (y_1, \ldots, y_n)$, all inference depends on $\overline{y} = 1/n \sum_{i=1}^n y_i$ and $s^2 = \sum_{i=1}^n (y_i - \overline{y})^2/(n-1)$. Denote

$$SS = \sum_{i=1}^n (y_i - \overline{y})^2 + \frac{nc}{n+c}(\overline{y} - \mu_0)^2 = (n-1)s^2 + \frac{nc}{n+c}(\overline{y} - \mu_0)^2$$

When the likelihood is normal, the problem is conjugate and the posterior for (μ,σ^2) is $\mathcal{NIG}(\mu_0^*,c^*,a^*,b^*)$, or equivalently, $\mathcal{NG}(\mu_0^*,c^*,a^*,b^*)$ for (μ,τ).

The updated parameters (from prior to the posterior) are shown in the following table:

Prior	Posterior
μ_0	$\mu_0^* = \frac{c}{n+c}\mu_0 + \frac{n}{n+c}\overline{y}$
c	$c^* = c+n$
a	$a^* = a + n/2$
b	$b^* = b + SS/2$

Posterior expectations (Bayes' estimators) and variances for μ,τ, and σ^2 are:

$$\begin{aligned}
&\mathbb{E}(\mu|\boldsymbol{y}) = \mu_0^* = \frac{c}{n+c}\mu_0 + \frac{n}{n+c}\overline{y},\\
&\mathbb{V}\mathrm{ar}\,(\mu|\boldsymbol{y}) = \frac{1}{n+c} \times \frac{SS+2b}{n+2a-2}, \qquad n > 2-2a,\\
&\mathbb{E}(\tau|\boldsymbol{y}) = \frac{n+2a}{SS+2b},\\
&\mathbb{V}\mathrm{ar}\,(\tau|\boldsymbol{y}) = \frac{2n+4a}{(SS+2b)^2},\\
&\mathbb{E}(\sigma^2|\boldsymbol{y}) = \frac{SS+2b}{n+2a-2}, \qquad n > 2-2a, \text{ and}\\
&\mathbb{V}\mathrm{ar}\,(\sigma^2|\boldsymbol{y}) = \frac{2(SS+2b)^2}{(n+2a-2)^2(n+2a-4)}, \qquad n > 4-2a.
\end{aligned}$$

Example 8.8. **Jeremy and NIG Prior.** Suppose that Jeremy took the IQ test 6 times. His scores $(101,98,114,105,108,111)$ are assumed to be a sample from a normal distribution with unknown mean μ and variance σ^2.

The prior on (μ,σ^2) is normal-inverse gamma with parameters $\mu_0 = 110$, $c = 1.5$, $a = 0.1$ and $b = 10$.

Using exact conjugate calculations, we find Bayes' estimators for μ and σ^2.

```
y = [ 101  98  114  105  108  111 ];
mu0 = 110; n=6; c=1.5; a=0.1;  b=10;
ybar = mean(y);
ss = (n-1) * var(y) + n*c/(n+c) * (ybar - mu0)^2;
%
muhat=  c/(n+c) * mu0  + n/(n+c) * ybar       %106.9333
varmuhat = 1/(n+c) * (ss + 2*b)/(n + 2*a -2)  %6.9989
stdmuhat = sqrt(varmuhat)                     %2.6456
tauhat = (n + 2 * a)/(ss + 2 * b)             %0.0281
vartauhat = 2 * (n  + 2 * a)/(ss  + 2 * b)^2  %2.5511e-04
```

```
stdtauhat = sqrt(vartauhat)                    %0.016
sigma2hat = (ss + 2 * b)/(n + 2*a - 2)         %52.4921
varsigma2hat = 2 * (ss + 2 * b)^2 /...
     ((n + 2*a - 2)^2 * (n + 2* a - 4))        %2.5049e+03
stdsigma2hat = sqrt(varsigma2hat)              %50.0492
```

Note that Bayes' estimator of μ is $\hat{\mu}_B = 106.9333$. The estimators of variance and precision are $\hat{\sigma}^2_B = 52.4921$ and $\hat{\tau}_B = 0.0281$. In addition to estimators of these parameters, Bayesian model gives us the estimators of their variances `varmuhat`, `varsigma2hat`, and `vartauhat` and their standard deviations `stdmuhat`, `stdsigma2hat`, and `stdtauhat`.

8.5 Prior Elicitation

Prior distributions are carriers of prior information that is coherently incorporated via Bayes' theorem into an inference. At the same time, parameters are unobservable, and prior specification is subjective in nature. The subjectivity of specifying the prior is a fundamental criticism of the Bayesian approach. Being subjective does not mean that the approach is nonscientific, as critics of Bayesian statistics often insinuate. On the contrary, vast amounts of scientific information coming from theoretical and physical models, previous experiments, and expert reports guides the specification of priors and merges such information with the data for better inference.

In arguing about the importance of priors in Bayesian inference, Garthwhite and Dickey (1991) state that "expert personal opinion is of great potential value and can be used more efficiently, communicated more accurately, and judged more critically if it is expressed as a probability distribution."

In the last several decades Bayesian research has also focused on priors that were noninformative and robust; this was in response to criticism that results of Bayesian inference could be sensitive to the choice of a prior.

For instance, in Examples 8.4 and 8.5 we saw that beta distributions are an appropriate family of priors for parameters supported in the interval $[0,1]$, such as a population proportion. It turns out that the beta family can express a wide range of prior information. For example, if the mean μ and variance σ^2 for a beta prior are elicited by an expert, then the parameters (a,b) can be determined by solving $\mu = a/(a+b)$ and $\sigma^2 = ab/[(a+b)^2(a+b+1)]$ with respect to a and b:

$$a = \mu\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right), \quad \text{and} \quad b = (1-\mu)\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right). \tag{8.4}$$

If a and b are not too small, the shape of a beta prior resembles a normal distribution and the bounds $[\mu - 2\sigma, \mu + 2\sigma]$ can be used to describe the

range of likely parameters. For example, an expert's claim that a proportion is unlikely to be higher than 90% can be expressed as $\mu + 2\sigma = 0.9$.

In the same context of estimating the proportion, Berry and Stangl (1996) suggest a somewhat different procedure:

(i) Elicit the probability of success in the first trial, p_1, and match it to the prior mean $\alpha/(\alpha+\beta)$.

(ii) Given that the first trial results in success, the posterior mean is $\frac{\alpha+1}{\alpha+\beta+1}$. Match this ratio with the elicited probability of success in a second trial, p_2, conditional upon the first trial's resulting in success. Thus, a system

$$p_1 = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad p_2 = \frac{\alpha+1}{\alpha+\beta+1}$$

is obtained that solves to

$$\alpha = \frac{p_1(1-p_2)}{p_2-p_1} \quad \text{and} \quad \beta = \frac{(1-p_1)(1-p_2)}{p_2-p_1}. \tag{8.5}$$

See Exercise 8.15 for an application.

If one has no prior information, many noninformative choices are possible, such as invariant priors, Jeffreys' priors, default priors, reference priors, and intrinsic priors, among others. Informally speaking, a noninformative prior is one which is dominated by the likelihood, or that is "flat" relative to the likelihood.

Popular noninformative choices are the flat prior $\pi(\theta) = C$ for the location parameter (mean) and $\pi(\theta) = 1/\theta$ for the scale/rate parameter. A vague prior for the population proportion is proportional to $p^{-1}(1-p)^{-1}$, $0 < p < 1$. This prior is sometimes called Zellner's prior and is equivalent of setting a flat prior on the $\text{logit}(p) = \log\frac{p}{1-p}$. The listed priors are not proper probability distributions, that is, they are not bonafide densities because their integrals are not finite. However, Bayes' theorem usually leads to posterior distributions that are proper densities and on which Bayesian analysis can be carried out.

Jeffreys' priors (named after Sir Harold Jeffreys, English statistician, geophysicist, and astronomer) are obtained from a particular functional of a density (Fisher information), and they are also examples of vague and noninformative priors. For a binomial proportion, Jeffreys' prior is proportional to $p^{-1/2}(1-p)^{-1/2}$, while for the rate of exponential distribution λ, Jeffreys' prior is proportional to $1/\lambda$. For a normal distribution, Jeffreys' prior on the mean is flat, while for the variance σ^2, it is proportional to $\frac{1}{\sigma^2}$.

Example 8.9. **Jeffreys' Prior on Exponential Rate Parameter.** If $X_1 = 1.7$, $X_2 = 0.6$, and $X_3 = 5.2$ come from an exponential distribution with a rate parameter λ, find the Bayes estimator if the prior on λ is $\frac{1}{\lambda}$.

The likelihood is $\lambda^3 e^{-\lambda\sum_{i=1}^3 X_i}$ and the posterior is proportional to

$$\frac{1}{\lambda} \times \lambda^3 e^{-\lambda \sum_{i=1}^3 X_i} = \lambda^{3-1} e^{-\lambda \sum X_i},$$

which is recognized as gamma $\mathcal{G}a\left(3, \sum_{i=1}^3 X_i\right)$. The Bayes estimator, as a mean of this posterior, coincides with the MLE, $\hat{\lambda} = \frac{3}{\sum_{i=1}^3 X_i} = \frac{1}{\overline{X}} = 1/2.5 = 0.4$.
✎

Effective Sample Size in Prior Elicitation. In the previous discussion we used the notion *noninformative*, as a prior attribute in quite informal manner. For example, uniform, Jeffreys, and Zellner priors on binomial proportions have all been called noninformative.

It is possible to calibrate the amount of information a prior is carrying by assigning a sample size value to it. Informally, the information in a prior is "worth" the information contained in a sample of size m. We will call m the effective sample size (ESS).

The ESS is inferred mainly on conjugate pairs of distributions by comparing hyperparameters of the prior and posterior, or prior and posterior means.

(i) When the model is binomial, and the prior is beta $\mathcal{B}e(a,b)$, the prior mean is $a/(a+b)$ and the posterior mean is $(a+X)/(a+b+n)$, so ESS $= a+b$ is adopted.

(ii) Gamma $\mathcal{G}a(a,b)$ prior on Poisson rate λ is conjugate and the Bayes rule a/b without data goes to $(\sum_i X_i + a)/(b+n)$ with the data, so ESS $= b$.

(iii) In gamma $\mathcal{G}a(a,b)$ prior on normal precision $\tau = 1/\sigma^2$, the Bayes rules are a/b and $(a+n/2)/(b+1/2\sum_i (X_i-\mu)^2)$, so ESS $= 2a$.

(iv) For the normal mean with normal prior, ESS is σ^2/ξ^2, where σ^2 is variance of the likelihood, and ξ^2 is the variance of the prior.

Sometimes the historic data used to elicit priors and determine ESS are not of the same quality, rigor, or importance as the data in the experiment that is under analysis, and we may want to discount the ESS by a factor between 0 and 1, say k. That leads to replacing the priors above with $\mathcal{B}e(ka,kb)$, $\mathcal{G}a(ka,kb)$, or in the normal case, replacing ξ^2 by $k\xi^2$.

For an example of use of ESS in prior elicitation, see Example 10.3.

An applied approach to prior selection was taken by Spiegelhalter et al. (1994) in the context of clinical trials. They recommended a *community of priors* elicited from a large group of experts. A crude classification of community priors is as follows:

(i) Vague priors – noninformative priors, in many cases leading to posterior distributions proportional to the likelihood.

(ii) Skeptical priors – reflecting the opinion of a clinician unenthusiastic about the new therapy, drug, device, or procedure. This may be a prior of a regulatory agency.

(iii) Enthusiastic or clinical priors – reflecting the opinion of the proponents of the clinical trial, centered around the notion that a new therapy, drug, device, or procedure is superior. This may be the prior of the industry involved or of clinicians running the trial.

For example, the use of a skeptical prior when testing for the superiority of a new treatment would be a conservative approach. In equivalence tests, both skeptical and enthusiastic priors may be used. The superiority of a new treatment should be judged by a skeptical prior, while the superiority of the old treatment should be judged by an enthusiastic prior.

8.6 Bayesian Computation and Use of WinBUGS

If the selection of an adequate prior is the major conceptual and modeling challenge of Bayesian analysis, the major implementational challenge is computation. When the model deviates from the conjugate structure, finding the posterior distribution and the Bayes rule is all but simple. A closed-form solution is more the exception than the rule, and even for such exceptions, lucky mathematical coincidences, convenient mixtures, and other tricks are needed to uncover the explicit expression.

If classical statistics relies on optimization, Bayesian statistics relies on integration. The marginal needed to normalize the product $f(x|\theta)\pi(\theta)$ is an integral

$$m(x) = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta,$$

while the Bayes estimator of $h(\theta)$ is a ratio of integrals,

$$\delta_{\pi}(x) = \int_{\Theta} h(\theta)\pi(\theta|x)d\theta = \frac{\int_{\Theta} h(\theta)f(x|\theta)\pi(\theta)d\theta}{\int_{\Theta} f(x|\theta)\pi(\theta)d\theta}.$$

The difficulties in calculating the above Bayes rule derive from the facts that (i) the posterior may not be representable in a finite form and (ii) the integral of $h(\theta)$ does not have a closed form even when the posterior distribution is explicit.

The last two decades of research in Bayesian statistics has contributed to broadening the scope of Bayesian models. Models that could not be handled before by a computer are now routinely solved. This is done by *Markov chain Monte Carlo* (MCMC) methods, and their introduction to the field of statistics revolutionized Bayesian statistics.

The MCMC methodology was first applied in statistical physics (Metropolis et al., 1953). Work by Gelfand and Smith (1990) focused on applications of MCMC to Bayesian models. The principle of MCMC is simple: one designs a Markov chain that samples from the target distribution. By simulat-

ing long runs of such a Markov chain, the target distribution can be well approximated. Various strategies for constructing appropriate Markov chains that simulate the desired distribution are possible: Metropolis–Hastings, Gibbs sampler, slice sampling, perfect sampling, and many specialized techniques. These are beyond the scope of this text, and the interested reader is directed to Robert (2001), Robert and Casella (2004), and Chen et al. (2000) for an overview and a comprehensive treatment.

In the examples that follow we will use WinBUGS for doing Bayesian inference when the models are not conjugate. Chapter 19 gives a brief introduction to the front end of WinBUGS. Three volumes of examples are a standard addition to the software; in the Examples menu of WinBUGS, see Spiegelhalter et al. (1996). It is recommended that you go over some of those examples in detail because they illustrate the functionality and modeling power of WinBUGS. A wealth of examples on Bayesian modeling strategies using WinBUGS can be found in the monographs of Congdon (2005, 2006, 2010, 2014), Lunn et al. (2013), and Ntzoufras (2009).

The following example is a WinBUGS solution of Example 8.2.

Example 8.10. **Jeremy's IQ in WinBUGS.** We will calculate a Bayes estimator for Jeremy's true IQ, θ, using simulations in WinBUGS. Recall that the model was $X \sim \mathcal{N}(\theta, 80)$ and $\theta \sim \mathcal{N}(100, 120)$. WinBUGS uses precision instead of variance to parameterize the normal distribution. Precision is simply the reciprocal of the variance, and in this example, the precisions are $1/120 = 0.00833$ for the prior and $1/80 = 0.0125$ for the likelihood. The WinBUGS code is as follows:

```
Jeremy in WinBUGS
model{
x ~ dnorm( theta, 0.0125)
theta ~ dnorm( 110, 0.008333333)
}
DATA
list(x=98)
INITS
list(theta=100)
```

Here is the summary of the MCMC output. The Bayes estimator for θ is rounded to 102.8. It is obtained as a mean of the simulated sample from the posterior.

	mean	sd	MC error	val2.5pc	median	val97.5pc	start	sample
theta	102.8	6.943	0.01991	89.18	102.8	116.4	1001	100000

Since this is a conjugate normal/normal model, the exact posterior distribution, $\mathcal{N}(102.8, 48)$, was easy to find, (Example 8.2). Note that in these simulations, the MCMC approximation, when rounded, coincides with the exact posterior mean. The MCMC variance of θ is $6.943^2 \approx 48.2$, which is close to the exact posterior variance of 48.

Example 8.11. **Uniform/Pareto Model in WinBUGS.** In Example 8.6, we found that a posterior distribution of θ, in a uniform $\mathcal{U}(0,\theta)$ model with a Pareto $\mathcal{P}a(6,2)$ prior, was Pareto $\mathcal{P}a(6,6)$. From the posterior, we found the mean, median, and mode to be 7.2, 6.7348, and 6, respectively. These are reasonable estimators of θ as location measures of the posterior.

```
Uniform with Pareto in WinBUGS
model{
for (i  in 1:n){
   x[i] ~ dunif(0, theta);
   }
theta ~ dpar(2,6)
}
DATA
list(n=4, x = c(2, 5, 0.5, 3) )
INITS
list(theta= 7)
```

Here is the summary of the WinBUGS output. The posterior mean was found to be 7.196 and the median 6.736. Apparently, the mode of the posterior was 6, as is evident from Figure 8.3. These approximations are close to the exact values found in Example 8.6.

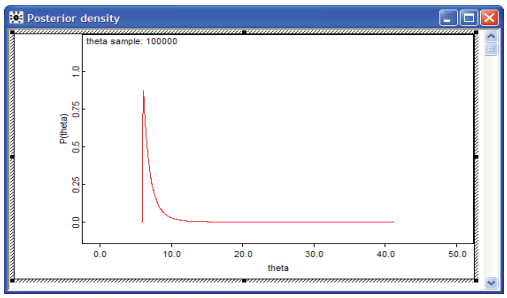


Fig. 8.3 Output from `Inference>Samples>density` shows MCMC approximation to the posterior distribution.

	mean	sd	MC error	val2.5pc	median	val97.5pc	start	sample
theta	7.196	1.454	0.004906	6.025	6.736	11.03	1001	100000

Example 8.12. **Jeremy, NIG Prior, and BUGS.** Using conjugate structure of the model in Example 8.8, we found the exact Bayes' estimator of μ as $\hat{\mu}_B = 106.9333$, and the estimators of variance and precision as $\hat{\sigma}^2_B = 52.4921$ and $\hat{\tau}_B = 0.0281$. In addition to estimators of these parameters, Bayesian

model produced the estimators of their standard deviations: `stdmuhat=2.6456`, `stdsigma2hat=50.0492`, and `stdtauhat=0.016`. The following WinBUGS script calculates these estimators by MCMC simulation:

```
model{
for (i in 1:n){
   y[i] ~ dnorm(mu, tau)}
   tauc <-  c*tau
   mu ~ dnorm(mu0, tauc)
   tau ~ dgamma(a, b)
   sigma2 <- 1/tau
}
DATA
list( n=6, c=1.5,  mu0=110, a=0.1, b=10,
y=c(101, 98, 114, 105, 108, 111))
INITS
list( tau=0.01, mu=100)
```

	mean	sd	MC error	val2.5pc	median	val97.5pc	start	sample
mu	106.9	2.646	0.002655	101.6	106.9	112.2	1001	1000000
sigma2	52.48	49.61	0.06046	14.92	39.75	166.2	1001	1000000
tau	0.02813	0.01599	1.764E-5	0.00601	0.02516	0.06701	1001	1000000

Zero-Tricks in WinBUGS. Although the list of built-in distributions for specifying the likelihood or the prior in WinBUGS is rich (page 952), sometimes we encounter densities that are not on the list. How do we set the likelihood for a density that is not built into WinBUGS?

There are several ways, the most popular of which is the so-called zero-trick. Let f be an arbitrary model and $\ell_i = \log f(x_i|\theta)$ the log-likelihood for the ith observation. Then

$$\prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} e^{\ell_i} = \prod_{i=1}^{n} \frac{(-\ell_i)^0 e^{-(-\ell_i)}}{0!} = \prod_{i=1}^{n} \mathbb{P}(Y_i = 0),$$

where Y_i are Poisson $\mathcal{P}oi(-\ell_i)$ random variables.

The WinBUGS code for a zero-trick can be written as follows:

```
for (i in 1:n){
zeros[i] <- 0
lambda[i] <- -llik[i] + 10000
  # Since lambda[i] needs to be positive as
  #    a Poisson rate, to ensure positivity
  #    an arbitrary constant C can be added.
  # Here we added C = 10000.
zeros[i] ~ dpois(lambda[i])
```

```
  llik[i] <- ... write the log-likelihood function here
  }
```

Example 8.13. **A Zero-Trick for Maxwell.** This example finds the Bayes estimator of parameter θ in a Maxwell distribution with a density of $f(x|\theta) = \sqrt{\frac{2}{\pi}}\,\theta^{3/2}\,x^2\,e^{-\theta x^2/2}$, $x \geq 0, \theta > 0$. The moment-matching estimator and the MLE were discussed in Example 7.4. For a sample of size $n = 3$, $X_1 = 1.4$, $X_2 = 3.1$, and $X_3 = 2.5$ the MLE of θ was $\hat{\theta}_{\text{MLE}} = 0.5051$. The same estimator was found by moment-matching when the second moment was matched. The Maxwell density is not implemented in WinBUGS and we will use a zero-trick instead.

```
#Estimation of Maxwell's theta
#Using a zero-trick
model{
  for (i in 1:n){
  zeros[i] <- 0
  lambda[i] <- -llik[i] + 10000
  zeros[i] ~ dpois(lambda[i])
  llik[i] <- 1.5 * log(theta)-0.5 * theta * pow(x[i],2)
}
     theta ~ dgamma(0.1, 0.1) #non-informative choice
}
DATA
list(n=3, x=c(1.4, 3.1, 2.5))
INITS
list(theta=1)
```

	mean	sd	MC error	val2.5pc	median	val97.5pc	start	sample
theta	0.5115	0.2392	8.645E-4	0.1559	0.4748	1.079	1001	100000

Note that the Bayes estimator with respect to a vague prior `dgamma(0.1, 0.1)` is 0.5115.

Example 8.14. **Zero-Tricks for Priors.** The preceeding examples showed how to set a likelihood that is not supported in WinBUGS. Setting unsupported priors via a zero-trick is similar to setting likelihoods. Since there are no observations when setting the prior for parameter θ, we start with `theta ~ dflat()`. The rest is analogous to zero-trick construction for the likelihood.

We illustrate setting of the normal likelihood and normal prior using zero-tricks in Jeremy's IQ from Example 8.2.

```
#Jeremy with Zero-Tricks
model{
#normal likelihood
    z1  <- 0
    z1    ~ dpois(lambda1)
    #lambda1: -log(likelihood) + constant
    lambda1 <-  log(sigma) + 0.5*pow((y - theta)/sigma, 2) + 1000
    #setting normal prior
  theta        ~ dflat()
  z2 <- 0
  z2 ~ dpois(lambda2)
  #lambda2: -log(prior) + constant
  lambda2 <- log(tau) + 0.5*pow((theta-mu)/tau, 2) + 1000
}
DATA
list(y = 98, mu = 110,  sigma = 8.944272,  tau=10.954451)
INITS
list(theta=100)
```

	mean	sd	MC error	val2.5pc	median	val97.5pc	start	sample
theta	102.8	6.966	0.0436	89.19	102.7	116.5	1001	100000

Note that we added constant 1000 to both $-\log(\text{likelihood})$ and $-\log(\text{prior})$ to ensure that `lambda1` and `lambda2` are nonnegative as rates in zero-trick Poisson distributions. In this case it was not necessary to add any constants since `log(sigma)` and `log(tau)` were both positive, but care is needed if either `tau` or `sigma` is small.

8.7 Bayesian Interval Estimation: Credible Sets

The Bayesian term for an interval estimator of a parameter is *credible set.* Naturally, the measure used to assess the credibility of an interval estimator is the posterior distribution. Students learning concepts of classical confidence intervals often err by stating that "the probability that a particular confidence interval $[L, U]$ contains parameter θ is $1-\alpha$." The correct statement seems more convoluted; one generates data from the underlying model many times and, for each generated data set, calculates the confidence interval. The proportion of confidence intervals covering the unknown parameter "tends to" $1-\alpha$. The Bayesian interpretation of a credible set C is arguably more natural: the probability of a parameter belonging to set C is $1-\alpha$. A formal definition follows.

