Chapter 1

Exploration of Biological Data

Classifying and Representing Data

At the heart of most scientific studies is the notion of repeated measurements, or counts, of traits among "individuals." The typical biological experiment involves "sampling" individuals and counting or measuring traits (or a combination of the two). Because traits are expected to vary to some degree across individuals they are typically referred to as **variates** or **variables**. Despite the huge diversity of traits measured by biologists there are some general approaches to classifying and representing data that are broadly applicable. We therefore begin by classifying the basic types of data generated from biological experiments and then focus on ways to represent data as abstract mathematical entities for use in a statistical analysis.

Discrete vs continuous variates

A typical biological experiment might involve DNA sequencing of the same gene from each of a set of sampled individuals. Another experiment might measure individual body weights for a sample of male undergraduates. In the DNA sequencing experiment, each distinct sequence that is observed can be uniquely indexed using a positive integer number (e.g., 0,1,2, etc) and such variables are referred to as **discrete variables**. In the experiment examining male undergraduate body weights, individual weights can be viewed as essentially unique; if one could very accurately measure body weight in kilograms (to say 12 decimal places) it is very likely that (even in a large sample of undergraduates) each individual would have a different weight. Thus, the weight measurements must be described using the real numbers (e.g., 1.41234..., 1.0, 6, etc) and such variables are referred to as **continuous variables**.

The relationship between two or more measurements of a continuous variable is defined by their relative positions on the real number line. Discrete variables, on the other hand, can be arbitrarily indexed and may not have any natural ordering. For example, the variable eye color (blue, green, etc) implies no obvious ordering among a set of observations (unless, of course, we redefine the colors according to the wavelength of light reflected). Discrete variables that have no natural rank order are referred to as **nominal** or **categorical** variables. Those variables for which the possible observations have a defined rank order (for example, the number of bristles on the abdomen of a fruit fly) are referred to as **ordinal** variables.

As is done in algebra, in statistics we use letters to represent variables. Statistical practice dictates that an upper case letter denotes a variable "class" (for example Y might be the symbol used for male undergraduate body weight) while a lower case letter denotes a particular "instance" of a variable (for example y might be the weight of a particular male undergraduate). Such formalism is uneccessary for the purposes of this text as the meaning of a symbol for a variable will usually be clear from the context; we use lower case letters to denote either classes, or instances, of random variables throughout. A collection of individuals for which a trait has been measured are represented as a vector. By convention, a bold letter is used to indicate a vector and an index is used to denote an individual instance of a variable. For example,

$$\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$$

might represent the variable body weight for a sample of n male undergraduates where y_1 is the weight of the first undergraduate, y_2 the weight of the second, and so on.

Univariate vs multivariate data

If only a single trait is measured for each individual, the data are said to be **univariate**. This was the case, for example, with the male undergraduate body weight measurements. If more than one trait is measured per individual, the data are said to be **multivariate**. For example, if one were to measure the heights of a sample of male undergraduates as well as the weights the data would be **bivariate** (a special instance of multivariate data). Sometimes, it is possible to think of the same data as either univariate or multivariate. For example, the DNA gene sequence data generated by the first experiment we considered might be thought of as univariate (each gene sequence from an individual is a single observation) or multivariate (the nucleotide at each site of each sequence from each individual is a single observation). The perspective that is taken depends on the interests of the experimenter.

A collection of individuals for which a multivariate trait has been measured are represented as a matrix. For example,

$$\mathbf{x} = \left\{ \begin{array}{ccccc} x_{11} & x_{12} & x_{13} & x_{14} & x_{15} \\ x_{21} & x_{22} & x_{23} & x_{24} & x_{25} \\ x_{31} & x_{32} & x_{33} & x_{34} & x_{35} \\ x_{41} & x_{42} & x_{43} & x_{44} & x_{45} \\ x_{51} & x_{52} & x_{53} & x_{54} & x_{55} \end{array} \right\}$$

might represent the variable "DNA nucleotide sequence" for an experiment sequencing a small region (5 nucleotides) for a sample of 5 individuals. In this case, the rows represent the individuals and the columns the nucleotide sites, so that x_{12} is the nucleotide (`T`, `C`, `A`, or `G`) present at site 2 from individual 1, and so on.

Quantifying Error

All measurements in biology, and other sciences, involve some degree of error. It is important to quantify such errors as otherwise any particular measurement will be uninterpretable. In practice, uncertainties are often implicit in statements about measurements. The statement that a gene is 312 kb (1kb = 1000 nucleotides) in length, for example, implies that the measurement is accurate within 1 kb, and so the actual length of the gene might range from 311.5 kb to 312.5 kb. In many circumstances, it is preferable to make an explicit statement about the uncertainty associated with a measurement because implicit error statements are prone to misinterpretation.

Measurements are often collected with the intention of either directly comparing them with other measurements that have been collected, or incorporating the measurements into calculations. In the former case, a procedure is needed for deciding when two or more measurements support the notion that the underlying variables are identical (apart from errors of measurement) or instead "significantly" different. In the latter case, a procedure is needed for establishing the error of the number that results from a calculation involving measurements with specified uncertainties.

The objective of this section is to describe methods for presenting measurements that concisely convey their uncertainties (e.g., significant digits and confidence ranges) as well as methods for comparing measurements (variables) with specified uncertainties and for predicting the uncertainty of the values obtained when evaluating functions of variables with specified uncertainties (e.g., sums, differences, products, etc). This is commonly referred to as "propagation of error" because the procedures carry the errors forward so that the error of the final number resulting from such a calculation can be determined.

Significant digit

The uncertainty of a measurement is often determined by the precision of the instrument being used. For example, using a ruler with increments of millimeters, if one is careful, it is possible to obtain measurements to the nearest millimeter. If an insect is measured to be 5 mm, the investigator might confidently conclude that the actual length lies somewhere between 4.5 mm and 5.5 mm. In expressing this measurement, the **significant digit** should be of the same order of magnitude as the **error of measurement**, in this case the first decimal place (since the range of error is 1 mm). Thus, one would write the measurement as 5 mm.

Roundoff error

If a number results from a calculation (such as division or multiplication), rather than a direct measurement, it is often neccessary to **round off** the number, truncating it to the correct significant digit. If a rounded estimate, denoted $x^{(d)}$, of the original variable x is accurate to the dth decimal place, the number has been correctly rounded if the **roundoff error**, ϵ, is

$$|\epsilon| = \left|x - x^{(d)}\right| \leq \frac{1}{2} \times 10^{-d}.$$

There are simple rules for rounding that guarantee a correct roundoff error. Suppose that we truncate the original number x at the dth decimal place. Let $x_R = x - x'$ be the remainder obtained by subtracting the truncated number $x'$from x. If $x_R < 1/2 \times 10^{-d}$ we "round down" leaving the significant digit (the last digit of the truncated number x') unaltered. Otherwise, if $x_R > 1/2 \times 10^{-d}$ we "round up" increasing the value of the signicant digit by one. If $x_R = 1/2 \times 10^{-d}$ then we need a new procedure to obtain the correct rounding error. Two approaches are commonly used: (1) either round up, or round down, as needed to insure that the significant digit is either always even, or always odd; (2) choose to either round up or round down at random, each with probability 1/2. Based on this inequality, the following rules of thumb can be applied to round numbers:

1. If the number following a significant digit is less than or equal to 4 leave the significant digit unchanged and truncate.
2. If the number following a significant digit is greater than 4, add 1 to the significant digit and truncate (except in the case outlined in 3 below).
3. If the number following the significant digit is 5, and all trailing numbers are 0, consistently apply one of the following rules: a) if the significant digit is even round up otherwise leave unchanged and truncate. b) with probability 1/2 add 1 to the significant digit otherwise leave unchanged and truncate.

Examples illustrating rounding off of numbers to different numbers of decimal places are given in Table 1.1. To avoid propagating roundoff errors when calculations are carried out using measurements it is best to wait until after the calculation is completed (retaining at least one additional significant digit during the calculations) before truncating (and rounding) the number to the desired significant digits. This was of much greater concern in the past when pocket calculators were commonly used for statistical calculations. Most computer packages for statistical analysis maintain a very large number of digits when carrying out calculations (usually at least 16 digits and often many more) and many mathematics packages now allow arbitrary precision calculations. Rounding errors are therefore less of a

Original number	Significant decimal	Rounded number
4.21778	2	4.22
4638	-2	4600
452.6131 . . .	3	452.613
6.12500	2	6.13
6.13500	2	6.13

Table 1.1: Examples of roundoff of numbers with different significant decimal places. Note that in the last two rows the strategy of either rounding up or down to insure an odd significant digit is used.

concern. A simple method to check for rounding errors in a multipart calculation is to rearrange the order of the terms (preserving rules of operator precedence) and recalculate the number. If changing the order in which operations are carried out influences the value of the final number, numerical inaccuracies due to rounding errors should be suspected.

Error analysis

The significant digit expresses only the order of magnitude of the measurement error. Sometimes the error interval needs to be expressed more precisely. The interval of plausible values for a variable x, with symmetrical measurement error δ, is written in the form $x \pm \delta$. The error of measurement, δ, should be rounded to one significant digit unless the leading digit is one (in which case two significant digits should be retained). For example, if $x = 0.21$ and $\delta = 0.022$ then we write $x = 0.21 \pm .02$. If $\delta = 0.012$ we instead write $x = 0.21 \pm .012$. Two significant digits are retained when the first significant digit is 1 because the trailing digits have a greater influence on the percent error due to rounding in that case. For example, if the true value of δ is 0.15 and we round to 0.1 the percent rounding error is $(0.15 - 0.1)/0.15 = 0.33...$, while if the true value is 0.95 and we round to 0.9 the percent rounding error is $(0.95 - 0.9)/0.95 = 0.053...$ which is 6 times smaller.

To compare two variables with the aim of determining whether they are identical (apart from measurement errors) one compares the smallest plausible value of the larger variable with the largest plausible value of the smaller variable. If there is no overlap then the evidence suggests they are different. For example, if $y_1 = 4.2 \pm .3$ and $y_2 = 5.6 \pm .4$ then we are confident that the largest plausible value for y_1 is 4.5 and the smallest plausible value for y_2 is 5.2. There is no overlap and we therefore reject the possibility that variables y_1 and y_2 are identical.

Often, biologists are interested in functions of variables. For example, the sum of a pair of variables, y_1 and y_2,

$$S = y_1 + y_2.$$

If y_1 and y_2 are measurements of the amount of rain falling on each of two days, for example, we might be interested in the total amount of rain that has fallen during a two day period. If δ_1 and δ_2 are the respective errors of measurement for variables y_1 and y_2 then the smallest plausible value of S is given by the sum of the smallest plausible values of each variable,

$$S_{min} = (y_1 - \delta_1) + (y_2 - \delta_2) = y_1 + y_2 - \delta_1 - \delta_2 = S - (\delta_1 + \delta_2),$$

and the largest plausible value of S is given by the sum of the largest plausible values of each variable,

$$S_{max} = (y1 + \delta_1) + (y_2 + \delta_2) = y_1 + y_2 + \delta_1 + \delta_2 = S + (\delta_1 + \delta_2).$$

The error of measurement for the sum is then

$$\delta_S = \delta_1 + \delta_2,$$

and the uncertainty of S can be concisely expressed as

$$S \pm \delta_S.$$

The error of measurement for a difference of two variables

$$D = y_1 - y_2$$

can be similarly derived. The smallest plausible value of D is given by the difference of the smallest plausible value of y_1 and the largest plausible value y_2,

$$D_{min} = (y_1 - \delta_1) - (y_2 + \delta_2) = (y_1 - y_2) - (\delta_1 + \delta_2) = D - (\delta_1 + \delta_2),$$

and the largest plausible value of D is given by the difference of the largest plausible value of y_1 and the smallest plausible value of y_2,

$$D_{max} = (y_1 + \delta_1) - (y_2 - \delta_2) = (y_1 - y_2) + (\delta_1 + \delta_2) = D + (\delta_1 + \delta_2).$$

In general, the error of measurement for any combination of sums and differences of n variables

$$Z = y_1 \pm y_2 \pm \cdots \pm y_n$$

is

$$\delta_Z = \sum_{i=1}^{n} \delta_i.$$

Before deriving the uncertainties of more complicated functions of measured variables, it is helpful to first define the concept of relative error. The **relative error** of a variable y is

$$\epsilon_y = \frac{\delta_y}{|y|},$$

where $|y|$ denotes the absolute value of y (see Appendix 1). This is also referred to as the "percentage uncertainty" of y. The relative error is useful for comparing the uncertainties of variables of very different magnitude. For example, when comparing the uncertainty of cranial volume measurements on mice and elephants the relative error may be more interesting than the absolute error because it is a dimensionless quantity – it remains constant, regardless of the units of measurement, the relative magnitudes of the quantities, etc. The uncertainty of a variable y can be compactly expressed using its relative error, ϵ_y, as

$$y \times (1 \pm \epsilon_y).$$

If an experimental measurement has relative errors of a few percent it can be viewed as quite accurate, whereas errors greater than 10 percent are probably unacceptable for most scientific studies. We now consider the uncertainty of a product of two variables

$$P = y_1 \times y_2.$$

The largest plausible value of the product will occur when both variables attain their largest plausible values,

$$\begin{aligned} P_{max} &= y_1 \times (1 + \epsilon_{y_1}) \times y_2 \times (1 + \epsilon_{y_2}) \\ &= y_1 y_2 \times (1 + \epsilon_{y_1})(1 + \epsilon_{y_2}) \\ &= y_1 y_2 \times (1 + \epsilon_{y_1} + \epsilon_{y_2} + \epsilon_{y_1}\epsilon_{y_2}). \end{aligned}$$

Now if the relative errors are quite small (less than 5%) then the product $\epsilon_{y_1}\epsilon_{y_2}$ will be extremely small (less than .25%) and may be safely neglected so that the largest plausible value is

$$P_{max} = y_1 y_2 \times (1 + \epsilon_{y_1} + \epsilon_{y_2})$$

Similarly, it can be shown that the smallest plausible value is

$$P_{min} = y_1 y_2 \times (1 - [\epsilon_{y_1} + \epsilon_{y_2}]).$$

Thus, the relative error of the product of two variables is given by the sum of their relative errors

$$P \times (1 \pm \epsilon_P),$$

where $\epsilon_P = \epsilon_{y_1} + \epsilon_{y_2}$. Similarly, we can derive the relative error of the ratio of two variables

$$R = \frac{y_1}{y_2}.$$

The largest plausible value of the ratio is achieved when the numerator takes on its largest plausible value and the denominator takes on its smallest plausible value,

$$R_{max} = \frac{y_1(1+\epsilon_{y_1})}{y_2(1-\epsilon_{y_2})} = \frac{y_1}{y_2} \times \left(\frac{1}{1-\epsilon_{y_2}} \times [1+\epsilon_{y_1}] \right).$$

The binomial theorem (see Appendix 1) specifies that $1/(1-\epsilon_{y_2})$ can be represented as the infinite sum,

$$\frac{1}{1-\epsilon_{y_2}} = 1 + \epsilon_{y_2} + \epsilon_{y_2}^2 + \epsilon_{y_2}^3 + \cdots$$

If we again assume that the relative error is small and discard higher order terms such as ϵ^2, ϵ^3, etc, in the above equation we can approximate R_{max} as

$$\begin{aligned} R_{max} &\approx \frac{y_1}{y_2} \times (1+\epsilon_{y_2}) \times (1+\epsilon_{y_1}) \\ &= \frac{y_1}{y_2} \times (1+\epsilon_{y_1}+\epsilon_{y_2}+\epsilon_{y_1}\epsilon_{y_2}) \\ &\approx \frac{y_1}{y_2} \times (1+[\epsilon_{y_1}+\epsilon_{y_2}]), \end{aligned}$$

and we can obtain R_{min} as

$$\begin{aligned} R_{min} &= \frac{y_1(1-\epsilon_{y_1})}{y_2(1+\epsilon_{y_2})} \\ &= \frac{y_1}{y_2} \times \left(\frac{1}{1+\epsilon_{y_2}} \times [1-\epsilon_{y_1}] \right), \end{aligned}$$

where we again make use of the binomial theorem,

$$\frac{1}{1+\epsilon_{y_2}} = \frac{1}{1-(-\epsilon_{y_2})} = 1 - \epsilon_{y_2} + \epsilon_{y_2}^2 - \epsilon_{y_2}^3 + \cdots$$

and neglect terms involving products and powers of the error terms to obtain the approximation for R_{min},

$$R_{min} \approx (1-\epsilon_{y_2}) \times (1-\epsilon_{y_1}) \approx (1-[\epsilon_{y_2}+\epsilon_{y_1}]).$$

Thus, the relative error of the ratio of two variables is given by the sum of the relative errors of each variable,

$$R(1 \pm \epsilon_R) \text{ where } \epsilon_R = \epsilon_{y_1} + \epsilon_{y_2}.$$

In general, the relative error of any combination of products and ratios of variables

$$X = \prod_{i=1}^{n} y_i^{\pm 1}$$

is equal to the sum of the relative errors,

$$\epsilon_X = \sum_{i=1}^{n} \epsilon_{y_i}.$$

The following general formula for an upper bound on the error can be applied to place upper and lower bounds on the confidence interval of an arbitrary function $f(\mathbf{x})$ of k variables,

$$\delta_{f(\mathbf{x})} \leq \left| \frac{\partial f(\mathbf{x})}{\partial x_1} \right| \delta_{x_1} + \cdots + \left| \frac{\partial f(\mathbf{x})}{\partial x_k} \right| \delta_{x_k}. \tag{1.1}$$

In the above equation we are taking partial derivatives of the function with respect to each of the variables (see Appendix 1). If we can assume that errors are symmetrically distributed around the estimated value and follow a normal distribution (see Chapter 2), the error can be further reduced to

$$\delta_{f(\mathbf{x})} \leq \sqrt{\left(\left| \frac{\partial f(\mathbf{x})}{\partial x_1} \right| \delta_{x_1} \right)^2 + \cdots + \left(\left| \frac{\partial f(\mathbf{x})}{\partial x_k} \right| \delta_{x_k} \right)^2}. \tag{1.2}$$

Considering our earlier example study of body weights of male undergraduates, the height-to-weight ratio might be a useful measure of obesity. The two variables, x_1 (height) and x_2 (weight) are related by the following function,

$$f(x_1, x_2) = \frac{x_1}{x_2},$$

and the partial derivatives are,

$$\frac{\partial f(x_1, x_2)}{\partial x_1} = \frac{1}{x_2}, \text{ and } \frac{\partial f(x_1, x_2)}{\partial x_2} = \frac{-x_1}{x_2^2},$$

so that an upper bound on the confidence interval is

$$\delta_{f(x_1,x_2)} \leq \frac{\delta_{x_1}}{x_2} + \frac{\delta_{x_2} x_1}{x_2^2} = \frac{x_1}{x_2}(\epsilon_{x_1} + \epsilon_{x_2}),$$

which agrees with the formula for the uncertainty of a ratio of variables derived earlier. Thus, if an individual's height is 180 ± 1 cm and his weight is 80.0 ± .1 kg, the predicted ratio (and confidence interval) are 2.25 ± .015 cm/kg. Note that we have retained an additional significant digit because the leading digit of the error term was 1. It is easy to verify the correctness of this result by considering the maximum and minimum values obtained by the ratio given the error bounds. For example, the maximum value is achieved when the height is 181 cm and the weight is 79.9 kg, which gives a ratio of 2.265 cm/kg, while the minimum value is achieved when the height is 179 cm and the weight is 81.1 kg, which gives a ratio of 2.235 cm/kg. These are, of course, the upper and lower bounds predicted from the formula. Example 1.1 illustrates the application of the general error formula of equation 1.2 to examine temporal distributions of corals using core sampling procedures.

EXAMPLE 1.1 Aronson et al. (2004) examined temporal species compositions of corals at sampling locations in two reef systems, one in northwestern Panama and the other in Belize. Reef cores were extracted and layers were characterized according to the dominant coral species. The upper and lower boundaries of the layers were dated using radiocarbon dating procedures. One question of interest was the relative duration of each layer measured in units of conventional radiocarbon years before 1950. The radiocarbon dating procedure provides a 95 % confidence interval (CI) for each age estimate and (assuming a normal distribution of errors) equation 1.2 can be used to derive the approximate CI for the duration of each layer. The duration is given by $D = T_{low} - T_{up}$, where $T_{up} \pm \delta_{T_{up}}$ is the radiocarbon date (and standard error) for a sample from the upper end of the layer and $T_{low} \pm \delta_{T_{low}}$ is the same quantity for a sample from the lower end. Applying our general formula gives $\delta_D = \sqrt{\delta^2_{T_{up}} + \delta^2_{T_{low}}}$. Data from for two species of coral sampled in northwestern Panama are presented below.

Core	Station	Bottom date	Top date	Duration of layer
A) *Agaricia tenuifolia* layers				
P00-1	A	1320 ± 60	950 ± 60	$370 \pm 80*$
P00-2	A	1030 ± 60	590 ± 60	$440 \pm 80*$
P01-36	C	700 ± 40	530 ± 40	$170 \pm 60*$
A) *Acropora cervicornis* layers				
P00-5	D	1560 ± 70	1530 ± 70	30 ± 100
P00-6	D	1300 ± 60	1030 ± 60	$270 \pm 80*$
P00-9	D	1450 ± 70	540 ± 70	$910 \pm 100*$

Considering the first row of the table (core sample P00-1), we see from columns 3 and 4 that $\delta_{T_{low}} = \delta_{T_{up}} = 60$ and therefore $\delta_D = \sqrt{60^2 + 60^2} = 84.9$. Rounding to one significant digit the error becomes $\delta_D = 80$ as given in column 5. If the confidence interval of the difference between dates does not include 0, the dates are significantly different (indicated by an asterisk).