Assume that set C is a subset of parameter space Θ. Then C is a *credible set* with credibility $(1-\alpha)100\%$ if

$$\mathbb{P}(\theta \in C|X) = \mathbb{E}(I(\theta \in C)|X) = \int_C \pi(\theta|x)d\theta \geq 1-\alpha.$$

If the posterior is discrete, then the integral is a sum, and

$$\mathbb{P}(\theta \in C|X) = \sum_{\theta_i \in C} \pi(\theta_i|x) \geq 1 - \alpha.$$

This is the definition of a $(1-\alpha)100\%$ credible set. For a fixed posterior distribution and a $(1-\alpha)100\%$ *credibility*, a credible set is not unique. We will consider two versions of credible sets: highest posterior density (HPD) and equal-tail credible sets.

HPD Credible Sets. For a given credibility level $(1-\alpha)100\%$, the shortest credible set has obvious appeal. To minimize size, the sets should correspond to the highest posterior probability density areas.

Definition 8.1. The $(1-\alpha)100\%$ HPD credible set for parameter θ is a set C, a subset of parameter space Θ of the form

$$C = \{\theta \in \Theta | \pi(\theta|x) \geq k(\alpha)\},$$

where $k(\alpha)$ is the largest constant for which

$$\mathbb{P}(\theta \in C|X) \geq 1 - \alpha.$$

Geometrically, if the posterior density is cut by a horizontal line at the height $k(\alpha)$, the credible set C is the projection on the θ-axis of the part of the line that lies below the density (Fig. 8.4).

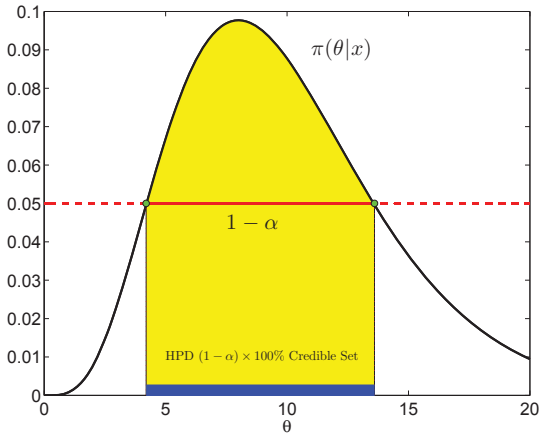


Fig. 8.4 Highest posterior density (HPD) $(1-\alpha)100\%$ credible set (*blue*). The *area in yellow* is $1-\alpha$.

Example 8.15. **Jeremy's IQ, Continued.** We are again back to Jeremy, the enthusiastic bioengineering student from Example 8.2 who used Bayesian inference in modeling his IQ test scores. For a score of X he was using a $\mathcal{N}(\theta, 80)$ likelihood, while the prior on θ was $\mathcal{N}(110, 120)$. After the score of $X = 98$ was recorded, the resulting posterior was normal $\mathcal{N}(102.8, 48)$.

Here, the MLE is $\hat{\theta} = 98$, and a 95% confidence interval is $[98 - 1.96\sqrt{80},\ 98 + 1.96\sqrt{80}] = [80.4692, 115.5308]$. The length of this interval is approximately 35. The Bayesian counterparts are $\hat{\theta} = 102.8$, and $[102.8 - 1.96\sqrt{48},\ \ 102.8 + 1.96\sqrt{48}] = [89.2207, 116.3793]$. The length of the 95% credible set is approx. 27. The Bayesian interval is shorter because the posterior variance is smaller than the likelihood variance; this is a consequence of the presence of prior information. Figure 8.5 shows the credible set (in blue) and the confidence interval (in red).

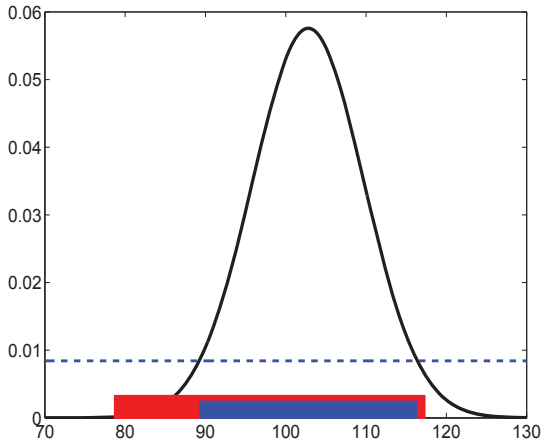


Fig. 8.5 HPD 95% credible set based on a density of $\mathcal{N}(102.8, 48)$ (*blue*). The interval in *red* is a 95% confidence interval based on the observation $X = 98$ and likelihood variance $\sigma^2 = 80$.

From the WinBUGS output table in Jeremy's IQ estimation example (page 351), the 95% credible set is $[89.18, 116.4]$.

	mean	sd	MC error	val2.5pc	median	val97.5pc	start	sample
theta	102.8	6.943	0.01991	**89.18**	102.8	**116.4**	1001	100000

Other posterior quantiles that lead to credible sets of different *credibility* levels can be specified in `Sample Monitor Tool` under `Inference>Samples` in Win-

BUGS. The credible sets from WinBUGS are HPD only if the posterior is symmetric and unimodal.

Equal-Tail Credible Sets. HPD credible sets may be difficult to find for asymmetric posterior distributions, such as gamma and Weibull, for example. Much simpler are *equal-tail credible sets* for which the tails have a probability of $\alpha/2$ each for a credibility of $1-\alpha$. An equal-tail credible set may not be the shortest set, but to find it, we need only $\alpha/2$ and $1-\alpha/2$ quantiles of the posterior. These two quantiles are the lower and upper bounds $[L,U]$:

$$\int_{-\infty}^{L} \pi(\theta|x)\,d\theta = \alpha/2, \qquad \int_{U}^{\infty} \pi(\theta|x)\,d\theta = 1-\alpha/2.$$

Note that WinBUGS gives posterior quantiles from which one can directly establish several equal-tail credible sets (95%, 90%, 80%, and 50%) by selecting appropriate pairs of percentiles in the `Sample Monitor Tool.`

Example 8.16. **Bayesian *Amanita muscaria*.** Recall that in Example 7.8 (page 302) observations were summarized by $\overline{X} = 10.098$ and $s^2 = 2.1702$, which are classical estimators of population parameters: mean μ and variance σ^2. We also obtained the 95% confidence interval for the population mean as $[9.6836, 10.5124]$ and the 90% confidence interval for the population variance as $[1.6074, 3.1213]$.

By assuming noninformative priors for the mean and variance, we use WinBUGS to find Bayesian counterparts of the estimators and confidence intervals. As we pointed out, the mean is a location parameter, and noninformative priors should be flat. WinBUGS allows for flat priors, `mu~dflat()`, but any prior with a large variance, or small precision, is a possibility. We take a normal prior with a variance of 10,000. The inverse gamma distribution is traditionally used for a prior on variance; thus, for precision as a reciprocal of variance, the gamma prior is appropriate. As we discussed earlier, gamma distributions with small parameters will have a large variance, thereby making the prior vague/noninformative. We selected `prec~dgamma(0.001, 0.001)` as a noninformative choice. This prior is noninformative because it is essentially flat; its variance is $0.001/(0.001)^2 = 1000$ (page 204). The WinBUGS program is simple:

```
model{
for ( i in 1:n ){
     amuscaria[i] ~ dnorm( mu, prec )
     }
     mu ~ dnorm(0, 0.00001)
     prec ~ dgamma(0.001, 0.001)
     sig2 <- 1/prec
}
DATA
```

```
list(n=51,amuscaria=c(10,11,12,9,10,11,13,12,10,11,11,13,9,10,
       9,10,8,12,10,11,9,10,7,11,8,9,11,11,10,12,10,8,7,11,12,
       10,9,10,11,10,8,10,10,8,9,10,13,9,12,9,9) )
INITS
list( mu =0, prec = 1 )
```

In WinBUGS' `Sample Monitor Tool` we asked for 2.5% and 97.5% posterior percentiles, which gives a 95% credible set and 5% and 95% posterior percentiles for the 90% credible set. The lower/upper bounds of the credible sets are given in boldface and the sets are [9.684,10.51] for the mean and [1.607,3.123] for the variance. The credible set for the mean is both HPD and equal-tail, but the credible set for the variance is only an equal-tail.

	mean	sd	MC error	val2.5pc	val5pc	val95pc	val97.5pc	start	sample
mu	10.1	0.2106	2.004E-4	**9.684**	9.752	10.44	**10.51**	1001	100000
prec	0.4608	0.09228	9.263E-5	0.2983	0.3202	0.6224	0.6588	1001	100000
sig2	2.261	0.472	4.716E-4	1.518	**1.607**	**3.123**	3.353	1001	100000

8.8 Learning by Bayes' Theorem

Bayesian statisticians often say: "Today's posterior is tomorrow's prior." This phrase captures the learning ability of Bayesian paradigm. As more data is acquired, Bayes' theorem updates our knowledge in a coherent manner.

We start with an example.

Example 8.17. **Leukemia Remission and 6-MP.** Freireich et al. (1963) conducted a remission maintenance therapy to compare 6-MP with placebo for prolonging the duration of remission in leukemia. From 42 patients affected with acute leukemia, but in a state of partial or complete remission, 21 pairs were formed. One randomly selected patient from each pair was assigned the maintenance treatment 6-MP, while the other patient received a placebo. Investigators monitored which patient stayed in remission longer. If that was a patient from the 6-MP treatment arm, this was recorded as a "success" (S); otherwise, it was a "failure" (F).

The results are given in the following table:

Pair	1	2	3	4	5	6	7	8	9	10
Outcome	S	F	S	S	S	F	S	S	S	S
11	12	13	14	15	16	17	18	19	20	21
S	S	S	F	S	S	S	S	S	S	S

The goal is to estimate p – the probability of success. Suppose we got information only on the first 10 subjects: 8 successes and 2 failures. When

the prior on p is uniform, and the likelihood binomial, the posterior is proportional to $p^8(1-p)^2 \times 1$, which is a beta $\mathcal{B}e(9,3)$.

Suppose now that the remaining 11 observations became available (10 successes and 1 failure). If the posterior from the first stage serves as a prior in the second stage, the updated posterior is proportional to $p^{10}(1-p)^1 \times p^8(1-p)^2$ which is a beta $\mathcal{B}e(19,4)$.

By sequentially updating the prior we arrive to the same posterior as if all observations were available at the first place (18 successes and 3 failures). With a uniform prior, this would lead to the same beta $\mathcal{B}e(19,4)$ posterior. The final posterior would be the same even if the updating was done observation by observation. This exemplifies the *learning ability* of Bayes' theorem.
✎

Suppose that observations $x_1,\ldots,x_n$ from the model $f(x|\theta)$ are available and that prior on θ is $\pi(\theta)$. Then the posterior is

$$\pi(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{\int f(\boldsymbol{x}|\theta)\pi(\theta)d\theta},$$

where $\boldsymbol{x} = (x_1,\ldots,x_n)$ and $f(\boldsymbol{x}|\theta) = \prod_{i=1}^n f(x_i|\theta)$.

Suppose an that additional observation x_{n+1} is collected. Then

$$\pi(\theta|\boldsymbol{x},x_{n+1}) = \frac{f(x_{n+1}|\theta)\pi(\theta|\boldsymbol{x})}{\int f(x_{n+1}|\theta)\pi(\theta|\boldsymbol{x})d\theta}.$$

Bayes' theorem updates inference in a natural way: the posterior based on previous observations serves as a new prior.

8.9 Bayesian Prediction

Up to now, we have been concerned with Bayesian inference about population parameters. We are often faced with the problem of predicting a new observation X_{n+1} after $X_1,\ldots,X_n$ from the same population have been observed. Assume that the prior for parameter θ is elicited. The new observation would have a likelihood of $f(x_{n+1}|\theta)$, while the observed sample $X_1,\ldots,X_n$ will lead to a posterior of θ, $\pi(\theta|X_1,\ldots,X_n)$.

Then, the *posterior predictive distribution* for X_{n+1} can be obtained from the likelihood after integrating out parameter θ using the posterior distribution,

$$f(x_{n+1}|X_1,\ldots,X_n) = \int_\Theta f(x_{n+1}|\theta)\,\pi(\theta|X_1,\ldots,X_n)\,d\theta,$$

where Θ is the domain for θ. Note that the marginal distribution also integrates out the parameter, but using the prior instead of the posterior, $m(x) = \int_\Theta f(x|\theta)\pi(\theta)\,d\theta$. For this reason, the marginal distribution is sometimes called the *prior predictive* distribution.

The prediction for X_{n+1} is the expectation $\mathbb{E}X_{n+1}$, taken with respect to the predictive distribution,

$$\hat{X}_{n+1} = \int_{\mathbb{R}} x_{n+1} f(x_{n+1}|X_1,\ldots,X_n)\,dx_{n+1},$$

while the *predictive variance*,

$$\int_{\mathbb{R}} (x_{n+1} - \hat{X}_{n+1})^2 f(x_{n+1}|X_1,\ldots,X_n)\,dx_{n+1},$$

can be used to assess the precision of the prediction.

Example 8.18. **Exponential Survival Time.** Consider the exponential distribution $\mathcal{E}(\lambda)$ for a random variable X representing survival time of patients affected by a particular disease. The density for X is $f(x|\lambda) = \lambda\exp\{-\lambda x\},\ x \geq 0$.

Suppose that the prior for λ is gamma $\mathcal{G}a(\alpha,\beta)$ with a density of $\pi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}\exp\{-\beta\lambda\},\ \lambda \geq 0$.

The likelihood, after observing a sample $X_1,\ldots,X_n$ from $\mathcal{E}(\lambda)$ population, is

$$\lambda e^{-\lambda X_1}\cdot\ \ldots\ \cdot\lambda e^{-\lambda X_n} = \lambda^n \exp\left\{-\lambda\sum_{i=1}^n X_i\right\},$$

and the posterior is proportional to

$$\lambda^{n+\alpha-1}\exp\{-(\sum_{i=1}^n X_i + \beta)\lambda\},$$

which can be recognized as a gamma $\mathcal{G}a(\alpha+n, \beta+\sum_{i=1}^n X_i)$ distribution and completed as

$$\pi(\lambda|X_1,\ldots,X_n) = \frac{(\sum_{i=1}^n X_i+\beta)^{n+\alpha}}{\Gamma(n+\alpha)}\lambda^{n+\alpha-1}\exp\{-(\sum_{i=1}^n X_i+\beta)\lambda\},\ \lambda\geq 0.$$

The predictive distribution for a new X_{n+1} is

$$f(x_{n+1}|X_1,\ldots,X_n) = \int_0^\infty \lambda \exp\{-\lambda x_{n+1}\}\, \pi(\lambda|X_1,\ldots,X_n)\, d\lambda$$
$$= \frac{(n+\alpha)(\sum_{i=1}^n X_i+\beta)^{n+\alpha}}{(\sum_{i=1}^n X_i+\beta+x_{n+1})^{n+\alpha+1}}, \qquad x_{n+1}>0.$$

Note that $X_{n+1}+\sum_{i=1}^n X_i+\beta$ is a Pareto $\mathcal{P}a(\sum_{i=1}^n X_i+\beta, n+\alpha)$, see page 212. The expected value for a new observation (a Bayesian prediction) is

$$\hat{X}_{n+1} = \int_0^\infty x_{n+1} f(x_{n+1}|X_1,\ldots,X_n)\, dx_{n+1} = \frac{\sum_{i=1}^n X_i+\beta}{n+\alpha-1}.$$

Also, the variance of the new observation is

$$\hat{\sigma}^2_{X_{n+1}} = \int_0^\infty (x_{n+1}-\hat{X}_{n+1})^2 f(x_{n+1}|X_1,\ldots,X_n)\, dx_{n+1}$$
$$= \frac{(\sum_{i=1}^n X_i+\beta)^2(n+\alpha)}{(n+\alpha-1)^2(n+\alpha-2)}.$$

For example, if $X_1=2.1$, $X_2=5.5$, $X_3=6.4$, $X_4=8.7$, $X_5=4.9$, $X_6=5.1$, and $X_7=2.3$ are the observations, and $\alpha=2$ and $\beta=1$, then $\hat{X}_8=9/2$ and $\hat{\sigma}^2_{X_8}=729/28=26.0357$. Figure 8.6 shows the posterior predictive distribution (solid blue line), observations (crosses), and prediction for the new observation (blue dot). The position of the mean of the data, $\overline{X}=5$, is shown as a dotted red line.

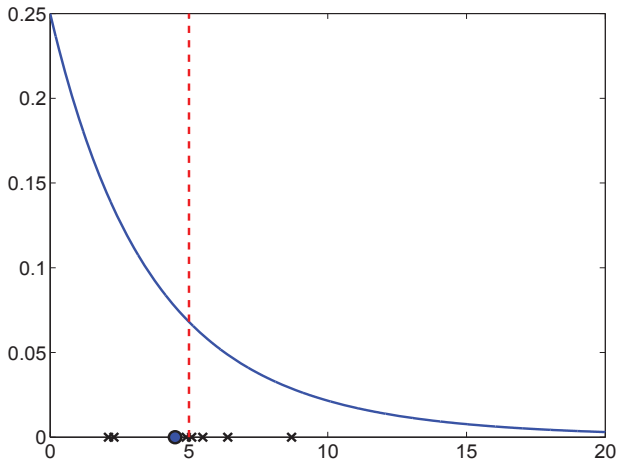


Fig. 8.6 Bayesian prediction (*blue dot*) based on the sample (*black crosses*) `X = [2.1, 5.5, 6.4, 8.7, 4.9, 5.1, 2.3]` from the exponential distribution $\mathcal{E}(\lambda)$. The parameter λ is given a gamma $\mathcal{G}a(2,1)$ distribution and the resulting posterior predictive distribution is shown as a *solid blue line*. The position of the sample mean is plotted as a *dotted red line*.

The prediction $\hat{X}_{n+1}$ can be found in an alternative manner that avoids the need for explicit posterior predictive distribution. The following holds:

$$\hat{X}_{n+1} = \int_{\Theta} \mu(\theta)\pi(\theta|X_1,\ldots,X_n)d\theta, \tag{8.6}$$

where $\mu(\theta) = \mathbb{E}^{X|\theta}X = \int xf(x|\theta)dx$ is the mean of X, as a function of the parameter.

When the parameter θ is in fact the the expectation, such as μ in $\mathcal{N}(\mu,\sigma^2)$ or λ in $\mathcal{P}oi(\lambda)$, the Bayes prediction for X_{n+1} is simply the posterior mean.

To find $\hat{X}_{n+1}$ from Example 8.18 by (8.6), note that $\mu(\lambda) = 1/\lambda$ and the posterior is $\mathcal{G}a(\alpha+n, \beta+\sum_{i=1}^n X_i)$. Thus,

$$\begin{aligned}
\hat{X}_{n+1} &= \int_0^\infty \frac{1}{\lambda} \times \frac{\lambda^{n-\alpha-1}(\beta+\sum_{i=1}^n X_i)^{n-\alpha}}{\Gamma(\alpha+n)} \exp\{-(\beta+\sum_{i=1}^n X_i)\lambda\}d\lambda \\
&= \frac{\beta+\sum_{i=1}^n X_i}{\alpha+n-1} \int_0^\infty \frac{\lambda^{(n-\alpha-1)-1}(\beta+\sum_{i=1}^n X_i)^{n-\alpha-1}}{\Gamma(\alpha+n-1)} \exp\{-(\beta+\sum_{i=1}^n X_i)\lambda\}d\lambda \\
&= \frac{\beta+\sum_{i=1}^n X_i}{\alpha+n-1},
\end{aligned}$$

after using the identity $\Gamma(a) = (a-1)\Gamma(a-1)$. To find the Bayesian prediction in WinBUGS, one simply samples a new observation from a likelihood that has updated parameters.

Example 8.19. **Predicting the Exponential.** The WinBUGS program below implements Example 8.18; the observations are read within the `for` loop. However, if a new variable is simulated from the same likelihood, this is done for the current version of the parameter λ, and the mean of simulations approximates the posterior mean of the new observation.

```
model{
for (i in 1:7){
    X[i] ~ dexp(lambda)
}
lambda ~ dgamma(2,1)
Xnew ~ dexp(lambda)
}
DATA
list(X = c(2.1,  5.5,  6.4,  8.7,  4.9,  5.1,  2.3))
INITS
list(lambda=1, Xnew=1)
```

The output is

	mean	sd	MC error	val2.5pc	median	val97.5pc	start	sample
Xnew	4.499	5.09	0.005284	0.1015	2.877	18.19	1001	100000
lambda	0.25	0.08323	8.343E-5	0.1142	0.2409	0.4378	1001	100000

Note that the posterior mean for `Xnew` is well approximated, $4.499 \approx 4.5$, and that the standard deviation `sd = 5.09` is close to $\sqrt{26.0357} = 5.1025$.

8.10 Consensus Means*

Suppose that several labs are reporting measurements of the same quantity and that a consensus mean should be calculated. This problem appears in interlaboratory studies, as well as in multicenter clinical trials and various meta-analyses. In this section we provide a Bayesian solution to this problem and compare it with some classical proposals.

Let $Y_{ij}, j = 1,\ldots,n_i;\ i = 1,\ldots,k$ be measurements made at k laboratories, where n_i measurements come from lab i. Let $n = \sum_i n_i$ be the total sample size.

We are interested in estimating the mean that would properly incorporate information coming from all the labs, called the *consensus mean.* Why is the solution not trivial, and what is wrong with the average $\overline{Y} = 1/n \sum_i \sum_j Y_{ij}$?

There is nothing wrong, under the proper conditions: (a) variabilities within the labs must be equal and (b) there must be no variability between the labs.

When (a) is relaxed, proper pooling of the lab sample means is done via a Graybill–Deal estimator:

$$\overline{Y}_{gd} = \frac{\sum_{i=1}^{k} \omega_i \overline{Y}_i}{\sum_{i=1}^{k} \omega_i}, \quad \omega_i = \frac{n_i}{s_i^2}.$$

When both conditions (a) and (b) are relaxed, there are many competing classical estimators. For example, the Schiller–Eberhardt estimator is given by

$$\overline{Y}_{se} = \frac{\sum_{i=1}^{k} \omega_i \overline{Y}_i}{\sum_{i=1}^{k} \omega_i}, \quad \omega_i = \frac{1}{s_i^2/n_i + s_b^2},$$

where s_b^2 is an estimator of the variance between the labs, $s_b^2 = \frac{(\bar{y}_{max} - \bar{y}_{min})^2}{12}$. The Mandel–Paule is the same as the Schiller–Eberhardt estimator but with s_b^2 obtained iteratively.

The Bayesian approach is conceptually simple. Individual means as random variables are generated from a single distribution. The mean of this distribution is the consensus mean. In somewhat convoluted wording, the consensus mean is the mean of a hyperprior placed on the individual means.

Example 8.20. **Selenium in Powdered Milk.** The data on selenium in nonfat milk powder `selenium.dat` are adapted from Witkovsky (2001). Four independent measurement methods are applied. The Bayes estimator of the consensus mean is 108.8.