Unknown Measurement Error

In some cases, the error of a measurement is unknown. For example, the precision of a novel measuring instrument may be uncertain. If the errors are not systematic, the uncertainty of measurements can be established by repeated measurement. **Systematic errors** are not equally distributed around the true value of a measurement. For example, a scale that always underestimates the true weight of an object has systematic error, whereas a scale that is equally likely to overestimate, or underestimate, the weight at each trial does not. Systematic errors can be difficult to detect and may require comparisons of results obtained using different measuring instruments, etc. If errors are not systematic, then one can estimate the minimum, and maximum, plausible values of a variable using the smallest, y_{min}, and largest, y_{max}, observations, respectively, in a sample of n repeated measurements with the inferred best value of the measurement to be the midpoint between these extremes $(y_{max} + y_{min})/2$.

Summarizing Data

The first steps in analyzing newly gathered data usually involve summarizing the data by various techniques to examine a number of very general properties. This includes graphical analysis aimed at examining the distribution of values for one or more variables in a sample, as well as the calculation of various functions of the data, referred to as **statistics**. The goal of this section will be to describe several general methods for summarizing data both graphically and through the use of summary statistics.

Frequency distributions

The first stage of analysis with any data is often to summarize the observations in the form of a frequency distribution. A **frequency distribution** tabulates the number of times that different values of a variable are observed in the sample. For categorical (nominal) variables, this will simply be a list of possible values of the variable and their frequencies (counts) in the sample. For example, suppose that 30 individuals were genotyped for a single nucleotide polymorphism (e.g., a specific site in a DNA sequence at which the nucleotide present is known to be vary among individuals). The outcome of the experiment is

$$\mathbf{y} = \{\texttt{A}, \texttt{T}, \texttt{T}, \texttt{G}, \texttt{T}, \texttt{T}, \texttt{T}, \texttt{A}, \texttt{A}, \texttt{G}, \texttt{T}, \texttt{A}, \texttt{T}, \texttt{T}, \texttt{T}, \texttt{A}, \texttt{T}, \texttt{A}, \texttt{T}, \texttt{T}, \texttt{G}, \texttt{T}, \texttt{T}, \texttt{A}, \texttt{A}, \texttt{T}, \texttt{A}, \texttt{T}, \texttt{C}, \texttt{T}\}.$$

Nucleotide	Frequency
G	3
C	1
A	9
T	17

Table 1.2: Frequency distribution of nucleotides at a single nucleotide polymorphism in a sample of $n = 30$ individuals.

The frequency distribution is presented in Table 1.2. The procedure for creating a frequency distribution using an ordinal variable is similar, except that the possible values are usually rank ordered. For example, suppose that an ornithologist counts the number of eggs in 26 cliff swallow nests. The outcome of the experiment is

$$\mathbf{w} = \{0,3,2,1,2,1,1,0,0,1,1,4,3,5,1,1,0,0,2,1,1,2,3,1,1,0\}.$$

The frequency distribution is presented in Table 1.3.

No. eggs	Frequency
0	6
1	11
2	4
3	3
4	1
5	1

Table 1.3: Frequency distribution of number of eggs per nest in a sample of $n = 26$ cliff swallows.

To construct a frequency distribution for a continuous variable, one rank-orders the observations, divides the range of possible observations into a series of intervals, and then counts the number of observations that fall into each interval. A rank ordering of the variables (in terms of increasing values) is a rearrangement such that for a vector of variables, $\mathbf{x}$, $x_i >= x_j$ if $i > j$. For example, suppose that in our earlier experiment measuring the body weight of a sample of $n = 19$ male undergraduates we obtain the following weights:

$$\mathbf{x} = \{79.4,74.2,90.1,78.9,87.4,79.6,72.0,78.7,68.4,77.6,$$

$$76.2, 63.1, 74.9, 73.1, 78.4, 67.2, 80.1, 78.9, 77.9\}.$$

Rank ordering the data points gives,

$$\begin{aligned}\mathbf{x} &= \{63.1, 67.2, 68.4, 72.0, 73.1, 74.2, 74.9, 76.2, 77.6, 77.9,\\ &\quad 78.4, 78.7, 78.9, 78.9, 79.4, 79.6, 80.1, 87.4, 90.1\}.\end{aligned}$$

If we choose intervals of width 5, we obtain the frequency distribution presented in table 1.4. In table 1.4, we have used the standard notation 60– as short-hand for the interval $60 \leq x < 65$, etc. To obtain the counts in each interval from the rank-ordered vector of weights given above, simply count the observations from left to right until the value of an observed variable exceeds the endpoint of the interval. The specified intervals must cover every possible observation, and be non-overlapping, so they are typically of the form $a \leq x < b$. Clearly, it is much easier to interpret the results of the experiment sampling body weights by examining the frequency distribution than the raw data. It is immediately clear, for example, that most individuals are in the 75 to 80 Kg range, and few are above 80 Kg.

Weight (Kg)	Number of male undergraduates
60–	1
65–	2
70–	4
75–	9
80–	1
85–	1
90–	1

Table 1.4: Frequency distribution of observed body weights (Kg) in a sample of $n = 19$ male undergraduate students.

If possible, it is preferable to use intervals of equal width in constructing a frequency distribution as this makes interpretation easier. For example, in the previous analysis if we had used intervals of width 2.5, rather than 5, for observations in the range $75 \leq x < 80$ it would have been less obvious that there is a large preponderance of weights in this range (without closely examining the interval widths). In some cases, where the distribution is very uneven it may be essential to use different interval widths. In the previous example, if we had sampled two additional individuals, one with weight 147.2 Kg and another with weight 165.1 Kg, rather than creating many additional empty intervals of width 5 in our frequency distribution (e.g., 95–, 100–, etc), it might be better to create a single additional interval of width $95 \leq x < 170$. If several intervals of varying width are used it is advisable to make the frequencies in the intervals comparable by dividing the number of observations in each interval by the width of the interval.

Histograms

A frequency distribution can be presented graphically in the form of a **histogram**. A histogram is a two dimensional plot with the horizontal axis representing the intervals and the vertical axis representing the counts in each interval. For each interval, a rectangle is constructed adjacent to the horizontal axis with width proportional to the interval width and height proportional to the number of observations in the interval. An example histogram is presented in Figure 1.1 that was constructed using the male undergraduate body weights summarized in Table 1.4. If intervals of unequal width are used in constructing a histogram this is most often accommodated in one of two ways: (1) make the width of the rectangles in the histogram proportional to the widths of the intervals while fixing the total area of each rectangle to be proportional to the number of observations in the interval; (2) make the total area of each rectangle inversely proportional to the interval width.

In constructing a histogram, an arbitrary decision must often be made about the interval widths. In general, intervals that are too narrow will contain few observations per interval and will not provide much insight beyond that available from examining the original data. Intervals that are too wide will obscure features of the distribution by combining adjacent intervals with very different numbers of observations. One strategy for choosing interval widths is to experiment with several widths and compare the appearances of the histograms choosing an interval that generates a relatively smooth distribution with at least several observations in each interval. Software packages usually also offer one or more choices of ”automated” methods for choosing interval width. For intervals of equal size, several rules have been proposed for choosing the number of intervals, I, to be used. Sturgis’ rule suggests that the interval size be chosen such that

$$I = \log_2 n + 1,$$

where $\log_2 n$ denotes the base 2 log of n (see Chapter 1) and Rice’s rule

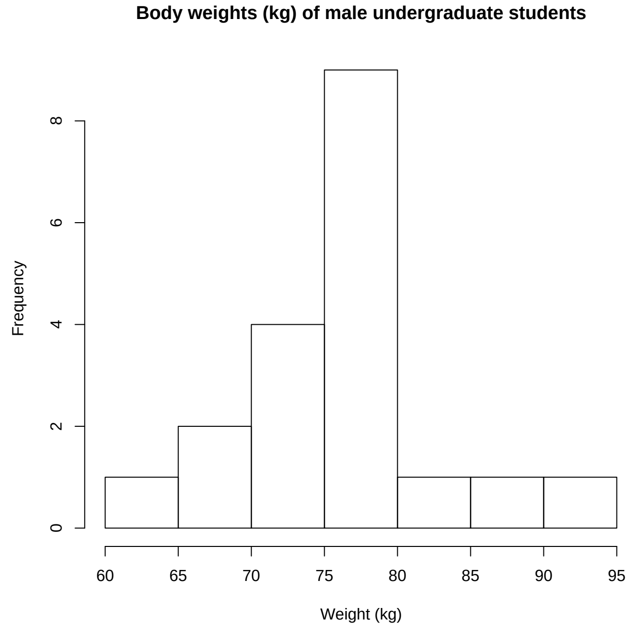


Figure 1.1: Frequency histogram of observed body weights (Kg) in a sample of $n = 19$ male undergraduate students.

specifies that it be chosen such that,

$$I = \sqrt[3]{n}.$$

For example, the number of intervals for a histogram presenting the body weights of undergraduates, according to Sturgis's rule (rounded to an integer number), is 5 and according to Rice's rule is 3. If the data are bivariate, a 3 dimensional histogram can be constructed using similar principles; in that case there are two horizonal axes and each histogram bar is a 3 dimensional rectangle with the width of each of two sides corresponding to each of the two variables and the height corresponding to the joint frequency of observations on the two variables. For example, if the two variables are male undergraduate body weight (in units of Kg) and height (in units of cm) a rectangle adjacent to the two intervals (weight, height) of (70–, 180–) would contain counts of individuals with weights in the range $70 \leq x < 75$ (Kg) and heights in the range $180 \leq y < 185$ (cm). [put box here with real data plotted as a histogram]

Summary statistics

A **summary statistic** is a simple function of the data that expresses information about properties of the sample distribution in a particularly convenient form, often as a single number. For example, a well-known summary statistic for univariate data is the sample mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{1.3}$$

where $\mathbf{x} = \{x_i\}$ is a vector of discrete (ordinal) or continuous variables. The sample mean conveys information about the "location" of the center of the sample distribution (see Figure 1.2). The mean of a sample is analogous to the expected "average" value of a random variable (see Chapter 2). Another important summary statistic is the **sample variance**. An **unbiased** estimator of the sample variance (see discussion below and Chapter 6 for an explanation of bias) is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\overline{x} - x_i)^2. \tag{1.4}$$

The sample variance provides a measure of the "dispersion" of the sample distribution. The square root of the sample variance is the sample standard deviation, $s = \sqrt{s^2}$, and is a useful measure of the spread of the distribution about the mean. A smaller s implies that the observations are more closely concentrated around the mean (see Figure 1.2). In some studies one may be interested in characterizing a population based on a sample. The classical strategy is to assume that a "true" population mean, μ, and variance, σ^2, exist and the objective is then to estimate these **parameters** from the sample data using the statistics given by equations 1.3 and 1.4 above.

Summary statistics abound in biology. Often they are based on simple "intuitive" functions of the data. An example of a summary statistic from molecular biology is the percent sequence divergence between a pair of aligned DNA sequences, $\mathbf{x} = \{x_i\}$ and $\mathbf{y} = \{y_i\}$, defined as

$$D = \frac{1}{n}\sum_{i=1}^{n}\{1 - I(x_i, y_i)\}, \tag{1.5}$$

where $I(x, y)$ is an "indicator function" (see Appendix 1) which takes the value 1 if $x = y$ and 0 otherwise. An example of a summary statistic from

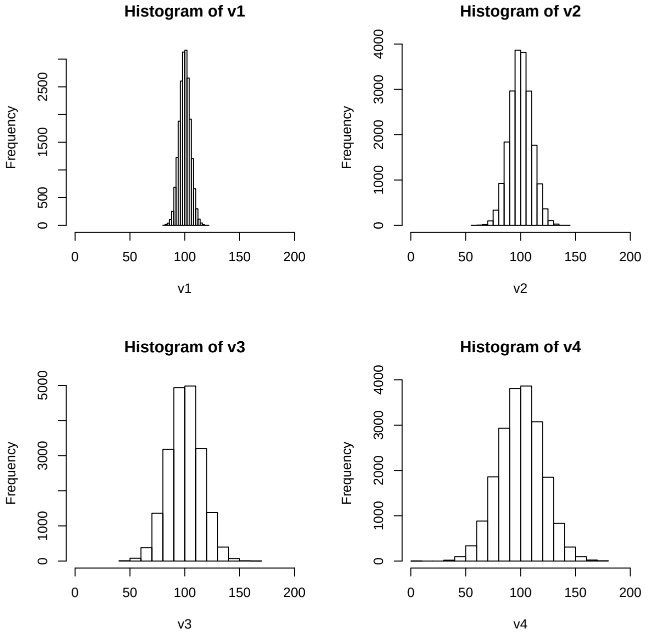


Figure 1.2: Frequency histograms of samples of 10, 000 individuals from each of 4 populations with identical means, $\overline{x} = 100$, but different standard deviations. The histograms at the upper left, upper right, lower left and lower right have standard deviations $\overline{\sigma} = 5$, $\overline{\sigma} = 10$, $\overline{\sigma} = 15$ and $\overline{\sigma} = 20$, respectively.

ecology is the Shannon index which is used to characterize the species diversity in an ecological community. The Shannon index is defined as

$$H = -\sum_{i=1}^{S} p_i \ln p_i, \tag{1.6}$$

where p_i is the proportion of species i relative to other species in the population, there are S species present in total, and ln denotes the natural logarithm (see Appendix 1). This index increases with an increase in the total number of species, or with change toward a more uniform distribution of species (e.g., more equal proportions). It is based on a statistical measure of the degree of disorder, or "surprise", in a data set known as **entropy**. [put box here with example calculating entropy for plots with invasive plant species]

Sampling Methods

A set of variates that will be the focus of a statistical analysis are often collected via a sampling experiment in which individuals are chosen for study in some well defined manner. It is extremely important that the **sampling experiment** be properly designed and carried out as otherwise the statements that a biologist would like to make about an experimental outcome may be invalid. For example, if one is interested in making statistical inferences about the distribution of body weights among male undergraduates as a whole, then one would not want to invite only individuals who are on the football team to participate as they would typically be heavier than most undergraduates. Clearly, we would want to generate a sample in such as way that the chosen subjects are representative of "typical" male undergraduates. This is usually achieved by use of **probability sampling** in which the set of possible samples is defined, as well as the probability of selecting each possible sample.

In most cases, sampling is carried out to allow statements to be made about properties of a large population by examining only a fraction of the individuals from the population. This is often much more efficient than sampling the entire population and may also be more accurate because careful measurements, etc, can be carried out on the sampled subset that would be impossible to collect for the entire population. The focus of this section will be on the design and execution of sampling studies aimed at characterising specified properties of a population. The properties, which include statistics such as the population mean and variance (defined below) are referred to as **population parameters**. This branch of statistics is known as "survey sampling" because essentially similar principles apply in designing, for example, political surveys, and other studies of opinion, behavior, etc, in human societies.

Sampling design

Before undertaking any sampling study careful attention should be given to the samping design. The objectives of the study should be defined first. What is the question that the investigator hopes to answer? The population to be studied should then be chosen such that it is representative of the populations to which the objectives apply. Next, the specific type of data to be collected should be defined, being sure that this data bears on the objectives of the study. Failing to collect the correct data to achieve the objectives, or collecting data that is unrelated to the objectives, can both jeopardize the study's chance of success. The degree of precision for the data measurements that is necessary to achieve the objectives should be determined as well as the measurement techniques that will achieve this level of accuracy. The **frame** of the study defines how the samples will be allocated. Will there be subsampling of groups within one or more larger units? If so, how will the groups be chosen? The sample selection methods then need to be defined. How will individuals be chosen for inclusion? The final stages might include preliminary tests of sampling and measurement methods to assess their feasibility, organization of the fieldwork, and design of the methods for data analysis.

Population parameters

We now focus on some particular trait of interest in a population of N individuals and suppose that individual i in the population has value x_i for the variate. We then define the **population mean** to be

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{1.7}$$

Similarly, we define the **population variance** to be

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2. \tag{1.8}$$

The **population standard deviation** is defined as the square root of the population variance, σ, and provides a measure of the spread of the frequency distribution of the trait about the mean. Another parameter that is often of interest is the **population total** for the trait which is the sum of trait values over all individuals,

$$\tau = \sum_{i=1}^{N} x_i = N\mu. \tag{1.9}$$

For example, if x_i is the number of offspring born to individual i in a given generation, then τ is the total number of births in the population. The sample mean and variance, described earlier in this Chapter, are often used as estimates of the population mean and variance. To evaluate how well these estimators work, we need to specify the sampling procedure.

Simple random sampling

When biologists speak of a **random sample** of individuals they do not mean a "haphazard" collection of individuals, but rather an organized collection scheme that incorporates, as one step, the random selection of individuals for inclusion in the sample. The most common random sampling scheme sequentially chooses individuals such that, at each stage, all individuals not yet in the sample are equally likely to be chosen for inclusion. This is referred to as **sampling without replacement**. If n individuals are sampled from a population of size N, the number of possible distinct sample configurations (see Appendix 1) is

$$ {}_N C_n = \binom{N}{n} = \frac{N!}{(N-n)!n!}, \tag{1.10}$$

and the probability that any particular sample configuration is chosen on a single random sample is $1/{}_N C_n$.

One way to generate such a sample using a computer is to label every individual in a population of N individuals using an integer index to create a vector $\mathbf{X} = \{1, 2, \ldots, N\}$ and then to generate a **random permutation** of the vector $\mathbf{X}'$ which chooses one of the $N!$ possible orderings of the integers, including that of the original vector $\mathbf{X}$ with equal probability. To generate a sample of n individuals, we then collect the individuals whose labels correspond to the first n elements of $\mathbf{X}'$.

As a simple example, consider an experiment in which we would like to sample $n = 2$ individuals from a total population of $N = 3$. In that case, we label the individuals from 1 to 3 and consider the $3! = 6$ possible permutations

$$\begin{aligned}
\mathbf{X}'(1) &= \{1,2,3\} \\
\mathbf{X}'(2) &= \{1,3,2\} \\
\mathbf{X}'(3) &= \{2,1,3\} \\
\mathbf{X}'(4) &= \{2,3,1\} \\
\mathbf{X}'(5) &= \{3,1,2\} \\
\mathbf{X}'(6) &= \{3,2,1\}
\end{aligned}$$

We then choose one of these permutations (each having an equal probability of being chosen) and the two sampled individuals are specified by the first two indexes of the permutation that we have chosen. For example, we would choose individuals 1 and 2 with probability $2/6 = 1/3$. Namely, by choosing either permutation 1 (with probability 1/6) or permutation 3 (with probability 1/6).

We now consider some statistical properties of the sample mean and variance under simple random sampling. A sample estimator of a population parameter is **unbiased** if its average value, taken over all possible sample configurations of size n, precisely equals the true value of the population parameter. The average of the sample mean over all possible sample configurations is

$$\begin{aligned} \frac{1}{{}_N C_n} \sum_{j=1}^{{}_N C_n} \overline{x}_j &= \frac{1}{{}_N C_n} \sum_{j=1}^{{}_N C_n} \left(\frac{1}{n} \sum_{i=1}^{n} x_{ij} \right), \\ &= \frac{\sum_{j=1}^{{}_N C_n} (x_{1j} + x_{2j} + \cdots + x_{nj})}{nN!/[(N-n)!n!]}, \end{aligned}$$

where x_{ij} denotes the variable observed in the ith individual of the jth sample configuration. Now let us assume that the population variates are $y_1, y_2, \ldots, y_N$. The number of sample configurations that contain some particular individual from the population y_k so that $x_{ij} = y_k$ for some i and j is

$${}_{N-1}C_{n-1} = \binom{N-1}{n-1} = \frac{(N-1)!}{(n-1)!(N-n)!}.$$

This result follows by noting that if we require that one of the sampled individuals is y_k then there are $N-1$ individuals that remain in the population from which to choose the remaining $n-1$ sampled individuals. Using this result we find that,

$$\sum_{j=1}^{{}_N C_n} (x_{1j} + x_{2j} + \cdots + x_{nj}) = \frac{(N-1)!}{(n-1)!(N-n)!} (y_1 + y_2 + \cdots + y_N),$$

and therefore

$$\frac{\sum_{j=1}^{{}_N C_n} (x_{1j} + x_{2j} + \cdots + x_{nj})}{nN!/[(N-n)!n!]} = \frac{(N-1)!(N-n)!n!}{(n-1)!(N-n)!N!n} \sum_{i=1}^{N} y_i = \frac{1}{N} \sum_{i=1}^{N} y_i.$$

Thus, the average of the sample mean over all possible sample configurations (under a simple random sampling scheme) is equal to the population mean and the sample mean is therefore unbiased. The average value of the sample variance can be similarly considered. However, the derivation is more complex and only the result is given here,

$$\frac{1}{{}_N C_n} \sum_{j=1}^{{}_N C_n} {s_j}^2 = \frac{1}{{}_N C_n} \sum_{j=1}^{{}_N C_n} \left[\frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2 \right] = \sigma^2 \left(\frac{N}{N-1} \right).$$

Thus, the sample variance tends to be an overestimate of the population variance unless the population size N is relatively large. For small populations, multiplying the sample variance by the factor $(N-1)/N$ will make the estimator unbiased. This is referred to as a **finite population correction**. Most populations are large enough (greater than about 100) that this correction is typically unimportant in practice.

Another commonly used random sampling scheme is **sampling with replacement**. In this case, n individuals are sequentially sampled from the population, but each sampled individual is returned to the population before the next is sampled so that individuals may be sampled more than once and each has a constant probability $1/N$ of being sampled at any stage. This might be the case in a mark-recapture study, for example, in which individuals in the population are equally likely to be captured and are immediately released (perhaps after banding, or recording a band number for a previously captured individual). Sampling with replacement can be viewed as equivalent to sampling from an infinitely large population with the frequency of individuals of type y_k equal to $1/N$ for all k. Thus, the correction term $(N-1)/N$ becomes 1,

$$\lim_{N\to\infty}\left(\frac{N-1}{N}\right)=1,$$

and the sample estimators of both the mean and variance are unbiased. Sampling with (and without) replacement are therefore equivalent procedures in the limit of large population size.