In the WinBUGS program below, the individual means `theta[i]` have a *t*-prior with location `mu`, precision `tau`, and 5 degrees of freedom. The choice of *t*-prior, instead of the usual normal, is motivated by robustness considerations.

```
model{
for (i in 1:n)
   {
   sel[i] ~ dnorm( theta[lab[i]], prec[lab[i]])
   }
for (i in 1:k)
    {
     theta[i] ~ dt(mu, tau,5) #individual means
     prec[i] ~ dgamma(0.0001, 0.0001)
     sigma2[i] <- 1/prec[i]
     }
mu ~ dt(0,0.0001,5) #consensus mean
tau ~ dgamma(0.0001,0.0001)
si2  <-1/tau
}


DATA
list(lab=c(1,1,1,1,1,1,1,1,    2,2,2,2,2,2,2,2,2,2,2,2,
3,3,3,3,3,3,3,3,3,3,3,3,3,3,          4,4,4,4,4,4,4,4),
sel = c(
115.7, 113.5, 103.3, 119.1, 114.2, 107.3,  91.2, 104.4,
108.6, 109.1, 107.2, 111.5, 100.6, 106.3, 105.9, 109.7,
                            111.1, 107.9, 107.9, 107.9,
107.6, 107.26,109.7, 109.7, 108.5, 106.5, 110.2, 108.3,
              110.5, 108.5, 108.8, 110.1, 109.4, 112.4,
118.7, 109.7, 114.7, 105.4, 113.9, 106.3, 104.8, 106.3),
                                            k=4, n=42)

INITS
list( mu=1, tau=1, prec=c(1,1,1,1), theta=c(1,1,1,1)  )
```

	mean	sd	MC error	val2.5pc	median	val97.5pc	start	sample
mu	108.8	0.6499	0.003674	107.6	108.9	110.0	5001	500000
si2	0.7252	9.456	0.02088	1.024E-4	0.01973	4.875	5001	500000
theta[1]	108.8	0.8593	0.003803	107.0	108.9	110.5	5001	500000
theta[2]	108.7	0.6184	0.004188	107.2	108.7	109.7	5001	500000
theta[3]	108.9	0.4046	0.00311	108.1	108.9	109.7	5001	500000
theta[4]	108.9	0.7505	0.003705	107.6	108.9	110.7	5001	500000

Next, we compare the Bayesian estimator with the classical Graybill–Deal and Schiller–Eberhardt estimators, 108.8892 and 108.7703, respectively. The Bayesian estimator falls between the two classical ones. A 95% credible set for the consensus mean is [107.6, 110].

```
lab1=[115.7, 113.5, 103.3, 119.1, 114.2, 107.3,  91.2, 104.4];
lab2=[108.6, 109.1, 107.2, 111.5, 100.6, 106.3, 105.9, 109.7,...
                                  111.1, 107.9, 107.9, 107.9];
lab3=[107.6, 107.26,109.7, 109.7, 108.5, 106.5, 110.2, 108.3,...
                    110.5, 108.5, 108.8, 110.1, 109.4, 112.4];
lab4=[118.7, 109.7, 114.7, 105.4, 113.9, 106.3, 104.8, 106.3];

m = [mean(lab1)  mean(lab2)  mean(lab3)  mean(lab4)];
s = [std(lab1)   std(lab2)   std(lab3)   std(lab4) ];
ni=[8 12 14 8]; k=length(m);

%Graybill-Deal Estimator
wei = ni./s.^2; %weights
m_gd = sum(m .* wei)/sum(wei)  %108.8892

%Schiller-Eberhardt Estimator
z = sort(m);
sb2 = (z(k)-z(1))^2/12;
wei = 1./(s.^2./ni + sb2);%weights
m_se = sum(m .* wei)/sum(wei)  %108.7703
```

Borrowing Strength and Vague Priors. As popularly stated, the model in Example 8.20 allows for *borrowing strength* in the estimation of both the means θ_i and the variances σ_i^2. Even if some labs have extremely small sample sizes (as low as $n = 1$), the lab variances can be estimated through pooling via a hierarchical model structure. The prior distributions above are *vague*, which is appropriate when prior information in the form of expert opinion or historic data is not available.

Analyses conducted using vague priors can be considered objective and are generally accepted by classical statisticians. When prior information is available in the form of a mean and variance of μ, it can be included by simply changing the mean and variance of its prior, in our case the normal distribution. It is well known that Bayesian rules are sensitive with respect to changes in hyperparameters in light-tailed priors (e.g., normal priors).

If more robustness is required, a t-distribution with a small number of degrees of freedom can be substituted for the normal prior. Via MCMC sampling in WinBUGS we get a full posterior distribution of μ as the ultimate summary information.

8.11 Exercises

8.1. **Exponential Lifetimes.** A lifetime X (in years) of a particular device is modeled by an exponential distribution with unknown rate parameter θ. The lifetimes of $X_1 = 5$, $X_2 = 6$, and $X_3 = 4$ are observed. Assume that an expert familiar with this type of device suggests that θ has an exponential distribution with a mean of 3.
(a) Write down the MLE of θ for those observations.
(b) Elicit a prior according to the expert assumptions.
(c) For the prior in (b), find the posterior. Is the problem conjugate?
(d) Find the Bayes estimator $\hat{\theta}_{Bayes}$ and compare it with the MLE from (a). Discuss.
(e) Check if the following WinBUGS program gives an estimator of λ close to the Bayes estimator in (d):

```
model{
for (i in 1:n){
  X[i] ~ dexp(lambda)
 }
lambda ~ dexp(1/3)
 #note that dexp is parameterized
 #in WinBUGS by the rate parameter
}

DATA
list(n=3, X=c(5,6,4))

INITS
list(lambda=1)
```

8.2. **Fibrinogen.** Fibrinogen is a soluble plasma glycoprotein, synthesized by the liver, that is converted by thrombin into fibrin during blood coagulation. Marnie takes a blood test and finds that her level of fibrinogen is 217 mg/dL. The test results are accurate up to a random error, which is normal with mean 0 and standard deviation of 9 mg/dL.
The normal range of fibrinogen in plasma is 150–400 mg/dL, and Marnie puts a uniform prior over this range, `dunif(150, 400)`.
(a) What is the Bayes estimator of the true level of fibrinogen given this uniform prior?

(b) Copy the `Inference>Samples>stats` output from WinBUGS. What is the 95% Credible Set for the parameter from (a)?
(c) What is the classical 95% CI for the parameter from (a)? (*Hint:* Sample Size = 1, σ known.) Compare the parameter estimates and 95% CI with Bayesian counterparts.

8.3. **Uniform/Pareto.** Suppose that $X = (X_1, \ldots, X_n)$ is a sample from $\mathcal{U}(0, \theta)$. Let θ have a Pareto $\mathcal{P}a(\theta_0, \alpha)$ prior. Show that the posterior distribution is $\mathcal{P}a(\max\{\theta_0, x_1, \ldots, x_n\}\ \alpha + n)$.

8.4. **Nylon Fibers.** Refer to Exercise 5.37, where times (in hours) between blockages of the extrusion process, T, had an exponential $\mathcal{E}(\lambda)$ distribution. Suppose that the rate parameter λ is unknown, but there are three measurements of interblockage times, $T_1 = 3$, $T_2 = 13$, and $T_3 = 8$.
(a) Estimate parameter λ using the moment-matching procedure. Write down the likelihood and find the MLE.
(b) What is the Bayes estimator of λ if the prior is $\pi(\lambda) = \frac{1}{\sqrt{\lambda}}$, $\lambda > 0$.
(c) Using WinBUGS find the Bayes estimator and 95% credible set if the prior is lognormal with parameters $\mu = 10$ and $\tau = \frac{1}{\sigma^2} = 0.0001$.
Hint: In (b) the prior is not a proper distribution, but the posterior is. Identify the posterior from the product of the likelihood from (a) and the prior.

8.5. **Gamma–Inverse Gamma.** Let $X \sim \mathcal{G}a\left(\frac{n}{2}, \frac{1}{2\theta}\right)$, so that X/θ is χ^2_n. Let $\theta \sim \mathcal{IG}(\alpha, \beta)$. Show that the posterior is $\mathcal{IG}(n/2 + \alpha, x/2 + \beta)$.
Hint: The likelihood is proportional to $\frac{x^{n/2-1}}{(2\theta)^{n/2}} e^{-x/(2\theta)}$ and the prior to $\frac{\beta^\alpha}{\theta^{\alpha+1}} e^{-\beta/\theta}$. Find their product and match the distribution for θ. There is no need to find the marginal distribution and apply Bayes' theorem since the problem is conjugate.

8.6. **Normal Precision–Gamma.** Suppose $X = -2$ was observed from a population distributed as $\mathcal{N}\left(0, \frac{1}{\theta}\right)$, and an analyst wishes to estimate the parameter θ. (Here θ is the reciprocal of the variance σ^2 and is called a *precision parameter*. Precision parameters are used in WinBUGS to parameterize the normal distribution). An MLE of θ does exist, but the analyst is tempted to estimate θ as $1/\hat{\sigma}^2$, which is troublesome since there is a single observation. Suppose the analyst believes that the prior on θ is $\mathcal{G}a(1/2, 1)$.
(a) What is the MLE of θ?
(b) Find the posterior distribution and the Bayes estimator of θ. If the prior on θ is $\mathcal{G}a(r, \lambda)$, can you represent the Bayes estimator as the weighted average (sum of weights = 1) of the prior mean and the MLE?
(c) Find a 95% equal-tail credible set for θ. Use MATLAB to evaluate the quantiles of the posterior distribution.

(d) Using WinBUGS, numerically find the Bayes estimator from (b) and credible set from (c).
Hint: The likelihod is proportional to $\theta^{1/2}e^{-\theta x^2/2}$ while the prior is proportional to $\theta^{r-1}e^{-\lambda\theta}$.

8.7. **Jeremy and a Variance from a Single Observation.** Jeremy believes that his IQ test scores follow a normal distribution with mean 110 and unknown variance σ^2. He takes a test and scores $X = 98$.
(a) Show that inverse gamma prior $\mathcal{IG}(r,\lambda)$ is the conjugate for σ^2 if the observation X is normal $\mathcal{N}(\mu,\sigma^2)$ with μ known. What is the posterior?
(b) Find a Bayes estimator of σ^2 and its standard deviation in Jeremy's model if the prior on σ^2 is an inverse gamma $\mathcal{IG}(3,100)$.
(c) Use WinBUGS to solve this problem and compare the MCMC approximations with exact values from (b).
Hint: Express the likelihood terms of precision τ with gamma $\mathcal{G}a(r,\lambda)$ prior, but then calculate and monitor $\sigma^2 = \frac{1}{\tau}$. See also Exercise 8.6.

8.8. **Negative Binomial–Beta.** If $X = (X_1,\ldots,X_n)$ is a sample from $\mathcal{NB}(m,\theta)$ and $\theta \sim \mathcal{B}e(\alpha,\beta)$, show that the posterior for θ is a beta $\mathcal{B}e(\alpha + mn, \beta + \sum_{i=1}^{n} x_i)$ distribution.

8.9. **Poisson–Gamma Marginal.** In Example 8.7 on page 344, show that the marginal distribution for $\sum_{i=1}^{n} X_i$ is a generalized negative binomial, $\mathcal{NB}(\alpha,\beta/(n+\beta))$.

8.10. **Exponential–Improper.** Find Bayes' estimator for θ if a single observation X was obtained from a distribution with a density of $f(x|\theta) = \theta\exp\{-\theta x\}$, $x > 0, \theta > 0$. Assume priors (a) $\pi(\theta) = 1$ and (b) $\pi(\theta) = 1/\theta$.

8.11. **Bayes' Estimator in a Discrete Case.** Refer to the likelihood and data in Exercise 7.5.
(a) If the prior for θ is

θ	1/12	1/6	1/4
Prob	0.3	0.3	0.4

find the posterior and the Bayes estimator.
(b) What would the Bayes estimator look like for a sample of size n?

8.12. **Histocompatibility.** A patient who is waiting for an organ transplant needs a histocompatible donor who matches the patient's human leukocyte antigen (HLA) type. For a given patient, the number of matching donors per 1,000 National Blood Bank records is modeled as Poisson with an unknown rate λ. If a randomly selected group of 1,000 records showed exactly one match, estimate λ in a Bayesian fashion.
For λ assume the following:
(a) Gamma $\mathcal{G}a(2,1)$ prior.
(b) Flat prior $\lambda = 1$, for $\lambda > 0$.

(c) Invariance prior $\pi(\lambda) = \frac{1}{\lambda}$, for $\lambda > 0$.
(d) Jeffreys' prior $\pi(\lambda) = \frac{1}{\sqrt{\lambda}}$, for $\lambda > 0$.
Note that the priors in (b)–(d) are not proper densities (the integrals are not finite); nevertheless, the resulting posteriors are proper.
Hint: In all cases (a)–(d), the posterior is gamma. Write the product $\frac{\lambda^1}{1!}\exp\{-\lambda\} \times \pi(\lambda)$ and match the gamma parameters. The first part of the product is the likelihood when exactly one matching donor was observed.

8.13. **Hemocytometer Counts Revisited.** Refer to Exercise 7.36.
(a) Elicit gamma prior $\mathcal{G}a(\alpha, \beta)$ on λ for which the effective sample size (ESS) is 100 and expectation is 6. (*Hint:* ESS $= \beta$; $\mathbb{E}^{\pi}\lambda = \alpha/\beta$).
(b) For the prior in (a), find an equal-tail credible set and compare it with confidence intervals from Exercise 7.36(b).

8.14. **Neurons Fire in Potter's Lab 2.** Data set  `neuronfires.mat` consisting of 989 firing times in a cell culture of neurons was analyzed in Exercise 7.3. From this data set, the count of firings in consecutive 20-ms time intervals was recorded:

20	19	26	20	24	21	24	29	21	17
23	21	19	23	17	30	20	20	18	16
14	17	15	25	21	16	14	18	22	25
17	25	24	18	13	12	19	17	19	19
19	23	17	17	21	15	19	15	23	22

It is believed that the counts are Poisson distributed with unknown parameter λ. An expert believes that the number of counts in the 20-ms interval should be about 15.
(a) What is the likelihood function for these 50 observations?
(b) Using the information the expert provided, elicit an appropriate gamma prior. Is such a prior unique?
(c) For the prior suggested in (b), find the Bayes estimator of λ. How does this estimator compare to the MLE?
(d) Suppose now that the prior is lognormal with a mean of 15 (e.g., one possible choice is $\mu = \log(15) - 1/2 = 2.2081$ and $\sigma^2 = 1$). Using WinBUGS, find the Bayes estimator for λ. Recall that WinBUGS uses the precision parameter $\tau = 1/\sigma^2$ instead of σ^2.

8.15. **Eliciting a Beta Prior I.** This exercise is based on an example from Berry and Stangl (1996). An important prognostic factor in the early detection of breast cancer is the number of axillary lymph nodes. The surgeon will generally remove between 5 and 30 nodes during a traditional axillary dissection. We are interested in making an inference about the proportion of all nodes affected by cancer and consult the surgeon in order to elicit a prior.

The surgeon indicates that the probability of a selected node testing positive is 0.05. However, if the first node tested positive, the second will be found positive with an increased probability of 0.2.
(a) Using equations (8.5), elicit a beta prior that reflects the surgeon's opinion.
(b) If, in a particular case, two out of seven nodes tested positive, what is the Bayes estimator of the proportion of affected nodes when the prior in (a) is adopted?

8.16. **Eliciting a Beta Prior II.** A natural question for the practitioner in the elicitation of a beta prior is to specify a particular quantile. For example, we are interested in eliciting a beta prior with a mean of 0.8 such that the probability of exceeding 0.9 is 5%. Find hyperparameters a and b for such a prior. *Hint:* See file  `belicitor.m`

8.17. **Eliciting a Weibull Prior.** Assume that the average recovery time for patients with a particular disease enters a statistical model as a parameter θ and that prior $\pi(\theta)$ needs to be elicited. Assume further that the functional form of the prior is Weibull $\mathcal{W}ei(r,\lambda)$, so the elicitation amounts to specifying hyperparameters r and λ. A clinician states that the first and third quartiles for θ are $Q_1 = 10$ and $Q_3 = 20$ (in days). Elicit the prior.
Hint: The CDF for the prior is $\Pi(\theta) = 1 - e^{-\lambda\theta^r}$, which with conditions on Q_1 and Q_3 leads to two equations $-e^{-\lambda\theta^r} = 0.75$ and $e^{-\lambda\theta^r} = 0.25$. Take the log twice to obtain a system of two equations with two unknowns r and $\log\lambda$.

8.18. **Bayesian Yucatan Pigs.** Refer to Example 7.23 (Yucatan Pigs). Using WinBUGS, find the Bayesian estimator of a and plot its posterior distribution.

8.19. **Eliciting a Normal Prior.** We elicit a normal prior $\mathcal{N}(\mu,\sigma^2)$ from an expert who can specify percentiles. If the 20th and 70th percentiles are specified as 2.7 and 4.8, respectively, how should μ and σ be elicited?
Hint: If x_p is the pth quantile (100%pth percentile), then $x_p = \mu + z_p\sigma$. A system of two equations with two unknowns is formed with z_ps as `norminv(0.20) = -0.8416` and `norminv(0.70) = 0.5244`.

8.20. **Is the Cloning of Humans Moral?** A recent Gallup poll estimates that about 88% of Americans oppose human cloning. Results are based on telephone interviews with a randomly selected national sample of $n =$ 1,000 adults, aged 18 and older. In these 1,000 interviews, 882 adults opposed the cloning of humans.
(a) Write a WinBUGS program to estimate the proportion p of people opposed to human cloning. Use a noninformative prior for p.
(b) Pretend that the original poll had $n = 1{,}062$ adults, whereby results for 62 adults are missing. Estimate the number of people opposed to cloning among the 62 missing in the poll.

8.21. **Poisson Observations with Truncated Normal Rate.** A sample average of $n = 15$ counting observations was found to be $\overline{X} = 12.45$. Assume that each count comes from a Poisson $\mathcal{P}oi(\lambda)$ distribution. Using WinBUGS, find the Bayes estimator of λ if the prior on λ is a normal $\mathcal{N}(0, 10^2)$ constrained to $\lambda \geq 1$.
Hint: $n\overline{X} = \sum X_i$ is Poisson $\mathcal{P}oi(n\lambda)$.

8.22. **Counts of Alpha Particles.** In Example 7.14 we analyzed data from the experiment of Rutherford and Geiger on counting α-particles.
The counts, given in the table below, can be well modeled by a Poisson distribution.

X	0	1	2	3	4	5	6	7	8	9	10	11	≥ 12
Freq	57	203	383	525	532	408	273	139	45	27	10	4	2

(a) Find sample size n and sample mean $\overline{X}$. In calculations for $\overline{X}$, take ≥ 12 as 12.
(b) Elicit a gamma prior for λ with rate parameter $\beta = 5$ and shape parameter α selected in such a way that the prior mean is 7.
(c) Find the Bayes estimator of λ using the prior from (b). Is the problem conjugate? Use the fact that $\sum_{i=1}^{n} X_i \sim \mathcal{P}oi(n\lambda)$.
(d) Write a WinBUGS script that simulates the Bayes estimator for λ and compare its output with the analytic solution from (c).

8.23. **Credible Sets for Alpha Particles.** A Bayesian version of Garwood's interval in (7.14) is

$$\left[\frac{1}{2(n+b)}\,\chi^2_{2(S+a,\alpha/2'}\ \frac{1}{2(n+b)}\,\chi^2_{2(S+a+1),1-\alpha/2}\right].$$

when the prior on λ is gamma $\mathcal{G}a(a,b)$.
(a) For gamma prior in Exercise 8.22 (b), find the Garwood interval that represents an equal-tail credible set.
(b) Compare the result in (a) with the credible set for λ from the WinBUGS output in Exercise 8.22 (d).

8.24. **Hemocytometer Counts Revisited.** In Exercise 7.36 the Poisson rate of counts, λ, was estimated and 95% CIs were found.
(a) Elicit gamma $\mathcal{G}a(\alpha,\beta)$ prior on λ. Assume that the effective sample size EES is 20, and the prior mean is 6.
(b) Using WinBUGS and the prior from (a) find a 95% credible set for λ, and compare it to those from 7.36(b).
(c) Repeat calculations from (b) using normal $\mathcal{N}(0, 10^2)$ prior on λ, constrained to $\lambda \geq 3$. (*Hint*: `lambda ~ dnorm(0, 0.01) I(3,)`.)

8.25. **Rayleigh Estimation by Zero Trick.** Referring to Exercise 7.11, find the Bayes estimator of σ^2 in a Rayleigh distribution using WinBUGS.

Since the Rayleigh distribution is not on the list of WinBUGS distributions, you may use a Poisson zero trick with a negative log-likelihood as `negloglik[i] <- C + log(sig2) + pow(r[i],2)/(2 * sig2)`, where `sig2` is the parameter and `r[i]` are observations.
Since σ is a scale parameter, it is customary to put an inverse gamma on σ^2. This can be achieved by putting a gamma prior on $1/\sigma^2$, as in

```
sig2 <- 1/isig2
isig2~dgamma(0.1, 0.1)
```

where the choice of `dgamma(0.1, 0.1)` is noninformative.

8.26. **Jack and Jill, Poisson, and Bayes' Rule Revisited.** In Exercise 5.19 we assumed that Jack does *exactly* 40% of the work. This may be just an approximation. We could instead elicit a prior on this proportion that is beta with mean 0.4, say $p \sim \mathcal{B}e(4,6)$.
Write a WinBUGS script that will use this prior, and estimate the probabilities in Exercise 5.19 (a) and (b). Are the results close?
Hint:

```
model {
y ~ dpois(lambda)
lambda <- pages * rate[index]
index <- T + 1  #1 or 2, 1 for Jill, 2 for Jack
T ~ dbern(p)
p ~ dbeta(4,6)
rate[1] <- 1/4
rate[2] <- 1
}
```

where number of errors `y` and number of pages `pages` are inputs.

8.27. **Predictions in a Poisson/Gamma Model.** For a sample $X_1, \ldots, X_n$ from a Poisson $\mathcal{P}oi(\lambda)$ distribution and a gamma $\mathcal{G}a(\alpha,\beta)$ prior on λ,
(a) Prove that the marginal distribution is Pólya (a negative binomial with noninteger r, page 185), and identify its parameters.
(b) Show that the posterior predictive distribution for X_{n+1} is also a Pólya. Identify its parameters and find the prediction $\hat{X}_4$ for $X_1 = 4$, $X_2 = 5$, and $X_3 = 4.2$, $\alpha = 2$, and $\beta = 1$.
(c) Calculate the posterior mean for the data in (b). According to (8.6), this posterior mean is $\hat{X}_4$. Do the results from (b) and (c) agree?
(d) Support your findings in (b) and (c) with a WinBUGS simulation.

8.28. **Estimating Chemotherapy Response Rates.** An oncologist believes that 90% of cancer patients will respond to a new chemotherapy treatment and that it is unlikely that this proportion will be below 80%. Elicit a beta prior that models the oncologist's beliefs.
Hint: $\mu = 0.9$, $\mu - 2\sigma = 0.8$, and use equations (8.4).
During the trial, of the 30 patients treated, 22 responded. What are the likelihood and posterior distribution.

(a) Using MATLAB, plot the prior, likelihood, and posterior in a single figure.
(b) Using WinBUGS, find the Bayes estimator of the response rate and compare it to the posterior mean.

MATLAB AND WINBUGS FILES AND DATA SETS USED IN THIS CHAPTER

http://statbook.gatech.edu/Ch8.Bayes/

BAint.m, belicitor.m, betaplots.m, HPDfigure.m, jeremy.m, nornorplot.m, ParetoUni.m, Predictive.m, selenium.m, [dir] matbugs

coin.odc, copd.odc, ExeTransplant.odc, histocompatibility.odc, jeremy.odc|txt, jeremyminimal.odc, metalabs1.odc, metalabs2.odc, muscaria.odc, neurons.odc, pareto.odc, poistrunorm.odc, predictiveexample.odc, rayleigh.odc, rutherford.odc, selenium.odc, zerotrickjeremy.odc, ztNN.odc, ztNN1.odc, ztcoshprior.odc, ztmaxwell.odc

selenium.dat

CHAPTER REFERENCES

Anscombe, F. J. (1962). Tests of goodness of fit. *J. Roy. Stat. Soc. B*, **25**, 81–94.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. Lond.*, **53**, 370–418.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York.

Berger, J. O. and Delampady, M. (1987). Testing precise hypothesis. *Stat. Sci.*, **2**, 317–352.

Berger, J. O. and Selke, T. (1987). Testing a point null hypothesis: the irreconcilability of *p*-values and evidence (with discussion). *J. Am. Stat. Assoc.*, **82**, 112–122.

Berry, D. A. and Stangl, D. K. (1996). Bayesian methods in health-related research. In: Berry, D. A. and Stangl, D. K. (eds.). *Bayesian Biostatistics*. Dekker, New York.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer, New York.

Congdon, P. (2005). *Bayesian Models for Categorical Data*. Wiley, Hoboken, NJ.

Congdon, P. (2006). *Bayesian Statistical Modelling*, 2nd ed. Wiley, Hoboken, NJ.

Congdon, P. (2010). *Hierarchical Bayesian Modelling*. Chapman & Hall/CRC, Boca Raton, FL.

Congdon, P. (2014). *Applied Bayesian Modelling*, 2nd ed. Wiley, Hoboken, NJ.

FDA (2010). Guidance for the use of Bayesian statistics in medical device clinical trials. Center for Devices and Radiological Health Division of Biostatistics, Rockville, MD. `http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071121.pdf`

Finney, D. J. (1947). The estimation from individual records of the relationship between dose and quantal response. *Biometrika*, **34**, 320–334.

Freireich, E. J., Gehan, E., Frei, E., Schroeder, L. R., Wolman, I. J., Anbari, R., Burgert, E. O., Mills, S. D., Pinkel, D., Selawry, O. S., Moon, J. H., Gendel, B. R., Spurr, C. L., Storrs, R., Haurani, F., Hoogstraten, B., and Lee, S. (1963). The effect of 6-Mercaptopurine on the duration of steroid-induced remissions in acute leukemia: a model for evaluation of other potentially useful therapy. *Blood*, **21**, 699–716.