Stratified random sampling

Often biological populations are highly heterogeneous for a trait at a large population scale but show much more homogeneity within appropriately chosen subpopulations. As well, individuals may be sparsely distributed in some areas and densely distributed in others so that different sampling strategies are needed. In such cases, it may be efficient to first define subpopulations, or **strata**, and then to randomly sample individuals within subpopulations. **Stratified random sampling** is carried out by first dividing the population of N individuals into K subpopulations, where N_j is the number of individuals in the jth subpopulation and

$$N=\sum_{j=1}^{K}N_j.$$

Within each subpopulation a simple random sample of individuals is collected where n_j is the number of individuals sampled from population j and

$$n = \sum_{j=1}^{K} n_j.$$

We define the sample mean of subpopulation j to be

$$\overline{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij},$$

where x_{ij} denotes the variate for sampled individual i from population j. The stratified sampling estimate of the overall population mean is

$$\overline{x} = \frac{1}{N} \sum_{j=1}^{K} N_j \overline{x}_j = \sum_{j=1}^{K} w_j \overline{x}_j, \tag{1.11}$$

where $w_j = N_j / N$ is the "weight" applied to the jth subpopulation mean and is the population proportion of individuals from that subpopulation. This makes intuitive sense since large populations should have a greater influence on the overall population mean than small ones. The weighting terms are unecessary if the proportion of individuals sampled from each subpopulation equals the proportion of the total population made up of individuals from that population,

$$\frac{n_j}{n} = \frac{N_j}{N}.$$

In that case, the estimator of the population mean becomes,

$$\overline{x} = \frac{1}{N} \sum_{j=1}^{K} N_j \overline{x}_j = \frac{1}{N} \sum_{j=1}^{K} N \left(\frac{n_j}{n} \right) \overline{x}_j = \sum_{j=1}^{K} \left(\frac{n_j}{n} \right) \frac{1}{n_j} \sum_{i=1}^{n_i} x_{ij} = \frac{1}{n} \sum_{i=1}^{K} \sum_{j=1}^{n_i} x_{ij},$$

where we have substituted $N \times (n_j / n)$ for N_j in equation 1.11 to obtain the result. This type of sampling strategy is referred to as **proportional allocation**. As is the case with each subpopulation mean, the overall population mean estimated by this stratified random sampling strategy is unbiased.

If a population that is highly heterogeneous for a trait can be divided into several smaller subpopulations, each of which is relatively homogeneous for the trait, stratified random sampling can greatly improve the accuracy of estimates of the population mean. We will return to the issue of the error of estimated population means when we discuss confidence intervals in Chapter 3.

Checking assumptions

The probabilistic sampling methods described in this section assume that individuals are equally likely to be sampled, regardless of what their trait value might be, and that the trait is measured without error. Non-systematic measurement errors can be incorporated as an independent source of variance. Violations of either assumption can lead to poor, or biased, estimates of population parameters. It is therefore important to minimize the risk that these assumptions are violated by careful sampling and accurate measurements. It is also advisable to test these assumptions whenever possible.

If individuals are equally likely to be sampled at each stage, there should be no detectable relationship between the order in which individuals are sampled and their trait values. If such a relationship were detected this could indicate changing sampling probabilities over the course of a sampling experiment. Tests for trends that can be applied to detect a violation of this assumption [ref]. Another implication of the first assumption is that if several independent samples are collected from a population there should be no significant difference between parameter estimates, such as the population mean, among samples. If such differences were detected this could reflect changes in sampling probabilities between sampling experiments. Tests for differences in population parameters among independent samples are described in Chapter 7.

Random measurement errors will increase the variance for a trait within a population but should not cause the mean to be biased unless the errors are "systematic." For example, tending to underestimate or overestimate the value of a trait on average. Many deviations from the assumptions outlined above may be very difficult to detect (e.g., a segment of the population that is systematically undersampled). For example, when examining the mean length of an insect in a population the investigator may be more likely to miss small insects than large ones.

Chapter 2

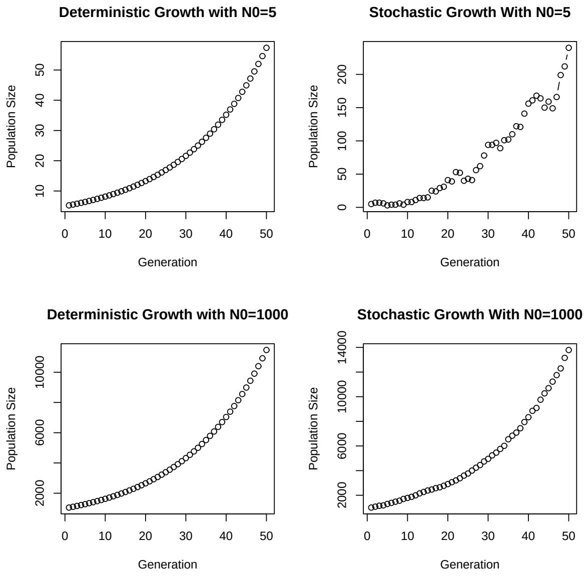
Modeling Biological Data

Modeling plays an important role in most areas of biology. To interpret the dynamics of a complicated biological system (such as a growing population), make useful predictions, etc, a mathematical model is normally needed. Two distinct types of models are typically employed: deterministic models and stochastic models. A **deterministic model** is one in which the outcomes are determined completely by the initial conditions. For example, a standard discrete generation model of population growth with a population increasing at rate r per generation specifies that, if the population is of size N_t at generation t, then it is of size $N_{t+1} = N_t(1+r)$ at generation $t+1$. Given any initial size N_0, the population size at generation t is

$$N_t = N_0(1+r)^t.$$

A **stochastic model** is one in which the outcomes are not completely determined by the initial conditions and possess some random element of variation. For example, a stochastic version of the population growth model described above might model the number of offspring born to each individual as a random quantity with $1+r$ offspring per individual on average. In this model, the population size after time t will vary randomly and the population may possibly even have gone extinct (e.g., $N_t = 0$) at or before time t.

If the population growth experiment were repeated many times (under the stochastic model), the average population size at time t across repeated experiments would be similar to that predicted by the deterministic model but the population size realized in any particular experiment would often be quite different. In general, stochasticity has a more important influence in modeling small populations. This is illustrated in Figure 2.1 in which population trajectories under the stochastic growth model described above are shown, as well as trajectories under the deterministic growth model with either a small initial population size ($N_0 = 5$) or a much larger initial population size ($N_0 = 1000$) and a growth rate of 5 percent per generation. Clearly, the trajectory is much closer between the stochastic and deterministic models when the larger population size is used. Deterministic models are often used in ecology when modeling the dynamics of large populations, under different environmental influences for example, or in population genetics when modeling the changes of gene frequency over time, in large populations, under the influence of natural selection.



Deterministic Growth with N0=5

Stochastic Growth With N0=5

Deterministic Growth with N0=1000

Stochastic Growth With N0=1000

Figure 2.1: Population growth trajectories under either a deterministic model with initial population sizes of $N_0 = 5$ (upper left) and $N_0 = 1000$ (lower left) or a stochastic model with initial population sizes of $N_0 = 5$ (upper right) and $N_0 = 1000$ (lower right).

In most biological experiments we are dealing with small sample sizes (and possibly small population sizes as well) and random effects play an important role in experimental outcomes so that the use of stochastic models is imperative to accurately describe experimental uncertainty. In this chapter, we formally introduce the notion of probability and describe how probability distributions can be used to model the outcomes of experiments taking account of stochastic effects.

Probability

Events and sample spaces

We begin our discussion of probability modeling by formally defining the experimental framework. An **event** is an outcome of an experiment. The **sample space** (sometimes called the "state space") is the set of all possible outcomes of an experiment. The "experiment" in this case does not have to be a controlled laboratory, or field, experiment and could simply be an observation of a naturally occuring (or unintentionally caused) process. Similarly, the experiment may already have occurred, as is the case with historical observations on fossil distributions, etc. However, in order for probabilities to be calculable, it is imperative that one be able to define all the possible outcomes of an experiment. An example of an experiment is as follows: a sample of 2 *Drosophila* flies are captured in an orchard and the flies are sorted according to sex. The three possible outcomes of this experiment are: (1) 1 male and 1 female, (2) 0 males and 2 females, (3) 2 males and 0 females.

The rules of set theory (see Appendix 1) apply to events on sample spaces. The sample space, Ω, is the complete set of possible events and any particular event, E, is a subset of the sample space $E \subset \Omega$. The union of two events $E_1 \cup E_2$ is the event that either E_1 or E_2 occurs or both E_1 and E_2 occur. The intersection of two events $E_1 \cap E_2$ is the event that both E_1 and E_2 occur. The empty set $\emptyset$ contains no events. Two events E_1 and E_2 are mutually exclusive if $E_1 \cap E_2 = \emptyset$. Consider again the experiment in which 2 *Drosophila* were captured. If E_1 is the event that at least 1 male is captured and E_2 is the event that at least 1 female is captured then $E_1 \cup E_2$ is the entire sample space, Ω, and $E_1 \cap E_2$ is the unique event "1 male and 1 female." Because experimental outcomes are typically "random" (meaning that no particular event is certain), we assign probabilities to events to describe the likelihood of different possible outcomes of an experiment.

Interpreting probabilities

The concept of probability is a familiar one. Most persons have an intuitive notion of probability and typically do not question the meaning of everyday statements such as "the probability of precipitation is 20 percent" or "the probability that politician A wins the election is greater than the probability politician B wins." However, the exact meaning ascribed to probabilities has been a subject of much debate among philosophers and statisticians. Many interpretations of probability have been suggested (see Barnett, 1999) but the two most widespread are the "frequentist" and "subjective" interpretations. The **frequentist** view of probability is based on the concept of the limiting frequency of outcomes in a very large number of repeated experiments. The probability of an event E is thus defined as

$$\Pr(E) = \lim_{n\to\infty} \sum_{i=1}^{n} \frac{I(E)}{n},$$

where $I(E)$ is an indicator function (see Appendix 1) that takes the value 1 if event E occurs and 0 otherwise. In simple terms, the long run frequency of occurrence of event E is $\Pr(E)$. The **subjective** view of probability is instead based on the concept of "degrees of belief." The probability $\Pr(E)$ measures the conviction of an individual that event E will occur in any particular experiment. The subjective interpretation of probabilities has often been viewed as the "Bayesian" interpretation. However, it is not specific to that inferential framework. Regardless of how one interprets probabilities, there is general agreement that certain mathematical rules (or axioms) must hold for numbers to be valid probabilities. We now consider the mathematical properties and rules for manipulating probabilities, the so-called probability calculus.

Rules of probability

The probability that event E occurs is defined as $\Pr(E)$. There are three basic rules (or "axioms") that probabilities must satistify. The first rule is that probabilities are positive numbers between 0 and 1. Thus, the probability of an event E, denoted by $\Pr(E)$, must satisfy

$$0 \leq \Pr(E) \leq 1.$$

The second rule is that some event must occur when an experiment is performed and therefore

$$\Pr(\Omega) = 1.$$

The third rule is that for any set of n mutually exclusive events $E_1, E_2, \ldots, E_n$, the probability that one of the events occurs is equal to the sum of the probabilities of the separate events,

$$\Pr\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \Pr(E_i).$$

This must hold for the limiting case of $n \to \infty$ as well. If we adopt a frequentist interpretation of probabilities these rules are guaranteed to hold. Thus, the frequencies of events must be between 0 and 1, the frequency with which some event occurs must be 1, and the frequency of a set of mutually exclusive events must equal the sum of the frequencies of the events. Subjective probabilities are normally restricted to be a set of "rational" beliefs that are assumed to satisfy these rules.

Probability calculations

All the rules for manipulating probabilities that will subsequently be described follow directly from the three rules stated above. We do not prove these relationships here but instead present a series of further rules for manipulating probabilities that are often useful for carrying out practical calculations. First, if we define E^c to be the complement of event E (the event that E does not occur) then

$$\Pr(E) + \Pr(E^c) = 1,$$

and

$$\Pr(E) = 1 - \Pr(E^c).$$

This rule is often useful in practical calculations when the probability of an event cannot be easily calculated but the probability of its complement can be. Second, if two events E_1 and E_2 are not mutually exclusive, then the probability of either E_1 or E_2 is

$$\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2).$$

Again consider the example in which two *Drosophila* flies are captured and sorted according to sex. If the population sex ratio is 50 : 50, then $\Pr(E_1) = \Pr(E_2) = 3/4$ and $\Pr(E_1 \cap E_2) = 1/2$. The complement of the event $E_1 \cap E_2$ is the event that either both flies are male or both are female which has probability $1/4 + 1/4 = 1/2$ (since the events all males or all females are mutually exclusive the probability of either event [the union] is the sum of the probabilities of each event). Thus, we can calculate the probability of the event $E_1 \cap E_2$ by applying the first rule and subtracting the probability of the complement $[E_1 \cap E_2]^c$from one,

$$\Pr(E_1 \cap E_2) = 1 - \Pr([E_1 \cap E_2]^c) = 1 - \frac{1}{2} = \frac{1}{2}.$$

We can also calculate the probability that either E_1 or E_2 occurs by applying the second rule,

$$\begin{aligned} \Pr(E_1 \cup E_2) &= \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2) \\ &= \frac{3}{4} + \frac{3}{4} - \frac{1}{2} = 1. \end{aligned}$$

This agrees with our earlier claim that the event of either at least one male or at least one female in the sample covers all possible experimental outcomes and therefore has probability one. In the above examples the probabilities calculated by applying these rules could also be calculated directly, but there are many situations where this is not the case.

Conditional probability

The conditional probability that event E_1 occurs, given that event E_2 has occurred, is

$$\Pr(E_1|E_2) = \frac{\Pr(E_1 \cap E_2)}{\Pr(E_2)}.$$

Considering our earlier example in which two *Drosophila* are sampled, the conditional probability that one or more males are sampled, given that one or more females are sampled, is

$$\Pr(E_1|E_2) = \frac{(1/2)}{(3/4)} = \frac{2}{3}.$$

Thus, the knowledge that at least one female is sampled decreases the probability that at least one male is sampled from 3/4 to 2/3.

Independent events

If the probability that event E_1 occurs is independent of whether event E_2 has occurred then

$$\Pr(E_1|E_2) = \Pr(E_1) = \frac{\Pr(E_1)\Pr(E_2)}{\Pr(E_2)},$$

which is equivalent to the condition that

$$\Pr(E_1 \cap E_2) = \Pr(E_1)\Pr(E_2).$$

In other words, if the events are independent then the probability both occur is simply the product of their marginal probabilities of occurrence.

Bayes theorem

Bayes theorem provides a mechanism for indirectly calculating conditional probabilities,

$$\Pr(E_1|E_2) = \frac{\Pr(E_2|E_1)\Pr(E_1)}{\Pr(E_2)}.$$

This allows an unknown conditional probability $\Pr(E_1|E_2)$ to be calculated if the conditional probability $\Pr(E_2|E_1)$ and the marginal probabilities $\Pr(E_1)$ and $\Pr(E_2)$ are known. Considering once more the *Drosophila* example, we earlier showed that the conditional probability that one or more males are sampled given that one or more females are sampled, $\Pr(E_1|E_2)$ is $2/3$, and the marginal probabilities of sampling one or more males (or females) are $\Pr(E_1) = \Pr(E_2) = 3/4$ and so the conditional probability of sampling one or more females, given that one or more males have been sampled, can be obtained by applying Bayes theorem,

$$\Pr(E_2|E_1) = \frac{\Pr(E_1|E_2)\Pr(E_2)}{\Pr(E_1)} = \frac{(2/3) \times (3/4)}{(3/4)} = \frac{2}{3}.$$

In this case, the conditional probability can also be calculated directly to verify that this result is correct. Many situations arise where a tricky conditional probability calculation can be greatly simplified by applying this formula. In Chapter 5 we will see that this theorem also forms the basis for an approach to parameter estimation known as Bayesian inference.

Random variables

A **random variable** is a quantity of interest whose potential values are a function of the events on the sample space (e.g., are determined by the outcome of an experiment). The probability that a random variable takes any particular value is then determined by the probabilities of events resulting in that value. For example, if we toss a coin three times the eight possible outcomes of the experiment are,

$$E_1 \quad = \quad \{H, H, H\},$$

$$\begin{aligned} E_2 &= \{H, H, T\}, \\ E_3 &= \{H, T, H\}, \\ E_4 &= \{T, H, H\}, \\ E_5 &= \{H, T, T\}, \\ E_6 &= \{T, H, T\}, \\ E_7 &= \{T, T, H\}, \\ E_8 &= \{T, T, T\}. \end{aligned}$$

However, we might be particularly interested in the number of heads obtained, denoted as x, not caring in which particular order the heads occur. In this case, there are only 4 possible values, $x \in \{0, 1, 2, 3\}$. Thus, several distinct events, such as $E_2 = \{H, H, T\}$, $E_3 = \{H, T, H\}$ and $E_4 = \{T, H, H\}$ produce the same value for the random variable, $x = 2$, and the probability that the variable assumes this value is then given by the sum of the probabilities of the events,

$$\Pr(x = 2) = \Pr(E_2) + \Pr(E_3) + \Pr(E_4).$$

For a fair coin this would be $3 \times (1/2)^3 = 3/8$. A **probability distribution** assigns probabilities to particular outcomes for a random variable. The way in which this is done differs for discrete versus continuous variables.

Modeling Discrete Data

Discrete variables usually result from an experiment involving counts of discrete traits. Because such experiments involve small samples and have uncertain outcomes we expect the discrete variables that we observe to vary from one experiment to the next. Often it is possible to model the observed outcomes for a variable of interest as a realization of a **discrete random variable** and thus to assign a probability to the outcome. Here, we consider mathematical models designed to emulate the inherent uncertainty of samples of discrete variables.

Probability mass functions

The **probability mass function**, abbreviated pmf, defines the probabilities that a discrete random variable takes particular values. For example, if the probability that a fly chosen at random is female is given by the proportion of females in the population, p, then the probability that x female flies are obtained in a random sample of n flies from a large population (or a sample taken with replacement) may be modeled using the pmf of the **binomial distribution** defined as,

$$\Pr(x|n,p) = \binom{n}{x} p^x (1-p)^{n-x}. \tag{2.1}$$

To clarify how this pmf arises, consider the outcome of a particular sampling experiment in which $n = 6$ flies are sampled with $x = 3$ being female. The vector of observations might be,

$$y^{(x=3)} = \{F, F, M, F, M, M\},$$

where the superscript $x = 3$ indicates that the observed sample contains 3 females in total, $y_1^{(x=3)} = F$ indicates that the first sampled fly was female, $y_3^{(x=3)} = M$ that the third sampled fly was male, etc. Assume that the samples are independent and that the population is sufficiently large (or sampling is with replacement) so that p is remains constant from one sample to the next. The probability of this particular sequence of sampling outcomes is

$$p \times p \times (1-p) \times p \times (1-p) \times (1-p) = p^3 \times (1-p)^3.$$

More generally, the probability of any particular sequence of x females and $n - x$ males is

$$p^x \times (1-p)^{n-x}.$$

Let z be number of possible distinct sequences of males and females in a sample of size n that contains x females in total. For example, with $x = 1$ and $n = 3$, we have the possible sequences $\{F, M, M\}$, $\{M, F, M\}$ and $\{M, M, F\}$ so that $z = 3$. Each such sequence has probability $p^x \times (1 - p)^{n-x}$ and the total probability of x (ignoring the order of females in the sample) is

$$\sum_{i=1}^{z} p^x \times (1-p)^{n-x} = z \times p^x \times (1-p)^{n-x} = \binom{n}{x} p^x (1-p)^{n-x},$$

where the term z is the number of distinct orderings of x females and $n - x$ males; this is just the number of ways to choose x individuals that are female from a total of n individuals (the binomial coefficient).

To illustrate how the binomial pmf can be used to calculate probabilities of particular values for a random variable, consider an experiment in which

$n = 10$ *Drosphila* are sampled and x is the number of females. Suppose that the sex ratio is $50 : 50$ so that $p = 0.5$. The probability of sampling 0 females is

$$\Pr(x = 0|p = 0.5, n = 10) = \binom{10}{0} 0.5^0 \times 0.5^{10} = 0.5^{10} \approx 0.001.$$

Probability mass functions are functions of probabilities of events on a state space and therefore must satisfy similar properties. To be valid, a pmf summed over all possible values of a random variable must equal one,

$$\sum_x \Pr(x) = 1,$$

and all probabilities must be positive numbers,

$$\Pr(x) \geq 0, \text{ for all } x.$$

Figure 2.2 plots the pmf of a binomial distribution with $n = 10$ and $p = 0.5$. The **cumulative distribution function**, abbreviated cdf, of a discrete

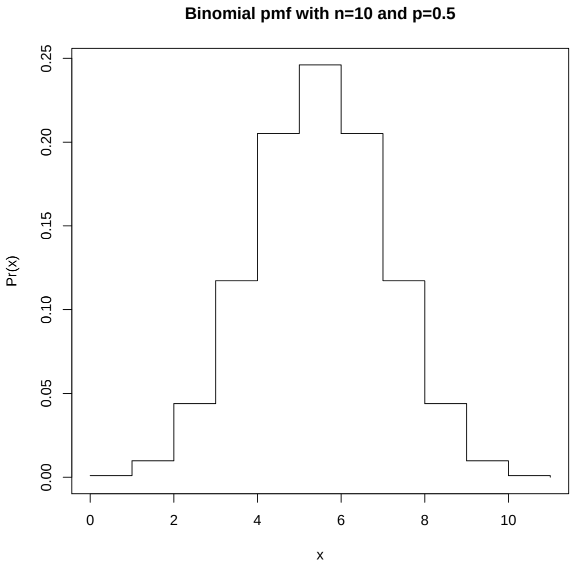


Figure 2.2: The probability mass function (pmf) of a binomial distribution with $n = 10$ and $p = 0.5$.

random variable is defined as

$$F(x) = \Pr(y \leq x) = \sum_{y=0}^{x} \Pr(y).$$

In words, the cdf of a random variable, by convention denoted using an upper case F, provides the probability that the variable is less than or equal to x. For the binomial distribution, for example, the cdf is,

$$F(x) = \sum_{y=0}^{x} \binom{n}{y} p^y (1-p)^{n-y} \text{ for all } 0 \leq x \leq n.$$

To illustrate the use of a cdf, consider again the *Drosophila* sampling experiment. Given $n = 10$ and $p = 0.5$ what is the probability that at least 5 female flies are sampled? This can be obtained from the cdf as follows,

$$\begin{aligned} \Pr(x > 4) &= 1 - \Pr(x \leq 4) \\ &= 1 - F_x(4) \\ &= 1 - (1 + 10 + 45 + 120 + 210)\left(\frac{1}{2}\right)^{10} \\ &= 1 - 386\left(\frac{1}{2}\right)^{10} \approx 0.623. \end{aligned}$$

Figure 2.3 plots the cdf for a binomial distribution with $n = 10$ and $p = 0.5$.