Garthwhite, P. H. and Dickey, J. M. (1991). An elicitation method for multiple linear regression models. *J. Behav. Decis. Mak.*, **4**, 17–31.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.*, **85**, 398–409.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, **44**, 187–192.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2013). *The BUGS Book. A Practical Introduction to Bayesian Analysis.* CRC, Boca Raton.

Martz, H. and Waller, R. (1985). *Bayesian Reliability Analysis*. Wiley, New York.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. Wiley, Hoboken, NJ.

Robert, C. (2001). *The Bayesian Choice: From Decision-Theoretic Motivations to Computational Implementation*, 2nd ed. Springer, New York.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. Springer, New York.

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *Am. Stat.*, **55**, 1, 62–71.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (1996). *BUGS Examples Volume 1*, ver. 0.5. Medical Research Council Biostatistics Unit, Cambridge, UK (PDF document).

Chapter 9
Testing Statistical Hypotheses

If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 percent point), or one in a hundred (the 1 percent point). Personally, the writer prefers to set a low standard of significance at the 5 percent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.

– Ronald Aylmer Fisher

WHAT IS COVERED IN THIS CHAPTER

- Basic Concepts in Testing: Hypotheses, Errors of the First and Second Kind, Rejection Regions, Significance Level, p-Value, Power
- Bayesian Approach to Testing
- Testing the Mean in a Normal Population: z and t Tests
- Testing the Variance in a Normal Population
- Testing the Population Proportion
- Multiple Testing, Bonferroni Correction, and False Discovery Rate

9.1 Introduction

The two main tasks of inferential statistics are parameter estimation and testing statistical hypotheses. In this chapter we will focus on the latter.

Although the expositions on estimation and testing are separate, the two inference tasks are highly related, as it is possible to conduct testing by inspecting confidence intervals or credible sets. Both tasks can be unified via the so-called decision-theoretic approach in which both the estimator and the selection of a hypothesis represent an optimal action given the model, observations, and loss function.

Generally, any claim made about one or more populations of interest constitutes a *statistical hypothesis*. These hypotheses usually involve population parameters, the nature of the population, the relationsips between the populations, and so on. For example, we could hypothesize that:

- The mean of a population, μ, is 2, or
- Two populations have the same variance, or
- A population is normally distributed, or
- The means in four populations are the same, or
- Two populations are independent.

Procedures leading to either the acceptance[1] or rejection of statistical hypotheses are called statistical tests.

We will discuss two approaches: the frequentist (classical) approach, which is based on the Neyman–Pearson lemma, and the Bayesian approach, which assigns probabilities to hypotheses directly.

The Neyman–Pearson lemma is technical (details can be found in Casella and Berger, 2001), and the testing procedure based on it will be formulated as an algorithm or a testing "recipe." In fact, this recipe is a mix of Neyman–Pearson's and Fisher's approaches since it takes the best from both: a framework for power analysis from the Neyman–Pearsonian approach and better sensitivity to the observations from the Fisherian method.

In the Bayesian framework, one simply finds and reports the probability that a particular hypothesis is true given the observations. The competing hypotheses are assigned probabilities, and those with the larger probability are favored. Frequentist tests do not assign probabilities to hypotheses directly but rather to the statistic on which the test is based. This point will be emphasized later, since p-values are often mistaken for probabilities of hypotheses.

We start by discussing the terminology and algorithm of the frequentist testing framework.

[1] The use of jargon such as *accept a hypothesis* in the testing context should be avoided. The equivalent but conservative wording for *accept* would be: *there is not enough statistical evidence to reject*. We will use the terms "reject" and "do not reject" when appropriate, leaving the careful wording to practicing statisticians who could be liable for the undesirable consequences of their straightforward recommendations.

9.2 Classical Testing Problem

9.2.1 Choice of Null Hypothesis

The usual starting point in statistical testing is the formulation of statistical hypotheses. There will be at least (in most cases, exactly) two competing hypotheses. The hypothesis that reflects the current *state of nature*, adopted standard, or believed truth is denoted by H_0 and is termed the *null hypothesis*. The competing hypothesis, H_1, is called the *alternative* or *research hypothesis*. Sometimes, the alternative hypothesis is denoted by H_a.

In the classical testing approach it is important to carefully select which of the two hypotheses is assigned to be H_0, since the subsequent testing procedure depends on this assignment. The following "rule" describes the choice of H_0 and hints at the reason why it is termed the null hypothesis.

Rule: We want to establish an assertion about a population with substantive support obtained from the data. The negation of the assertion is taken to be the *null* hypothesis H_0, and the assertion itself is taken to be the research or alternative hypothesis H_1. In this context, the term *null* can be interpreted as a void research hypothesis.

The following example illustrates several hypothetical testing scenarios.

Example 9.1. **Hypothetical Testing Scenarios.** (a) A biomedical engineer wants to determine whether a new chemical agent provides a faster reaction than the agent currently in use. The new agent is more expensive, so the engineer would not recommend it unless its faster reaction is supported by experimental evidence. The reaction times are observed in several experiments prepared with the new agent. If the reaction time is denoted by the parameter θ, then the two hypotheses can be expressed in terms of that parameter. It is assumed that the reaction speed of the currently used agent is known, $\theta = \theta_0$. Null hypothesis H_0: The new agent is not faster ($\theta = \theta_0$). Alternative hypothesis H_1: The new agent is faster ($\theta > \theta_0$).

(b) A state labor department wants to determine if the current rate of unemployment varies significantly from the forecast of 8% made 2 months ago. Null hypothesis H_0: The current rate of unemployment is 8%. Alternative hypothesis H_1: The current rate of unemployment differs from 8%.

(c) A biomedical company claims that a new treatment is more effective than the standard treatment for prolonging the lives of terminal cancer patients. The standard treatment has been in use for a long time, and from reports in medical journals, the mean survival period is known to be 5.2 years. Null hypothesis H_0: The new treatment is as effective as the standard one, that is, the survival time θ is equal to 5.2 years. Alternative

hypothesis H_1: The new treatment is more effective than the standard one, that is, $\theta > 5.2$.

(d) Katz et al. (1990) examined the performance of 28 students taking the SAT who answered multiple-choice questions without reading the referred passages. The mean score for the students was 46.6 (out of 100), with a standard deviation of 6.8. The expected score in random guessing is 20. Null hypothesis H_0: The mean score is 20. Alternative hypothesis H_1: The mean score is larger than 20.

(e) A pharmaceutical company claims that its best-selling painkiller has a mean effective period of at least 6 hours. Experimental data found that the average effective period was actually 5.3 hours. Null hypothesis H_0: The best-selling painkiller has a mean effective period of 6 hours. Alternative hypothesis H_1: The best-selling painkiller has a mean effective period of less than 6 hours.

(f) A pharmaceutical company claims that its generic drug has a mean AUC response equivalent to that of the innovative (brand name) drug. The regulatory agency considers two drugs bioequivalent if the population means in their AUC responses differ for no more than δ. Null hypothesis H_0: The difference in mean responses in AUC between the generic and innovative drugs is either smaller than $-\delta$ or larger than δ. Alternative hypothesis H_1: The absolute difference in the mean responses is smaller than δ; that is, the generic and innovative drugs are bioequivalent.

When H_0 is stated as $H_0 : \theta = \theta_0$, the alternative hypothesis can be any of

$$\theta < \theta_0, \quad \theta \neq \theta_0, \quad \theta > \theta_0.$$

The first and third alternatives are one-sided, while the middle one is two-sided. Usually, the context of the problem indicates which one-sided alternative is appropriate. For example, if the pharmaceutical industry claims that the proportion of patients allergic to a particular drug is $p = 0.01$, then either $p \neq 0.01$ or $p > 0.01$ is a sensible alternative in this context, especially if the observed proportion $\hat{p}$ exceeds 0.01.

In the context of the bioequivalence trials, the research hypothesis H_1 states that the difference between the responses is tolerable, as in (f). There $H_0 : \mu_1 - \mu_2 < -\delta$ or $\mu_1 - \mu_2 > \delta$ and the alternative is $H_1 : -\delta \leq \mu_1 - \mu_2 \leq \delta$.

9.2.2 *Test Statistic, Rejection Regions, Decisions, and Errors in Testing*

Famous and controversial Cambridge astronomer Sir Fred Hoyle (1915–2001) once said: "I don't see the logic of rejecting data just because they seem incredible." The calibration of the credibility of data is done with respect to some theory or model; instead of rejecting data, the model should be questioned.

Suppose that a hypothesis H_0 and its alternative H_1 are specified, and a random sample from the population under research is obtained. As in the estimation context, an appropriate statistic is calculated from the random sample. Testing is carried out by evaluating the realization of this statistic. If the realization appears unlikely under the assumption stipulated by H_0, H_0 is rejected, since the experimental support for H_0 is lacking.

If a null hypothesis is rejected when it is actually true, then a *type I error*, or *error of the first kind*, is committed. If, however, an incorrect null hypothesis is not rejected, then a *type II error*, or *error of the second kind*, is committed. It is customary to denote the probability of a type I error as α and the probability of a type II error as β.

This is summarized in the table below:

	Decide H_0	Decide H_1
True H_0 probability	Correct action $1-\alpha$	Type I error α
True H_1 probability	Type II error β	Correct action power $= 1-\beta$

We will also use the notation $\alpha = P(H_1|H_0)$ to denote the probability that hypothesis H_1 is decided when in fact H_0 is true. Analogously, $\beta = P(H_0|H_1)$.

A good testing procedure minimizes the probabilities of errors of the first and second kind. However, minimizing both errors simultaneously, for a fixed sample size, is impossible. Controlling the errors is a trade-off; when α decreases, β increases, and vice versa. For this and other practical reasons, α is chosen from among several typical values: 0.01, 0.05, and 0.10.

Sometimes within testing problems there is no clear dichotomy: the *established truth* versus the *research hypothesis*, and both hypotheses may seem to be research hypotheses. For instance, the statements "The new drug is safe" and "The new drug is not safe" are both research hypotheses. In such cases H_0 is selected in such a way that the type I error is more severe than the type II error. If the hypothesis "The new drug is not safe" is chosen as H_0, then the type I error (rejection of a true H_0, "use unsafe drug") is more serious (at least for the patient) than the type II error (keeping a false H_0, "do not use a safe drug").

That is another reason why α is fixed as a small number; the probability of a more serious error should be controlled. The practical motivation for

fixing a few values for α was originally the desire to keep the statistical tables needed to conduct a given test brief. This reason is now outdated since the "tables" are electronic and their brevity is not an issue.

9.2.3 Power of the Test

Recall that $\alpha = \mathbb{P}(\text{reject } H_0 | H_0 \text{ true})$ and $\beta = P(\text{reject } H_1 | H_1 \text{ true})$ are the probabilities of first- and second-type errors. For a specific alternative H_1, the probability $\mathbb{P}(\text{reject } H_0 | H_1 \text{ true})$ is the *power* of the test.

$$\text{Power} = 1 - \beta \quad (= \mathbb{P}(\text{reject } H_0 | H_1 \text{ true}))$$

In plain terms, the power is measured by the probability that the test will reject a false H_0. To find the power, the alternative must be specific. For instance, in testing $H_0 : \theta = 0$, the alternative $H_1 : \theta = 2$ is specific but $H_1 : \theta > 0$ is not. A specific alternative is needed for the evaluation of the probability $\mathbb{P}(\text{reject } H_0 | H_1 \text{ true})$. The specific null and alternative hypotheses lead to the definition of *effect size*, a quantity that researchers want to set as a sensitivity threshold for a test.

Usually, the power analysis is prospective in nature. One plans the sample size and specifies the parameters in H_0 and H_1. This allows for the calculation of an error of the second kind β and the power as $1 - \beta$. This prospective power analysis is desirable and often required. In the real world of research and drug development, for example, no regulating agency will support a proposed clinical trial if the power analysis was not addressed.

Test protocols need sufficient sample sizes for the test to be sensitive enough to discrepancies from the null hypotheses. However, the sample sizes should not be unnecessarily excessive because of financial and ethical considerations (expensive sampling, experiments that involve laboratory animals). Also, overpowered tests may detect the effects of sizes irrelevant from a clinical or engineering standpoint.

The calculation of the power after data are observed and the test was conducted, known as *retrospective power*, is controversial (Hoenig and Heisey, 2001). After the sampling is done, more information is available. If H_0 was not rejected, the researcher may be interested in knowing if the sampling protocol had enough power to detect effect sizes of interest. Inclusion of this new information in the power calculation and the perception that the goal of retrospective analysis is to justify the failure of a test to reject the null hypothesis lead to the controversy referred to earlier. Some researchers argue that retrospective power analysis should be conducted in cases where H_0 was rejected "in order not to declare H_1 true if the test was underpowered." However, this argument only emphasizes the need for

the power analysis to be done beforehand. Calculating effect sizes from the collected data may also lead to a low retrospective power of well-powered studies.

9.2.4 Fisherian Approach: p-Values

A lot of information is lost by reporting only that the null hypothesis should or should not be rejected at some significance level. Reporting a measure of support for H_0 is much more desirable. For this measure of support, the p-value is routinely reported despite controversy surrounding its meaning and use. The p-value approach was favored by Fisher, who criticized the Neyman–Pearsonian approach for reporting only a fixed probability of errors of the first kind, α, no matter how strong the evidence against H_0 was. Fisher also criticized the Neyman–Pearsonian paradigm for its need of an alternative hypothesis and for a power calculation that depends on unknown parameters.

A p-value is the probability of obtaining a value of the test statistic as extreme or more extreme (from the standpoint of the null hypothesis) than that actually obtained, given that the null hypothesis is true.

Equivalently, the p-value can be defined as the lowest significance level at which the observed statistic would be significant.

Advantage of Reporting p-Values. When a researcher reports a p-value as part of their research findings, users can judge the findings according to the significance level of their choice.

Decisions from a p-value:
- The p-value is less than α: reject H_0.
- The p-value is greater than α: do not reject H_0.

In the Fisherian approach, α is not connected to the error probability; it is a significance level against which the p-value is judged. The most frequently used value for α is 5%, though values of 1% or 10% are sometimes used as well. The recommendation of $\alpha = 0.05$ is attributed to Fisher (1926), whose "one-in-twenty" quote is provided at the beginning of this chapter. Although philosophically the p-values and error probabilities are quite different, there is a link. Since under H_0 the p-value is uniformly distributed on $[0,1]$, the probability of rejecting H_0 when $p < 0.05$ is equivalent to the statement that true H_0 was rejected with probability not exceeding 0.05.

A hypothesis may be rejected if the p-value is less than 0.05; however, a p-value of 0.049 is not the same evidence against H_0 as a p-value of 0.000001. Also, it would be incorrect to say that for any non-small p-value the null hypothesis is *accepted.* A large p-value indicates that the model stipulated under the null hypothesis is merely consistent with the observed data and that there could be many other such consistent models. Thus, the appropriate wording would be that the null hypothesis is *not rejected.* This point is further elaborated in Section 10.9.

Many researchers argue that the p-value is strongly biased against H_0 and that the evidence against H_0 derived from p-values not substantially smaller than 0.05 is rather weak. In Section 9.4 we discuss the calibration of p-values against Bayes factors and errors in testing.

The p-value is often confused with the probability of H_0, which it does not represent. As we stated, it is the probability that the test statistic will be more extreme than observed when H_0 is true. If the p-value is small, then an unlikely statistic has been observed that casts doubt on the validity of H_0.

9.3 Bayesian Approach to Testing

In frequentist tests, it was customary to formulate H_0 as $H_0 : \theta = 0$ versus $H_1 : \theta > 0$ instead of $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$, as one might expect. The reason was that we calculated the p-value under the assumption that H_0 is true, and this is why a precise null hypothesis was needed.

Bayesian testing is conceptually straightforward: The hypothesis with a higher posterior probability is favored. There is nothing special about the "null" hypothesis, and for a Bayesian, H_0 and H_1 are interchangeable.

Assume that Θ_0 and Θ_1 are two nonoverlapping sets for parameter θ. We assume that $\Theta_1 = \Theta_0^c$, although arbitrary nonintersecting sets Θ_0 and Θ_1 are easily handled. Let $\theta \in \Theta_0$ be the statement of the null hypothesis H_0 and let $\theta \in \Theta_1 = \Theta_0^c$ be the same for the alternative hypothesis H_1:

$$H_0 : \ \theta \in \Theta_0 \qquad H_1 : \ \theta \in \Theta_1.$$

Bayesian tests amount to a comparison of posterior probabilities of Θ_0 and Θ_1, the regions corresponding to the two competing hypotheses. If $\pi(\theta|x)$ is the posterior distribution, then the hypothesis corresponding to the smaller of

$$p_0 = \mathbb{P}(H_0|X) = \int_{\Theta_0} \pi(\theta|x)d\theta,$$
$$p_1 = \mathbb{P}(H_1|X) = \int_{\Theta_1} \pi(\theta|x)d\theta,$$

is rejected. Here $\mathbb{P}(H_i|X)$ is the notation for the posterior probability of hypothesis H_i, $i = 0,1$.

Conceptually, this approach differs from frequentist testing, where the p-value measures the agreement of data with the model postulated by H_0, but not the probability of H_0.

Example 9.2. **A Bayesian Test for Jeremy's IQ.** We return to Jeremy (Examples 8.2 and 8.10) and consider the posterior for the parameter θ, $\mathcal{N}(102.8, 48)$. Jeremy claims he had a bad day, and his true IQ is at least 105. The posterior probability of $\theta \geq 105$ is

$$p_0 = \mathbb{P}(\theta \geq 105|X) = \mathbb{P}\left(Z \geq \frac{105 - 102.8}{\sqrt{48}}\right) = 1 - \Phi(0.3175) = 0.3754,$$

less than $1/2$, so his claim is rejected in favor of $\theta < 105$.
✎

We represent the prior and posterior odds in favor of the hypothesis H_0, respectively, as

$$\frac{\pi_0}{\pi_1} = \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)} \quad \text{and} \quad \frac{p_0}{p_1} = \frac{\mathbb{P}(H_0|X)}{\mathbb{P}(H_1|X)}.$$

The *Bayes factor* in favor of H_0 is the ratio of the corresponding posterior to prior odds:

$$B_{01}^{\pi}(x) = \frac{\mathbb{P}(H_0|X)}{\mathbb{P}(H_1|X)} \Big/ \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)} = \frac{p_0/p_1}{\pi_0/\pi_1}. \tag{9.1}$$

In the context of Bayes' rule in Chapter 3 we discussed the Bayes factor (page 103). Its meaning here is analogous: the Bayes factor updates the prior odds of hypotheses to their posterior odds, after an experiment was conducted.

Example 9.3. **Jeremy Continued.** In the context of Example 9.2, the posterior odds in favor of H_0 are $\frac{0.3754}{1-0.3754} = 0.4652$, less than 1.
✎

When the hypotheses are simple (i.e., $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$) and the prior is just the two-point distribution $\pi(\theta_0) = \pi_0$ and $\pi(\theta_1) = \pi_1 = 1 - \pi_0$, then the Bayes factor in favor of H_0 becomes the likelihood ratio:

$$B_{01}^{\pi}(x) = \frac{\mathbb{P}(H_0|X)}{\mathbb{P}(H_1|X)} \Big/ \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)} = \frac{f(x|\theta_0)\pi_0}{f(x|\theta_1)\pi_1} \Big/ \frac{\pi_0}{\pi_1} = \frac{f(x|\theta_0)}{f(x|\theta_1)}.$$

If the prior is a mixture of two priors, ξ_0 under H_0 and ξ_1 under H_1, then the Bayes factor is the ratio of two marginal (prior-predictive) distributions generated by ξ_0 and ξ_1. Thus, if $\pi(\theta) = \pi_0\xi_0(\theta)\mathbf{1}(\theta \in \Theta_0) + \pi_1\xi_1(\theta)\mathbf{1}(\theta \in \Theta_1)$, then

$$B_{01}^{\pi}(x) = \frac{\frac{\int_{\Theta_0} f(x|\theta)\pi_0\xi_0(\theta)d\theta}{\int_{\Theta_1} f(x|\theta)\pi_1\xi_1(\theta)d\theta}}{\frac{\pi_0}{\pi_1}} = \frac{m_0(x)}{m_1(x)}.$$

As noted earlier, the Bayes factor measures the relative change in prior odds once the evidence is collected. Table 9.1 offers practical guidelines for Bayesian testing of hypotheses depending on the value of the log-Bayes factor (Jeffreys, 1961, Appendix B). One could use $B_{01}^{\pi}(x)$, but then $a < \log B_{10}(x) \leq b$ becomes $-b \leq \log B_{01}(x) < -a$. Negative values of the log-Bayes factor are handled by using symmetry and appropriately changed wording.

Table 9.1 Treatment of H_0 according to log-Bayes factor values: Jeffreys' scale (Jeffreys, 1961, page 432)

Value (log 10)	Evidence against H_0 is
$0 \leq \log_{10} B_{10}(x) \leq 0.5$	**Poor**
$0.5 < \log_{10} B_{10}(x) \leq 1$	**Substantial**
$1 < \log_{10} B_{10}(x) \leq 1.5$	**Strong**
$1.5 < \log_{10} B_{10}(x) \leq 2$	**Very strong**
$\log_{10} B_{10}(x) > 2$	**Decisive**

Suppose $X|\theta \sim f(x|\theta)$ is observed and we are interested in testing

$$H_0 : \theta = \theta_0 \quad v.s. \quad H_1 : \theta <, \neq, > \theta_0.$$

If the priors on θ are continuous distributions, Bayesian testing of precise hypotheses in the manner we just discussed is impossible. With continuous priors, and subsequently continuous posteriors, the probability of a singleton $\theta = \theta_0$ is always 0, and the precise hypothesis is always rejected.

The Bayesian solution is to adopt a prior where singleton θ_0 has a probability of π_0 and the rest of the probability is spread on $\Theta\backslash\{\theta_0\}$ by a distribution $\xi(\theta)$ that is the prior under H_1. Thus, the prior on θ is a mixture of the point mass at θ_0 with a weight π_0 and a continuous density $\xi(\theta)$ on $\Theta\backslash\{\theta_0\}$, with a weight of $\pi_1 = 1 - \pi_0$. One can show that the marginal density for X is

$$m(x) = \pi_0 f(x|\theta_0) + \pi_1 m_1(x),$$

where

$$m_1(x) = \int_{\theta\in\Theta\backslash\{\theta_0\}} f(x|\theta)\xi(\theta)d\theta. \tag{9.2}$$

The posterior probability of the null hypothesis uses this marginal distribution and is equal to

$$\begin{aligned}\pi(\theta_0|x) &= \frac{f(x|\theta_0)\pi_0}{m(x)} = \frac{\pi_0 f(x|\theta_0)}{\pi_0 f(x|\theta_0) + \pi_1 m_1(x)} \\ &= \left(1 + \frac{\pi_1}{\pi_0}\cdot\frac{m_1(x)}{f(x|\theta_0)}\right)^{-1}. \end{aligned} \tag{9.3}$$

Example 9.4. **Improvement of Surgical Procedure.** In a disease in which the postoperative mortality is usually 10%, a surgeon devises a novel surgical technique. He implements the technique on 15 patients and has no fatalities.

A Bayesian wants to test a precise null hypothesis

$$H_0 : \theta = 0.1 \quad \text{versus} \quad H_1 : \theta < 0.1.$$

and adopts prior

$$\pi(\theta) = \pi_0 \cdot \mathbf{1}(\theta = 0.1) + \pi_1 \cdot 10 \cdot \mathbf{1}(0 \le \theta < 0.1),$$

with equal prior probabilities of the hypotheses $\pi_0 = \pi_1 = 1/2$. What is the posterior probability of H_0? What is the Bayes factor B_{01}?

Here, the number of fatalities is binomial $\mathcal{B}in(15,\theta)$, the observed number of fatalities is $x = 0$, $\theta_0 = 0.1$, the likelihood is $f(x|\theta) = \binom{15}{x}\theta^x(1-\theta)^{15-x}$, and ξ from (9.2) is uniform on $[0,0.1)$. Note also that the parameter space is $\Theta = [0,1]$ and that $\Theta_0 = \{0.1\}$ and $\Theta_1 = [0,0.1)$. Then,

$$m_1(0) = \int_0^{0.1} \binom{15}{0}\theta^0(1-\theta)^{15}\cdot 10\, d\theta = 10/16\cdot(1-0.9^{16}) = 0.5092,$$

and by (9.3)

$$\pi(\theta_0|x) = \left[1 + \frac{0.509186}{0.1^0(1-0.1)^{15}}\right]^{-1} = 0.2879.$$

Since $\pi_0/\pi_1 = 1$, the Bayes factor $B_{01} = p_0/p_1 = 0.2879/(1-0.2879) = 0.4043$. The logarithm for basis 10 of B_{01} is approximately -0.39, that is, $\log_{10} B_{10} = 0.39$. Thus, the evidence against H_0 is poor (Table 9.1), or as Jeffreys (1961) phrases it: "not worth more than a bare mention."

The surgeon's claim is not substantiated by the evidence. Even if one finds the exact frequentist p-value, which in this case is $\mathbb{P}(X \leq 0) = 0.9^{15} = 0.2059$ (see Exercise 9.23), the null hypothesis is not rejected at any reasonable significance level.

There is an alternate way of testing the precise null hypothesis in a Bayesian fashion. One could test the hypothesis $H_0 : \theta = \theta_0$ against the two-sided alternative by credible sets for θ. If θ_0 belongs to a 95% credible set for θ, then H_0 is not rejected. One-sided alternatives can be accommodated as well by one-sided credible sets. This approach is natural and mimics testing by confidence intervals; however, the posterior probabilities of hypotheses are not calculated.

Testing Using WinBUGS. WinBUGS generates samples from the posterior distribution. Testing hypotheses is equivalent to finding the relative frequencies of a posterior sample falling in competing regions Θ_0 and Θ_1. For example, if

$$H_0 : \theta \leq 1 \quad \text{versus} \quad H_1 : \theta > 1$$

is tested in the WinBUGS program, the command `ph1<-step(theta-1)` will calculate the proportion of the simulated chain falling in Θ_1, that is, satisfying $\theta > 1$. The `step(x)` is equal to 1 if $x \geq 0$ and 0 if $x < 0$.

9.4 Criticism and Calibration of p-Values*

In a provocative article, Ioannidis (2005) states that *many published research findings are false* because statistical significance by a particular team of researchers is found. Ioannidis lists several reasons: "...a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser pre-selection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are in-

volved in a scientific field in case of statistical significance." Certainly great responsibility for an easy acceptance of research (alternative) hypotheses can be attributed to the p-values. There are many objections to the use of raw p-values for testing purposes.

Since the p-value is the probability of obtaining the statistic as large or more extreme than observed, when H_0 is true, the p-values measure how consistent the data are with H_0, and they may not be a measure of support for a particular H_0.

Misinterpretation of the p-value as the error probability leads to a strong bias against H_0. What is the posterior probability of H_0 in a test for which the reported p-value is p? Berger and Sellke (1987) and Sellke et al. (2001) show that the minimum Bayes factor (in favor of H_0) for a null hypothesis having a p-value of p is $-e\ p\log p$. The Bayes factor transforms the prior odds π_0/π_1 into the posterior odds p_0/p_1, and if the prior odds are 1 (H_0 and H_1 equally likely *a priori*, $\pi_0 = \pi_1 = 1/2$), then the posterior odds of H_0 are not smaller than $-e\ p\log p$ for $p < 1/e \approx 0.368$:

$$\frac{p_0}{p_1} \geq -ep\log p, \quad p < 1/e,\ p_0 + p_1 = 1.$$

By solving this inequality with respect to p_0, we obtain a posterior probability of H_0 as

$$p_0 \geq \frac{1}{1 + (-e\ p\log p)^{-1}},$$

which also has a frequentist interpretation as a type I error, $\alpha(p)$. Now, the effect of bias against H_0, when judged by the p-value, is clearly visible. The type I error, α, always exceeds $(1 + (-e\ p\log p)^{-1})^{-1}$.

It is instructive to look at specific numbers. Assume that a particular test yielded a p-value of 0.01, which led to the rejection of H_0 with decisive evidence. However, if *a priori* we do not have a preference for either H_0 or H_1, the posterior odds of H_0 always exceed 12.53%. The frequentist type I error or, equivalently, the posterior probability of H_0 is never smaller than 11.13% – certainly not strong evidence against H_0.

Figure 9.1 (generated by SBB.m) compares a p-value (dotted line) with a lower bound on the Bayes factor (red line) and a lower bound on the probability of a type I error α (blue line).

```
%SBB.m
sbb = @(p)  - exp(1) * p .* log(p);
alph = @(p) 1./(1 + 1./(-exp(1)*p.*log(p)) );
 %
pp = 0.0001:0.001:0.15
plot(pp, pp, ':', 'linewidth',lw)
hold on
```

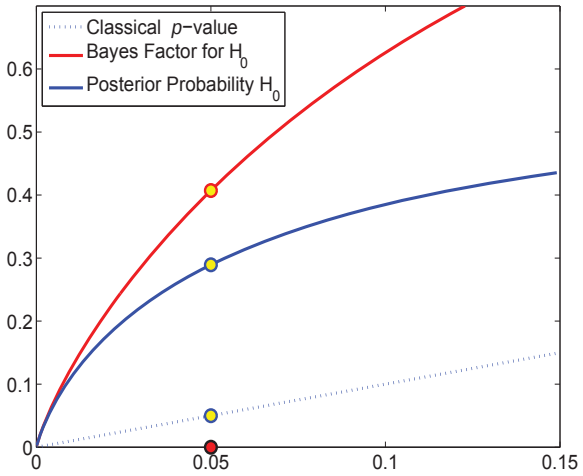


Fig. 9.1 Calibration of p-values. A p-value (*dotted line*) is compared with a lower bound on the Bayes factor (*red line*) and a lower bound on the frequentist type I error α (*blue line*). The bound on α is also the lower bound on the posterior probability of H_0 when the prior probabilities for H_0 and H_1 are equal. For the p-value of 0.05, the type I error is never smaller than 0.2893, while the Bayes factor in favor of H_0 is never smaller than 0.4072.

```
plot(pp, sbb(pp), 'r-','linewidth',lw)
plot(pp, alph(pp), '-','linewidth',lw)
```

The interested reader is directed to Berger and Sellke (1987), Schervish (1996), and Goodman (1999a,b, 2001), among many others, for a constructive criticism of p-values.

We start description of some important testing procedures by first discussing testing for the normal mean.

9.5 Testing the Normal Mean

Testing the normal mean is arguably the most important and fundamental statistical test. In this testing, we will distinguish between two cases depending on whether the population variance is known in advance (z-test) or not known (t-test). We will start with the case of known variance. Scenarios in which the population mean is unknown but the population variance would be known are not common, but not unrealistic. For example, a particular measuring equipment generating data has well-known precision characteristics specified by the factory but is not well calibrated.

9.5.1 z-Test

Let us assume that we are interested in testing the null hypothesis $H_0 : \mu = \mu_0$ on the basis of a sample $X_1, \ldots, X_n$ from a normal distribution $\mathcal{N}(\mu, \sigma^2)$, where the variance σ^2 is assumed known.

We know (page 248) that $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ and that $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ has the standard normal distribution if the null hypothesis is true, that is, if $\mu = \mu_0$. This statistic, Z, is used to test H_0, and the test is called a z-test. Statistic Z is compared to quantiles of the standard normal distribution.

The test can be performed using either (i) the rejection region or (ii) the p-value.

(i) The rejection region depends on the level α and the alternative hypothesis. For one-sided hypotheses, the tail of the rejection region follows the direction of H_1. For example, if $H_1 : \mu > 2$ and the level α is fixed, the rejection region is $[z_{1-\alpha}, \infty)$. For the two-sided alternative hypothesis $H_1 : \mu \neq \mu_0$ and significance level of α, the rejection region is two-sided, $(-\infty, z_{\alpha/2}] \cup [z_{1-\alpha/2}, \infty)$. Since the standard normal distribution is symmetric about 0 and $z_{\alpha/2} = -z_{1-\alpha/2}$, the two-sided rejection region is sometimes given as $(-\infty, -z_{1-\alpha/2}] \cup [z_{1-\alpha/2}, \infty)$.

The test is now straightforward. If statistic Z, calculated from the observations $X_1, \ldots, X_n$, falls within the rejection region, the null hypothesis is rejected. Otherwise, we say that hypothesis H_0 is not rejected.

(ii) As discussed earlier, the p-value gives a more refined analysis in testing than the "reject–do not reject" decision rule. The p-value is the probability of the rejection-region-like area cut by the observed Z (and, in the case of a two-sided alternative, by $-Z$ and Z) where the probability is calculated by the distribution specified by the null hypothesis.

The following table summarizes the z-test for $H_0 : \mu = \mu_0$ and $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$:

Alternative	α-level rejection region	p-value (MATLAB)
$H_1 : \mu > \mu_0$	$[z_{1-\alpha}, \infty)$	`1-normcdf(z)`
$H_1 : \mu \neq \mu_0$	$(-\infty, z_{\alpha/2}] \cup [z_{1-\alpha/2}, \infty)$	`2*normcdf(-abs(z))`
$H_1 : \mu < \mu_0$	$(-\infty, z_{\alpha}]$	`normcdf(z)`

9.5.2 Power Analysis of a z-Test

The power of a test is found against a specific alternative, $H_1 : \mu = \mu_1$. In a z-test, the variance σ^2 is known and μ_0 and μ_1 are specified by their respective H_0 and H_1.

The power is the probability that a z-test of level α will detect the effect of size e and, thus, reject H_0. The effect size is defined as $e = \frac{|\mu_0 - \mu_1|}{\sigma}$. Usually, μ_1 is selected such that effect e has a medical or engineering relevance.

Power of the z-test for $H_0 : \mu = \mu_0$ when μ_1 is the actual mean.
• One-sided test:

$$1 - \beta = \Phi\left(z_\alpha + \frac{|\mu_0 - \mu_1|}{\sigma/\sqrt{n}}\right) = \Phi\left(-z_{1-\alpha} + \frac{|\mu_0 - \mu_1|}{\sigma/\sqrt{n}}\right).$$

• Two-sided test:

$$\begin{aligned} 1 - \beta &= \Phi\left(-z_{1-\alpha/2} + \frac{(\mu_0 - \mu_1)}{\sigma/\sqrt{n}}\right) + \Phi\left(-z_{1-\alpha/2} + \frac{(\mu_1 - \mu_0)}{\sigma/\sqrt{n}}\right) \\ &\approx \Phi\left(-z_{1-\alpha/2} + \frac{|\mu_0 - \mu_1|}{\sigma/\sqrt{n}}\right). \end{aligned}$$

Typically the sample size is selected prior to the experiment. For example, it may be of interest to decide how many respondents to interview in a poll or how many tissue samples to process. We already selected sample sizes in the context of interval estimation to achieve a given interval size and confidence level.

In a testing setup, consider a problem of testing $H_0 : \mu = \mu_0$ using $\overline{X}$ from a sample of size n. Let the alternative have a specific value μ_1, i.e., $H_1 : \mu = \mu_1 (> \mu_0)$. Assume a significance level of $\alpha = 0.05$. How large should n be so that the power $1 - \beta$ is 0.90?

Recall that the power of a test is the probability that a false null will be rejected, $\mathbb{P}(\text{reject } H_0 | H_0 \text{ false})$. The null is rejected when $\overline{X} > \mu_0 + 1.645 \cdot \frac{\sigma}{\sqrt{n}}$. We want the power of 0.90 leading to $\mathbb{P}(\overline{X} > \mu_0 + 1.645 \cdot \frac{\sigma}{\sqrt{n}} | \mu = \mu_1) = 0.90$, that is,

$$\mathbb{P}\left(\frac{\overline{X} - \mu_1}{\sigma/\sqrt{n}} > \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + 1.645\right) = 0.9.$$

Since $\mathbb{P}(Z > -1.282) = 0.9$, it follows that $\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} = 1.282 - 1.645 \Rightarrow n = \frac{8.567 \cdot \sigma^2}{(\mu_1 - \mu_0)^2}$.

In general terms, if we want to achieve the power $1 - \beta$ within the significance level of α for the alternative $\mu = \mu_1$, we need $n \geq \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2}{(\mu_0 - \mu_1)^2}$ observations. For two-sided alternatives α is replaced by $\alpha/2$.

The sample size for fixed α, β, σ, μ_0, and μ_1 is

$$n = \frac{\sigma^2}{(\mu_0 - \mu_1)^2}(z_{1-\alpha} + z_{1-\beta})^2,$$

where σ is either known, estimated from a pilot experiment, or elicited from experts. If the alternative is two-sided, then $z_{1-\alpha}$ is replaced by $z_{1-\alpha/2}$. In this case, the sample size is approximate.

If σ is not known and no estimate exists, one can elicit the *effect size*, $e = |\mu_0 - \mu_1|/\sigma$, directly. This number is the distance between the competing means in units of σ. For example, for $e = 1/2$ we would like to find a sample size such that the difference between the true and postulated mean equal to $\sigma/2$ is detectable with a probability of $1 - \beta$.

9.5.3 Testing a Normal Mean When the Variance Is Not Known: t-Test

To test a normal mean when the population variance is unknown, we use the t-test. We are interested in testing the null hypothesis $H_0 : \mu = \mu_0$ against one of the alternatives $H_1 : \mu >, \neq, < \mu_0$ on the basis of a sample $X_1, \ldots, X_n$ from the normal distribution $\mathcal{N}(\mu, \sigma^2)$, where the variance σ^2 is unknown.

If $\overline{X}$ and s are the sample mean and standard deviation, then under H_0, which states that the true mean is μ_0, the statistic $t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$ has a t-distribution with $n - 1$ degrees of freedom; see arguments on page 297.

The test can be performed either using (i) the rejection region or (ii) the p-value. The following table summarizes the test.

Alternative	α-level rejection region	p-value (MATLAB)
$H_1 : \mu > \mu_0$	$[t_{n-1,1-\alpha}, \infty)$	`1-tcdf(t, n-1)`
$H_1 : \mu \neq \mu_0$	$(-\infty, t_{n-1,\alpha/2}] \cup [t_{n-1,1-\alpha/2}, \infty)$	`2*tcdf(-abs(t),n-1)`
$H_1 : \mu < \mu_0$	$(-\infty, t_{n-1,\alpha}]$	`tcdf(t, n-1)`

It is sometimes argued that the z-test and the t-test are an unnecessary dichotomy and that only the t-test should be used. The population variance in the z-test is assumed "known," but this can be too strong an assumption. Most of the time when μ is not known, it is unlikely that the researcher would have definite knowledge about the population variance. Also, the t-test is more conservative and robust to deviations from normality than

the z-test. However, the z-test has an educational value since the testing process and power analysis are easily formulated and explained. Moreover, when the sample size is large, say, larger than 100, the $z-$ and t-tests are practically indistinguishable, due to the Central Limit Theorem.

Example 9.5. **The Moon Illusion.**  Kaufman and Rock (1962) stated that the commonly observed fact that the moon near the horizon appears larger than does the moon at its zenith (highest point overhead) could be explained on the basis of the greater *apparent* distance of the moon when at the horizon. The authors devised an apparatus that allowed them to present two artificial moons, one at the horizon and one at the zenith. Subjects were asked to adjust the variable horizon moon to match the size of the zenith moon, and vice versa. For each subject the ratio of the perceived size of the horizon moon to the perceived size of the zenith moon was recorded. A ratio of 1.00 would indicate no illusion, whereas a ratio other than 1.00 would represent an illusion. For example, a ratio of 1.50 would mean that the horizon moon appeared to have a diameter 1.50 times that of the zenith moon. Evidence in support of an illusion would require that we reject $H_0 : \mu = 1.00$ in favor of $H_1 : \mu > 1.00$.

Obtained ratio: 1.73 1.06 2.03 1.40 0.95 1.13 1.41 1.73 1.63 1.56

For these data,

```
x = [1.73, 1.06, 2.03, 1.40, 0.95, 1.13, 1.41, 1.73, 1.63, 1.56];
n = length(x)
t = (mean(x)-1)/(std(x)/sqrt(n))
  % t= 4.2976
crit = tinv(1-0.05, n-1)
  % crit=1.8331. RR = (1.8331, infinity)
pval = 1-tcdf(t, n-1)
  % pval = 9.9885e-004 < 0.05
```

As evident from the MATLAB output, the data do not support H_0, and H_0 is rejected.

A Bayesian solution implemented in WinBUGS is provided next. Each parameter in a Bayesian model should be assigned a prior distribution. Here we have two parameters, the mean μ, which is the population ratio, and σ^2, the unknown variance. The prior on μ is normal with mean 0 and variance $1/0.00001 = 100{,}000$. We also restricted the prior to be on the nonnegative domain (since negative ratios are not possible) by WinBUGS option `mu~dnorm(0,0.00001)I(0,).` Such a large variance makes the normal prior essentially flat over $\mu \geq 0$. This means that our prior opinion on μ is vague, and the adopted prior is noninformative.

The prior on the precision, $1/\sigma^2$, is gamma with parameters 0.0001 and 0.0001. As we argued in Example 8.16, this selection of hyperparameters

makes the gamma prior essentially flat, and we are not injecting any prior information about the variance.

```
model{
for (i in 1:n){
X[i] ~ dnorm(mu, prec)
}
 mu ~ dnorm(0, 0.00001) I(0, )
 prec ~ dgamma(0.0001, 0.0001)
 sigma <- 1/sqrt(prec)
 #TEST
 prH1 <- step(mu - 1)
 }
 DATA
 list(n=10, X=c(1.73, 1.06, 2.03, 1.40, 0.95,
              1.13, 1.41, 1.73, 1.63, 1.56) )

 INITS
 list(mu = 0, prec = 1)
```

	mean	sd	MC error	val2.5pc	median	val97.5pc	start	sample
mu	1.463	0.1219	1.26E-4	1.219	1.463	1.707	1001	100000
prH1	0.999	0.03115	3.188E-5	1.0	1.0	1.0	1001	100000
sigma	0.3727	0.101	1.14E-4	0.2344	0.354	0.6207	1001	100000

Note that the MCMC output in the previous example produced $\mathbb{P}(H_0) = 0.001$ and $\mathbb{P}(H_1) = 0.999$ and the Bayesian solution agrees with the classical. Moreover, the posterior probability of hypothesis H_0 of 0.001 is quite close to the p-value of 0.000998, which is often the case when the priors in the Bayesian model are noninformative. Note also that posterior probability of H_1 was estimated by the relative frequency of `step(mu-1)`, that is, by the proportion of cases in which `mu-1` resulted as positive in MCMC simulations.

Example 9.6. **Hypersplenism and White Blood Cell Count.** Hypersplenism is a disorder that causes the spleen to rapidly and prematurely destroy blood cells. In the general population the count of white blood cells per mm^3 is normal with a mean of 7,200 and standard deviation of $\sigma = 1{,}500$.

It is believed that hypersplenism decreases the leukocyte count. In a sample of 16 persons affected by hypersplenism, the mean white blood cell count was found to be $\overline{X} = 5{,}213$. The sample standard deviation was $s = 1{,}682$.

Using WinBUGS, find the posterior probability of H_1 and estimate the mean and variance in the affected population. The program in WinBUGS will operate on the summaries $\overline{X}$ and s since the original data are not available. The sample mean is normal and the precision (reciprocal of the variance) of the mean is n times the precision of a single observation. In this

case, knowledge of the population standard deviation σ will guide the setting of an informative prior on the precision. To keep the numbers manageable, we will express the counts in 1,000's, and $\overline{X}$ and s will be coded as 5.213 and 1.682, respectively. Since $s = 1.682, s^2 = 2.8291$, and $prec = 0.3535$, it is tempting to set the prior on the precision as `precx~dgamma(0.3535,1)` or `precx~dgamma(3.535,10)` since the mean of these priors will match the observed precision. However, this would be a "data-built" prior in the spirit of the empirical Bayes approach. We will use the fact that in the population σ was 1.5 and we will elicit the prior `precx~dgamma(4.444,10)` since $1/1.5^2 = 0.4444$.

```
model {
precxbar <- n * precx
xbar ~ dnorm(mu, precxbar)
mu ~   dnorm(0, 0.0001) I(0, )
   # sigma = 1.5, s^2 = 2.25, prec = 0.4444
   # X gamma(a,b) -> EX=a/b, Var X = a/b^2
precx ~ dgamma(4.444, 10 )
indh1 <- step(7.2 - mu)
sigx <- 1/sqrt(precx)
}


DATA
list(xbar = 5.213, n=16)

INITS
list(mu=1.000, precx=1.000)
```

	mean	sd	MC error	val2.5pc	median	val97.5pc	start	sample
indh1	0.9997	0.01643	3.727E-5	1.0	1.0	1.0	1001	200000
mu	5.212	0.4263	9.842E-4	4.367	5.212	6.064	1001	200000
sigx	1.644	0.4486	0.001081	1.032	1.561	2.749	1001	200000

Note that the posterior probability of H_1 is 0.9997 and this hypothesis is a clear winner.

9.5.4 *Power Analysis of a t-Test*

When an experiment is planned, the data are not available. Even if the variance is unknown, as in the case of a t-test, it would be elicited. Alternatively, the absolute difference $|\mu_0 - \mu_1|$ that we want to consider as significant can be expressed in units of standard deviation, so an explicit knowledge of σ may not be necessary. Thus, at the pre-experimental stage, the power analysis applicable to the z-test is also applicable to the prospective t-test.

Once the data are available and the test is performed, the sample mean and sample variance are available, and it becomes possible to assess the power retrospectively. We have already discussed controversies surrounding retrospective power analyses.

In a retrospective evaluation of the power, it is not recommended to replace $|\mu_0 - \mu_1|$ by $|\mu_0 - \overline{X}|$, as is sometimes done, but to simply update the elicited σ^2 with the observed variance. When σ is replaced by s, the expressions for calculating the power involve t and noncentral t-distributions. Here is an illustration.

Example 9.7. **Power in the t-Test.** Suppose that we are testing $H_0 : \mu = 10$ versus $H_1 : \mu > 10$, at a level $\alpha = 0.05$. A sample of size $n = 20$ gives $\overline{X} = 12$ and $s = 5$. We are interested in finding the power of the test against the alternative $H_1 : \mu = 13$.

The exact power is $\mathbb{P}(t \in RR | t \sim nct(df = n-1, ncp = (\mu_1 - \mu_0)\sqrt{n}/\sigma))$, since under H_1, t has a noncentral t-distribution with $n-1$ degrees of freedom and a noncentrality parameter $\frac{(\mu_1 - \mu_0)\sqrt{n}}{\sigma}$. "RR" denotes the rejection region.

```
n=20;  mu0 = 10; s=5; mu1= 13; alpha=0.05;
pow1 = nctcdf( -tinv(1-alpha, n-1), n-1,-abs(mu1-mu0)*sqrt(n)/s)
% or pow1=1-nctcdf(tinv(1-alpha, n-1),n-1,abs(mu1-mu0)*sqrt(n)/s)
% pow1 = 0.8266
%
pow = normcdf(-norminv(1-alpha) + abs(mu1-mu0)*sqrt(n)/s)
% or pow = 1-normcdf(norminv(1-alpha)-abs(mu1-mu0)*sqrt(n)/s)
% pow = 0.8505
```

For a large sample size, the power calculated as in the z-test approximates the exact power, but from the "optimistic" side, that is, by always overestimating it. In this MATLAB script we find a power of approx. 85%, which in an exact calculation (as above) drops to 82.66%.

For the two-sided alternative $H_1 : \mu \neq 10$, the exact power decreases,

```
pow2 = nctcdf(tinv(1-alpha/2, n-1), n-1,-abs(mu1-mu0)*sqrt(n)/s)  ...
          -nctcdf(tinv(1-alpha/2, n-1), n-1, abs(mu1-mu0)*sqrt(n)/s)
%pow2 =0.7210
```

When calculation of the noncentral t CDF is not available, a good approximation for the power is

$$1 - \Phi\left(\frac{t_{n-1,\alpha} - |\mu_1 - \mu_0|\sqrt{n}/s}{\sqrt{1 + \frac{t^2_{n-1,1-\alpha}}{2(n-1)}}} \right).$$

In our example,

```
1-normcdf((tinv(1-alpha,n-1)- ...
(mu1-mu0)/s * sqrt(n))/sqrt(1 + (tinv(1-alpha,n-1))^2/(2*n-2)))
%ans =    0.8209
```

The summary of retrospective power calculations for the t-test us listed below:

Power of the t-test for $H_0 : \mu = \mu_0$, when μ_1 is the actual mean.