Expectation and variance

The **expectation** (also called the expected value) of a discrete random variable with pmf $\Pr(x)$ is defined as

$$\mathbb{E}(x) = \sum_x x\Pr(x), \tag{2.2}$$

where the sum is over the domain of $\Pr(x)$. The expectation is often referred to as the mean of random variable x and denoted μ_x. The mean of a pmf is analogous to the mean of a population if we consider the pmf to model the frequency distribution of a variable in a population; it predicts the location of the center of the distribution. The variance of x is defined as

$$\text{Var}(x) = \sum_x (x - \mu_x)^2 \Pr(x). \tag{2.3}$$

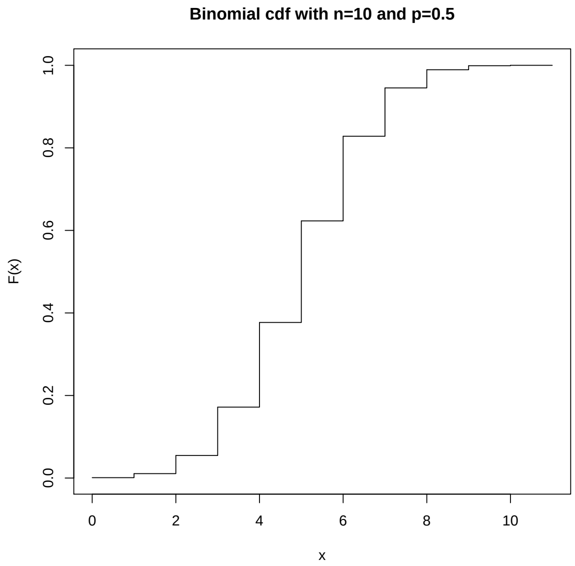


Figure 2.3: The cumulative distribution function (cdf) of a binomial distribution with $n = 10$ and $p = 0.5$.

The variance of a distribution is analogous to the population variance and measures the spread of the distribution about the mean. Considering the binomial pmf given in equation 2.1, the expectation is

$$\mathbb{E}(x) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} = np,$$

and the variance is

$$\text{Var}(x) = \sum_{x=0}^{n} (x - np)^2 \binom{n}{x} p^x (1-p)^{n-x} = np(1-p).$$

For example, a binomial distribution with $n = 10$ and $p = 0.5$ will have $\mu_x = 10 \times 0.5 = 5$ and $\text{Var}(x) = 10 \times (0.5)^2 = 10 \times 0.25 = 2.5$.

Common discrete distributions

Two additional discrete distributions important for biological modeling are worth mentioning at this point. The **Poisson distribution** has pmf

$$\Pr(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda} \text{ for all } x = 0, 1, 2, ... \tag{2.4}$$

and is a useful model of the distribution of rare events in a large sample. The Poisson distribution can be derived as the limit of a binomial distribution as $n \to \infty$, $p \to 0$ and $np \to \lambda$. The expectation and variance of the Poisson distribution are equal, with $\mu_x = \text{Var}(x) = \lambda$. Figure 2.4 shows a plot of a Poisson pmf with parameter $\lambda = 5$. Example 2.1 illustrates the use of the Poisson distribution as a model for the distribution of substitutions in DNA sequences.

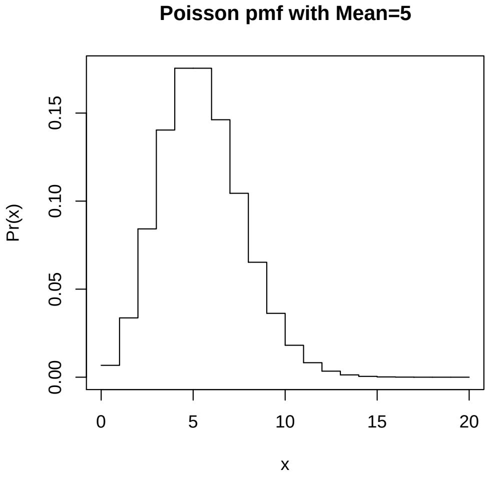


Figure 2.4: The probability mass function (pmf) of a Poisson distribution with $\lambda = 5$.

The **geometric distribution** has pmf

$$\Pr(x|p) = (1-p)^{x-1}p \text{ for all } x = 1, 2, ... \tag{2.5}$$

and is often used to model the number of trials needed until a particular event first occurs, where the probability of the event occurring at each trial

EXAMPLE 2.1 The so-called infinite sites model has been used to model the probability distribution of DNA substitutions between pairs of evolutionarily related sequences. This approach assumes that the mutation rate is very low and the number of sites in a sequence is very large so that the distribution of the number of DNA substitutions (segregating sites) between two sequences that diverged T_2 generations ago follows a Poisson distribution with parameter $2\mu s T_2$ where μ is the per site mutation rate and s is the number of sites. The factor 2 appears because there are two branches (each of length T_2) separating the sequences from their common ancestral sequence. The pmf of the number of segregating sites, M, under this model is

$$\Pr(M|T_2) = \frac{e^{-2\mu s T_2}(2\mu s T_2)^M}{M!}$$

This is a straightforward application of the Poisson approximation to the binomial distribution with $n = s \rightarrow \infty$, $p = 2\mu T_2 \rightarrow 0$ and $np = -2\mu s T_2$.

is p. The term $(1-p)^{x-1}$ is the probability the event does not occur in the first $x-1$ trials and p is the probability that the event occurs at trial x. The expectation of a geometric random variable is $\mu = 1/p$ and the variance is $\text{Var} = (1-p)/p^2$. Figure 2.5 shows a plot of the pmf for a geometric distribution with parameter $p = 0.1$.

Modeling Continuous Data

We have already considered continuous variables as a type of data resulting from a biological experiment in which we measure traits. Because of small sample sizes, etc, as mentioned earlier, we expect experiments sampling continuous variables to have variable outcomes. Often it is possible to model such outcomes as realizations of **continuous random variables** and the assign probabilities that variables reside within a particular interval of values. Here we consider mathematical models of continuous random variables. Recall that measurements of continuous traits are never exact and the true value lies within an interval determined by the significant digit. Analogously, we cannot assign a probability to a particular value of a continuous random variable but instead calculate the probability the

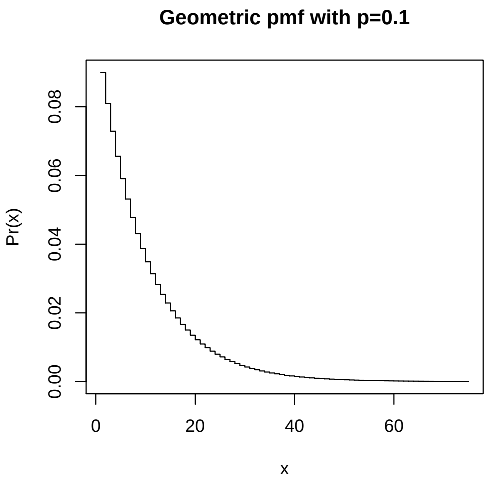


Figure 2.5: The probability mass function (pmf) of a Geometric distribution with $p = 0.1$.

variable lies within a given interval.

Probability density functions

The **probability density function**, abbreviated pdf, defines the probability density at any point on the real number line for a continuous random variable. The integral of the probability density over an interval defines the amount of probability concentrated in that region. The probability density function satisfies

$$F(x) = \int_{-\infty}^{x} f(t)dt, \tag{2.6}$$

where for random variables that are strictly positive, negative values are defined to have probability density 0. For simplicity, we will subsequently evaluate integrals only over the domain for which the density function is non-zero. The probability density can be equivalently defined as the derivative of the cumulative distribution function

$$f(x) = \frac{d}{dx}F(x). \tag{2.7}$$

The probability that a random variable takes a value in the interval $[a, b]$ is

$$\Pr(a \leq x \leq b) = \int_a^b f(x)dx = F(b) - F(a).$$

A pdf must satisfy the conditions that probability densities are strictly positive

$$f(x) \geq 0 \text{ for all } x,$$

and integrating over the real numbers yields a total probability of 1,

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

A simple example of a continuous distribution is the **uniform distribution** which is often used to model ignorance about the value of a variable (apart from its being located in a particular interval). The uniform distribution has a rectangular-shaped pdf that assigns constant probability density to all values within an interval $[a, b]$ and density 0 to values outside this interval,

$$f(x|a, b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise.} \end{cases}$$

It is straightforward to check that this is a valid pdf by checking the two conditions, stated earlier, that a pdf must satisfy. The pdf is clearly strictly positive and

$$\int_a^b \frac{1}{b-a} dx = \left. \frac{x}{b-a} \right|_a^b = \frac{b}{b-a} - \frac{a}{b-a} = 1.$$

The cdf of the uniform distribution is easily calculated,

$$F(x) = \int_a^x \frac{1}{b-a} dx = \frac{x-a}{b-a} = x\left(\frac{1}{b-a}\right) - \frac{a}{b-a}.$$

Thus, the cdf is a simple linear function of x with slope $1/(b-a)$. The pdf and cdf of a uniform distribution on the interval $[5, 10]$ are shown in Figure 2.6. To calculate the probability that a uniform $[5, 10]$ random variable lies in the interval $[5.1, 5.2]$ we take

$$\Pr(5.1 \leq x \leq 5.2) = F(5.2) - F(5.1) = \frac{0.2 - 0.1}{5} = \frac{1}{50}.$$

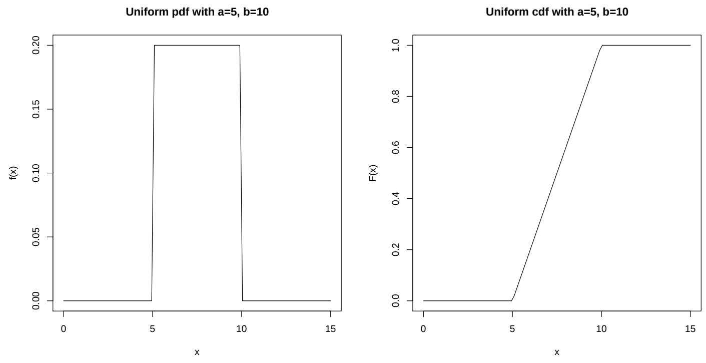


Figure 2.6: The probability density function (pdf) (at left) and cumulative distribution function (cdf) (at right) of a Uniform distribution with $a = 5$ and $b = 10$.

Expectation and variance

The expectation and variance of a continuous random variable are defined in a manner similar to that of discrete random variables, except that one replaces sums with integrals in the expressions. The expectation (expected value) of a continuous random variable x with pdf $f(x)$ is

$$\mathbb{E}(x) = \int_{-\infty}^{\infty} x\, f(x)\, dx = \mu_x, \tag{2.8}$$

and the variance is

$$\text{Var}(x) = \int_{-\infty}^{\infty} (x - \mu_x)^2\, f(x)\, dx.$$

To illustrate, we again consider the uniform distribution. In that case, the expectation is

$$\mathbb{E}(x) = \int_a^b x \left(\frac{1}{b-a}\right) dx = \left.\frac{x^2}{2(b-a)}\right|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}.$$

Thus, the expected value is the midpoint of the interval. For example, if $a = 2$ and $b = 3$, the expectation is $\mathbb{E}(x) = (3+2)/2 = 2.5$. The variance of a uniform distribution is

$$\text{Var}(x) = \int_a^b \left(x - \frac{b+a}{2}\right)^2 \frac{1}{b-a} dx = \frac{(b-a)^2}{12}.$$

A simple derivation of this result will be provided later using expectations of the moments of the distribution.

Common continuous distributions

Two additional continuous distributions with broad utility for biological modeling are the exponential distribution and the normal distribution. We first consider the **exponential distribution**. The pdf of an exponential random variable, x, is

$$f(x|\beta) = \frac{e^{-x/\beta}}{\beta} \text{ for all } x \in (0, \infty).$$

The exponential distribution has expectation $\mu = \beta$ and variance $\mathrm{Var} = \beta^2$. Example 2.2 below illustrates how the exponential distribution arises in the context of population genetics to model the distribution of the time until a pair of genes chosen at random from a population first share a common ancestor.

Another continuous distribution with widespread applications in biology is the **normal distribution**. The normal distribution has been used to model error distributions of measurements, the distribution of quantitative traits (traits influenced by many genes), the sampling distribution of the population mean, etc. The pdf of a normal random variable, x, is

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

The pdf of a normal random variable with $\mu = 10$ and $\sigma = 3$ is plotted in Figure 2.7. The normal distribution is symmetrical and unimodal. The normal distribution with $\mu = 0$ and $\sigma = 1$ is a **standard normal distribution**. Any random variable, x, that has a normal distribution with mean μ and variance σ^2 can be transformed to be a standard normal random variable z using the transformation

$$z = \frac{x - \mu}{\sigma}.$$

In a bygone era, normal variables were usually transformed to standard normal variables to allow cumulative probabilities and so on to be looked up using a single table (for the standard normal distribution). Modern computers and software packages have made this uneccessary.

EXAMPLE 2.2 The coalescent theory models the ancestral genealogical relationships of a sample of chromosomes under a model of random mating. Let $2N$ be the total number of chromosomes in a diploid population. Two chromosomes are sampled from the population. The probability that the chromosomes coalesce (have a shared parental origin) at the previous generation $(T = 1)$ is $1/(2N)$. The probability that they have different parental chromosomes (do not coalesce) is $1 - 1/(2N)$. The process of sampling parents occurs independently in each generation and so the probability that no coalescence occurs by generation $T - 1$ and then a coalescence occurs at generation T is

$$\Pr(T) = \frac{1}{2N}\left(1 - \frac{1}{2N}\right)^{(T-1)},$$

which is a geometric distribution with parameter $1/(2N)$ so the expected waiting time until coalescence is $2N$ generations. One can approximate the coalescent as a continuous time process by transforming the time scale to units of $2N$ generations, namely $T' = T/(2N)$. The probability of no coalescence by time T' is then

$$\left(1 - \frac{1}{2N}\right)^{(2NT')}$$

One generation on this new timescale is $1/2N$ time units and if we take the limit as $N \to \infty$ time appears continuous and the duration of 1 generation tends to zero. The limiting probability of no coalescence by time T on this new timescale is then

$$\lim_{N\to\infty}\left(1 - \frac{1}{2N}\right)^{(2NT')} = e^{-T'},$$

which is an exponential random variable with parameter 1. Thus, we wait 1 unit of time ($2N$ generations) on average for a coalescence event to occur. Similarly, it can be shown that the variance of the waiting time to coalescence is $4N^2$ generations.

Expectations of Functions

Often we are interested in the expected (or average) value of a function of a random variable, $g(x)$, rather than simply the expectation of the variable x

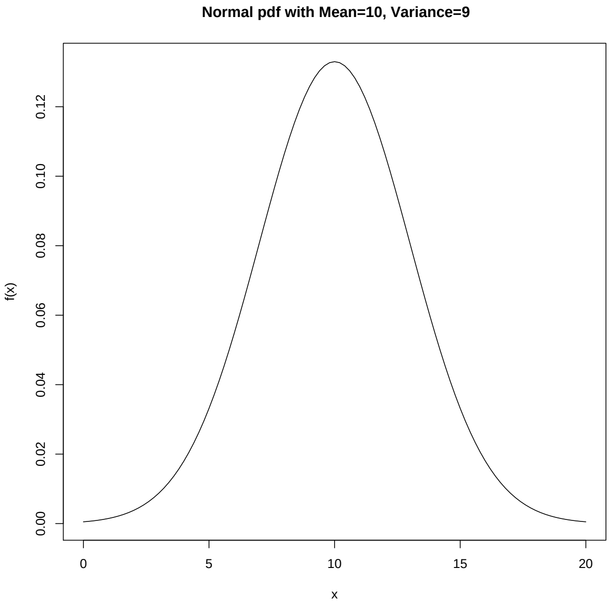


Figure 2.7: The probability density function (pdf) of a Normal distribution with $\mu = 10$ and $\sigma = 3$.

itself (the variance, calculated earlier, is an example). Here, we describe the procedure for calculating the expectation of a function of a random variable and derive solutions for several straightforward yet important functions. We then go on to consider a particular class of polynomial functions of random variables called "moments" that play an important role in distribution theory.

The expectation of a function $g(x)$ of the discrete random variable x, with pmf $\Pr(x)$, is defined as

$$\mathbb{E}[g(x)] = \sum_x g(x) \Pr(x). \tag{2.9}$$

Similarly, the expectation of $g(x)$ for a continuous random variable x with pdf $f(x)$ is defined as

$$\mathbb{E}[g(x)] = \int_{-\infty}^{\infty} g(x) \, f(x) \, dx. \tag{2.10}$$

For example, if we define $g(x) = (x - \mu)^2$ then $\mathbb{E}[g(x)] = \text{Var}(x)$. For simplicity, the results will be derived here for a discrete random variable, but the results apply for continuous random variables as well. Perhaps the simplest function of a variable is the constant function $f(x) = c$. The expectation of a constant function is

$$\mathbb{E}(c) = \sum_x c\Pr(x) = c \sum_x \Pr(x) = c,$$

where we have made use of the fact that the sum over the pmf equals 1. A slightly more complicated function is the linear function $f(x) = a\,x + c$. The expectation of a linear function is

$$\mathbb{E}(a\,x + c) = \sum_x (a\,x + c)\Pr(x) = a \sum_x x\Pr(x) + c \sum_x \Pr(x) = a\mathbb{E}(x) + c.$$

Finally, we consider a sum (or difference) of independent variables (later we will show that the same result holds for sums and difference of variables that are not independent). Consider the bivariate function $g(x, y) = x + y$. The expectation, taken with respect to both variables x and y, is

$$\begin{aligned}\mathbb{E}(x+y) &= \sum_x \sum_y (x+y)\Pr(x)\Pr(y) \\ &= \sum_x \sum_y x\Pr(x)\Pr(y) + \sum_x \sum_y y\Pr(x)\Pr(y) \\ &= \sum_x x\Pr(x) \sum_y \Pr(y) + \sum_y y\Pr(y) \sum_x \Pr(x) \\ &= \mathbb{E}(x) + \mathbb{E}(y).\end{aligned}$$

Thus, "the expectation of a sum equals the sum of expectations." An analogous result can be established for differences by changing the sign in the above equations. This result generalizes to sums and differences of any number of variables. Note that for non-linear functions,

$$\mathbb{E}[g(x)] = \sum_x g(x)\Pr(x) \neq g[\sum_x x\Pr(x)].$$

In words, for a non-linear function"the expectation of the function does not equal the function of the expectation." This is a common source of mistakes. For example, if one is interested in the expectation (or mean) of $\log(x)$ this should be calculated as $\mathbb{E}[\log(x)]$ and not $\log(\mathbb{E}[x])$.

We now derive a formula that, in many cases, simplifies calculation of the variance of a distribution,

$$\mathbb{E}[(x-\mu)^2] \quad = \quad \mathbb{E}(x^2 - 2\mu x + \mu^2)$$

$$\begin{aligned}
&= \mathbb{E}(x^2) - \mathbb{E}(2\mu x) + \mathbb{E}(\mu^2) \\
&= \mathbb{E}(x^2) - 2\mu\mathbb{E}(x) + \mu^2 \\
&= \mathbb{E}(x^2) - 2\mu^2 + \mu^2 \\
&= \mathbb{E}(x^2) - \mu^2 \\
&= \mathbb{E}(x^2) - \mathbb{E}(x)^2.
\end{aligned}$$

In words, "the variance equals the expectation of the square minus the square of the expectation." To illustrate the utility of this result we now derive the variance for a uniform random variable. We showed previously that $\mu = (b+a)/2$, and

$$\mathbb{E}(x^2) = \int_a^b x^2 \frac{1}{b-a}\, dx = \frac{x^3}{3}\frac{1}{b-a}\bigg|_a^b = \frac{(b+a)^2 - ab}{3},$$

so that

$$\begin{aligned}
\mathbb{E}(x^2) - \mathbb{E}(x)^2 &= \frac{(b+a)^2 - ab}{3} - \left(\frac{b+a}{2}\right)^2 \\
&= \frac{(b+a)^2 - 4ab}{12} \\
&= \frac{(b-a)^2}{12}.
\end{aligned}$$

Moments of distributions

Polynomial functions of random variables have historically been used to summarize the shape of a distribution, to compare different distributions, and to estimate parameters of distributions (see Chapter 3). The first moment, or mean, is already familiar,

$$\mathbb{E}(x) = \sum_x x \,\mathrm{Pr}(x),$$

as is the second moment which we used earlier to derive a formula for calculating the variance,

$$\mathbb{E}(x^2) = \sum_x x^2 \,\mathrm{Pr}(x).$$

In fact, older books may refer to the variance as the second moment about the mean. The kth moment is defined as

$$\mathbb{E}(x^k) = \sum_x x^k \,\mathrm{Pr}(x).$$

We will revisit these expressions in Chapter 3 when we consider the "methods of moments" approach to estimating parameters.

Modeling Multivariate Data

We have previously considered situations where multiple counts, or measurements, are collected for each individual in a sample. For example, weights and heights of individuals. For finite samples, the outcomes of such sampling will vary and we will focus on methods for modeling the combined measurements (or counts) of two or more traits for each individual as a **multivariate random variable**. We can again use a probability distribution as a model of the sampling experiment. Here, we focus on multivariate probability distributions, illustrating their use in modeling multivariate random variables.