• One-sided test:

$$1 - \beta = 1\text{-}nctcdf\left(t_{n-1,1-\alpha},\ n-1,\ \frac{|\mu_1 - \mu_0|}{s/\sqrt{n}}\right).$$

• Two-sided test:

$$1 - \beta = nctcdf\left(t_{n-1,1-\alpha/2},\ n-1,\ \frac{-|\mu_1 - \mu_0|}{s/\sqrt{n}}\right) - nctcdf\left(t_{n-1,1-\alpha/2},\ n-1,\ \frac{|\mu_1 - \mu_0|}{s/\sqrt{n}}\right).$$

Here $nctcdf(x,df,\delta)$ is the CDF of a noncentral t-distribution, with df degrees of freedom and noncentrality parameter δ, evaluated at x. In MATLAB this function is `nctcdf(x,df,delta)`, see page 264

Example 9.8. **Sample Size in t-Test.** In Example 9.7 we were testing $H_0 : \mu = 10$ versus $H_1 : \mu > 10$, at a level $\alpha = 0.05$, where, for sample size $n = 20$ and $s = 5$, we found the power against the alternative $H_1 : \mu = 13$ to be 82.66%. What sample size is needed to increase this power to 95% in a future one-sided test with the same alternative, α and s?

```
mu0 = 10; mu1= 13; s=5; alpha=0.05; beta=0.05;
a = @(n) nctcdf( -tinv(1-alpha, n-1), n-1,-abs(mu1-mu0)*sqrt(n)/s)-(1-beta);
ssize=fzero(a, 20)  %31.4694
```

Thus, the sample of size 32 would ensure power of 95% in repeating the test from Example 9.7.

9.6 Testing the Multivariate Normal Mean*

Testing in the domain of multivariate data generalizes well-known univariate techniques. Conducting the univariate inference on the components of an observed data vector is not adequate since it ignores the covariance structure of observations. This naïve approach can lead to various biases. For example, the tests for individual component means $H_0 : \mu_1 = 3$ and $H_0' : \mu_2 = -1$ may not be significant, while the test $H_0'' : (\mu_1, \mu_2) = (3, -1)$ may turn out to be significant. This is because the evidence may accumulate across the components. On the other hand, in some situations a test on an individual component may turn significant, while the multivariate test involving that component may not be significant due to, again, the interplay with other components. In addition to this "borrowing of strength" from component to component, controlling the family-wise error of first kind is built in, whereas it could represent a problem when components are tested individually.

In this section we look at the multivariate extensions of a t-test, Hotelling's T-square test.

9.6.1 T-Square Test

Assume that a p-dimensional sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is coming from multivariate normal distribution,

$$\boldsymbol{X}_i \sim \mathcal{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu}$ is the parameter of interest, and the population covariance matrix $\boldsymbol{\Sigma}$ is unknown.

For some fixed $\boldsymbol{\mu}_0$, the testing $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ is based on T^2 statistics,

$$T^2 = n(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0)' \, \boldsymbol{S}^{-1} \, (\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0),$$

where $\overline{\boldsymbol{X}}$ and $\boldsymbol{S}$ are sample mean and sample covariance matrix. This statistic is sometimes called the Hotelling T-square in honor of Harold Hotelling, one of the pioneers in multivariate statistical inference. When H_0 is true, the scaled statistic $\frac{n-p}{p(n-1)} T^2$ follows an F-distribution with p and $n - p$ degrees of freedom.

The null hypothesis is rejected if $T^2 \geq \frac{p(n-1)}{n-p} F_{p,n-p,1-\alpha}$, where $F_{p,n-p,1-\alpha}$ is the $(1-\alpha)$ quantile of F-distribution with p and $n-p$ degrees of freedom.

A $100(1-\alpha)\%$ confidence region for $\boldsymbol{\mu}$ consists of all such $\boldsymbol{\mu}$ for which

$$(\overline{\boldsymbol{X}} - \boldsymbol{\mu})' \boldsymbol{S}^{-1} (\overline{\boldsymbol{X}} - \boldsymbol{\mu}) \leq \frac{p(n-1)}{n(n-p)} F_{p,n-p,1-\alpha}.$$

Remark. If $p = 1$, we recover the standard t-statistic and CI. Indeed, note that for $t = \frac{\overline{X}-\mu_0}{s/\sqrt{n}}$,

$$t^2 = \left(\frac{\overline{X} - \mu_0}{s/\sqrt{n}} \right)^2 = n(\overline{X} - \mu_0)(s^2)^{-1}(\overline{X} - \mu_0),$$

which is the one-dimensional counterpart of T^2. The inference is also recovered since t and F distributions are connected, i.e., distributions t_n^2 and $F_{1,n}$ coincide. The confidence regions become standard t-confidence intervals as well, since for the quantiles, $(t_{n,1-\alpha/2})^2 = F_{1,n,1-\alpha}$.

A simultaneous $100(1-\alpha)\%$ confidence interval for all linear combinations $\boldsymbol{a}'\boldsymbol{\mu} = a_1\mu_1 + a_2\mu_2 + \cdots + a_p\mu_p$ is

$$\left[\boldsymbol{a}'\overline{\boldsymbol{X}} - \sqrt{\frac{p\,(n-1)}{n-p}\, F_{p,n-p,1-\alpha}}\ \sqrt{\frac{1}{n}\, \boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}}\ , \right.$$
$$\left. \boldsymbol{a}'\overline{\boldsymbol{X}} + \sqrt{\frac{p\,(n-1)}{n-p}\, F_{p,n-p,1-\alpha}}\ \sqrt{\frac{1}{n}\, \boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}} \right].$$

These simultaneous bounds are true for *any* number of arbitrary vectors $\boldsymbol{a}$. By properly choosing vector $\boldsymbol{a}$, various linear combinations of component means can be monitored.

Example 9.9. **Hook-Billed Kites.** Data set bird.dat|mat|xlsx was analyzed by Johnson and Wichern (2002) and contains bivariate measurements on $n = 45$ female hook-billed kites. The data set contains three columns: bird

number, tail length X_1, and wing length X_2. A bivariate normal distribution is assumed for $(X_1, X_2)'$. We are interested in testing $H_0 : \boldsymbol{\mu} = (190, 275)'$ versus $H_1 : \boldsymbol{\mu} \neq (190, 275)'$.

For this data set the sample mean is $\overline{\boldsymbol{X}} = (193.6222, 279.7778)'$ and the sample covariance matrix is $\boldsymbol{S} = \begin{bmatrix} 120.6949 & 122.3460 \\ 122.3460 & 208.5404 \end{bmatrix}$. MATLAB script `bird.m` performs the test and explores the relationship between individual and simultaneous testing.

```
%bird.m
load 'bird.mat'
x1 = bird(:,2); x2 = bird(:,3);
X=[x1 x2];
[n p]=size(X);
Xbar =transpose(mean(X))  %[193.6222;  279.7778]
S = cov(X)
    % 120.6949  122.3460
    % 122.3460  208.5404
mu0 = [190; 275];
T2 = n * (Xbar - mu0)'* inv(S) * (Xbar - mu0) %5.5431
F = (n-p)/(p*(n-1)) * T2  %2.7086
pval = 1-fcdf(F, p, n-p)  %0.078
```

We fail to reject H_0 at 5% significance level. However, if t-tests are performed on the individual components, the tests are significant.

```
%bird.m continued
t1 = (Xbar(1)-mu0(1))/sqrt(S(1,1)/n)  %2.2118
t2 = (Xbar(2)-mu0(2))/sqrt(S(2,2)/n)  %2.2194
p1 = 2*tcdf(-abs(t1), n-2)  %0.0323
p2 = 2*tcdf(-abs(t2), n-2)  %0.0318
```

If instead of `mu0=[190; 275]` we tested for `mu0=[192; 283]`, the significance statements will be reversed. The T-square test will be significant, whereas the individual t-tests will not be significant. This situation was alluded to in the introduction of this section. The reasons for this discrepancy are illustrated in Figure 9.2.

Here, the 95% simultaneous confidence ellipse for population mean $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ is plotted together with individual 95% confidence intervals for μ_1 and μ_2.

The green dot corresponds to `mu0=[190; 275]` falls outside individual intervals and the corresponding componentwise t-tests are both significant. However, this point falls inside the confidence ellipse and the T-square test is not significant.

If the test is about `mu0=[192; 283]`, then this point (red dot) falls outside the ellipse, but inside the individual confidence intervals.

```
%bird.m modified
mu0=[192; 283];
```

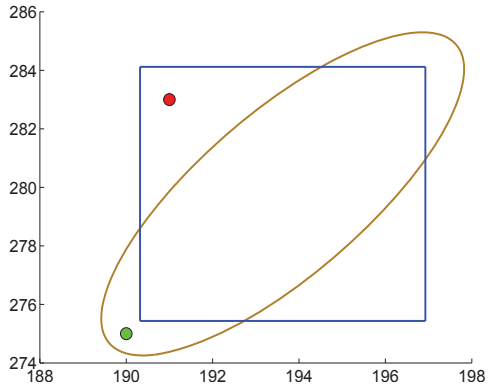


Fig. 9.2 Comparison of simultaneous and individual 95% confidence sets. The confidence ellipse contains `mu0=[190; 275]` (green dot); thus the individual tests are significant but not the multivariate. For `mu0=[192; 283]` (red dot), the significance results are reversed.

```
T2 = n * (Xbar - mu0)'* inv(S) * (Xbar - mu0) %13.5909
F = (n-p)/(p*(n-1)) * T2   %6.6410
pval = 1-fcdf(F, p, n-p)   %0.0031
%
t1 = (Xbar(1)-mu0(1))/sqrt(S(1,1)/n)  %0.9905
t2 = (Xbar(2)-mu0(2))/sqrt(S(2,2)/n)  %-1.4968
p1 = 2*tcdf(-abs(t1), n-2)  %0.3275
p2 = 2*tcdf(-abs(t2), n-2)  %0.1417
```

9.6.1.1 Power Analysis for T-Square Test

Suppose that we need to find the power of T-square test for testing $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 = (0.3, 0.3)'$ against the alternative $H_1 : \boldsymbol{\mu} = \boldsymbol{\mu}_1 = (0.4, 0.4)'$ if the sample size of $n = 930$ is planned, and elicited covariance matrix is $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}$.

The effect size

$$D = \sqrt{(\mu_1 - \mu_0)' \Sigma^{-1} (\mu_1 - \mu_0)}.$$

is a multivariate analogue of Cohen's $d = |\mu_1 - \mu_0|/\sigma$, while the noncentrality parameter for F statistic (connected with T^2 via $F = (n-p)/p \; T^2/(n-1), df = (p, n-p)$) is $\lambda = n \cdot D^2$.

```
%Power of T^2 test
sigma=[1 0.2; 0.2 1];
```

```
n=930;
D2=[0.1; 0.1]' * inv(sigma) * [0.1; 0.1]
%D2 = 0.01667
%effect is D = sqrt(D2) = 0.1291
lambda = n * D2     %lambda = 15.5

power=1-ncfcdf( finv(1-0.05, 2, 930-2), 2, 930-2, lambda)
%power =  0.9501
```

Next, we will find the sample size so that effect $D = 0.2$ is found significant with the power of $1 - \beta = 0.90$ for $p = 2$ and $\alpha = 0.05$.

```
ssize=ceil(fzero(@(n) ncfcdf( finv(1-0.05, 2, n-2), ...
                      2, n-2, n*0.2^2)-(1-0.90), 1000))
%ssize = 320
```

The MATLAB script `powerT2.m` contains the calculations.

9.6.2 *Test for Symmetry*

In a multivariate context, tests for the equality of component means are called *tests of symmetry*. Let $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)'$ be the mean of $\mathcal{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from which a sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ is obtained. Assume that $p \geq 2$.

The hypothesis of symmetry

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p,$$

can be expressed as

$$H_0 : \boldsymbol{C}\boldsymbol{\mu} = \boldsymbol{0} \quad \text{versus} \quad H_1 : \boldsymbol{C}\boldsymbol{\mu} \neq \boldsymbol{0}$$

where $\boldsymbol{C}$ is *any* $(p-1) \times p$ matrix, of rank $p-1$ (rows are linearly independent), such that

$$\boldsymbol{C}\boldsymbol{1} = \boldsymbol{0}, \quad \text{for } \boldsymbol{1} = (1, 1, \ldots, 1)'.$$

Popular choices for $\boldsymbol{C}$ are,

$$\boldsymbol{C} = \begin{pmatrix} 1 & -1 & 0 \ldots & 0 & 0 \\ 0 & 1 & -1 \ldots & 0 & 0 \\ 0 & 0 & 1 \ldots & 0 & 0 \\ \vdots & & & & \\ 0 & 0 & 0 \ldots & -1 & 0 \\ 0 & 0 & 0 \ldots & 1 & -1 \end{pmatrix} \quad \text{or} \quad \boldsymbol{C} = \begin{pmatrix} 1 & -1 & 0 \ldots & 0 & 0 \\ 1 & 0 & -1 \ldots & 0 & 0 \\ 1 & 0 & 0 \ldots & 0 & 0 \\ \vdots & & & & \\ 1 & 0 & 0 \ldots & -1 & 0 \\ 1 & 0 & 0 \ldots & 0 & -1 \end{pmatrix}.$$

The test is based on

$$T^2 = n\ \overline{\boldsymbol{X}}'\ \boldsymbol{C}'(\boldsymbol{C}\boldsymbol{S}\boldsymbol{C}')^{-1}\boldsymbol{C}\ \overline{\boldsymbol{X}}.$$

In this case,

$$\frac{n-(p-1)}{(p-1)\ (n-1)}\ T^2 \sim F_{p-1,n-(p-1)},$$

which is used for the inference.

Example 9.10. **Cork Boring Data Revisited.** Consider data from Exercise 2.23 consisting of the weights of cork boring for 28 trees. We will test for the equality of component means (four directions: north, east, south, and west). In the MATLAB file below, we show that for two valid choices of C (rows sum up to 0) the value of the T^2 statistic remains the same.

```
%Rao's Cork Data
X =[   72 66 76 77;   60 53 66 63;   56 57 64 58;  41 29 36 38; ...
       32 32 35 36;   30 35 34 26;   39 39 31 27;  42 43 31 25; ...
       37 40 31 25;   33 29 27 36;   32 30 34 28;  63 45 74 63; ...
       54 46 60 52;   47 51 52 43;   91 79 100 75; 56 68 47 50; ...
       79 65 70 61;   81 80 68 58;   78 55 67 60;  46 38 37 38; ...
       39 35 34 37;   32 30 30 32;   60 50 67 54;  35 37 48 39; ...
       39 36 39 31;   50 34 37 40;   43 37 39 50;  48 54 57 43];
  [n p]=size(X);
  Xbar = mean(X)';  S=cov(X);
  %N E S W
  C=[ 1  -1  -1  1;       0  0  1  -1;      1   0  -1   0 ];
T2 = n * Xbar' * C'  * inv(C * S * C')* C  * Xbar      %20.7420
pval = 1-fcdf( (n-p+1)/((p-1)*(n-1)) * T2, p-1, n-p+1) %0.0023

%invariance wrt C
C1 =[1 -1 0 0;    1 0 -1 0;    1 0 0 -1];
T2 = n * Xbar' * C1'  * inv(C1 * S * C1')* C1  * Xbar  %20.7420
```

9.7 Testing the Normal Variances

When we discussed the estimation of the normal variance (Section 7.4.2), we argued that the statistic $(n-1)s^2/\sigma^2$ had a χ^2-distribution with $n-1$ degrees of freedom. The test for the normal variance is based on this statistic and its distribution.

Suppose we want to test $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 >, \neq, <, \sigma_0^2$. The test statistic is

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}.$$

The testing procedure at the α level can be summarized by

Alternative	α-level rejection region	p-value (MATLAB)
$H_1: \sigma > \sigma_0$	$[\chi^2_{n-1,1-\alpha}, \infty)$	`1-chi2cdf(chi2,n-1)`
$H_1: \sigma \neq \sigma_0$	$[0, \chi^2_{n-1,\alpha/2}] \cup [\chi^2_{n-1,1-\alpha/2}, \infty)$	`2*chi2cdf(min(chi2,1/chi2),n-1)`
$H_1: \sigma < \sigma_0$	$[0, \chi^2_{n-1,\alpha}]$	`chi2cdf(chi2,n-1)`

The power of the test against the specific alternative is the probability of the rejection region evaluated as if H_1 were a true hypothesis. For example, if $H_1: \sigma^2 > \sigma_0^2$, and specifically if $H_1: \sigma^2 = \sigma_1^2,\ \ \sigma_1^2 > \sigma_0^2$, then the power is

$$1-\beta = \mathbb{P}\left(\frac{(n-1)s^2}{\sigma_0^2} \geq \chi^2_{1-\alpha,n-1} \Big| H_1 \right) = \mathbb{P}\left(\frac{(n-1)s^2}{\sigma_1^2} \cdot \frac{\sigma_1^2}{\sigma_0^2} \geq \chi^2_{1-\alpha,n-1} \Big| H_1 \right)$$
$$= \mathbb{P}\left(\chi^2 \geq \frac{\sigma_0^2}{\sigma_1^2} \chi^2_{1-\alpha,n-1} \right),$$

or in MATLAB:

```
power=1-chi2cdf(sigmasq0/sigmasq1*chi2inv(1-alpha,n-1),n-1).
```

For the one-sided alternative in the opposite direction and for the two-sided alternative, the procedure for finding the power is analogous. The sample size necessary to achieve a preassigned power can be found by trial and error or by using MATLAB's function `fzero`.

Example 9.11. **LDL-C Levels.** A new handheld device for assessing cholesterol levels in blood is presented for approval to the FDA. The variability of measurements obtained by the device for people with normal levels of LDL cholesterol is one of the measures of interest. A calibrated sample of size $n = 224$ of serum specimens with a fixed 130-level of LDL-C is measured by the device. The variability of measurements is assessed.

(a) If $s^2 = 2.47$ was found, test the hypothesis that the population variance is 2 (as achieved by a clinical computerized Hitachi 717 analyzer, with enzymatic, colorimetric detection schemes) against the one-sided alternative. Use $\alpha = 0.05$.

(b) Find the power of this test against the specific alternative, $H_1: \sigma^2 = 2.5$.

(c) What sample size ensures the power of 90% in detecting the effect $\sigma_0^2/\sigma_1^2 = 0.8$ as significant.

```
n = 224; s2 = 2.47; sigmasq0 = 2; sigmasq1 = 2.5; alpha = 0.05;
%(a)
chisq = (n-1)*s2 /sigmasq0
  %test statistic chisq = 275.4050.
  %The alternative is H_1: sigma2 > 2
chi2crit = chi2inv( 1-alpha, n-1 )
  %one sided upper tail RR = [258.8365, infinity)
pvalue = 1 - chi2cdf(chisq, n-1) %pvalue = 0.0096
%(b)
power = 1-chi2cdf(sigmasq0/sigmasq1 * chi2inv(1-alpha, n-1), n-1 )
                          %power = 0.7708
%(c)
ratio = sigmasq0/sigmasq1  %0.8
pf = @(n) 1-chi2cdf( ratio * chi2inv(1-alpha, n-1), n-1 ) - 0.90;
ssize = fzero(pf, 300)     %342.5993 approx 343
```

9.8 Testing the Proportion

When discussing the CLT, and in particular the de Moivre theorem, we saw that the binomial distribution can be well approximated with the normal if n is large and $np(1-p) > 5$.

Suppose that we observe n Bernoulli $\mathcal{B}er(p)$ random variables $Y_1, Y_2, \ldots, Y_n$, with p to be tested. The sum $X = Y_1 + \cdots + Y_n$ is $\mathcal{B}in(n,p)$ and sample proportion of Y's, $\hat{p} = \frac{X}{n}$ is the MLE of p. By the CLT, sample proportion $\hat{p}$ has an approximately normal distribution with mean p and variance $p(1-p)/n$. This approximate normality will be used to construct the test.

Suppose that we are interested in testing $H_0 : p = p_0$ versus one of the three possible alternatives. When H_0 is true, the test statistic

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

has approximately a standard normal distribution. The testing procedure is summarized in the following table:

Alternative	α-level rejection region	p-value (MATLAB)
$H_1 : p > p_0$	$[z_{1-\alpha}, \infty)$	`1-normcdf(z)`
$H_1 : p \neq p_0$	$-\infty, z_{\alpha/2}] \cup [z_{1-\alpha/2}, \infty)$	`2*normcdf(-abs(z))`
$H_1 : p < p_0$	$(-\infty, z_{\alpha}]$	`normcdf(z)`

Using the normal approximation one can derive that the power against the specific alternative $H_1 : p = p_1$ is

$$1 - \beta = \Phi\left[\frac{\sqrt{n}|p_1 - p_0| - z_{1-\alpha}\sqrt{p_0(1-p_0)}}{\sqrt{p_1(1-p_1)}}\right],$$

for the one-sided test. In the case of two-sided alternative, $z_{1-\alpha}$ is replaced by $z_{1-\alpha/2}$. The sample size needed to find the effect $|p_0 - p_1|$ significant $(1-\beta)100\%$ of the time (i.e., the one-sided test would have a power of $1-\beta$) is

$$n = \frac{\left(\sqrt{p_0(1-p_0)}\, z_{1-\alpha} + \sqrt{p_1(1-p_1)}\, z_{1-\beta}\right)^2}{(p_0 - p_1)^2}.$$

For the two sided alternative, $z_{1-\alpha}$ is replaced by $z_{1-\alpha/2}$. Note that specifying only $|p_1 - p_0|$ is not sufficient for sample size determination; both p_0 and p_1 need to be specified.

Example 9.12. **Proportion of Hemorrhagic-Type Strokes among American Indians.** The study described in the American Heart Association's news release of September 22, 2008, included 4,507 members of 13 American Indian tribes in Arizona, Oklahoma, and North and South Dakota. It found that American Indians have a stroke rate of 679 per 100,000, compared to 607 per 100,000 for African Americans and 306 per 100,000 for Caucasians. None of the participants, ages 45 to 74, had a history of stroke when they were recruited for the study from 1989 to 1992. Almost 60% of the volunteers were women.

During more than 13 years of follow-up, 306 participants suffered a first stroke, most of them in their mid-60s when it occurred. There were 263 strokes of the ischemic type, caused by a blockage that cuts off the blood supply to the brain, and 43 hemorrhagic (bleeding) strokes.

It is believed that in the general population one in five of all strokes is hemorrhagic.

(a) Test the hypothesis that the proportion of hemorrhagic strokes in the population of American Indians that suffered a stroke is lower than the national proportion of 0.2.

(b) What is the power of the test in (a) against the alternative $H_1 : p = 0.15$?

(c) What sample size ensures a power of 90% in detecting $p = 0.15$, if H_0 states $p = 0.2$?

Since $306 \times 0.2 > 10$, a normal approximation can be used.

```
z = (43/306 - 0.2)/sqrt(0.2 *(1- 0.2)/306)
% z = -2.6011

pval = normcdf(z)
% pval = 0.0046


%(b)
p0=0.2; p1=0.15; alpha=0.05; n=306;
power = normcdf((sqrt(n)*abs(p1-p0) - ...
        norminv(1-alpha)*sqrt(p0*(1-p0)))/sqrt(p1*(1-p1)) )
%0.7280

%(c)
beta = 0.1;
n=( sqrt(p0*(1-p0)) * norminv(1-alpha) + ...
    sqrt(p1*(1-p1)) * norminv(1-beta) )^2/(p1-p0)^2
%497.7779   approx 498
```

9.8.1 Exact Test for Population Proportions

In the previous section we used a normal approximation to the binomial distribution to test the population proportion via the familiar z-test. Since we assume a binomial model for the data, it is possible (and in the case of small $np(1-p)$, e.g., < 5, necessary) to test for the proportion in an exact manner.

Here we operate not with $\hat{p} = X/n$ but with X that, under $H_0 : p = p_0$, has binomial $\mathcal{B}in(n, p_0)$ distribution. Thus, the statistic X takes a value k with probability

$$p_{0,n,k} = \binom{n}{k} p_0^k (1-p_0)^{n-k},\ k = 0, 1, \ldots, n.$$

For the one-sided alternative, say $H_1 : p < p_0$, we find k^* that is the maximum k for which $\mathbb{P}(X \le k) \le \alpha$. The hypothesis H_0 is rejected for X less than or equal to k^*, that is, the rejection region is $X \in \{0, 1, \ldots, k^*\}$. The level of this test is $\alpha^* = \mathbb{P}(X \le k^*)$. For the alternative $H_1 : p > p_0$ the critical region is $X \ge k^*$, where k^* is the minimum k for which $\mathbb{P}(X \ge k) \le \alpha$.