Joint probability distributions

Let x and y be random variables defined on the same sample space Ω. If the variables are discrete, their joint pmf is defined as $\Pr(x, y)$. If they are continuous, their joint pdf is defined as $f(x, y)$. These bivariate distributions are special cases of more general multivariate distributions, which may involve any number of variables. A simple example of a continuous bivariate pmf is the bivariate standard normal density,

$$f(x,y|\rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right\}, \qquad (2.11)$$

where ρ is the pearson correlation coefficient, which we will consider in detail below. The parameter ρ describes the degree of association of the two variables. If x and y are independent, for example, then $\rho = 0$ and $f(x, y|\rho = 0) = f(x) \times f(y)$ where x and y are both standard normal random variables,

$$f(x)f(y) = f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \times \frac{1}{\sqrt{2\pi}} e^{-y^2/2} = \frac{1}{2\pi} e^{-(x^2+y^2)/2},$$

which agrees with the result obtained by setting $\rho = 0$ in equation 2.11 above.

More generally, the joint pmf of k discrete random variables is defined as $\Pr(x_1, x_2, \ldots, x_k)$ and the joint pdf of k continuous random variables is defined as $f(x_1, x_2, \ldots, x_k)$. As is the case with univariate distributions, valid multivariate pmfs and pdfs must satisfy certain conditions. For discrete random variables, if $\Pr(x_1, x_2, \ldots, x_k)$ is a joint pmf then

$$\Pr(x_1, x_2, \ldots, x_k) \geq 0 \text{ for all } x_i \in \Omega,$$

and

$$\sum_{x_1} \cdots \sum_{x_k} \Pr(x_1, \ldots, x_k) = 1.$$

For continuous random variables, if $f(x_1, x_2, \ldots, x_k)$ is a joint pdf then,

$$f(x_1, x_2, \ldots, x_k) \geq 0 \text{ for all } x_i \in \Omega,$$

and

$$\int_{x_1} \cdots \int_{x_k} \Pr(x_1, \ldots, x_k) dx_k \cdots dx_1 = 1.$$

We now consider a particularly important multivariate discrete distribution known as the **multinomial distribution**. This distribution is a generalization of the binomial distribution considered earlier. Suppose that k distinct types of a discrete random variable exist in a population and a random sample of n individuals is collected (with replacement). The probability of sampling type i is p_i, with $\mathbf{p} = \{p_i\}$ and the number of counts of type i in the sample is x_i. The joint pmf of a sample is then,

$$\Pr(x_1, x_2, \ldots, x_k | n, \mathbf{p}) = \binom{n}{x_1, x_2, \ldots, x_k} \prod_{i=1}^{k} p_i^{x_i}, \tag{2.12}$$

where

$$\binom{n}{x_1, x_2, \ldots, x_k} = \frac{n!}{x_i! x_2! \cdots x_k!},$$

is the **multinomial coefficient** (see Appendix 1) which is a generalization of the binomial coefficient and specifies the number of disinct ordered samples that would produce the same sample counts for the k variables.

To illustrate, we consider the multinomial distribution as a model of the process of sampling alleles in a population genetic study. If there are $k = 3$ alleles present at a particular locus in a haploid population with allele frequencies $p_1 = 0.1$, $p_2 = 0.6$ and $p_3 = 1 - p_1 - p_2 = 0.3$ and we sample $n = 3$ chromosomes, the probability of the sample $x_1 = 0$, $x_2 = 2$ and $x_3 = 1$, where x_1 is the number of copies of allele 1, etc, is

$$\Pr(x_1 = 0, x_2 = 2, x_3 = 1) = \frac{3!}{0!2!1!} 0.1^0 (0.6)^2 0.3^1 = 3(0.6)^2 0.3 = 0.324.$$

Thus, it appears quite probable that we observe the sample $x_1 = 0$, $x_2 = 2$ and $x_3 = 1$, given the allele frequencies in the population.

Marginal distributions

Conceptually, the **marginal distribution** of x_i is the univariate distribution of the variable that is obtained by focusing exclusively on x_i and ignoring the outcomes for the other variables of the joint density. In the experiment described above in which we are examining counts of alleles in a population genetic sample, for example, if we were to focus our attention only on the number of copies of allele 1 in the sample, x_1, lumping the remaining alleles into a single category (not allele 1), then we would be considering the marginal distribution of x_1. For a discrete random variable, the marginal pmf of x_i is defined as

$$\Pr(x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_k} \Pr(x_1, x_2, \ldots, x_k),$$

where we are summing over the domain of every variable except x_i. Similarly, for a continuous random variable, x_i, the marginal pdf is

$$f(x_i) = \int_{x_1} \cdots \int_{x_{i-1}} \int_{x_{i+1}} \cdots \int_{k} f(x_1, x_2, \ldots, x_k)\, dx_k \cdots dx_{i+1} dx_{i-1} \cdots dx_1.$$

To illustrate, the marginal pmf for variable x_1 from a multinomial distribution with $k = 3$ is

$$\begin{aligned} \Pr(x_1|n, \mathbf{p}) &= \sum_{x_2=0}^{n-x_1} \binom{n}{x_1, x_2, n - x_1 - x_2} p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{n-x_1-x_2} \\ &= \binom{n}{x_1} p_1^{x_1} (1 - p_1)^{n-x_1}, \end{aligned}$$

which is the familiar binomial distribution. In general, the marginal distribution of the ith variable, x_i, from a multinomial distribution of dimension k is a binomial distribution with parameters n and p_i. Considering the bivariate standard normal density $f(x, y|\rho)$, the marginal density $f(x)$ of the variable x is

$$\begin{aligned} f(x) &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right\} dy \\ &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}, \end{aligned}$$

which is a univariate standard normal random variable.

We are now prepared to revisit the problem of calculating the expectation of a sum of variables $\mathbb{E}(x+y)$ but now allow the variables to be dependent. The expectation is

$$\begin{aligned}
\mathbb{E}(x+y) &= \sum_x \sum_y (x+y)\Pr(x,y) \\
&= \sum_x x \sum_y \Pr(x,y) + \sum_y y \sum_x \Pr(x,y) \\
&= \sum_x x\Pr(x) + \sum_y y\Pr(y) \\
&= \mathbb{E}(x) + \mathbb{E}(y). \qquad (2.13)
\end{aligned}$$

Thus, our earlier result that the expectation of a sum is the sum of the expectations holds for dependent variables as well. Example 2.3 below derives the marginal distribution of the number of mutations between two sampled sequences by combining the probability density of the coalescent time and the infinite sites model of mutation presented in earlier examples 2.1 and 2.2.

Marginal expectation and variance

If $\mathbf{x} = \{x_1, x_2, \dots, x_k\}$ is a multivariate discrete random variable with x_i to be the counts for variable i, the marginal expectation (mean) of x_i is defined as

$$\begin{aligned}
\mathbb{E}(x_i) &= \sum_{x_1} \cdots \sum_{x_k} x_i \Pr(x_1, x_2, \dots, x_k) \\
&= \sum_{x_i} x_i \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_k} \Pr(x_1, \cdots, x_k) \\
&= \sum_{x_i} x_i \Pr(x_i).
\end{aligned}$$

The marginal variance of x_i is

$$\mathrm{Var}(x_i) = \sum_{x_1} \cdots \sum_{x_k} (\mu_{x_i} - x_i)^2 \Pr(x_1, \dots, x_k) = \sum_{x_i} (\mu_{x_i} - x_i)^2 \Pr(x_i).$$

To illustrate, the marginal expectation of x_i for a multinomial distribution is

$$\mathbb{E}(x_i) = \sum_{x_1} \cdots \sum_{x_k} x_i \binom{n}{x_1, \dots, x_k} \prod_{i=1}^{k} p_i^{x_i} = np_i,$$

and the marginal variance is $\mathrm{Var}(x_i) = np_i(1-p_i)$.

Example 2.3 In example 2.2 we derived the probability density of the waiting time for a sample of two alleles to coalesce to a common ancestor,

$$f(T_2) = e^{-T_2},$$

where we assume here that the species is diploid and time is thus scaled in units of $2N$ generations. We also saw in example 2.1 that the number of segregating sites, M, for a pair of sequences of length s that diverged $T_2 \times 2N$ generations ago (under an infinite sites model) is

$$\Pr(M|T_2) = \frac{e^{-2\mu s T_2}(2\mu s T_2)^M}{M!},$$

where we have conditioned on the coalescence time T_2. The marginal probability that there are M segregating sites for a sample of 2 sequences is then

$$\begin{aligned} \Pr(M) &= \int_0^\infty \Pr(M|T_2) f(T_2) dT_2 \\ &= \frac{1}{M!}\int_0^\infty e^{-T_2(1+2\mu s)}(2\mu s T_2)^M dT_2 \\ &= \left(\frac{2\mu s}{1+2\mu s}\right)^M \frac{1}{1+2\mu s} \\ &= \left(\frac{\theta}{1+\theta}\right)^M \frac{1}{1+\theta}, \end{aligned} \tag{2.14}$$

where we have substituted $\theta = 2\mu s$. Note that if we transform back to a timescale of generations $\mu = 2N\mu'$ and $\theta = 4N\nu$ where $\nu = \mu' s$ is the per gene mutation rate on a generational timescale; that is the typical definition of θ found in the population genetics literature. Equation 2.14 is the probability mass function of a geometric distribution with parameter $p = 1/(1+\theta)$. The expectation of the geometric distribution is $(1-p)/p$ and therefore $\mathbb{E}(M) = \theta$.

Covariance and correlation

The **covariance** of a pair of random variables measures the degree of association between them. The covariance of the random variables x and y is defined as

$$\begin{aligned}\mathrm{Cov}(x, y) &= \mathbb{E}[(x-\mu_x)(y-\mu_y)] \\ &= \mathbb{E}(xy) - \mathbb{E}(x)\mathbb{E}(y),\end{aligned}$$

where $\mathbb{E}(x) = \mu_x$ and $\mathbb{E}(y) = \mu_y$ are the marginal expectations of x and y, respectively. If x and y are discrete random variables,

$$\mathbb{E}(xy) = \sum_x \sum_y xy\mathrm{Pr}(x, y).$$

If x and y are continuous random variables,

$$\mathbb{E}(xy) = \int_x \int_y xyf(x, y)dydx.$$

The covariance is positive if the variables tend to vary in the same direction. For example, if when x is large, y tends also to be large and when x is small y tends also to be small. The covariance is negative if they tend to vary in opposite directions. For example, when x is large, y tends to be small, and so on. If x and y are independent, then $\mathrm{Cov}(x, y) = 0$ because $\mathbb{E}(xy) = \mathbb{E}(x)\mathbb{E}(y)$. This is demonstrated for the case of discrete variables but is true in general,

$$\begin{aligned}\mathbb{E}(xy) &= \sum_x \sum_y xy\mathrm{Pr}(x, y) \\ &= \sum_x \sum_y xy\mathrm{Pr}(x)\mathrm{Pr}(y) \text{ (assuming independence)} \\ &= \sum_x x\mathrm{Pr}(x) \sum_y y\mathrm{Pr}(y) \\ &= \mathbb{E}(x)\mathbb{E}(y).\end{aligned}$$

The scale of the covariance depends on the units of measurement and can be difficult to interpret. A more useful measure of association is the correlation coefficient defined as

$$\rho_{xy} = \frac{\mathrm{Cov}(x, y)}{\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}}, \tag{2.15}$$

where $\mathrm{Var}(x)$ and $\mathrm{Var}(y)$ are the marginal variances of x and y, respectively. The units cancel in the above equation and ρ is a dimensionless quantity that varies on the interval -1 to $+1$.

To illustrate, we derive the correlation coefficient for a pair of variables x_i and x_j from a multinomial distribution. Earlier, we noted that $\mathbb{E}(x_i) = np_i$ and $\text{Var}(x_i) = np_i(1 - p_i)$. An additional result we will need is,

$$\mathbb{E}(x_i x_j) = \sum_{x_1} \cdots \sum_{x_k} x_i x_j \binom{n}{x_1, \ldots, x_k} \prod_{h=1}^{k} p_h^{x_h} = n(n-1)p_i p_j.$$

The covariance of x_i and x_j is then

$$\text{Cov}(x_i, x_j) = \mathbb{E}(x_i x_j) - \mathbb{E}(x_i)\mathbb{E}(x_j) = n(n-1)p_i p_j - n^2 p_i p_j = -np_i p_j,$$

and the correlation coefficient is

$$\rho_{x_i x_j} = \frac{\text{Cov}(x_i, x_j)}{\sqrt{\text{Var}(x_i)\text{Var}(x_j)}} = \frac{-np_i p_j}{\sqrt{np_i(1-p_i)np_j(1-p_j)}} = \frac{-\sqrt{p_i p_j}}{\sqrt{(1-p_i)(1-p_j)}}.$$

An interesting feature of the above result is that the correlation between variables does not depend on the sample size n or the number of types k. To illustrate, we return to our earlier population genetic example in which chromosomes are sampled from a haploid population with three alleles present in frequencies $p_1 = 0.1$, $p_2 = 0.6$ and $p_3 = 0.3$. The expected correlation between the number of copies of allele 1, x_1, and allele 2, x_2, is $-\sqrt{0.1 \times 0.6}/\sqrt{0.9 \times 0.4} = -.408$. Thus, for fixed n if x_1 is increased, x_2 tends to decrease and vice versa.

By rearranging the equation for the correlation coefficient we can rewrite the covariance as $\text{Cov}(x, y) = \rho_{xy}\sigma_x\sigma_y$. The covariance structure for a multivariate distribution is often represented in the form of a **variance-covariance matrix**,

$$\mathbf{V} = \left\{ \begin{array}{cccc} \sigma_{x_1}^2 & \rho_{x_1 x_2}\sigma_{x_1}\sigma_{x_2} & \cdots & \rho_{x_1 x_k}\sigma_{x_1}\sigma_{x_k} \\ \rho_{x_2 x_1}\sigma_{x_2}\sigma_{x_1} & \sigma_{x_2}^2 & \cdots & \rho_{x_2 x_k}\sigma_{x_2}\sigma_{x_k} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{x_k x_1}\sigma_{x_k}\sigma_{x_1} & \rho_{x_k x_2}\sigma_{x_k}\sigma_{x_2} & \cdots & \sigma_{x_k}^2 \end{array} \right\},$$

where the rows and the columns are both indexed by $x_1, \ldots, x_k$. This matrix is symmetrical and so we only need to display the values for off-diagonals either above, or below, the diagonal elements. For a k-dimensional distribution there will be $(k-1)!$ covariance terms and k variance terms.

Simple Model Fitting

The choice of a probability distribution to model a biological process is often arbitrary, and several plausible models may be available. A criterion is therefore needed to decide when a model provides an adequate description of the data. Here we introduce a general set of techniques known as **goodness of fit tests** designed for this purpose. The fundamental test we will focus on is the χ^2 test which measures the fit of a sample to a multinomial distribution with k categories. Although hypothesis tests are first introduced here, we defer a complete treatment of the subject until Chapter 5.

χ^2 test of goodness of fit

The basic approach of the χ^2 test is to choose a set of intervals to bin the observations (as when constructing a frequency distribution) and then measure the discrepancy between the observed counts in each interval and the marginal expected counts, using as a test statistic their squared difference. Earlier, we noted that the marginal expectation of the counts for the ith category from a multinomial distribution with k categories is $\mathbb{E}(x_i) = np_i$. The χ^2 test statistic is then defined as

$$T = \sum_{i=1}^{k} \frac{(x_i - np_i)^2}{np_i}. \tag{2.16}$$

The limiting distribution of this test statistic (as the sample size becomes large) is a continuous distribution known as the χ^2 distribution with $k-1$ **degrees of freedom**, abbreviated df. The degrees of freedom of the test are determined by the number of independent observations. For example, if $k = 2$ there are two categories of observations, but the counts in the two categories must sum to n (which is fixed in the experiment) so there is only one independent observation and therefore 1 df. A χ^2 distribution with 1 df is shown in Figure 2.8. The null hypothesis specifies that the probabilities that observations occur in the intervals $1, 2, \ldots, k$ are $p_1, p_2, \ldots, p_k$. The alternative hypothesis allows one, or more, of the probabilities to be different from these values. To determine whether the observed discordance between observed and expected counts is significant we consider the probability, α, that a value at least as large as T would be observed under the null hypothesis. Assuming that T is distributed according to a χ^2 distribu-

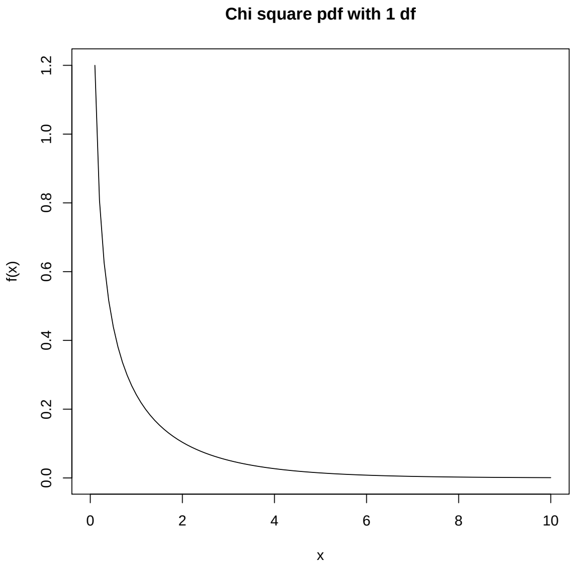


Figure 2.8: Probability density function (pdf) of a χ^2 distribution with 1 df.

tion with $k-1$ df the significance of the test is

$$\alpha = 1 - F_{\chi^2(k-1)}(T)$$

where $F_{\chi^2(k-1)}$ denotes the cdf of a χ^2 distribution with $k-1$ df. The level of significance for which the null hypothesis is rejected is arbitrary, with $\alpha = .05$ often viewed as a significant result and $\alpha = .001$ as a highly significant one.

To illustrate, consider a sample of n flies from a population of *Drosophila* that are sorted according to sex. Under the null hypothesis the sex ratio is even, so the probability of a female, or male, on each sample is $p = 0.50$. Let x be the number of females observed in the sample. The test statistic is

$$T = \frac{(x - n/2)^2}{n/2},$$

and the distribution of T is approximately χ^2 with 1 degree of freedom. If a sample of $n = 100$ flies were collected, for example, and $x = 60$ females were observed ($T = 4$,$\alpha = .046$), the null hypothesis (that the sex ratio is equal) would be rejected at the .05 significance level. If $x = 59$ females were instead observed ($T = 3.24, \alpha = .072$) then we would fail to reject the null hypothesis.

Our final example will use the χ^2 test to examine how well a sample of data conform to a particular discrete distribution, the Poisson distribution. Let x be a Poisson random variable. We choose 4 categories for our observations: $x = 0$, $x = 1$, $x = 2$, and $x \geq 3$. This choice is arbitrary; because the Poisson distribution includes observations on the integer numbers from 0 to ∞ we obviously cannot include the entire collection of possible observations. Moreover, if we consider too many categories we will have few observations in each and this can lead to a violation of the assumption that the distribution of the test statistic is χ^2. However, too few categories can also be problematic causing the test to be less discriminating and to accept the null hypothesis more often. A useful rule of thumb for choosing suitable numbers of categories, k, and sample size, n, is that $np_i > 5$ for all $i = 1, \ldots, k$.

The probability that observations fall into each of the 4 categories in the above model-fitting example are given by the Poisson probabilities

$$\begin{aligned}
p_1 &= \Pr(x = 0) = e^{-\lambda} \\
p_2 &= \Pr(x = 1) = \lambda e^{-\lambda} \\
p_3 &= \Pr(x = 2) = \lambda^2 e^{-\lambda}/2 \\
p_4 &= \Pr(x \geq 3) = 1 - e^{-\lambda} - \lambda e^{-\lambda} - \lambda^2 e^{-\lambda}/2,
\end{aligned}$$

and the test statistic is

$$T = \sum_{i=1}^{4} \frac{(x_i - np_i)^2}{np_i},$$

where x_i is the number of observations in category i. If λ is specified this test has $k - 1 = 4 - 1 = 3$ df. However, in most cases λ will not be known and must be estimated from the data (in this case we might use the sample mean as an estimate of λ) reducing the degrees of freedom by 1 so that the test statistic is approximately distributed as a χ^2 with 2 df. In general, if there are k categories and m free parameters to be estimated from the data the df $= k - m - 1$.

It is no longer strictly necessary to assume a χ^2 distribution to determine the significance of T. With modern computers we can easily simulate the distribution of the test statistic under the null hypothesis. In the above case in which we are fitting observations to a Poisson distribution, for example, we would simulate a large number of samples, each of size n, under a Poisson distribution with parameter λ and calculate the test statistic T for each simulated data set. The frequency of simulated T values that equal, or exceed, the observed value T calculated for the original data then provides an estimate of the significance level α of the observed T. Simulation methods for determining the significance of test statistics will be discussed in greater detail in Chapter 9.

Chapter 3

Point Estimates and Confidence

A central problem in statistics is how to infer the plausible values of one, or more, unknown parameters, θ, given a probability model $f(\mathbf{x}|\theta)$ and a sample of data, $\mathbf{x} = x_1, x_2, \ldots, x_n$. The procedures for **parametric inference** we consider here assume that a particular family of probability distributions can be specified that adequately model the physical problem and our main focus will be procedures to infer the unknown parameters. It is conventional to refer to a statistic for predicting the value of a parameter as an **estimator** and a particular value of the estimator (e.g., for a given set of data **x**) as an **estimate**.

It is typically assumed that the sample data represent independent and identically distributed (iid) random variables drawn from a density (or probability mass) function $f(x|\theta)$, so that the joint probability density (mass) of the sampled data is

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

Many of the classical results that will be described in this chapter for deriving sampling distributions, and so on, rely on the assumption that sampled variables are iid, as do results concerning the large sample (asymptotic) properties of maximum likelihood estimators, etc. The assumption of iid random variables is equivalent to assuming random sampling with replacement from a population with frequency distribution $f(x_i|\theta)$. If the population size is large relative to the sample size the difference between sampling with (or without) replacement is negligible and the iid assumption is justified.