One of the difficulties in exact testing is that the significance level α^* can take only discrete values, since X is a discrete statistic, and none of these discrete values may match or even be close to the preassigned significance level α, say 0.05.

For the two-sided alternative, $H_0 : p \neq p_0$, the rejection region is $\{X \leq k_1^*\} \cup \{X \geq k_2^*\}$, where k_1^*, k_2^* are selected such that $\mathbb{P}(X \leq k_1^*) + \mathbb{P}(X \geq k_2^*) \leq \alpha$. The pair k_1^*, k_2^* is not unique, however, the choice where the probabilities of the two tails are similar (close to $\alpha/2$) is preferred.

It would be helpful to look at some numbers. For example, assume that in $n = 27$ trials we found X successes and are interested in testing $H_0 : p = 0.3$ at $\alpha = 0.05$. Under H_0, statistic $X \sim \mathcal{B}in(27, 0.3)$.

If the alternative is $H_1 : p > 0.3$, then the test with critical region $\{X \geq 14\}$ would have the level of `1-binocdf(13-1, 27, 0.3)=0.0359`. For the alternative $H_1 : p < 0.3$, the critical region $\{X \leq 3\}$ would have the level of `binocdf(3, 27, 0.3) = 0.0202`, while the test with critical region $\{X \leq 4\}$ would have the level of `binocdf(4, 27, 0.3) = 0.0591`, thus slightly exceeding 0.05. The exact $\alpha = 0.05$ level test is not possible here, so the test with $k^* = 3$ will be used since $0.0202 < 0.05$. We note that the exact tests could be randomized so that any α is achieved, but this theory is beyond the scope of this text.

If, for instance, $X = 5$ is observed, H_0 is not rejected since $X > k^* = 3$. The p-value is `binocdf(5, 27, 0.3) = 0.1358.`

For the two-sided alternative, $H_1 : p \neq 0.3$, and $\alpha = 0.05$, the values for k_1^* and k_2^* are 3 and 14, respectively, since `1-binocdf(14-1, 27, 0.3) = 0.0143`, and again X is not in rejection region. For this alternative, the achieved significance level is $0.0202 + 0.0143 = 0.0345 < 0.5$. The p-value is `2*min( binocdf(5, 27, 0.3), 1- binocdf(5-1, 27, 0.3)) = 0.2716`, so H_0 is not rejected.

These results are summarized in the table below where $p_{0,n,i} = \binom{n}{i} p_0^i (1 - p_0)^{n-i}$ are probabilities of $X = i$ under H_0.

Alternative	Critical region	p-value (MATLAB)
$H_1 : p < p_0$	$X \leq k^* = \max k : \sum_{i=0}^{k} p_{0,n,i} \leq \alpha$	`binocdf(X,n,p0)`
$H_1 : p \neq p_0$	$X \leq k_1^* = \max k : \sum_{i=0}^{k} p_{0,n,i} \leq \alpha/2$, or $X \geq k_2^* = \min k : \sum_{i=k}^{n} p_{0,n,i} \leq \alpha/2$	`2* min(binocdf(X,n,p0), 1-binocdf(X-1,n,p0))`
$H_1 : p > p_0$	$X \geq k^* = \min k : \sum_{i=k}^{n} p_{0,n,i} \leq \alpha$	`1-binocdf(X-1,n,p0)`

Example 9.13. **Proportion of Hemorrhagic Strokes: Exact Test.** In a follow-up study discussed in Example 9.12, out of 306 participants suffering a stroke, 43 of the strokes were of hemorrhagic type, and the rest of the ischemic type. We tested hypotheses $H_0 : p = 0.2$ versus $H_1 : p < 0.2$ at $\alpha = 0.05$ level using the normal approximation, and found a p-value of 0.0046.

The results for the exact test are summarized in the annotated MATLAB code below:

```
pvalue = binocdf(43, 306, 0.20)        %0.0044
k=binoinv(0.05, 306, 0.2)              %k=50
```

```
kstar = k-1;                          %RRegion X <= k*;  k*=49
alphastar = binocdf(kstar, 306, 0.2)  %alpha*=0.0445<0.05
pow=binocdf(kstar, 306, 0.15)         %power against H1: p=0.15
                                      %pow = 0.7220
```

Note that the exact p-value (0.0044) is quite close to the p-value obtained by the normal approximation (0.0046). The achieved significance level α^* is $0.0445 < 0.05$. Note also that the power is 0.7220, which is slightly less than the power found using the normal approximation. In general, power analyses based on the normal approximation are more "optimistic."
✎

Exact Sample Size in Testing the Proportion. Let $p_{1,n,k} = \binom{n}{k} p_1^k (1-p_1)^{n-k}$ be the probabilities of $X = k$ under the precise alternative $H_1 : p = p_1$.

The power of an α-level test of $H_0 : p = p_0$ versus $H_1 : p = p_1$ for sample size n is

$$\sum_{k=0}^{n} \left[p_{1,n,k}\, \mathbf{1}\left(\sum_{i=k}^{n} p_{0,n,i} \le \alpha \right) \right], \text{ when } H_1 : p = p_1 > p_0,$$

$$\sum_{k=0}^{n} \left[p_{1,n,k}\, \mathbf{1}\left(\sum_{i=0}^{k} p_{0,n,i} \le \alpha \right) \right], \text{ when } H_1 : p = p_1 < p_0, \text{ and}$$

$$\sum_{k=0}^{n} \left[p_{1,n,k}\, \mathbf{1}\left(2 \cdot \min\left\{ \sum_{i=0}^{k} p_{0,n,i}, \sum_{i=k}^{n} p_{0,n,i} \right\} \le \alpha \right) \right], \text{ when } H_1 : p = p_1 \neq p_0.$$

Here, $\mathbf{1}$ is an indicator, and $p_{0,n,i} = \binom{n}{i} p_0^i (1-p_0)^{n-i}$ are binomial probabilities of $X = i$ under the null hypothesis. The sample size is now determined by increasing n until the power reaches the preassigned level of $1 - \beta$.

Example 9.14. **Proportion of Hemorrhagic Strokes: Exact Power and Sample Size.** In Example 9.13, we tested hypotheses $H_0 : p = 0.2$ versus $H_1 : p < 0.2$ at $\alpha = 0.05$ level using the exact binomial test. We also found the exact power, against the one-sided specific alternative $p = 0.15$, to be 0.7220.

Here, we repeat the power calculation in a more systematic fashion and also find the sample size necessary to achieve the power of 90% in a prospective test of the same hypotheses, at $\alpha = 0.05$ level.

```
n = 306; p0 = 0.2; p1 = 0.15; alpha = 0.05;
kargs = 0:n;
u =  binocdf(kargs, n, p0)  <= alpha;        %indicator
exactpower = sum( binopdf(kargs, n, p1).*u ) %0.7220
%sample size
beta = 0.1; %preset power of 90%
exactpower = 0; n = 10;
```

```
while exactpower < 1-beta
     n=n+1;
     kargs = 0:n;
     ind = binocdf(kargs, n, p0)  <= alpha;   %indicator
     exactpower = sum( binopdf(kargs, n, p1).* ind ) ;
end
 disp(['samplesize = ' num2str(n)])
%samplesize = 501
```

Thus, a sample of size 501 would be required to achieve the desired power. Note that in Example 9.12, a sample size of 498 was found using normal approximation. Show that, if the test is two-sided, for the same α, p_0, p_1, and n, the power would be 0.6078. Show also that, for the two-sided testing, with the same α, p_0, p_1 and $1-\beta = 90\%$, the necessary sample size would be 619. For the two-sided test the indicator is `ind = 2*min(binocdf(kargs, n, p0),1-binocdf(kargs-1, n, p0)) <= alpha;`

9.8.2 Bayesian Test for Population Proportions

A Bayesian test for binomial proportion was already discussed in Example 9.4 on page 387.

In its simplest form, a Bayesian test requires a prior on population proportion p. In Example 9.4 the prior was uniform on $[0,0.1]$ with a point mass at $p = 0.1$.

In the context of Example 9.13, a beta prior with parameters 1 and 4 is elicited, so that the prior mean $\mathbb{E}^{\pi} p = 1/(1+4) = 0.2$ matches the mean under H_0. The following simple WinBUGS script conducts the test

$$H_0 : p \leq 0.2 \quad \text{versus} \quad H_1 : p > 0.2.$$

```
model{
X ~ dbin(p, n)
p ~ dbeta(1,4)
pH1 <- step(0.2-p)
     }
DATA
list(n=306, X=43)
#Generate Inits
```

The output variable `pH1` gives the posterior probability of H_1.

	mean	sd	MC error	val2.5pc	median	val97.5pc	start	sample
p	0.1415	0.01974	1.9158E-5	0.1051	0.1407	0.1823	1001	1000000
pH1	0.9967	0.05694	5.675E-5	1.0	1.0	1.0	1001	1000000

Since population proportions are in $[0,1]$, typical prior on p is beta. Discussion on eliciting beta priors can be found on page 348. The following example uses Zellner's prior on p. Zellner's prior is in fact a flat prior on $\text{logit}(p)$ and it was also discussed on page 348.

Example 9.15. **eBay Story.** You decided to purchase a new Orbital Shaking Incubator for your research lab on eBay. A single seller is offering this item. The seller has positive feedback from 223 out of 230 responders.

(a) What is the 95% credible set for the population satisfaction rate with this seller, p?

(b) Test hypotheses (i) $H_0' : p \leq 0.98$ vs. $H_1' : p > 0.98$ and (ii) $H_0'' : 0.96 \leq p \leq 0.99$ vs. $H_1'' = (H_0'')^c$.

```
model{
    Positives ~ dbin(p,n)
    # Zellner's  1/[p (1-p)] improper prior
    # set as flat prior on logit
    logit(p) <- eta
    eta ~ dflat()
    pH1prime <- step(0.98-p)
    pH1second <- 1-step(p-0.96)*step(0.99-p)
     }
DATA
list(n=230, Positives=223)
INITS
list(eta=0)
```

The output variables `pH1prime` and `pH1second` give the posterior probabilities of corresponding H_1's.

	mean	sd	MC error	val2.5pc	median	val97.5pc	start	sample
eta	3.533	0.3975	4.07E-4	2.823	3.508	4.379	1001	1000000
p	0.9696	0.01129	1.153E-5	0.9439	0.9709	0.9876	1001	1000000
pH1prime	0.8222	0.3823	3.77E-4	0.0	1.0	1.0	1001	1000000
pH1second	0.1956	0.3967	4.065E-4	0.0	0.0	1.0	1001	1000000

The 95% credible set for p is $[0.9439, 0.9876]$. The classical 95% Wald's confidence interval in this case is $[0.9474, 0.9918]$, which is slightly shifted right. The posterior for p is slightly skewed to the left, indicating that symmetry of likelihood assumed in normal approximation biases the interval; see Figure 9.3.

Note that H_1' and H_0'' have higher posterior probabilities, 0.8222 and $1 - 0.1956$, and should be favored.

The following example emphasizes the conditional nature of Bayesian inference and its conformity to the *likelihood principle*, which states that *all* information about the experimental results are summarized only in the likelihood.

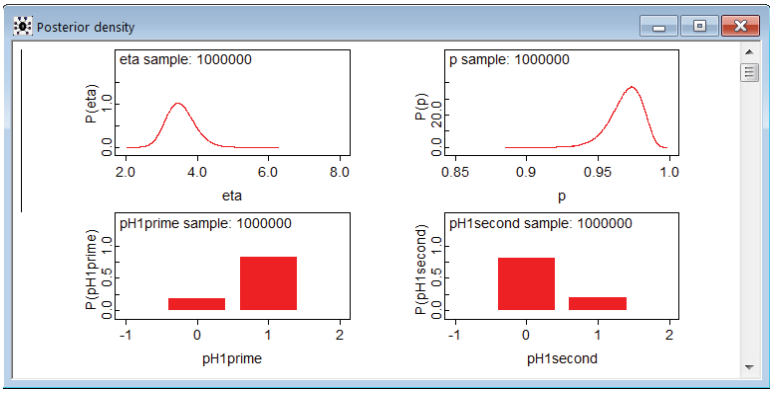


Fig. 9.3 Output from ebaystory0.odc. Posterior distribution for p appears slightly skewed to the left indicating that Wald type confidence intervals are biased. The bottom two bar-plots represent the posterior probabilities of hypotheses H_0', H_1' and H_0'', H_1'', respectively.

Example 9.16. **Savage's Disparity.** A Bayesian inference is based on data observed and not on data that could possibly be observed, or on the manner in which the sampling was conducted. This is the crux of the likelihood principle.

This is not the case in classical testing, and the argument first put forth by Jimmie Savage at the Purdue Symposium in 1962 emphasizes the difference.

Suppose a coin is flipped 12 times and 9 heads and 3 tails are obtained. Let p be the probability of heads. We are interested in testing whether the coin is fair against the alternative that it is more likely to come heads up, or

$$H_0 : p = 1/2 \quad \text{versus} \quad H_1 : p > 1/2.$$

The p-value for this test is the probability that one observes 9 or more heads if the coin is fair, that is, when H_0 is true.

Consider the following two scenarios:

(a) Suppose that the number of flips $n = 12$ was decided a priori. Then the number of heads X is binomial and under H_0 (fair coin) the p-value is

$$\mathbb{P}(X \geq 9) = 1 - \sum_{k=0}^{8} \binom{12}{k} p^k (1-p)^{12-k} = 1 - \texttt{binocdf}(8,12,0.5) = 0.0730.$$

At a 5% significance level H_0 is *not rejected*.

(b) Suppose that the flipping is carried out until 3 tails have appeared. Let us call tails "success" and heads "failures." Then, under H_0, the number of failures (heads) Y is a negative binomial $\mathcal{NB}(3,1/2)$ and the p-value is

$$\mathbb{P}(Y \geq 9) = 1 - \sum_{k=0}^{8} \binom{3+k-1}{k} (1-p)^3 p^k = 1 - \texttt{nbincdf}(8,3,1/2) = 0.0327.$$

At a 5% significance level H_0 *is rejected*.

Thus, two Fisherian tests recommend opposite actions for the same data simply because of how the sampling was conducted.

Note that in both (a) and (b) the likelihoods are proportional to $p^9(1-p)^3$, and for a fixed prior on p there is no difference in any Bayesian inference.

Edwards et al. (1963, p. 193) note "... the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience."

9.9 Multiplicity in Testing, Bonferroni Correction, and False Discovery Rate

Recall that when testing a single hypothesis H_0, a type I error is made if it is rejected, when it is actually true. The probability of making a type I error in a test is usually controlled to be smaller than a certain level of α, typically equal to 0.05.

When there are several null hypotheses, $H_{01}, H_{02}, \ldots, H_{0m}$, and all of them are tested simultaneously, one may want to control the type I error at some level α as well. In this scenario, a type I error is then made if at least one true hypothesis in the family of hypotheses being tested is rejected. Because it pertains to the family of hypotheses, this significance level is called the familywise error rate (FWER).

If the hypotheses in the family are independent, then

$$\text{FWER} = 1 - (1 - \alpha_i)^m,$$

where FWER and α_i are overall and individual significance levels, respectively.

For arbitrary, possibly dependent, hypotheses, the Bonferroni inequality (page 415) translates to

$$\text{FWER} \leq m\alpha_i.$$

Suppose $m = 15$ tests are conducted simultaneously. For an individual α_i of 0.05, the FWER is $1 - 0.95^{15} = 0.5367$. This means that the chance of claiming a significant result when there should not be one is larger than 1/2. For possibly dependent hypotheses, the upper bound of FWER increases to 0.75.

Bonferroni Correction: To control FWER $\leq \alpha$, one should reject all H_{0i} among $H_{01}, H_{02}, \ldots, H_{0m}$ for which the p-value is found smaller than α/m.

Thus, if for $n = 15$ arbitrary hypotheses we want an overall significance level of FWER ≤ 0.05, then the individual test levels should be set to $0.05/15 = 0.0033$.

Testing for significance with gene expression data from DNA microarray experiments involves simultaneous comparisons of hundreds or thousands of genes, and controlling the FWER by the Bonferroni method would require very small individual α_is. Yet, setting such small α levels decreases the power of individual tests and many false H_0 are not rejected. Therefore the Bonferroni correction is considered by many practitioners as overly conservative. Some call it a "panic approach."

Remark. If, in the context of interval estimation, k simultaneous interval estimates are desired with an overall confidence level $(1-\alpha)100\%$, then each interval can be constructed with a confidence level $(1-\alpha/k)100\%$, and the Bonferroni inequality would ensure that the overall confidence is at least $(1-\alpha)100\%$.

Bonferroni–Holm Method. The Bonferroni–Holm method is an iterative procedure in which individual significance levels are adjusted to increase power and still control the FWER. One starts by ordering the p-values of all tests for $H_{01}, H_{02}, \ldots, H_{0m}$ and then compares the smallest p-value to α/m. If that p-value is smaller than α/m, then one should reject that hypothesis and compare the second ranked p-value to $\alpha/(m-1)$. If this hypothesis is rejected, one should proceed to the third ranked p-value and compare it with $\alpha/(m-2)$. This should be continued until the hypothesis with the smallest remaining p-value cannot be rejected. At this point the procedure stops and all hypotheses that have not been rejected at previous steps are retained.

Let $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$ correspond to ordered p-values $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$. For a given α, find minimum k such that

$$p_{(k)} > \frac{\alpha}{m+1-k}.$$

Reject hypotheses $H_{(1)}, \ldots, H_{(k-1)}$, and keep $H_{(k)}, \ldots, H_{(m)}$.

To better see this, let us assume that five hypotheses are to be tested with a FWER of 0.05. The five p-values are 0.09, 0.01, 0.04, 0.012, and 0.004. The smallest of these is 0.004. Since this is less than 0.05/5, hypothesis four is rejected. The next smallest p-value is 0.01, which is also smaller than 0.05/4. So this hypothesis is also rejected. The next smallest p-value is 0.012, which is smaller than 0.05/3, and this hypothesis is rejected. The next smallest p-value is 0.04, which is not smaller than 0.05/2. Therefore, the hypotheses with p-values of 0.004, 0.01, and 0.012 are rejected while those with p-values of 0.04 and 0.09 are not rejected.

False Discovery Rate. The false discovery rate paradigm (Benjamini and Hochberg, 1995) considers the proportion of falsely rejected null hypotheses (false discoveries) among the total number of rejections.

Controlling the expected value of this proportion, called the false discovery rate (FDR), provides a useful alternative that addresses low-power problems of the traditional FWER methods when the number of tested hypotheses is large. The test statistics in these multiple tests are assumed to be independent or positively correlated. Suppose that we are looking at the result of testing m hypotheses, among which m_0 are true. In the table that follows, V denotes the number of false rejections, and the FWER is $\mathbb{P}(V \geq 1)$:

	H_0 not rejected	H_0 rejected	Total
H_0 true	U	V	m_0
H_1 true	T	S	m_1
Total	W	R	m

If R denotes the number of rejections (declared significant genes, discoveries), then V/R, for $R > 0$, is the proportion of false rejected hypotheses. The FDR is

$$\mathbb{E}\left(\frac{V}{R}\middle| R > 0\right)\mathbb{P}(R > 0).$$

Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ be the ordered, observed p-values for the m hypotheses to be tested. Algorithmically, the FDR method finds k such that

$$k = \max\left\{i | p_{(i)} \leq (i/m)\alpha\right\}. \tag{9.4}$$

The FDR is controlled at the α level if the hypotheses corresponding to $p_{(1)}, \ldots, p_{(k)}$ are rejected. If no such k exists, no hypothesis from the family is rejected. When the test statistics in the multiple tests are possibly negatively correlated as well, the FDR is modified by replacing α in (9.4) with $\alpha/(1 + 1/2 + \cdots + 1/m)$. The following MATLAB script (FDR.m) finds the critical p-value $p_{(k)}$. If $p_{(k)} = 0$, then no hypothesis is rejected.

```
function pk = FDR(p,alpha)
```

```
%Critical p-value pk for FDR <= alpha.
%All hypotheses with p-value less than or equal
%to pk are rejected.
%if pk = 0 no hypothesis is to be rejected
m = length(p);     %number of hypotheses
po = sort(p(:));   %ordered p-values
i = (1:m)';        %index
pk = po(max(find(  po < i./m * alpha)));
%critical p-value
if ( isempty(pk)==1 )
    pk=0;
end
```

Suppose that we have 1,000 hypotheses and all hypotheses are true. Then their p-values represent a random sample from the uniform $\mathcal{U}(0,1)$ distribution. About 50 hypotheses would have a p-value of less than 0.05. However, for reasonable FDR levels (0.05–0.2) $p_{(k)} = 0$, as it should be since we do not want false discoveries.

```
p = rand(1000,1);
[FDR(p, 0.05), FDR(p, 0.2), FDR(p, 0.6), FDR(p, 0.92)]
%ans =     0     0    0.0022    0.0179
```

9.10 Exercises

9.1. **Public Health.** A manager of public health services in an area downwind of a nuclear test site wants to test the hypothesis that the mean amount of radiation in the form of strontium-90 in the bone marrow (measured in picocuries) for citizens who live downwind of the site does not exceed that of citizens who live upwind from the site. It is known that "upwinders" have a mean level of strontium-90 of 1 picocurie. Measurements of strontium-90 radiation for a sample of $n = 16$ citizens who live downwind of the site were taken, giving $\overline{X} = 3$. The population standard deviation is $\sigma = 4$. Assume normality and use a significance level of $\alpha = 0.05$.
(a) State H_0 and H_1.
(b) Calculate the appropriate test statistic.
(c) Determine the critical region of the test.
(d) State your decision.
(e) What would constitute a type II error in this setup? Describe this in one sentence.

9.2. **Testing IQ.** We wish to test the hypothesis that the mean IQ of the students in a school system is 100. Using σ = 15, α = 0.05, and a sample of 25 students the sample value $\overline{X}$ is computed. For a two-sided test find:
(a) The range of $\overline{X}$ for which we would not reject the hypothesis.

(b) If the true mean IQ of the students is 105, find the probability of falsely not rejecting $H_0 : \mu = 100$.
(c) What are the answers in (a) and (b) if the alternative is one-sided, $H_1 : \mu > 100$?

9.3. **Bricks.** A purchaser of bricks suspects that the quality of bricks is deteriorating. From past experience, the mean crushing strength of such bricks should be 400 pounds. A sample of $n = 100$ bricks yields a mean of 395 pounds and standard deviation of 20 pounds.
(a) Test the hypothesis that the mean quality has not changed against the alternative that it has deteriorated. Choose $\alpha = 0.05$.
(b) What is the p-value for the test in (a)?
(c) Suppose that the producer of the bricks contests your findings in (a) and (b). Their company suggests that you construct the 95% confidence interval for μ with a total length of no more than 4. What sample size is needed to construct such a confidence interval?

9.4. **Soybeans.** According to advertisements, a strain of soybeans planted on soil prepared with a specific fertilizer treatment has a mean yield of 500 bushels per acre. Fifty farmers planted the soybeans. Each used a 40-acre plot and reported the mean yield per acre. The mean and variance for the sample of 50 farms are $\overline{x} = 485$ and $s^2 = 10{,}045$. Use the p-value for this test to determine whether the data provide sufficient evidence to indicate that the mean yield for the soybeans is different from that advertised.

9.5. **Great White Shark.**  One of the most feared predators in the ocean is the great white shark *Carcharodon carcharias*. Although it is known that the great white shark grows to a mean length of 14 ft. (record: 23 ft.), a marine biologist believes that the great white sharks off the Bermuda coast grow significantly longer due to unusual feeding habits. To test this claim, a number of full-grown great white sharks are captured off the Bermuda coast, measured, and then set free. However, because the capture of sharks is difficult, costly, and very dangerous, only five are sampled. Their lengths are 16, 18, 17, 13, and 20 ft.
(a) What assumptions must be made in order to carry out the test?
(b) Do the data provide sufficient evidence to support the marine biologist's claim? Formulate the hypotheses and test at a significance level of $\alpha = 0.05$. Provide solutions using both the rejection-region approach and the p-value approach.
(c) Find the power of the test against the specific alternative $H_1 : \mu = 17$.
(d) What sample size is needed to achieve the power of 0.90 in testing the preceding hypothesis if $\mu_1 - \mu_0 = 2$ and $\alpha = 0.05$. Pretend that the described experiment was a pilot study to assess the variability in data and adopt $\sigma = 2.5$.