There are two broad approaches to the problem of parametric inference: point estimation and interval estimation. A **point estimate** is a prediction of the true value of θ and an **interval estimate** is a prediction of the plausible range of values of θ. A point estimate may appear more precise, being a single number, but the probability that the point estimate is precisely equal to the true parameter value is zero (because θ is a real number), whereas the interval estimate has non-zero probability of including the true parameter value. Several different criteria for optimality of estimators are commonly used, as well as several methods for obtaining optimal estimators. In this book, we will focus on three widely used approaches known as classical (or frequentist), likelihood and Bayesian inference.

Method of moments

One of the oldest methods for obtaining estimators of parameters, developed by Karl Pearson in the early 1900s, is the method of moments. If there are k unknown parameters then a set of k equations are constructed by setting the first k moments of the probability distribution to equal the first k sample moments. Thus,

$$\begin{aligned} m_1 &= \mathbb{E}(x^1|\theta_1,\ldots,\theta_k), \\ &\vdots \\ m_k &= \mathbb{E}(x^k|\theta_1,\ldots,\theta_k). \end{aligned}$$

These equations are simultaneously solved to obtain a set of estimators of the parameters θ that are functions of the sample moments (also called the empirical moments). The kth sample moment is defined as

$$m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

For example, the first sample moment is the familiar sample mean,

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

To illustrate the method, consider a simple example in which n diploid individuals are sampled from a population and genotyped for a co-dominant genetic marker. The total counts of chromosomes bearing allele A in the genotype of individual i is $x_i \in \{0,1,2\}$. Assuming random mating and random sampling from a large population (or sampling with replacement) the probability distribution of x_i is a binomial distribution,

$$\Pr(x_i|p,2) = \binom{2}{x_i} p^{x_i}(1-p)^{2-x_i}, \tag{3.1}$$

where p is the population frequency of allele A and $x_i \in \{0,1,2\}$. The mean number of A alleles per genotype is

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

and the first moment of the binomial distribution is

$$\begin{aligned} \mathbb{E}(x_i) &= \sum_{j=0}^{2} x_i \Pr(x_i|p,2), \\ &= 2p(1-p) + 2p^2, \\ &= 2p. \end{aligned}$$

Setting $m_1 = \mathbb{E}(x_i)$ and solving for p gives

$$\hat{p} = \frac{1}{2n}\sum_{i=1}^{n} x_i.$$

Thus, the moment estimator of the population frequency of allele A is simply the overall frequency of the allele in the sample, which is quite intuitive. We will see later that the sample frequency is also the maximum likelihood estimator of the population allele frequency.

As another example, consider a sample of n iid random variables $\mathbf{x} = x_1, \ldots, x_n$ from a normal distribution with mean μ and variance σ^2. The expected value of x_i is

$$\mathbb{E}(x_i) = \mu,$$

and the expectation of the second moment is,

$$\mathbb{E}(x_i^2) = \sigma^2 + \mu^2,$$

for all $i = 1, 2, \ldots, n$. Setting the first empirical moment equal to the first moment of the distribution gives the moment estimator of the mean to be the sample mean,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Setting the second empirical moment equal to the second moment of the distribution and substituting the estimator of μ into the formula gives,

$$\begin{aligned}
\mathbb{E}(x_i^2) &= \frac{1}{n}\sum_{i=1}^{n} x_i^2, \\
\sigma^2 + \mu^2 &= \frac{1}{n}\sum_{i=1}^{n} x_i^2, \\
\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2, \\
\hat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 .
\end{aligned}$$

This moment estimator of the variance is intuitive but is biased for small sample sizes unless multiplied by $n/(n-1)$ (see discussion below). Example 3.1 below derives a moment estimator for the population genetic parameter θ using the number of segregating sites between a pair of DNA sequences observed at multiple loci.

Sampling distributions of estimators

In the frequentist framework, the parameter of a probability distribution, θ, has a fixed value for a given experiment and each sample is a collection of n random variables $\mathbf{x} = x_1, \ldots, x_n$ drawn from the probability distribution specified by $f(\mathbf{x}|\theta)$. One can then consider the probability distribution of an estimator of θ, $\hat{\theta} = g_n(\mathbf{x})$, if many hypothetical samples, each of size n, were drawn from $f(\mathbf{x}|\theta)$. This conceptual distribution is known as the **sampling distribution** of the statistic $g_n(\mathbf{x})$ and provides a picture of how variable the statistic (or estimate of θ) will be from one sample to another, summarizing its uncertainty. We will also write the sampling distribution as $f_n(\hat{\theta}|\theta)$. Later, we will see how the sampling distribution can be used to derive a confidence interval for an estimator. The sampling distribution of the method of moments estimator of allele frequency, $\hat{p}$, derived in the previous section is shown in Figure 3.1 for samples of $n = 10$, 20, 50 or 100 individuals. This distribution was estimated using simulation methods that are described later in this chapter and assuming a true parameter value of $p = 0.5$.

The sampling distribution of a statistic $g_n(\mathbf{x})$ under the density $f(\mathbf{x}|\theta)$ normally depends on the parameter, θ, which is unknown (and is, in fact,

EXAMPLE 3.1 In Chapter 2 we derived the probability distribution for the number of DNA substitutions between a pair of sequences (assuming random mating and an infinite sites model of substitution) as

$$\Pr(M) = \left(\frac{\theta}{1+\theta}\right)^M \frac{1}{1+\theta},$$

where $\theta = 4N\nu$ and $\nu = \mu' s$ is the per gene mutation rate on a generational timescale. The expectation of this distribution is $\mathbb{E}(M) = \theta$. To derive a moment estimator of θ we assume that 2 random sequences are sampled from a population for each of l neutral loci with equal per gene mutation rates. This assumption is required to insure that the samples across loci are independent and identically distributed random variables from the distribution given in equation 2.14. The mean number of segregating sites in the sample (averaging across loci) is

$$\overline{M} = \frac{1}{l}\sum_{i=1}^{l} M_i,$$

where M_i is the number of segregating sites for sequences at locus i. The moment estimator of θ is obtained by setting the mean equal to the expectation and solving for θ which is particularly simple in this example,

$$\hat{\theta} = \overline{M}.$$

We have used a $\hat{\theta}$ symbol to denote an estimator of the parameter θ. Thus, as one might expect a larger number of segregating sites implies either a larger mutation rate or a larger population size.

what we are trying to estimate). How can we therefore obtain the sampling distribution for an estimator of an unknown parameter? One solution to this problem is to identify a function of both $g_n(\mathbf{x})$ and θ that has a distribution that is independent of θ. Simple examples include $g_n(\mathbf{x}) - \theta$ and $g_n(\mathbf{x})/\theta$ for so-called “location” and “scale” parameters, respectively. Such functions are called **pivots**. Another solution is to predict the sampling distribution by assuming that the sample estimate of the parameter, $\hat{\theta}$, is close to the true value of θ and that the sampling distribution can therefore be approximated as the distribution of $g_n(\mathbf{x})$ under $f(\mathbf{x}|\hat{\theta})$. In general,

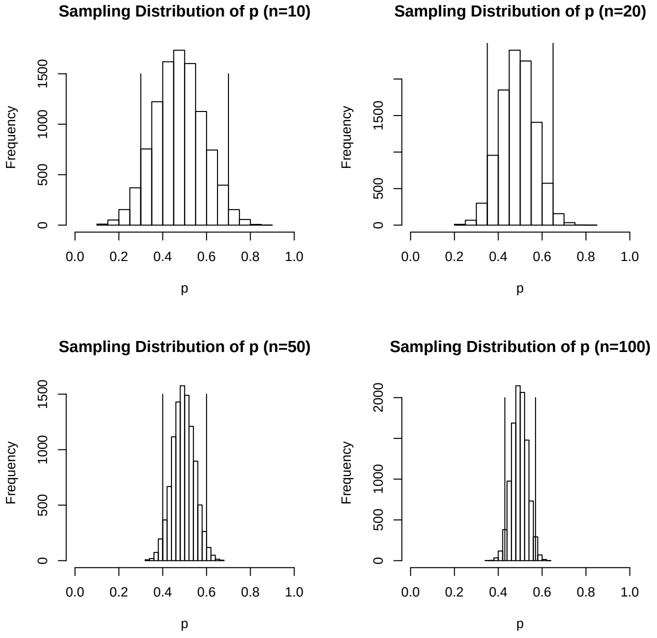


Figure 3.1: The sampling distribution of the estimator $\hat{p} = x/2n$ for 4 different sample sizes of $n = 10$ (upper left), $n = 20$ (upper right), $n = 50$ (lower left) and $n = 100$ (lower right). Each sampling distribution was estimated by simulating 10, 000 samples, each of size n, from the distribution of equation 3.1 with $p = 0.5$. The upper and lower limits of the 95 percent confidence intervals are indicated by vertical lines.

deriving the sampling distribution of a statistic $g_n(\mathbf{x})$ under a probability distribution $f(\mathbf{x}|\theta)$ is a difficult transformation of variables problem and exact analytical results for sampling distributions often do not exist. Before fast computers became widely available that allowed distributions to be generated by simulation, statisticians expended considerable effort deriving sampling distributions. Often they focused on "asymptotic" approximations for sampling distributions of commonly used statistics in the limit of large sample size, n. Arguably the most important asymptotic sampling distribution result is that for the sampling distribution of the sample mean in the limit of large n (see below).

It is sometimes possible to derive a sampling distribution for a statistic based on the difference $g_n(\mathbf{x}) - \theta$ that does not depend on the parameter,

θ, of the model as the following example illustrates. Let $\mathbf{x} = x_1, x_2, \ldots, x_n$ be a sample of random variables from a normal distribution with unknown mean μ and known variance σ^2 with the pdf for the ith variable being,

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

The sample mean $\overline{x}$ will be used as an estimator of μ. It can be shown that the pdf of the sampling distribution of $\overline{x}$ is

$$f_n(\overline{x}) = \sqrt{\frac{n}{2\pi\sigma^2}} e^{-\frac{n}{2\sigma^2}(\overline{x}-\mu)^2}, \; -\infty < x_i < \infty,$$

which is a normal distribution with mean μ and variance σ^2/n. Note that μ only appears in the term $(\overline{x} - \mu)$ and so the probability density of the magnitude of the difference of the estimator $\overline{x}$ from the true value μ does not depend on the true value of the parameter. This allows a confidence interval for an estimate of μ to be derived from the sampling distribution that does not depend on the true value of the parameter μ apart from determining the "location" of the distribution on the x axis (e.g., the position of the mode of the sampling distribution; see below).

Properties of optimal point estimators

In principle, any function of the data may be used as an estimator of a parameter. Some estimators, such as the sample mean as an estimate of the population mean, are intuitively sensible, whereas others (such as using a constant value of 4 as an estimate of the population mean) are not. From a practical perspective, it makes sense to use some information from the specified family of probability distributions to derive reasonable estimators and to use as much relevant information from the data as possible. The method of moments, maximum likelihood and Bayesian posterior means (or modes) are all examples of formal frameworks for deriving point estimators with reasonable properties. In some cases, it is possible to show that any estimator derived by one of these approaches satisfies certain mathematical properties that are desirable.

Here, we consider some mathematical properties that can be used to assess the quality of an estimator. We first consider two "finite sample" properties, bias and mean square error both of which are influenced by sample size. The **bias** of an estimator $g_n(\mathbf{x})$ of a parameter θ is defined as

$$\text{Bias} = \mathbb{E}[g_n(\mathbf{x})] - \theta,$$

where for a discrete random variable

$$\mathbb{E}[g_n(\mathbf{x})] = \sum_{\mathbf{x}} g_n(\mathbf{x})\Pr(\mathbf{x}|\theta),$$

is the expectation over all possible distinct samples of size n. The definition for a continuous random variable is similar but with the sum replaced by an integral. An estimator is said to be **unbiased** if its expectation equals the true value of the parameter (so that the Bias equals zero). In terms of the sampling distribution of the estimator, this is equivalent to its average value equaling the true parameter value. The mean square error (MSE) of an estimator is defined as

$$\begin{aligned} \text{MSE} &= \mathbb{E}[(g(\mathbf{x})-\theta)^2], \\ &= \text{Var}[g(\mathbf{x})] + \text{Bias}^2. \end{aligned}$$

An estimator with lower MSE deviates less, on average, from the true parameter value and is therefore preferable to an estimator with higher MSE. The above formulation suggests the fundamental trade-off between bias and variance in developing estimators. If the MSE is fixed then in order to reduce bias one must accept more variance and vice versa. Bias and variance are illustrated in Figure 3.2. The example of the sampling distribution for an estimator of the mean of a normal distribution given above shows that the sample mean is an unbiased estimator of μ in this case because $\mathbb{E}[\overline{x}] = \mu$. The MSE is then completely determined by the variance which is σ^2/n and the MSE of the estimator strictly decreases with increasing n. Most reasonable estimators have this property. As an example of a biased estimator, consider the method of moments estimator of the sample variance derived above,

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2.$$

Taking the expectation of this estimator over the probability density of $\mathbf{x}$ gives

$$\begin{aligned} \mathbb{E}(\hat{\sigma^2}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(x_i^2) + \mathbb{E}(\overline{x}^2), \\ &= \frac{1}{n}n[\text{Var}(x_i) + \mathbb{E}(x_i)^2] + \text{Var}(\overline{x}) + [\mathbb{E}(\overline{x})]^2, \\ &= \sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2\right), \end{aligned}$$

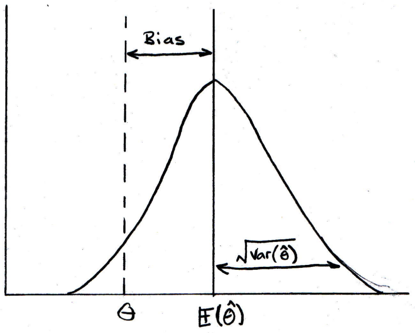


Figure 3.2: The relationship between the bias and variance of an estimator in determining its mean squared error. Dashed line indicated true value of parameter θ, solid line indicates expected value of estimator. The mean square error is determined by the squared deviations of estimator around true value of θ.

$$= \sigma^2 \left(1 - \frac{1}{n}\right). \qquad (3.2)$$

Note that on the third line of equation 3.2 we have used the result that the sampling distribution of $\overline{x}$ is normal with parameters $\mathbb{E}(\overline{x}) = \mu$ and $\text{Var}(\overline{x}) = \sigma^2/n$. Thus, $\mathbb{E}(\hat{\sigma^2}) \neq \sigma^2$ and we must multiply the methods of moment estimator $\hat{\sigma^2}$ by the term $n/(n-1)$ to obtain the unbiased estimator given previously.

It is interesting to note that in the case of a sample from a normal distribution the unbiased estimator s^2 has uniformly larger mean square error than the biased method of moments estimator. To see this, first express the method of moments estimator as a function of the unbiased estimator,

$$\begin{aligned} \hat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2, \\ &= \left(\frac{n-1}{n}\right)\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right], \end{aligned}$$

$$= \left(\frac{n-1}{n}\right) s^2.$$

The sampling variance of the methods of moments estimator is then

$$\mathrm{Var}(\hat{\sigma^2}) = \mathrm{Var}\left[\left(\frac{n-1}{n}\right) s^2\right] = \left(\frac{n-1}{n}\right)^2 \mathrm{Var}(\mathrm{s}^2),$$

and the bias is

$$\begin{aligned}\mathrm{Bias}(\hat{\sigma^2}) &= \mathbb{E}(\hat{\sigma^2}) - \sigma^2, \\ &= \sigma^2\left(\frac{n-1}{n}\right) - \sigma^2, \\ &= -\frac{\sigma^2}{n}.\end{aligned}$$

Combining these results, we can write the mean square error as

$$\begin{aligned}\mathrm{MSE}(\hat{\sigma^2}) &= \mathrm{Var}(\hat{\sigma^2}) + \mathrm{Bias}(\hat{\sigma^2})^2, \\ &= \left(\frac{n-1}{n}\right)^2 \mathrm{Var}(s^2) + \frac{(\sigma^2)^2}{n^2}. \qquad (3.3)\end{aligned}$$

If the variables have a normal distribution, the sampling variance (and mean square error) of the unbiased estimator of variance is

$$\mathrm{Var}(s^2) = \frac{2(\sigma^2)^2}{n-1}, \qquad (3.4)$$

and substituting equation 3.4 into 3.3 we obtain

$$\begin{aligned}\mathrm{MSE}(\hat{\sigma^2}) &= \frac{2(n-1)(\sigma^2)^2}{n^2} + \frac{(\sigma^2)^2}{n^2}, \\ &= \frac{(2n-1)(\sigma^2)^2}{n^2}. \qquad (3.5)\end{aligned}$$

To examine the relative efficiency of the two estimators we consider the ratio of their mean square errors,

$$\frac{\mathrm{MSE}(\hat{\sigma^2})}{\mathrm{MSE}(s^2)} = \frac{\left[\frac{2n-1}{n^2}\right](\sigma^2)^2}{\left[\frac{2}{n-1}\right](\sigma^2)^2} = \frac{(n-1/2)(n-1)}{n^2}.$$

For any finite n this ratio is less than 1, and it only approaches 1 asymptotically with increasing n. Therefore, if a variable has a normal probability distribution, the biased method of moments estimator is always better than the unbiased estimator in terms of its mean square error. Nonetheless, the unbiased estimator of the variance is used in virtually all software packages.

An important asymptotic property of reasonable estimators is consistency. Intuitively, an estimator is consistent if its sampling distribution becomes increasingly concentrated near the true value of the parameter with increasing n. More formally, we say that an estimator is **consistent** if it converges in probability to θ with increasing n,

$$g_n(\mathbf{x}) \rightarrow^P \theta.$$

The function $g_n(\mathbf{x})$ converges in probability to θ if

$$\Pr(|g_n(\mathbf{x}) - \theta| < \epsilon) \rightarrow 1 \text{ as } n \rightarrow \infty,$$

for every $\epsilon > 0$. The estimate $\overline{x}$ of μ for a sample of iid random variables from a normal distribution (presented above) is consistent. This follows from the fact that

$$\mathbb{E}[\overline{x}] = 0 \text{ and } \lim_{n\rightarrow\infty} \mathrm{Var}(\overline{x}) = \lim_{n\rightarrow\infty} \sigma^2/n = 0.$$

The increasing concentration of the sampling distribution of $\overline{x}$ around μ with increasing n is illustrated in Figure 3.3. Consistent estimators are not unique and some may have much better statistical properties for finite n, so consistency should be viewed as a minimal requirement for estimators. The optimality criteria presented here are arguably the most widely used. However, be aware than many alternatives exist although a more comprehensive treatment is beyond the scope of this book.

Frequentist confidence intervals

The frequentist approach to construct a confidence interval (CI) for a parameter estimate makes use of the sampling distribution of the estimator. The parameter θ is treated as a fixed quantity and we consider the sampling distribution of an estimator $\hat{\theta}$ (the probability density of $\hat{\theta}$ given θ) denoted as $f_n(\hat{\theta}|\theta)$. To specify an interval estimator for the range of plausible values of the parameter θ, we choose an upper bound U and a lower bound L, such that $L \leq U$. The $100(1-\alpha)$ percent confidence interval for an estimator $\hat{\theta} = g_n(\mathbf{x})$ (with equal tail probabilities) has a lower bound satisfying $F(L|\theta) = \alpha/2$ and an upper bound satisfying $F(U|\theta) = 1 - \alpha/2$ where F is

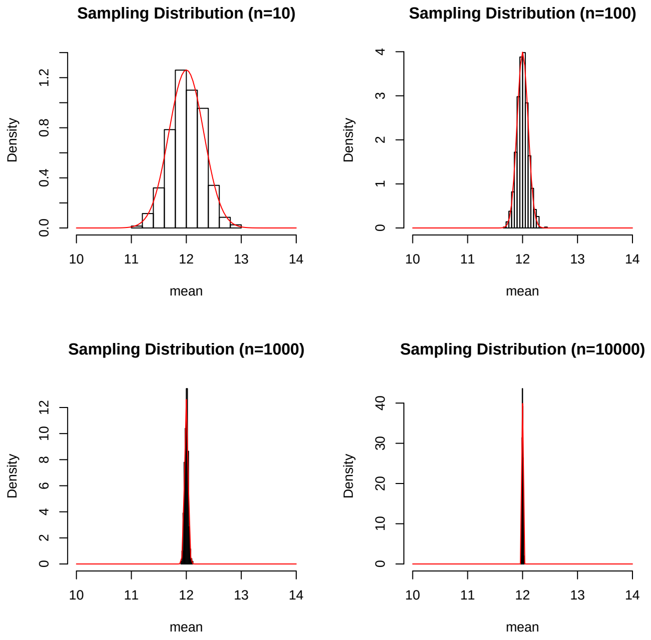


Figure 3.3: Sampling distributions of the sample mean, $\overline{x}$ for samples of n iid normal random variables with $\mu = 12$ and $\sigma^2 = 1$. The distribution inferred from 1,000 simulated data sets, each of size n, is represented by the histogram and the predicted analytical sampling distribution (with $\mu = 1$ and $\sigma = 1/\sqrt{n}$) is shown as the red line. The four plots are for samples of $n = 10$ (upper left), $n = 100$ (upper right), $n = 1000$ (lower left) and $n = 10000$ (lower right).

the cdf of the sampling distribution of $g_n(\mathbf{x})$, so that $F(L|\theta) = \Pr(g_n(\mathbf{x}) \leq L)$ and so on. The bounds U and L are referred to as the $1 - \alpha/2$ and $\alpha/2$ percentiles of the sampling distribution.