(e) Provide a Bayesian solution using WinBUGS with noninformative priors on μ and $1/\sigma^2$ (precision). Compare with results from (b) and discuss.

9.6. **Serum Sodium Levels.** A data set compiled by Queen Elizabeth Hospital, Birmingham, and referenced in Andrews and Herzberg (1985), provides the results of analysis of 20 samples of serum measured for their sodium content. The average value for the method of analysis used is 140 ppm.

140	143	141	137	132	157	143	149	118	145
138	144	144	139	133	159	141	124	145	139

Is there evidence that the mean level of sodium in this serum is different from 140 ppm?

9.7. **Weight of Quarters.** The US Department of the Treasury claims that the procedure it uses to mint quarters yields a mean weight of 5.67 g with a standard deviation of 0.068 g. A random sample of 30 quarters yielded a mean of 5.643 g. Use an $\alpha = 0.05$ significance to test the claim that the mean weight is 5.67 g.
(a) What alternatives make sense in this setup? Choose one sensible alternative and perform the test.
(b) State your decision in terms of rejection region.
(c) Find the p-value and confirm your decision from (b).
(d) Would you change the decision if α were 0.01?

9.8. **Dwarf Plants.** A genetic model suggests that three-fourths of the plants grown from a cross between two given strains of seeds will be of the dwarf variety. After breeding 200 of these plants, 136 were of the dwarf variety.
(a) Does this observation strongly contradict the genetic model?
(b) Construct a 95% confidence interval for the true proportion of dwarf plants obtained from the given cross.
(c) Answer (a) and (b) using Bayesian arguments and WinBUGS.

9.9. **Eggs in a Nest.** The average number of eggs laid per nest each season by the Eastern Phoebe bird is a parameter of interest. A random sample of 70 nests was examined and the following results were obtained (Hamilton, 1990):

Number of eggs/nest	1	2	3	4	5	6
Frequency f	3	2	2	14	46	3

Test the hypothesis that the true average number of eggs laid per nest by the Eastern Phoebe bird is equal to five versus the two-sided alternative. Use $\alpha = 0.05$.

9.10. **Penguins.** A researcher is interested in testing whether the mean height of Emperor penguins (*Aptenodytes forsteri*) from a small island is less than $\mu = 45$ in., which is believed to be the average height for the whole Emperor penguin population. The heights were measured of 14 randomly selected adult birds from the island with the following results:

41	44	43	47	43	46	45	42	45	45	43	45	47	40

State the assumptions and hypotheses. Perform the test at the level $\alpha = 0.05$.

9.11. **Hypersplenism and White Blood Cell Count.** In Example 9.6, the belief was expressed that hypersplenism decreased the leukocyte count, so a Bayesian test was conducted. In a sample of 16 people affected by hypersplenism, the mean white blood cell count per mm^3 was found to be $\overline{X} = 5{,}213$. The sample standard deviation was $s = 1{,}682$.
(a) With this information, test $H_0 : \mu = 7{,}200$ versus the alternative $H_1 : \mu < 7{,}200$ using both the rejection region and the p-value. Compare the results with the WinBUGS output.
(b) Find the power of the test against the alternative $H_1 : \mu = 5{,}800$.
(c) What sample size is needed if, in a repeated study, a difference of $|\mu_1 - \mu_0| = 600$ is to be detected with a power of 80%? Use the estimate $s = 1{,}682$.

9.12. **Jigsaw.** An experiment with a sample of 18 nursery-school children involved the elapsed time required to put together a small jigsaw puzzle. The times in minutes were as follows:

3.1	3.2	3.4	3.6	3.7	4.2	4.3	4.5	4.7
5.2	5.6	6.0	6.1	6.6	7.3	8.2	10.8	13.6

(a) Calculate the 95% confidence interval for the population mean.
(b) Test the hypothesis $H_0 : \mu = 5$ against the two-sided alternative. Take $\alpha = 10\%$.

9.13. **Anxiety.** A psychologist has developed a questionnaire for assessing levels of anxiety. The scores on the questionnaire range from 0 to 100. People who obtain scores of 75 and greater are classified as *anxious.* The questionnaire has been given to a large sample of people who have been diagnosed with an anxiety disorder, and scores are well described by a normal model with a mean of 80 and a standard deviation of 5. When given to a large sample of people who do not suffer from an anxiety disorder, scores on the questionnaire can also be modeled as normal with a mean of 60 and a standard deviation of 10.
(a) What is the probability that the psychologist will misclassify a nonanxious person as anxious?

(b) What is the probability that the psychologist will erroneously label a truly anxious person as nonanxious?

9.14. **Aptitude Test.** An aptitude test should produce scores with a large amount of variation so that an administrator can distinguish between people with low aptitude and those with high aptitude. The standard test used by a certain university has been producing scores with a standard deviation of 5. A new test given to 20 prospective students produced a sample standard deviation of 8. Are the scores from the new test significantly more variable than scores from the standard? Use $\alpha = 0.05$.

9.15. **Rats and Mazes.** Eighty rats selected at random were taught to run through a new maze. All rats eventually succeeded in learning the maze, and the number of trials to perfect their performance was normally distributed with a sample mean of 15.4 and sample standard deviation of 2. Long experience with populations of rats trained to run a similar maze shows that the number of trials to attain success is normally distributed with a mean of 15.
(a) Is the new maze harder for rats to learn than the older one? Formulate the hypotheses and perform the test at $\alpha = 0.01$.
(b) Report the p-value. Would the decision in (a) be different if $\alpha = 0.05$?
(c) Find the power of this test for the alternative $H_1 : \mu = 15.6$.
(d) Assume that the experiment above was conducted to assess the standard deviation, and the result was 2. Design a sample size for a new experiment that will detect the difference $|\mu_0 - \mu_1| = 0.6$ with a power of 90%. Here $\alpha = 0.01$, and μ_0 and μ_1 are postulated means under H_0 and H_1, respectively.

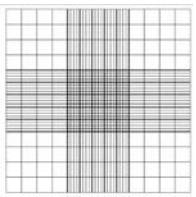
9.16. **Hemopexin in DMD Cases I.** Refer to data set  `dmd.dat|mat|xls` from Exercise 2.19. The measurements of `hemopexin` are assumed normal.
(a) Form a 95% confidence interval for the mean response of hemopexin `h` in a population of all female DMD carriers (`carrier=1`).
Although the level of pyruvate kinase seems to be the strongest single predictor of DMD, it is an expensive measure. Instead, we will explore the level of hemopexin, a protein that protects the body from oxidative damage. The level of hemopexin, in a general population of women of comparable age, is believed to be 85.
(b) Test the hypothesis that the mean level of hemopexin in the population of woman DMD carriers significantly exceeds 85. Use $\alpha = 5\%$. Report the p-value as well.
(c) What is the power of the test in (b) against the alternative $H_1 : \mu_1 = 89$.
(d) The data for this exercise come from a study conducted in Canada. If you wanted to replicate the test in the United States, what sample size would guarantee a power of 99% if H_0 were to be rejected whenever the difference from the true mean was 4, ($|\mu_0 - \mu_1| = 4$)? A small pilot

study conducted to assess the variability of hemopexin level estimated the standard deviation as $s = 12$.
(e) Find the posterior probability of the hypothesis $H_1 : \mu > 85$ using WinBUGS. Use noninformative priors. Also, compare the 95% credible set for μ that you obtained with the confidence interval in (a).
Hint: The commands

```
%file dmd.mat should be on path
load 'dmd.mat';  hemo = dmd( dmd(:,6)==1, 3);
```

will distill the levels of hemopexin in carrier cases.

9.17. **Haden's Data.**



In the past, the blood counts were performed manually using the hemocytometers with microscopic grid scoring. By properly diluting blood, counting all cells in specified squares, and multiplying by the proper conversion factor, the number of cells per cubic millimeter can be approximated.
The Coulter principle[2] led to the availability of Coulter counters and thereafter, the development of sophisticated automated blood-cell analyzers. The level of sophistication has been rising ever since.
The data set in MATLAB file `haden.m` comes from Haden (1923, Tables 1, 2, p. 770). It provides red blood cell count for 40 healthy men aged 18–50.

4.27	4.32	4.40	4.52	4.56	4.58	4.64	4.70	4.72	4.73
4.80	4.80	4.80	4.80	4.84	4.87	4.89	4.93	4.97	4.98
4.99	5.00	5.02	5.05	5.09	5.09	5.10	5.15	5.16	5.20
5.20	5.20	5.26	5.28	5.36	5.46	5.49	5.50	5.57	5.62

(a) Find 95% CI for the population mean.
(b) Test the hypothesis that the population mean from which Haden's sample was taken is 5.1, versus the alternative that it is less than 5.1.

$$H_0 : \mu = 5.1 \quad \text{versus} \quad H_1 : \mu < 5.1$$

Find both the rejection region and the p-value.
(c) What is the power of this test against the alternative $H_1 : \mu = 4.9$?
(d) You are to determine the sample size for Haden's project so that a 0.05 level, two sided test rejects the null hypothesis with probability 0.95 whenever the true mean differs from 5.1 by more than 0.1. By assuming that the population variance is $\sigma^2 = 0.16$, determine the sample size that achieves the required power.

[2] The Coulter principle states that particles pulled through an orifice by an electric current produce a change in electrical impedance that is proportional to the size of the particle traversing the orifice. This is based on the principle that cells are relatively poor conductors of electricity in relation to the diluent fluid.

9.18. **Retinol and a Copper-Deficient Diet.** The liver is the main storage site of vitamin A and copper. Inverse relationships between copper and vitamin A liver concentrations have been suggested. In Rachman et al. (1987) the consequences of a copper-deficient diet on liver and blood vitamin A storage in Wistar rats was investigated. Nine animals were fed a copper-deficient diet for 45 days from weaning. Concentrations of vitamin A were determined by isocratic high-performance liquid chromatography using UV detection. Rachman et al. (1987) observed in the liver of the rats fed a copper-deficient diet a mean level of retinol, in micrograms/g of liver, was $\overline{X} = 3.3$ and $s = 1.4$. It is known that the normal level of retinol in a rat liver is $\mu_0 = 1.6$.
(a) Find the 95% confidence interval for the mean level of liver retinol in the population of copper-deficient rats. Recall that the sample size was $n = 9$.
(b) Test the hypothesis that the mean level of retinol in the population of copper-deficient rats is $\mu_0 = 1.6$ versus a sensible alternative, either one-sided or two-sided, at the significance level $\alpha = 0.05$. Use both rejection region and p-value approaches.
(c) What is the power of the test in (b) against the alternative $H_1 : \mu = \mu_1 = 2.4$? Comment.
(d) Suppose that you are designing a new, larger study in which you are going to assume that the variance of observations is $\sigma^2 = 1.4^2$, as the limited nine-animal study indicated. Find the sample size so that the power of rejecting H_0 when an alternative $H_1 : \mu = 2.1$ is true is 0.80. Use $\alpha = 0.05$.
(e) Provide a Bayesian solution using WinBUGS.

9.19. **Rubidium.** Meltzer et al. (1973) demonstrated that there is a large variability in the amount of rubidium excreted each day, even when the amount of potassium ingested is controlled. However, when the rubidium excretion is computed as a ratio to potassium excretion, this variability is markedly diminished. Meltzer et al. concluded that the factors that normally control potassium flux operate at the same time to control rubidium flux.
The data consists of measurements on 17 hospitalized patients and represent the mean of naturally occurring rubidium-to-potassium ratio, in hundreds of mEq of *Ru* to mEq of *K*.

0.028	0.032	0.031	0.041	0.028
0.039	0.042	0.036	0.037	0.029
0.048	0.037	0.037	0.044	0.039
0.029	0.038			

Two published studies state that the ratio in healthy subjects is approx $\mu_0 = 0.036$.

(a) Assuming the normality of the ratio, test the hypothesis that population mean μ does not significantly differ from μ_0. Use $\alpha = 0.05$.
(b) How does your finding in (a) agree with the 95% CI for the population mean ratio? Is μ_0 in the confidence interval?

9.20. **Aniline.** Organic chemists often purify organic compounds by a method known as *fractional crystallization.* An experimenter wanted to prepare and purify 5 grams of aniline. It is postulated that 5 grams of aniline would yield 4 grams of acetanilide. Ten 5-gram quantities of aniline were individually prepared and purified.
(a) Test the hypothesis that the mean dry yield differs from 4 grams if the mean yield observed in a sample was $\overline{X} = 4.21$. The population is assumed normal with known variance $\sigma^2 = 0.08$. The significance level is set to $\alpha = 0.05$.
(b) Report the p-value.
(c) For what values of $\overline{X}$ will the null hypothesis be rejected at the level $\alpha = 0.05$?
(d) What is the power of the test for the alternative $H_1 : \mu = 3.6$ at $\alpha = 0.05$?
(e) If you are to design a similar experiment but would like to achieve a power of 90% versus the alternative $H_1 : \mu = 3.6$ at $\alpha = 0.05$, what sample size would you recommended?

9.21. **DNA Random Walks.** DNA random walks are numerical transcriptions of a sequence of nucleotides. The imaginary walker starts at 0 and goes one step up ($s = +1$) if a purine nucleotide (A, G) is encountered, and one step down ($s = -1$) if a pyramidine nucleotide (C, T) is encountered. Peng et al. (1992) proposed identifying coding/noncoding regions by measuring the irregularity of associated DNA random walks. A standard irregularity measure is the Hurst exponent H, an index that ranges from 0 to 1. Numerical sequences with H close to 0 are irregular, while the sequences with H close to 1 appear more smooth.
Figure 9.4 shows a DNA random walk in the DNA of a spider monkey (*Ateles geoffroyi*). The sequence is formed from a noncoding region and has a Hurst exponent of $H = 0.61$.
A researcher wishes to design an experiment in which n nonoverlapping DNA random walks of a fixed length will be constructed, with the goal of testing to see if the Hurst exponent for noncoding regions is 0.6.
The researcher would like to develop a test so that an effect $e = |\mu_0 - \mu_1|/\sigma$ will be detected with a probability of $1 - \beta = 0.9$. The test should be two-sided with a significance level of $\alpha = 0.05$. Previous analyses of noncoding regions in the DNA of various species suggest that exponent H is approximately normally distributed with a variance of approximately $\sigma^2 = 0.03^2$. The researcher believes that $|\mu_0 - \mu_1| = 0.02$ is a biologically meaningful difference. In statistical terms, a 5%-level test for $H_0 : \mu = 0.6$ versus the alternative $H_1 : \mu = 0.6 \pm 0.02$ should have a power

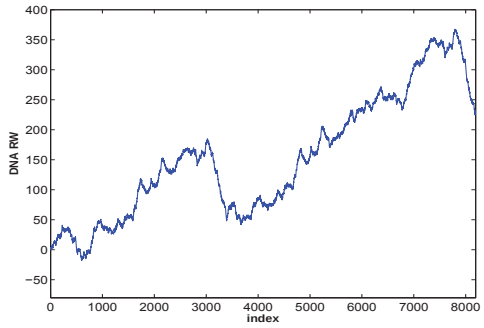


Fig. 9.4 A DNA random walk formed by a noncoding region from the DNA of a spider monkey. The Hurst exponent is 0.61.

of 90%. The preexperimentally assessed variance $\sigma^2 = 0.03^2$ leads to an effect size of $e = 2/3$.
(a) Argue that a sample size of $n = 24$ satisfies the power requirements. The experiment is conducted, and the following 24 values for the Hurst exponent are obtained:

```
 H =[0.56  0.61   0.62   0.53   0.54  0.60   0.56   0.59  ...
     0.60  0.60   0.62   0.60   0.58  0.57   0.61   0.64  ...
     0.60  0.61   0.58   0.59   0.55  0.59   0.60   0.65 ];
%   [mean(H)  std(H)]  %%% 0.5917    0.0293
```

(b) Using the t-test, test H_0 against the two-sided alternative at the level $\alpha = 0.05$ using both the rejection-region approach and the p-value approach.
(c) What is the retrospective power of your test? Use the formula with a noncentral t-distribution and s found from the sample.

9.22. **Binding of Propofol.** Serum protein binding is a limiting factor in the access of drugs to the central nervous system. Disease-induced modifications of the degree of binding may influence the effect of anaesthetic drugs.
The protein binding of *propofol*, an intravenous anaesthetic agent that is highly bound to serum albumin, has been investigated in patients with chronic renal failure. Protein binding was determined by the ultrafiltration technique using an Amicon Micropartition System, MPS-1.
The mean proportion of unbound propofol in healthy individuals is 0.96, and it is assumed that individual proportions follow a beta distribution, $\mathcal{B}e(96,4)$. Based on a sample of size $n = 87$ patients with chronic renal failure, the average proportion of unbound propofol was found to be 0.93 with a sample standard deviation of 0.12.
(a) Test the hypothesis that the mean proportion of unbound propofol in a population of patients with chronic renal failure is 0.96 versus the

one-sided alternative. Use $\alpha = 0.05$ and perform the test using both the rejection-region approach and the p-value approach. Would you change the decision if $\alpha = 0.01$?
(b) Even though the individual measurements (proportions) follow a beta distribution, the normal theory could be used in (a). Why?

9.23. **Improvement of Surgical Procedure.** Refer to Example 9.4.
(a) What is the probability of the surgeon having no fatalities in treating 15 patients if the mortality rate is 10%?
(b) The surgeon claims that his new surgical technique significantly improves the survival rate. Is his claim justified? Conduct the test and report the p-value. Note that np_0 here is small, so the z test based on normal approximation may not be accurate.
(c) What is the minimum number of patients the surgeon needs to treat without a single fatality in order to convince you that his procedure is a significant improvement over the old technique? Specify your criteria and justify your answer.
(d) Conduct the test in a Bayesian manner as in Example 9.4. Find the posterior probability of H_0 if the prior ξ on $[0, 0.1)$ is $\xi(\theta) = 200\ \theta$.

9.24. **Cancer Therapy.** Researchers in cancer therapy often report only the number of patients who survive for a specified period of time after treatment rather than the patients' actual survival times. Suppose that 40% of the patients who undergo the standard treatment are known to survive 5 years. A new treatment is administered to 200 patients, and 92 of them are still alive after a period of 5 years.
(a) Formulate the hypotheses for testing the validity of the claim that the new treatment is more effective than the standard therapy.
(b) Test with $\alpha = 0.05$ and state your conclusion; use the rejection-region method.
(c) Perform the test by finding the p-value.
(d) What is the power of the test in (a) against the alternative $H_1 : p = 0.5$?
(e) What sample size is needed so that effect $p_1 - p_0 = 0.1$ is found significant in the $\alpha = 0.05$ level testing with the power of 90%? As before, $p_0 = 0.4$.

9.25. **Is the Cloning of Humans Moral?** The Gallup Poll estimates that 88% of Americans believe that cloning humans is morally unacceptable. Results are based on telephone interviews with a randomly selected national sample of $n = 1,000$ adults, aged 18 and older.
(a) Test the hypothesis that the true proportion is 0.9, versus the two-sided alternative, based on the Gallup data. Use $\alpha = 0.05$.
(b) Does 0.9 fall in the 95% confidence interval for the proportion?
(c) What is the power of this test against the alternative $H_1 : p = 0.85$?

9.26. **Smoking Illegal?** In a recent Gallup poll of Americans, fewer than a third of respondents thought smoking in public places should be made illegal, a significant decrease from the 39% who thought so in 2001.
The question used in the poll was: *Should smoking in all public places be made totally illegal*? In the poll, 497 people responded and 154 answered yes. Let p be the proportion of people in the US voting population supporting the idea that smoking in public places should be made illegal.
(a) Test the hypothesis $H_0 : p = 0.39$ versus the alternative $H_1 : p < 0.39$ at the level $\alpha = 0.05$.
(b) What is the 90% confidence interval for the unknown population proportion p?

9.27. **Spider Monkey DNA.** An 8,192-long nucleotide sequence segment taken from the DNA of a spider monkey (*Ateles geoffroyi*) is provided in the file dnatest.m.
(a) Find the relative frequency of adenine $\hat{p}_A$ as an estimator of the overall population proportion, p_A.
(b) Find a 99% confidence interval for p_A and test the hypothesis $H_0 : p_A = 0.2$ versus the alternative $H_1 : p_A > 0.2$. Use $\alpha = 0.05$.

MATLAB AND WINBUGS FILES AND DATA SETS USED IN THIS CHAPTER

http://statbook.gatech.edu/Ch9.Testing/

bayestestprecise.m, bird.m, ConfidenceEllipse.m, corkraotest.m, dnarw.m, dnatest.m, exactpowerprop.m, FDR.m, hemopexin1.m, hemoragic.m, hypersplenism.m, LDLCLevels.m, moon.m, powerT2.m, powers.m, SBB.m

bird.odc, hemopexin.odc, hemorrhagic.odc, hypersplenism.odc, moonillusion.odc, retinol.odc, shark.odc, spikes.odc, systolic.odc

bird.dat|mat|xlsx, dnadat.mat|txt, haden.mat, spid.dat

CHAPTER REFERENCES

Andrews, D. F. and Herzberg, A. M. (1985). *Data. A Collection of Problems from Many Fields for the Student and Research Worker*. Springer, New York.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of p-values and evidence (with discussion). *J. Am. Stat. Assoc.*, **82**, 112–122.

Casella, G. and Berger, R. (2001). *Statistical Inference*, 2nd ed. Duxbury Press, Belmont, CA.

Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychol. Rev.*, **70**, 193–242.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

Fisher, R. A. (1926). The arrangement of field experiments. *J. Ministry Agricult.*, **33**, 503–513.

Goodman, S. (1999a). Toward evidence-based medical statistics. 1: The *p*-value fallacy. *Ann. Intern. Med.*, **130**, 995–1004.

Goodman, S. (1999b). Toward evidence-based medical statistics. 2: The Bayes factor. *Ann. Intern. Med.*, **130**, 1005–1013.

Goodman, S. (2001). Of *p*-values and Bayes: a modest proposal. *Epidemiology*, **12**, 3, 295–297

Haden, R. L. (1923). Accurate criteria for differentiating anemias. *Arch. Intern. Med.*, **31**, 5, 766–780.

Hamilton, L. C. (1990). *Modern Data Analysis: A First Course in Applied Statistics.* Brooks/Cole, Pacific Grove, CA.

Hoenig, J. M. and Heisey, D. M. (2001). Abuse of power: the pervasive fallacy of power calculations for data analysis. *Am. Statist.*, **55**, 1, 19–24.

Ioannidis, J. P. (2005). Why most published research findings are false. PLoS Med 2(8): e124. doi:10.1371/journal.pmed.0020124.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Oxford University Press, Oxford, UK.

Johnson, R. A. and Wichern,D. W. (2002). *Applied Multivariate Statistical Analysis*, 5th ed. Prentice Hall, NY.

Kaufman, L. and Rock, I. (1962). The moon illusion, I. *Science*, **136**, 953–961.

Katz, S., Lautenschlager, G. J., Blackburn, A. B., and Harris, F. H. (1990). Answering reading comprehension items without passages on the SAT. *Psychol. Sci.*, **1**, 122–127.

Meltzer, H. L., Lieberman, K. W., Shelley, E. M., Stallone, F., and Fieve, R. R. (1973). Metabolism of naturally occurring *Rb* in the human: the constancy of urinary Rb-K. *Biochem Med.*, **7**, 2, 218–225. PubMed PMID: 4704456.

Peng, C. K., Buldyrev, S. V., Goldberger, A. L., Goldberg, Z. D., Havlin, S., Sciortino, E., Simons, M., and Stanley, H. E. (1992). Long-range correlations in nucleotide sequences. *Nature*, **356**, 168–170.

Rachman, F., Conjat, F., Carreau, J. P., Bleiberg-Daniel, F., and Amedee-Maneseme, O. (1987). Modification of vitamin A metabolism in rats fed a copper-deficient diet. *Int. J. Vitamin Nutr. Res.*, **57**, 247–252.

Schervish, M. (1996). *P*-values: what they are and what they are not. *Am. Stat.*, **50**, 203–206.

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001). Calibration of *p* values for testing precise null hypotheses. *Am. Stat.*, **55**, 62–71.