The CI depends on the value of the parameter θ (which is unknown) and so we use the sampling distribution $f(\mathbf{x}|\hat{\theta})$, with $\hat{\theta}$ estimated from the observed data. The CI is then a function of the data, $\mathbf{x}$, which determines our estimate $\hat{\theta}$ and thus the value of θ that is used to generate the sampling distribution, and also specifies the sample size n. In some cases, a pivot function, $Q(\overline{x}, \theta)$ can be found that has a sampling distribution that is independent of θ. The sampling distribution and CI for the genotype sampling experiment described above are shown in Figure 3.1. The upper and lower limits of the $100(1-\alpha)$ percent confidence interval are determined by $\hat{\theta}$ and n and are therefore functions of the sample data, $\mathbf{x}$. We can thus treat the upper and lower bounds of the CI as functions of the data, $U(\mathbf{x})$ and $L(\mathbf{x})$. It is clear from this representation that the CIs will have a sampling distribution and the probability that the CI includes the parameter value θ (which is fixed) is

$$\begin{aligned} 1-\alpha &= \mathbb{E}[I(L(\mathbf{x}) < \theta < U(\mathbf{x}))] \\ &= \int_{\mathbf{x}} I(L(\mathbf{x}) < \theta < U(\mathbf{x})) f(\mathbf{x}|\theta) d\mathbf{x}. \end{aligned} \tag{3.6}$$

where the expectation is taken over the probability distribution of the sampled data and I is an indicator function that takes a value of 1 if the inequality is true and zero otherwise. One can interpret this value as being the expected proportion of datasets for which the CI "covers" the true value of the parameter θ. The significance value of a CI is therefore sometimes also called its coverage level. This is not to be interpreted as the probability that the parameter is located in the interval because the parameter value is fixed under the frequentist paradigm. Instead, we think of the random interval as having some probability of covering the true parameter value θ (the CI is treated as a random variable). We will see shortly that the probability associated with a Bayesian credible set for a parameter can be interpreted as the probability that the parameter lies in the interval. This is possible because Bayesian inference views parameters as random quantities which are essentially random variables themselves.

To illustrate, consider a sample of iid random variables $\mathbf{x} = x_1, x_2, \ldots, x_n$ from a normal distribution with unknown mean μ_x and known variance σ_x^2. We noted previously that the sampling distribution of the estimator $\overline{x}$ of the sample mean is also a normal distribution, but with mean μ_x and variance σ_x^2/n. A $100(1-\alpha)$ percent CI may be derived from this sampling distribution as follows. The transformed sample mean $(\overline{x}-\mu_x)/\sigma_x$ has a standard normal pdf and therefore if $F(.)$ is the cdf for a standard normal pdf we can choose L and U to satisfy $F(L) = \alpha/2$ and $F(U) = 1-\alpha/2$ to obtain,

$$\Pr\left(L \leq \frac{(\overline{x}-\mu)\sqrt{n}}{\sigma} \leq U\right) = 1-\alpha.$$

Because the normal distribution is symmetrical about the mean $L = -U$. By convention, we denote $U = z_{\alpha/2} = -L$. The value $z_{\alpha/2}$ is the $\alpha/2$ percentile for a standard normal distribution. The $100(1-\alpha)$ percent CI for μ_x is then obtained by rearranging terms to give

$$\mu_x \in \left(\overline{x} - z_{\alpha/2} \frac{\sigma_x}{\sqrt{n}}, \overline{x} + z_{\alpha/2} \frac{\sigma_x}{\sqrt{n}} \right). \tag{3.7}$$

The term $\sigma_x / \sqrt{n}$ is termed the standard deviation of $\overline{x}$, denoted as $\text{SD}_{\overline{x}}$. If σ_x is unknown and we instead use $\sqrt{s^2}$ as an estimator of σ_x, the result is the standard error of $\overline{x}$,

$$\text{SE}_{\overline{x}} = \frac{s_x}{\sqrt{n}}.$$

If μ and σ^2 are both unknown and we instead transform the sample mean using

$$\frac{\sqrt{n}(\overline{x} - \mu_x)}{s_x},$$

then the transformed sample mean has a sampling distribution that is a t distribution with $n-1$ df. The $100(1-\alpha)$ percent confidence interval for the mean is then

$$\mu_x \in \left(\overline{x} - t_{\alpha/2,n-1} \frac{s_x}{\sqrt{n}}, \overline{x} + t_{\alpha/2,n-1} \frac{s_x}{\sqrt{n}} \right), \tag{3.8}$$

where $t_{\alpha/2,n-1}$ satisfies $F_{n-1}(t_{\alpha/2,n-1}) = \alpha/2$ where $F_{n-1}(.)$ is the cdf for the t distribution with $n-1$ df. The $100(1-\alpha)$ percent CI is often written in the simplified form,

$$\overline{x} \pm t_{\alpha/2,n-1} \text{SE}_{\overline{x}}.$$

The t distribution is a special case of a more general distribution known as the gamma (the t distribution has one parameter, whereas the gamma has two). The parameter that defines the shape of the t distribution is the number of degrees of freedom $n-1$. The pdf of the t distribution is

$$f_{n-1}(t) = \frac{\Gamma[n/2]}{\Gamma[(n-1)/2]\sqrt{(n-1)\pi}} \left(1 + \frac{t^2}{n-1} \right)^{-n/2}.$$

This is the exact sampling distribution and is therefore appropriate for use with small samples. However, the underlying variables must follow a normal distribution which is a rather restrictive requirement. For $n \geq 4$ the variance of the t distribution is $\text{Var}(t) = (n-1)/(n-3)$ which is greater than the variance of a standard normal distribution, reflecting the additional uncertainty due to the unknown σ^2. The sampling distribution for the variance is also available in this restricted case. The transformed sample variance

$$\frac{(n-1)\hat{\sigma}_x^2}{\sigma_x^2},$$

is approximately χ^2 distributed with $n-1$ df. Thus, a $100(1-\alpha)$ percent confidence interval for the estimated variance is

$$\sigma_x^2 \in \left(\frac{(n-1)\hat{\sigma}_x^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)\hat{\sigma}_x^2}{\chi^2_{1-\alpha/2,n-1}} \right),$$

where $F_{n-1}(\chi^2_{\alpha/2,n-1}) \leq \alpha/2$ and $F_{n-1}(.)$ is the cdf for a χ^2 distributed random variable with $n-1$ df.

Asymptotic distribution theory

Asymptotic distribution theory describes the sampling distribution of statistics (such as the sample mean) in the limit of large n. This can lead to simplifying formulas and robust conclusions that do not depend on the specific form of the probability distribution that has generated the data (which is usually unknown). Thus, fewer assumptions may be needed to apply the methods. An important asymptotic result is that the t distribution converges to a standard normal density with increasing n. This can be seen by noting that

$$\lim_{n\to\infty} \left(1 + \frac{t^2}{n-1}\right)^{-n/2} = e^{-t^2/2},$$

and by applying Stirling's approximation for the Γ function,

$$\Gamma(y) \approx \sqrt{\frac{2\pi}{y}} \left(\frac{y}{e}\right)^y,$$

so that by substitution we obtain,

$$\begin{aligned} \frac{\Gamma[n/2]}{\Gamma[(n-1)/2]\sqrt{(n-1)\pi}} &\approx \frac{\sqrt{\frac{2\pi}{(n/2)}} \left(\frac{n/2}{e}\right)^{n/2}}{\sqrt{\frac{2\pi}{((n-1)/2)}} \left(\frac{(n-1)/2}{e}\right)^{(n-1)/2} \sqrt{\pi}}, \\ &= \left(\frac{n}{n-1}\right)^{(n-1)/2} \frac{1}{\sqrt{2\pi e}}, \end{aligned}$$

and taking limits,

$$\lim_{n\to\infty}\left(\frac{n}{n-1}\right)^{(n-1)/2}=\sqrt{e},$$

so that the limiting density of t is

$$\lim_{n\to\infty}f_{n-1}(t)=\frac{1}{\sqrt{2\pi}}e^{-t^2/2}.$$

This confirms that a t distribution can be used to describe the sampling distribution of the mean for both large (and small) samples from a normal distribution.

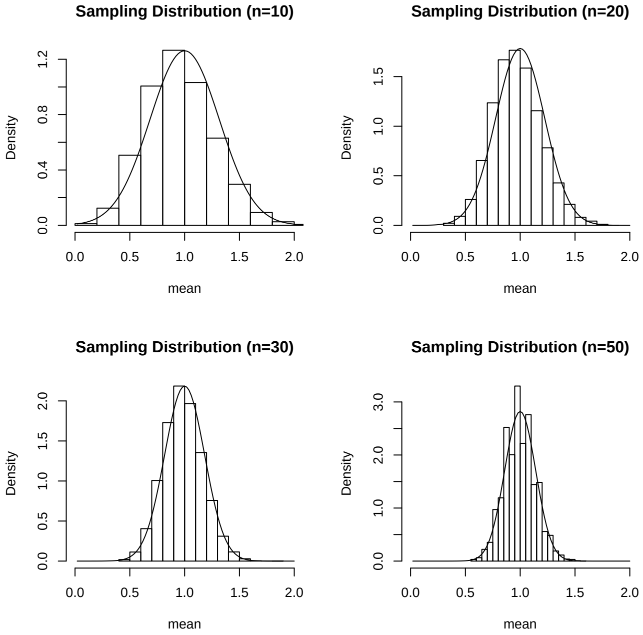
One of the most important asymptotic results is that the sampling distribution (in the limit of large n) of the sample mean (suitably transformed) is a normal distribution, regardless of the specific distribution that the variables are drawn from. This follows from the famous central limit theorem (CLT) which specifies that the distribution of a sum of n iid random variables (multiplied by $\sqrt{n}$) follows a normal distribution in the limit of large n.

The sample mean is calculated by taking a sum of random variables and multiplying by a constant $1/n$. If the variables are iid then the limiting distribution of this statistic, according to the CLT, is also a normal distribution but with mean μ_x and variance σ_x^2/n (see Figure 3.4). Thus, the sampling distribution of $\overline{x}$ for iid variables $\mathbf{x}$ from an arbitrary distribution (with a few reasonable constraints, such as finite variance) is exactly the same as the distribution of $\overline{x}$ when the samples are from a normal distribution. In other words, the sampling distribution of the transformed mean

$$\frac{\sqrt{n}(\overline{x}-\mu_x)}{s_x}=\frac{\overline{x}-\mu_x}{SE_{\overline{x}}},$$

is, for large n, a standard normal density. In the limit of large n a consistent estimator such as s^2 leads to a precise estimate of σ^2 and so we can use the sample variance s^2 in place of the true variance (which is unknown) to construct a $100(1-\alpha)$ percent CI using the formulation presented in equation 3.8 above.

As an example, the normal approximation for the distribution of the statistic $\hat{p}$ derived earlier (which is simply the sample mean of the number of copies of allele A per chromosome) is shown in Figure 3.5. To generate the figure we have used $\sigma_x^2=p(1-p)/2$ which is the expected variance for this distribution. It is clear from the figure that the approximation is very



Sampling Distribution (n=10)

Sampling Distribution (n=20)

Sampling Distribution (n=30)

Sampling Distribution (n=50)

Figure 3.4: Sampling distributions of the sample mean, $\overline{x}$ for samples of n iid Poisson random variables with $\lambda = 1$. The distribution inferred from $10,000$ simulated data sets, each of size n, is represented by the histogram and the normal approximation to the sampling distribution (with $\mu = 1$ and $\sigma = 1/\sqrt{n}$) is shown as the solid line. The four plots are for samples of $n = 10$ (upper left), $n = 20$ (upper right), $n = 30$ (lower left) and $n = 50$ (lower right).

close to the observed sampling distribution (obtained by simulation) when $n \geq 30$.

The results presented above are restricted to some simple statistics (such as the sample mean and variance) and, in some cases, to samples generated under particular distributions (such as the normal distribution). We now consider some alternative methods that use computer simulation (the parametric bootstrap) and can generate the sampling distribution of a test statistic for random variables under any sample size and from virtually any probability distribution. We finish up by describing a nonparametric method (the nonparametric bootstrap) for generating the sampling distribution of a test statistic via simulated sampling without explicitly specify-

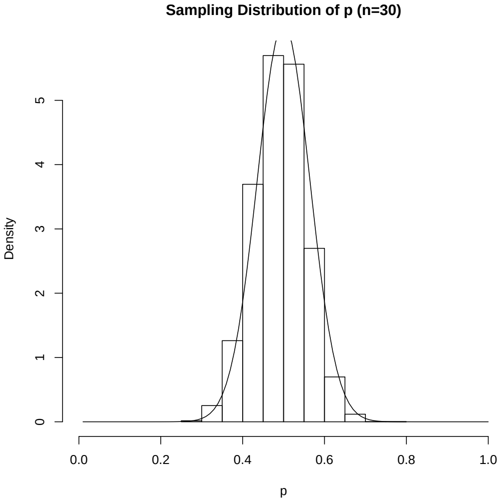


Figure 3.5: Sampling distributions of the estimator $\hat{p} = x/2n, \overline{x}$ for a sample of $n = 30$ from the density defined by equation 3.1 with $p = 0.5$. The distribution inferred from 10,000 simulated data sets, each of size n, is represented by the histogram and the normal approximation to the sampling distribution (with $\mu = p = 0.5$ and $\sigma = p(1-p)/\sqrt{2n}$) is shown as the solid line.

ing the underlying probability distribution that has generated the random variables.

Parametric bootstrap

The parametric bootstrap uses variables simulated on a computer to predict the sampling distribution of a statistic and obtain quantities such as CIs. Let $f(x|\theta)$ be a probability density (mass) function for a sample of $\mathbf{x} = x_1, x_2, \ldots, x_n$ iid random variables. Let $g_n(\mathbf{x}) = \hat{\theta}$ be an estimator of θ. If θ were known, we could generate observations from the sampling distribution $f(\hat{\theta}|\theta)$ by simulating n independent random variables from $f(x|\theta)$ and calculating $g_n(\mathbf{x})$. By repeating this procedure many times, we could simulate the expected frequency distribution of $\hat{\theta}$. This in turn could be used to predict, for example, the $100(1-\alpha)$ percent confidence interval (with equal tail probabilities) for $\hat{\theta}$ by setting the upper and lower limits equal to the empirical quantiles for $\alpha/2$ and $(1-\alpha/2)$. Namely, the threshold values that a percentage $\alpha/2$ of simulated values are less than and a percentage $(1-\alpha/2)$ of simulated values are greater than, respectively.

We again face the problem that the sampling distribution of our estimator depends on the true value of the parameter. If the estimator is consistent, and the sample size is sufficiently large, we may not go too far wrong if we simply simulate from the sampling distribution of the statistic using the point estimate $\theta = \hat{\theta}$ calculated using the original data to parameterize the density function, $f(x|\hat{\theta})$. This is the basic idea underlying the parametric bootstrap (PB). This form of bootstrap resampling is "parametric" because we have assumed that the probability distribution generating the observed data is specified, apart from one or more unknown parameters. To clarify the procedure, we outline an algorithm in pseudocode below.

1. Calculate $g_n(\mathbf{x}) = \hat{\theta}$ using the original data
2. Simulate R samples $\mathbf{y} = \{y_i^n\}$ where y_i^n is a vector of n iid random variables simulated from $f(x|\hat{\theta})$ for the ith simulation.
3. Construct a vector $\mathbf{q} = \{q_i\}$, where $q_i = g_n(y_i^n)$ and rank order the elements of $\mathbf{q}$
4. Set $L = q_i$ where $i/R \leq \alpha/2 \leq (i+1)/R$ and set $U = q_j$ where $(j-1)/R \leq 1-\alpha/2 \leq j/R$
5. The approximate $100(1-\alpha)$ percent CI is $[L, U]$

The error of this approximation to the sampling distribution is inversely proportional to the sample size. Advanced techniques exist for further reducing the error of a PB but are beyond the scope of this book.

As an example applying the PB suppose that $n = 10$ variables are sampled from a normal distribution with unknown mean and variance as given below,

$$\begin{aligned}\mathbf{x} \quad = \quad & (-0.398, -1.447, -0.269, -0.179, -0.640, \\ & -0.605, 1.599, 0.439, -0.863, -0.653)\end{aligned}$$

The estimates of the mean and variance are $\overline{x} = -0.302$ and $s^2 = 0.683$. To construct a 95 percent CI for μ using the bootstrap we simulated $R = 10000$ random variables from a normal distribution with mean and variance as calculated above. In this case, 2.5 percent of the sample means for simulated datasets are less than -0.813 and 2.5 percent are greater than 0.202

and so the 95% CI for μ is $\mu \in [-0.81, 0.20]$. If we instead use a t distribution to generate the 95% CI we obtain $\mu \in [-0.89, 0.29]$. In this case, we know the sampling distribution and the use of the PB is uneccessary, but the comparison reveals an interesting pattern. The CI obtained using the PB is narrower than that obtained using the t distribution (which is the exact sampling distribution) because we have ignored the uncertainty about σ^2 (which was estimated by s^2 and then treated as fixed in our simulations). If we generate a CI using the normal approximation presented earlier (which also assumes that σ^2 is known) we obtain $\mu \in [-0.81, 0.21]$ which is very similar to the result obtained using the parametric bootstrap. Thus, the bootstrap can be expected to provide more accurate CIs when sample sizes are larger. The error of the PB approximation is in this case proportional to $1/10 = 0.1$ which is roughly the difference we see between the exact CI and the CI estimated using the PB. Of course, the PB will be most useful in cases where an exact sampling distribution is unavailable but we are able to simulate from the probability density that is assumed to have generated the sample data.

Nonparametric bootstrap

The basic idea underlying the nonparametric bootstrap (NPB) is to use the sample data to infer an "empirical" distribution function defined as,

$$\hat{F}(y) = \sum_{i=1}^{n} \frac{I[y_j \leq y]}{n},$$

where $I[.]$ is the indicator function, evaluating to 1 if the enclosed condition is true and 0 otherwise. Samples can be simulated from the empirical distribution to estimate the sampling distribution of an arbitrary statistic $T(y)$. The approach is "nonparametric" because we have not specified any particular functional form for $F(.)$. A straightforward way to simulate from $\hat{F}(.)$ is to resample datasets from the original data. Each simulated dataset is constructed by sampling with replacement from the original observations. This is equivalent to assigning a probability $1/n$ to every observation in the original data.

To illustrate the use of a NPB to construct a 95% CI for μ we apply the method to the dataset of $n = 10$ observations presented in the previous section. We generate $R = 10000$ datasets by resampling (with replacement) from the original data. The statistic estimator $\overline{x}$ is calculated for each resampled dataset and the lower 2.5 percent and upper 97.5 prerent quantiles are taken as the lower and upper limits, respectively, for the CI. The resulting CI is $\mu \in [-0.74, 0.24]$ which is slightly narrower than the CI obtained using the PB. The error of the simple NPB outlined here is also proportional to the inverse of the sample size. This error can be reduced using modified bootstrap methods and the development of such methods is currently an active area of research in theoretical statistics.

Chapter 4

Maximum Likelihood

The method of maximum likelihood was developed by R. A. Fisher in the 1920s. The basic principle is to view the probability distribution as a function of the parameters, rather than the data. This is intuitively reasonable because once the experiment has been carried out the data are effectively fixed. Of course, when treated as a function of the parameters the pdf (or pmf) no longer specifies a probability distribution (i.e., it does not integrate to one, when integrating over the parameters with the data fixed, etc). Fisher called this new function the **likelihood function** to emphasize the fact that it was not a probability distribution but nonethless provided information about plausible values for the parameters. The **maximum likelihood estimate** of the parameters is obtained by maximizing the likelihood over the parameters; effectively finding the set of parameter values that maximize the probability of the observed data.

The likelihood

We now provide a more formal description of the maximum likelihood method of inference. We present the theory for continuous pdfs but the formulation for discrete pmfs is essentially similar. Let $\mathbf{x} = x_1, x_2, \ldots, x_n$ be a vector of random variables with joint pdf

$$f(x_1, x_2, \ldots, x_n|\theta).$$

The likelihood function is defined as

$$L(\theta|x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\theta).$$

If $\mathbf{x} = x_1, \ldots, x_n$ are iid with density $f(x_i|\theta)$ then

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta).$$

It is usual to maximize the natural logarithm of the likelihood rather than the likelihood itself. There are several reasons for this, most importantly that products of probabilities rapidly become very tiny numbers as n increases, making it difficult to carry out precise calculations, whereas sums of logs are more manageable. The log-likelihood function for a sample of n iid random variables is defined as

$$l(\theta|\mathbf{x}) = \sum_{i=1}^{n} \log\left[f(x_i|\theta)\right].$$

The log-likelihood function of μ for a sample of $n = 100$ random variables generated under a standard normal distribution (fixing $\sigma = 1$ in the likelihood) is shown in Figure 4.1

Maximum likelihood estimator

The maximum likelihood estimator (MLE) of θ is defined as

$$\hat{\theta} = \max_{\theta} l(\theta|x_1, x_2, \ldots, x_n).$$

To illustrate the method in practice, we provide a simple example estimating the parameter p of a binomial distribution. The likelihood is

$$L(p|x) = \binom{n}{x} p^x (1-p)^{n-x} = p^x (1-p)^{n-x},$$

where we have dropped the binomial coefficient from the equation because it does not depend on the parameter p and is therefore an irrelevant constant. The log-likelihood function is

$$l(p|x) = x \log(p) + (n-x)\log(1-p).$$

The log-likelihood function is plotted in Figure 4.2 for $n = 100$ and $x = 32$. It is straightforward to maximize this simple one-dimensional function using calculus. The partial derivative of the log-likelihood function (with respect to parameter p) gives the slope of a tangent line parallel to the curve

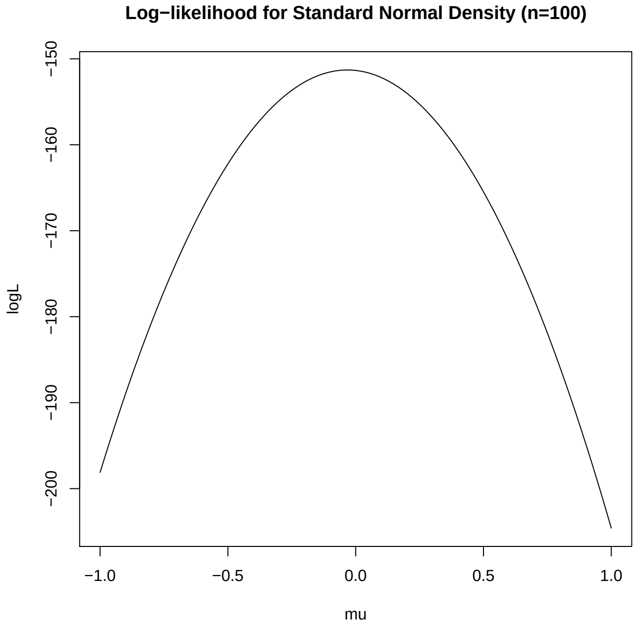


Figure 4.1: Plot of the log-likelihood as a function of μ (with $\sigma = 1$) for a simulated sample of $n = 100$ random variables from a standard normal distribution.

defined by the log-likelihood function at any point p. At the maximum, the partial derivative is equal to zero. Thus, we can find a maximum by calculating the partial derivative of the log-likelihood function (with respect to p), setting this equal to zero, and solving the resulting equation for p. The partial derivative is

$$\frac{\partial}{\partial p} l(p|x) = \frac{x}{p} - \frac{n-x}{1-p}.$$

This is equal to zero if

$$\frac{x}{p} = \frac{n-x}{1-p}.$$

Solving this equation for p gives,

$$\hat{p} = \frac{x}{n},$$

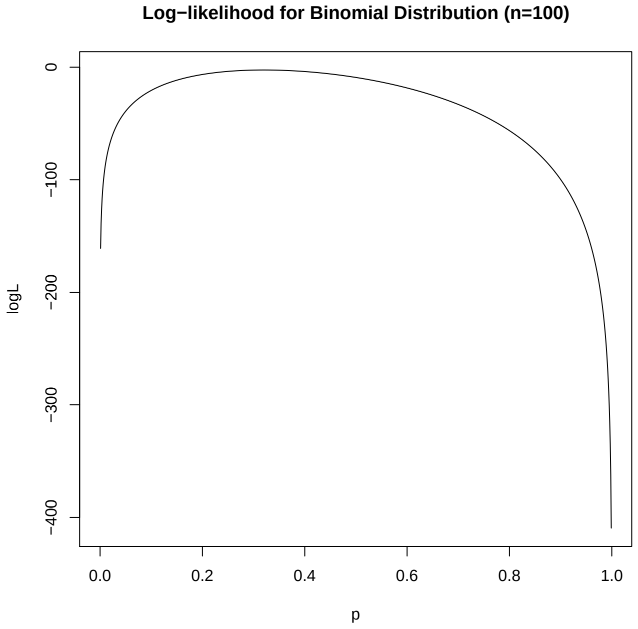


Figure 4.2: Plot of the log-likelihood as a function of p for a sample of $n = 100$ and observed random variable $x = 32$ for a binomial distribution.

so the maximum likelihood estimate of the chance of success in a binomial distribution is just the observed proportion of successes (as was the case for the method of moments estimator derived earlier). We have not been very careful in this example. A point at which the derivative is zero could also be a minimum of the log-likelihood, or there could be a maximum at one of the boundaries of the parameter space and the partial derivative will not be zero in that case. It is evident from plotting the function, however, that the point at which the derivative is zero is a maximum in this example (see Figure 4.2). In general, additional checks are needed to insure that such points are indeed global maxima.

For more complex likelihood functions analytical maximization is rarely possible and one must resort to maximizing the likelihood using numerical techniques. Often numerical routines are designed to minimize, rather than maximize, functions and we therefore minimize the negative log-likelihood instead (which is equivalent to maximizing the log-likelihood).

Large sample confidence intervals

The confidence interval for a MLE is essentially similar to the frequentist confidence interval described earlier. The asymptotic distribution of the MLE (for large n) is a normal distribution with mean $\mu = \theta$ and variance $\sigma^2 = 1/I(\theta)$, where

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} l(\theta|\mathbf{x})\right],$$

and the expectation is taken over the probability distribution $f(\mathbf{x}|\theta)$. This quantity is known as **Fisher's information**. The sampling distribution depends on the parameter θ, which is assumed to be fixed as in classical frequentist inference, and we again approximate θ by $\hat{\theta}$. The asymptotic $100(1-\alpha)$ percent confidence interval is then

$$\theta \in \left(\hat{\theta} - z_{\alpha/2}\frac{1}{\sqrt{I(\hat{\theta})}}, \hat{\theta} + z_{\alpha/2}\frac{1}{\sqrt{I(\hat{\theta})}}\right).$$

To illustrate, we revisit the binomial distribution. The second derivative (with respect to p) of the log-likelihood is

$$\frac{\partial^2}{\partial p^2} l(p|\mathbf{x}) = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}.$$

Substituting in the above result and taking the negative of the expectation with respect to variable x gives,

$$-\mathbb{E}\left[\frac{\partial^2}{\partial p^2} l(p|\mathbf{x})\right] = \frac{\mathbb{E}(x)}{p^2} + \frac{n-\mathbb{E}(x)}{(1-p)^2} = \frac{np}{p^2} + \frac{n-np}{(1-p)^2} = \frac{n}{p(1-p)}.$$

The asymptotic variance of the MLE $\hat{p} = x/n$ is therefore,

$$\sigma^2 = \frac{p(1-p)}{n}.$$

This is also the exact variance which we can derive directly,

$$\mathrm{Var}(\hat{p}) = \mathrm{Var}(x/n) = \frac{1}{n^2}\mathrm{Var}(x) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}.$$

The $100(1-\alpha)$ percent confidence interval of the estimate of p is

$$p \in \left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right),$$

where $\Pr(z > z_\alpha) = \alpha$ and z is a standard normal random variable.

Multivariate maximum likelihood

If θ is vector (or matrix, etc) of many parameters then the maximization is carried out over all the parameters jointly to find the global maximum of the log-likelihood function. For example, if there are K parameters $\theta_1, \ldots, \theta_K$, we maximize

$$\hat{\theta}_1, \ldots, \hat{\theta}_K = \max_{\theta_1, \ldots, \theta_K} l(\theta_1, \ldots, \theta_K | \mathbf{x}).$$

In simple situations, it may be possible to maximize the likelihood directly using analytical techniques to find a solution (in terms of the K parameters) to the set of K equations obtained by setting each of the K partial derivatives equal to zero

$$\begin{aligned} \frac{\partial}{\partial \theta_1} l(\theta_1, \ldots, \theta_K | \mathbf{x}) &= 0 \\ &\vdots \\ \frac{\partial}{\partial \theta_K} l(\theta_1, \ldots, \theta_K | \mathbf{x}) &= 0 \end{aligned} \tag{4.1}$$

There are certain technical requirements that must be satisfied for the solution to these equations to be a maxima (or a global maxima) which are beyond the scope of this book. Under certain conditions, the joint sampling distribution of the MLEs is multivariate normal with a vector of means $\theta_1, \ldots, \theta_K$ and variance-covariance matrix

$$V^{-1} = -\mathbb{E}\left\{ \frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\theta_1, \ldots, \theta_K | \mathbf{x}) \right\},$$

where V^{-1} denotes the inverse of the variance-covariance matrix. This is analogous to the asymptotic result for a single parameter likelihood that the sampling distribution of the MLE is a normal distribution. Again certain conditions must be satisfied for this to hold which are beyond the scope of this book.

To illustrate, consider an iid sample $\mathbf{x} = x_1, \ldots, x_n$ from a normal distribution with μ and σ unknown. Our objective will be to obtain MLEs for these two parameters. The likelihood is

$$L(\mu, \sigma | \mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{(x_i - \mu)^2/(2\sigma^2)} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2)\sum(x_i - \mu)^2/\sigma^2}$$

and the log-likelihood is

$$l(\mu, \sigma|\mathbf{x}) = -\frac{n}{2}(\log 2\pi + \log\sigma^2) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}.$$

The partial derivatives with respect to μ and σ^2 are

$$\begin{aligned}
\frac{\partial}{\partial\mu} l(\mu, \sigma^2|\mathbf{x}) &= \frac{n}{\sigma^2}(\overline{x}-\mu) \\
\frac{\partial}{\partial\sigma^2} l(\mu, \sigma^2|\mathbf{x}) &= \frac{n}{2\sigma^4}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - \sigma^2\right)
\end{aligned}$$

so that the MLEs obtained by setting the partial derivatives equal to zero are $\hat{\mu} = \overline{x}$ and $\hat{\sigma}^2 = (1/n)\sum(x_i-\mu)^2$. The inverse of the variance-covariance matrix is

$$V^{-1} = -\mathbb{E}\begin{pmatrix} \frac{\partial^2 l(\mu,\sigma^2|\mathbf{x})}{(\partial\mu)^2} & \frac{\partial^2 l(\mu,\sigma^2|\mathbf{x})}{\partial\mu\partial\sigma^2} \\ \frac{\partial^2 l(\mu,\sigma^2|\mathbf{x})}{\partial\sigma^2\partial\mu} & \frac{\partial^2 l(\mu,\sigma^2|\mathbf{x})}{(\partial\sigma^2)^2} \end{pmatrix}$$

where

$$-\mathbb{E}\left(\frac{\partial^2 l(\mu,\sigma^2|\mathbf{x})}{(\partial\mu)^2}\right) = -\mathbb{E}\left(\frac{-n}{\sigma^2}\right) = \frac{n}{\sigma^2},$$

and

$$\begin{aligned}
-\mathbb{E}\left(\frac{\partial^2 l(\mu,\sigma^2|\mathbf{x})}{\partial\sigma^2\partial\mu})\right) &= -\mathbb{E}\left(\frac{\partial^2 l(\mu,\sigma^2|\mathbf{x})}{\partial\mu\partial\sigma^2}\right) \\
&= -\mathbb{E}\left(-\frac{n}{\sigma^4}(\overline{x}-\mu)\right) \\
&= \frac{n}{\sigma^4}(\mathbb{E}(\overline{x})-\mu) = 0,
\end{aligned}$$

and

$$-\mathbb{E}\left(\frac{\partial^2 l(\mu,\sigma^2|\mathbf{x})}{(\partial\sigma^2)^2}\right) = \frac{n-1}{2\sigma^4}.$$

The asymptotic sampling distributions of $\hat{\mu}$ and $\hat{\sigma}^2$ are therefore independent normal densities with means μ and σ^2, respectively, and variances σ^2/n and $(2\sigma^4)/(n-1)$, respectively. The matrix V can also be obtained using simulation (see Chapter 6). Using 10,000 simulated datasets, each of size $n = 100$ with $\mu = 5$ and $\sigma^2 = 9$ we obtained $\text{Var}(\hat{\mu}) = .089$, $\text{Var}(\hat{\sigma}^2) = 1.596$ and $\text{Cov}(\hat{\mu}, \hat{\sigma^2}) = -.00008$. The predicted values from the formulas are $\text{Var}(\hat{\mu}) = .09$, $\text{Var}(\hat{\sigma}^2) = 1.636$ and $\text{Cov}(\hat{\mu}, \hat{\sigma^2}) = 0$. Thus, the values are in quite close agreement and the asymptotic approximation appears to be satisfactory in this case.

The example given above is exceptionally simple. Most multidimensional likelihood functions are considerably more complex and often it is necessary to resort to a numerical multidimensional optimization of the log-likelihood. Most statistical packages have some generic algorithms available for multidimensional optimization. A final complication that can arise for multidimensional likelihoods are constraints among parameters. **Constrained maximum likelihood** attempts to maximize the likelihood function subject to one or more constraints on the parameters. Although constraints may be present for one-dimensional likelihood functions the most complicated types of constraints are found in the multidimensional context.

To illustrate how constraints arise and to present one approach for constrained maximization we will consider the multinomial distribution introduced earlier. The pmf of the multinomial distribution is

$$\Pr(\mathbf{x}|\mathbf{p}, n) = \binom{n}{x_1, x_2, \ldots, x_K} \prod_{i=1}^{K} p_i^{x_i}.$$

The constraint, in this case, is that the category probabilities sum to 1,

$$\sum_{i=1}^{K} p_i = 1.$$

The log-likelihood is

$$l(p_1, \ldots, p_K|\mathbf{x}) = \sum_{i=1}^{K} x_i \log p_i - \sum_{i=1}^{K} \log x_i!$$

To impose the constraint we introduce an additional term in the log-likelihood called a **Lagrange multiplier** which in this case is

$$\lambda \left(\sum_{i=1}^{n} p_i - 1 \right).$$

The partial derivative with respect to λ is

$$\frac{\partial}{\partial \lambda} \lambda \left(\sum_{i=1}^{n} p_i - 1 \right) = \sum_{i=1}^{n} p_i - 1.$$

Therefore, setting this partial derivative equal to zero imposes the constraint. The log-likelihood with the Lagrange multiplier term added is

$$l(p_1, \ldots, p_K, \lambda|\mathbf{x}) = \sum_{i=1}^{K} x_i \log p_i - \sum_{i=1}^{K} \log x_i! + \lambda \left(\sum_{i=1}^{n} p_i - 1 \right).$$

Taking partial derivatives with respect to the category probabilities gives

$$\frac{\partial}{\partial p_i} l(p_1, \ldots, p_k, \lambda|\mathbf{x}) = \frac{x_i}{p_i} + \lambda.$$

Setting the partial derivative equal to zero we obtain

$$x_i = -\lambda p_i,$$

and summing over the categories (index i) on both sides of the equation gives $n = -\lambda$. Solving for p_i then gives the set of MLEs,

$$\hat{p}_i = \frac{x_i}{n}$$

which, in this case, are identical to the maximum likelihood estimator of p_i for the marginal binomial distribution.

EM algorithm

Problems in genetics frequently involve variables that are not directly observed. Examples include the heterozygous genotype at a dominant genetic locus, or the combinations of alleles on particular chromosomes (known as the haplotype phase) when diploid genotypes are surveyed for several linked markers. The **Expectation-Maximization algorithm** (EM algorithm) is a method for obtaining maximum likelihood estimates of parameters in models with unobserved variables. One of the earliest applications of the EM algorithm was a study by Fisher and Balmukand (1928) estimating the degree of linkage between two loci using data from a genetic cross. The term EM algorithm originated with the paper by Dempster et al. (1977) which provides a more general mathematical description of such algorithms.

The EM algorithm is an iterative procedure for updating parameter estimates that ultimately converges to the maximum likelihood estimates. Let $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$ be the complete set of observations and let $\mathbf{y}_+ =$

$\{y_1, y_2, \ldots, y_r\}$ be the observed variables and $\mathbf{y}_- = \{y_{r+1}, y_{r+2}, \ldots, y_n\}$ the unobserved variables. The pdf of the complete data is $f(\mathbf{y}|\theta)$, where θ is a set of one or more model parameters. Initial values are assigned to the parameters, θ_0, and the algorithm proceeds in two steps: (1) the expectation step derives the expected value of the log-likelihood for the current parameter value,

$$l_0 = \mathbb{E}[\log f(\mathbf{y}|\theta_0)], \tag{4.2}$$

with the expectation taken over the unobserved variables, conditioned on the observed data. (2) the maximization step derives the maximum likelihood estimates of the parameters using the expected log-likelihood derived in step (1) as the likelihood to be maximized to obtain a new estimate θ_1,

$$\theta_1 = \max_{\theta} l_0. \tag{4.3}$$

This procedure is repeated until the estimated parameters converge to the MLEs within some prespecified level of accuracy ϵ so that at iteration t,

$$|\theta_t - \theta_{t-1}| < \epsilon. \tag{4.4}$$

To illustrate, consider a sample of n individuals genotyped at two linked single nucleotide polymorphism (SNP) loci. We are interested in using the sampled two-locus genotypes to infer the population frequencies of the unobserved haplotypes. SNP genotyping studies generate counts of individual multilocus genotypes which provide information about genotype frequencies at a pair of loci and not necessarily haplotype frequencies. The problem is that the phase is uncertain for individuals that are heterozygous at both loci. For example, an individual with the two-locus genotype $A/a, B/b$ might have either of the diplotypes $A - B, a - b$ or $A - b, a - B$. The basic data from a two locus genotyping study (with two alleles at each locus) are the counts of numbers of individuals with each of the 9 possible combinations of 2 locus genotypes shown in Table 4.1. The population haplotype frequency parameters and (unobservable) haplotype counts are shown in Table 4.2. If we were able to directly observe the haplotypes (rather than the genotype combinations) the probability of the observed counts would follow a multinomial distribution,

$$\Pr(n_{AB}, n_{Ab}, n_{aB}, n_{ab}) = \binom{2N}{n_{AB} n_{Ab} n_{aB} n_{ab}} p_{AB}^{n_{AB}} p_{Ab}^{n_{Ab}} p_{aB}^{n_{aB}} p_{ab}^{n_{ab}}, \tag{4.5}$$

where $2N = n_{AB} + n_{Ab} + n_{aB} + n_{ab}$ and $p_{AB} + p_{Ab} + p_{aB} + p_{ab} = 1$, The log-likelihood (score) function is

$$L = C + \sum_{i,j} n_{ij} \log(p_{ij}). \tag{4.6}$$

	BB	Bb	bb	Total
AA	N_{11}	N_{12}	N_{13}	$N_{1\cdot}$
Aa	N_{21}	N_{22}	N_{23}	$N_{2\cdot}$
aa	N_{31}	N_{32}	N_{33}	$N_{3\cdot}$
Total	$N_{\cdot 1}$	$N_{\cdot 2}$	$N_{\cdot 3}$	N

Table 4.1: SNP genotype counts for 9 possible genotype combinations at two biallelic loci.

Haplotype	$A-B$	$A-b$	$a-B$	$a-b$
Frequency	p_{AB}	p_{Ab}	p_{aB}	p_{ab}
Count	n_{AB}	n_{Ab}	n_{aB}	n_{ab}

Table 4.2: Population haplotype frequency parameters and and counts for 4 possible allele combinations at two biallelic loci.

where C is an irrelevant constant that is not a function of the frequency parameters. By differentiating the log-likelihood function with respect to the frequency parameters (using Lagrangian multipliers to constrain the frequency parameters to sum to 1), setting the partial derivatives to equal zero and simultaneously solving the resulting equations the maximum likelihood estimates of the frequency parameters are found to be (cf. earlier example on constained maximum likelihood),

$$\hat{p}_{ij} = \frac{n_{ij}}{2N}. \tag{4.7}$$

Assuming HWE the expected haplotype counts (as a function of the genotype counts) are

$$\begin{aligned}
n_{AB} &= 2N_{11} + N_{12} + N_{21} + \left(\frac{p_{AB}p_{ab}}{p_{Ab}p_{aB} + p_{AB}p_{ab}} \right) N_{22}, \\
n_{ab} &= 2N_{33} + N_{23} + N_{32} + \left(\frac{p_{Ab}p_{aB}}{p_{Ab}p_{aB} + p_{AB}p_{ab}} \right) N_{22}, \\
n_{Ab} &= 2N_{13} + N_{12} + N_{23} + \left(\frac{p_{Ab}p_{aB}}{p_{Ab}p_{aB} + p_{AB}p_{ab}} \right) N_{22}, \\
n_{aB} &= 2N_{31} + N_{21} + N_{32} + \left(\frac{p_{Ab}p_{aB}}{p_{Ab}p_{aB} + p_{AB}p_{ab}} \right) N_{22},
\end{aligned}$$

Note that $p_{AB}p_{ab}/(p_{Ab}p_{aB} + p_{AB}p_{ab})$ is the probability that an individual with heterozygous genotype Aa/Bb has diplotype $A - B/a - b$ assuming that haplotypes combine in individuals according to HWE proportions (random mating). We can reduce the number of parameters in these equations by noting that the maximum likelihood estimates of allele frequencies at each locus are

$$\begin{aligned} \hat{p}_A &= \hat{p}_{AB} + \hat{p}_{Ab} \\ &= (2N_{11} + N_{12} + N_{21} + 2N_{13} + N_{12} + N_{23} + N_{22})/2N \\ &= \frac{N_{1\cdot} + N_{2\cdot}/2}{N}, \end{aligned}$$

and

$$\begin{aligned} \hat{p}_B &= \hat{p}_{AB} + \hat{p}_{aB} \\ &= (2N_{11} + N_{12} + N_{21} + 2N_{31} + N_{21} + N_{32} + N_{22})/2N \\ &= \frac{N_{\cdot 1} + N_{\cdot 2}/2}{N}, \end{aligned}$$

and that

$$\hat{p}_{Ab} = \hat{p}_A - \hat{p}_{AB}, \ \hat{p}_{aB} = \hat{p}_B - \hat{p}_{AB}, \text{ and } \hat{p}_{ab} = 1 - \hat{p}_{AB} - \hat{p}_A - \hat{p}_B.$$

We now substitute these expressions for the 3 haplotype frequency parameters (those other than p_{AB}) in terms of allele frequencies and $\hat{p}_{AB}$ into the formula for n_{AB} to obtain a predictor of the expectation of n_{AB} given $\hat{p}_{AB}$,

$$\begin{aligned} &\mathbb{E}(n_{AB}|\hat{p}_{AB}) = 2N_{11} + N_{12} + N_{21} + \\ &\frac{\hat{p}_{AB}(1 + \hat{p}_{AB} - \hat{p}_A - \hat{p}_B)N_{22}}{(\hat{p}_A - \hat{p}_{AB})(\hat{p}_B - \hat{p}_{AB}) + \hat{p}_{AB}(1 + \hat{p}_{AB} - \hat{p}_A - \hat{p}_B)}. \end{aligned} \tag{4.8}$$

The only unspecified variable in this formula is $\hat{p}_{AB}$ which is the MLE of the parameter that we wish to obtain. We now make use of the EM algorithm to obtain the MLE. The Expectation step in the EM algorithm calculates the expected value of the count n_{AB} given a specified value of $\hat{p}_{AB}$ (using equation 4.8 above) and the Maximization step uses the maximum likelihood estimator to obtain a revised estimate of $\hat{p}_{AB}$ using $\mathbb{E}(n_{AB}|\hat{p}_{AB})$ in place of n_{AB} in the MLE formula (equation 4.7). The procedure is started with an arbitrary initial value for $\hat{p}_{AB}$. Th expectation-maximization steps are performed repeatedly, each time replacing the current value of $\hat{p}_{AB}$ with the MLE value obtained at the previous iteration, until the estimate converges to the MLE. At the ith iteration, the maximization step is

$$\hat{p}_{AB}(i) = \frac{\mathbb{E}(n_{AB}|\hat{p}_{AB}(i-1))}{2N}.$$